Flying Under the Radar: Ghosts and the Income Tax*

Brian Erard (B. Erard & Associates, LLC), Patrick Langetieg (Internal Revenue Service),
Mark Payne (Internal Revenue Service) and Alan Plumley (Internal Revenue Service)

Abstract

The tax compliance literature is primarily focused on taxpayers who fail to accurately report their taxes when they file their returns. In this paper, our focus is on "ghosts" – individuals who do not even file a tax return. To learn more about this relatively understudied population, we examine a combination of US administrative data and matched Census survey data. Our results indicate that 10-12 percent of US households with a federal filing requirement fail to file a timely income tax return. Approximately 40 percent of such households do eventually file a late return. However, the tax gap associated with those who never file is substantial, amounting to an estimated $18-20 billion each year. To gain new insights into what drives individuals to become ghosts, we employ a novel econometric methodology (calibrated probit analysis). We find that the failure to file a timely return is negatively associated with age and income, but positively associated with having a high filing burden and being married. Taxpayers with income near the filing threshold are also less likely to file on time, particularly if they are not eligible for a refundable tax credit. We also find evidence of regional variation in filing compliance.

* The opinions expressed herein are those of the authors and do not necessarily reflect the views of the IRS.

1. Introduction

Although there is an ever-growing body of literature on tax compliance, much of its attention continues to be cast on tax filers and their reporting behavior. A relatively neglected group is those individuals who elect to take the ultimate compliance shortcut of not filing a tax return at all. Such individuals are sometimes referred to as "ghosts" by academics and policy makers – a term that seems fitting given both their propensity to fly beneath a tax administration's radar and their absence from many models of tax compliance behavior.

In this paper, we present estimates of the share of US taxpayers who fail to file their required federal individual income tax returns as well as the taxes that they owe. Although ghosts are commonly associated with tax regimes in developing countries, our results show that they also pose a challenge for the US Internal Revenue Service (IRS). Following our presentation of estimates regarding the scope of the nonfiler problem in the US, we provide a discussion of the economic incentives to become a ghost. This discussion serves as the basis for our novel empirical analysis of the drivers of filing compliance. The paper concludes with a discussion of potential avenues for future research.

2. Voluntary Filing Rate

In the US, households generally are not required to file a federal individual income tax return if their gross income is below a specified filing threshold.[1] Of course, many do so in any case either to receive a refund of overpaid taxes or to claim a refundable credit. The focus of this study, however, is on the filing behavior of those with a legal filing obligation.

The IRS defines the Voluntary Filing Rate (VFR) as the share of all required returns that are filed by their due date (typically, April 15 of the subsequent calendar year, or October 15 for those requesting an extension). Table 1 presents recent IRS estimates of the VFR for 2008 through 2010, which coincides with years for which we have been able to develop the nonfiler tax gap estimates presented below in Section 3.[2] As indicated in Table 1, the VFR estimates fluctuate modestly over this period, ranging from 88.1 to 89.9 percent. Among those who fail to file a required return on time, roughly three-fifths never file at all. Furthermore, among those who do file a late return, roughly 20 percent do so only after receiving a notification from the IRS reminding them of the requirement to file a return.

[Insert Table 1 about here]

3. Nonfiler Tax Gap

We have employed two alternative methodologies to estimate the tax gap among households that fail to file a timely required return. Taxpayers often make tax prepayments in the form of employer withholding and estimated tax payments, and they are also sometimes eligible for refundable credits. The tax gap for those who fail to submit a tax return by the due date therefore only refers to the net tax balance that is owed as of that date. Although we refer to this concept as the "nonfiler tax gap", it actually represents the net tax balance among late filers as well as those who have never filed their required returns. Below, we provide a brief overview of each estimation approach.[3]

As its name suggests, the *Administrative Data Method* relies primarily on administrative tax information. To identify individuals who have never filed a required return, the starting point is the population of individuals for whom the IRS has received a third-party information report for the year in question. For such individuals, administrative information on income, tax prepayments, age, gender, and state of residence is gathered. The following steps are then followed:

- Remove individuals who filed an income tax return for the year in question; also remove the spouse (If present) if the filing status on the return is married joint.
- Impute a filing status to the remaining members of the population.[4]
- Impute self-employment earnings, adjustments, deductions, and credits for the filing unit.
- Apply a computer algorithm to test whether the filing unit had a legal filing obligation; if not, exclude the unit from further analysis.
- Apply a tax calculator to compute total tax liability net of tax prepayments and refundable credits for the remaining filing units.
- Aggregate the net tax balance due computed in the previous step across all filing units with a positive net balance to estimate the tax gap associated with never-filers.

Late filers are straightforward to identify from administrative records. Details from their tax returns as well as associated third-party information reports are compiled for each member of the late-filer population. The following steps are then followed to arrive at an estimate of the tax gap attributable to late filing.

- Adjust the income and prepayment amounts reported on the late-filed returns to account for any omitted amounts that are revealed on the third-party documents.
- Apply a computer algorithm to assess whether the taxpayer had a legal filing obligation; if there was no filing requirement, the late-filed return is excluded from further analysis.
- Apply a tax calculator to compute total tax liability net of tax prepayments and refundable credits for each of the remaining late-filers.
- Aggregate the net tax balance due computed in the previous step across all taxpayers with a positive net tax balance to estimate the tax gap associated with late filers.

The *Census Method* relies primarily on data obtained by matching federal income tax returns and third-party information reports for a given year against the *Annual Social and Economic Supplement* of the *Current Population Survey* (CPS-ASEC). The CPS-ASEC is a large and statistically representative annual survey of the US non-institutional population. The survey elicits very detailed information on the socio-economic characteristics of individuals and their households, including income earned during the prior calendar year from a wide range of sources. For the survey years used in this paper, this information was collected for more than 200,000 individuals from approximately 95,000 households in each year. Census has assigned a "Protected Identification Key" (PIK) that links a CPS-ASEC respondent to his or her IRS records (including selected information from tax returns and third-party information reports) with a high degree of accuracy. Potential nonfilers are then identified from the sub-population of CPS-ASEC respondents who have a PIK but who did not file a tax return for the year in question.[5] These potential nonfilers include late filers who filed after December 31 of the calendar year in which the return was due as well as respondents who never filed.

The following steps are applied to the identified population of potential nonfilers to estimate the nonfiler tax gap:

- Reweight the sample to account for records that could not be matched using a PIK.
- Since CPS-ASEC respondents do not always provide accurate reports of their income from various sources, adjust the reported amounts to reflect any omitted income that is revealed on third-party information reports.
- Impute additional income from sources not subject to third-party reporting that tend to be understated on the survey (such as self-employment earnings).
- Impute tax filing statuses using CPS-ASEC records on the marital statuses and ages of the household members.

- Impute tax deduction, credit, and prepayment amounts based on a combination of survey and matched tax administrative data.
- Apply a computer algorithm to identify whether a return was required to be filed; if not, exclude the filing unit.
- Apply a tax calculator to evaluate the net tax balance due after tax prepayments and refundable credits.
- Compute the weighted net tax balance due from the previous step across all filing units with a positive net balance to obtain a preliminary estimate of the nonfiler tax gap..
- Since the potential nonfiler sample excludes late filers who filed between the return due date and the end of the calendar year in which the return was due, an estimate of the tax gap associated with these "early late filers" is incorporated into the final tax gap estimate based on the *Administrative Data Method*.

Table 2 presents our estimates of the nonfiler tax gap based on both the *Administrative Data Method* and the *Census Method.* Both methods yield similar results, attributing a three-year average net tax balance of about $30 billion as of the return due date among those taxpayers with a positive balance due who failed to file a timely return. Although late filers did eventually report and pay most of their tax balances, an estimated $17.6 to $19.9 billion in taxes went unpaid in each of these years by households that never filed their required tax returns.

[Insert Table 2 about here]

As shown in Table 3, the nonfiler tax gap is highly concentrated with 64 percent of the overall gap attributable to the top balance due decile in tax year 2010.

[Insert Table 3 about here]

Our nonfiler tax gap estimates that were presented in Table 2 consider only taxpayers who have a positive net tax balance after prepayments and refundable credits. However, many taxpayers who fail to file a timely required return are actually entitled to a refund. Indeed, on net, late filers as a group ultimately received an aggregate net refund of $6.4 billion for tax year 2010. In contrast, those who never filed their required tax year 2010 returns (including both nonfilers with a balance due and nonfilers entitled to a refund) had an estimated aggregate net tax balance for that year of $13.8 billion. Overall, then, the federal tax revenue savings associated with taxpayers who fail to claim refunds to which they are entitled only partially offsets the losses attributable to nonfilers who fail to pay their positive tax balances.

4. Theoretical Insights

Under the standard economic model of tax compliance, it is taken for granted that a taxpayer will file a return. The question is not *whether* the taxpayer will make a tax report, but rather *how much* will he choose to report. Under this framework, the taxpayer approaches his reporting decision as he would an ordinary gamble, balancing the prospect of retaining a greater income share under successful tax evasion against the risk of audit and penalty. In particular, the taxpayer chooses an amount *X* of income to report to maximize expected utility (EU):

$$EU = (1-p)U[Y - tX] + pU[Y - tY - \theta t(Y - X)],$$

where p is the audit risk, t is the tax rate (assumed here to be proportional), Y is the taxpayer's true income that should be reported, X is the amount actually reported on the return, θ represents the penalty rate on the unreported tax amount, and $U[z]$ represents the taxpayer's utility associated with net wealth *z.* In more sophisticated versions of this model, the audit probability for a given tax report might be strategically chosen by the tax authority, true income might depend on an individual's labor

supply decision, the tax rate might vary with the level of income, taxpayer perceptions of risk may deviate from actual risk, audits may not identify all instances of underreporting, and various additional pecuniary and non-pecuniary factors might influence decisions. However, this simple specification captures the basic insight that taxpayers have an incentive to underreport an additional dollar of income so long as the terms of the risk-reward tradeoff are favorable for one's marginal utility.

To account for ghosts, Erard and Ho (2001) have extended this framework to consider what would happen if no return were filed at all. In this extended framework, an individual separately considers his potential utility under scenarios where he does and does not file. Under the filing scenario, the individual must choose not only how much income X to report, but also how much tax to prepay through withholding or estimated tax payments. In addition, there is a compliance cost *c* associated with filing. The expected utility under this scenario EU_F is:

$$EU_F = (1-p)U[Y - tX - \gamma(\overline{W} - W) - c] + pU[Y - tY - \theta t(Y - X) - \gamma(\overline{W} - W) - c],$$

where $\overline{W}$ is the minimum required tax prepayment amount and γ is the penalty rate on under-withholding .[6] Based on this scenario, the taxpayer would choose to make the minimum tax prepayment $\overline{W}$ if he were to file. He would then select the value of X that maximizes the above expression, conditional on $W = \overline{W}$. At this report, the net marginal expected utility associated with reporting an additional dollar of income would be equal to zero.

Under the nonfiling scenario, the taxpayer has only to choose how much tax to prepay (W), recognizing that his expected utility ($EU_{\bar{F}}$) depends on the risk of enforcement and the accompanying tax and penalty assessments:

$$EU_{\bar{F}} = (1-q)U[Y - W] + qU[Y - tY - f(tY - W) - c],$$

where q is the risk of nonfiling enforcement and f is the penalty rate on the unpaid tax balance. If apprehended, the nonfiler would be required to submit a tax return, so the above expression accounts for the filing burden *c* in the event of nonfiler enforcement. At the optimal choice of W under this nonfiling scenario, the net expected marginal utility associated with prepaying an additional dollar of tax would be equal to zero.

To decide whether to file a return, the individual separately computes his maximum expected utility under the filing and nonfiling scenarios and chooses the option that produces a higher result. To understand what drives this decision, it is instructive to consider the base case where the filing burden *c* is equal to zero, the risks of enforcement under the two scenarios are equal ($p = q$), and the penalty rates are the same ($\theta = f$). In this case, Erard and Ho (2001) show that the maximum expected utility under the two scenarios is exactly the same, so that the individual is indifferent about whether to file a return. If the individual were to file, he would make a tax prepayment of $W = \overline{W}$ and he would make an optimal income report X^*. Alternatively, the individual could achieve the same expected utility by making a tax prepayment of $W = tX^*$ but not filing a return at all. Thus, the magnitude of the filing burden and the relative rates of risk and penalty are what drive filing behavior in this model. Individuals are more likely to be ghosts when the filing burden *c* is high, the risk of nonfiler enforcement *q* is low relative to the audit rate *p*, and the penalty rate facing nonfilers *f* is low relative to that faced by filers θ.

5. Estimating the Drivers of Filing Compliance

In this section we combine some of the theoretical insights from the previous section with a novel econometric methodology to examine the drivers of filing compliance. Typically, economists examine the drivers of a participation decision (such as filing a tax return) using a qualitative response model, such as probit or logit. However, the application of such a model requires detailed information on the relevant characteristics of a representative sample of both participants and non-participants. Although

the IRS has very detailed information on filers of individual income tax returns, it lacks comparably detailed information on nonfilers. To address this informational gap, we supplement IRS information on the characteristics of filers with publicly available CPS-ASEC survey information on the characteristics of both filers and nonfilers in the general population. Although this latter data source does not identify which respondents are filers and which are nonfilers, we are nonetheless able to estimate a qualitative response model using the calibrated probit methodology developed by Erard (2018).[7]

Calibrated Probit Methodology

Intuitively, whereas a standard probit analysis relies on the differences between the characteristics of participants (in this case, filers) and non-participants (non-filers) in a single representative sample to infer what drives behavior, the calibrated probit methodology relies on differences in characteristics between a representative sample of participants and a supplementary sample that, when weighted, is representative of the overall population of participants and non-participants. The calibrated probit model is estimated by solving the following optimization problem:

$$\max_{\beta} \sum_{i=1}^{n_f} ln[\Phi(\beta' X_i)]$$

$$s.t. \quad \sum_{j=1}^{n_o} w_j \Phi(\beta' X_j) = N_f.$$

where n_f represents the number of filers in our filer-only data sample, n_o is the size of our supplementary sample of filers and nonfilers, *X* is a vector of explanatory variables, β is a vector of coefficients to be estimated, $\Phi(\mathrm{z})$ represents the value of the standard normal cumulative distribution function evaluated at z, w refers to sample weights in the supplementary sample, and N_f is the population number of returns filed. The solution to this problem is the value of β that maximizes the predicted likelihood of filing among households in the filer-only data sample subject to the constraint

that the weighted aggregate predicted number of filers in the supplementary sample based on this solution is consistent with the actual number of filers in the population. In other words, the estimated value of β is calibrated to be consistent with the population count of filed required returns (hence the name "calibrated probit analysis").We estimate our model using the NLPQN subroutine in SAS/IML[@].[8]

Data Description

Our IRS data source on filers is the Individual Returns Transaction File (IRTF), a detailed extract of return information from all federal individual income tax returns filed for a given tax year. Many households have no legal filing obligation because their income is below the filing threshold and they do not meet certain other filing criteria, such as a need to report self-employment tax or taxes on tip income. Some of these households do file, however, to claim refunds of withheld earnings or to claim a refundable tax credit, such as the Earned Income Credit. Since our focus is on filing compliance, we restrict our IRTF sample to households with a legal filing obligation. This is achieved by applying an algorithm to check whether a given return satisfies any of the various conditions that trigger a filing requirement (such as gross income above the relevant filing threshold or net self-employment earnings in excess of $400). Our supplementary sample of filers and nonfilers is drawn from the CPS-ASEC. In past research, we have found that certain income sources are understated in this survey. Therefore, in order to more accurately identify households with a legal filing obligation, we follow Erard, Langetieg, Payne, and Plumley (2014) in imputing additional income across the sample (in many cases, based on third-party information return data). To assign household members to tax returns, we also impute tax filing status. The CPS-ASEC is a stratified random sample; however, the stratification criteria are not publicly available. A desirable feature of our econometric methodology is that we are able to effectively control for the stratified nature of the sample simply by applying the sample weights.

For both data sources, we have large cross-sectional samples for tax year 2010. The data include a simple random sample of approximately 113,000 filed required returns the IRTF and a stratified random sample of approximately 76,000 required returns from the CPS-ASEC.

Estimation Results

In this section, we present the results of our calibrated probit analysis of the drivers of filing compliance. Because our estimation methodology relies on a comparison of variables from two separate data sources (IRTF and CPS-ASEC), it is important to restrict the set of regressors to those variables that are comparably measured in the two sources. So, for instance, while the IRTF provides information on whether a taxpayer is owed a refund or has a balance due (which is likely to be relevant to the filing decision), comparable information is not available in the CPS-ASEC. It also would be desirable to include some indicators of filing status as explanatory variables. However, a non-trivial number of taxpayers claim the incorrect status on their return. For instance, the percentage of filers claiming head of household status greatly exceeds the estimated percentage of required returns with this status based on the CPS-ASEC. Instead of filing status indicators, we include an indicator for marital status in our specifications. Similarly, we would like to account for Earned Income Credit eligibility in our analysis. However, a claims-based measure from the IRTF would be misleading, owing to a nontrivial number of Earned Income Credit claimants who are not truly eligible. Ultimately, we have selected the following explanatory variables for our analysis, which we believe are measured reasonably comparably across our two data sources:

CONST: Constant term.

AGE 65: Dummy for primary or secondary filer age 65 or over.

MARRIED: Dummy for married taxpayer.

CHILD3UP: Dummy variable for having 3 or more dependent children.

LN(GROSSINC): Natural log of gross income, where gross income is computed as the sum of the positive amounts of wages and salaries, interest, taxable dividends, pensions, rents, unemployment compensation, taxable social security benefits, alimony, and gross self-employment earnings.

NO TAX STATE: Dummy variable for residence in a state with no individual income tax (AK, FL, NH, NV, SD, TN, TX, WA, WY).

SEFILREQ: Dummy variable for a filing requirement triggered by having net earnings from farm and nonfarm self-employment in excess of $433.

NEARTHRESH: Dummy variable for gross income less than 1.10 times the filing threshold, where the filing threshold for non-dependent joint filers is applied for married joint filing status and the filing threshold for single filers is applied to all other non-dependent filers. The lower statutory thresholds are applied to single and married dependent filers.

LN(BURDEN): Natural log of taxpayer burden.

LN(BURDEN)*NEAR THRESHOLD: Interaction between LN(BURDEN) and NEARTHESH.

MIDATL: Dummy variable for residence in the Mid-Atlantic division.

EASTNC: Dummy variable for residence in the East North Central division.

WESTNC: Dummy variable for residence in the West North Central division.

SOUTHATL: Dummy variable for residence in the South Atlantic division.

EASTSC: Dummy variable for residence in the East South Central division.

WESTNC : Dummy variable for residence in the West South Central division.

MOUNTAIN : Dummy variable for residence in the Mountain division.

PACIFIC : Dummy variable for residence in the Pacific division.

The omitted Census division is New England. Table 4 breaks down the average of the weighted mean values of our explanatory variables separately for the tax year 2010 IRTF and CPS-ASEC samples. The results show that filers were relatively more likely to have three or more dependent children, to reside in states without an income tax, and to live in the Mid-Atlantic and East North Central divisions in that year. They were relatively less likely to be married, receive non-trivial self-employment earnings, or

reside in the Mountain or Pacific divisions. On average, their gross earnings were somewhat lower than the overall population of households with a filing requirement, and they faced a somewhat lower filing burden.

[Insert Table 4 about here]

The results of estimation are summarized in Table 5. To assist with interpretation, the estimated marginal effects and their t-statistics are included. The results indicate that elderly taxpayers are relatively more likely to file than younger taxpayers. All else equal, the likelihood of filing is 8.1 percentage points higher when the individual is age 65 or over. Gross income is also positively associated with filing compliance. All else equal, a filing unit with $90,000 in gross income is about 3.5 percentage points more likely to file than a more typical filing unit with a gross income of $45,000.

Filing compliance is also linked to marital status. Compared to an unmarried individual with similar characteristics, a married individual is about 6.2 percentage points less likely to file. As well, there is evidence of regional variation in filing compliance, with higher compliance among those residing in the Mid-Atlantic division and lower compliance among those residing in the Mountain and Pacific divisions.

Individuals with gross income close to the filing threshold may perceive that they can fly beneath IRS's radar, since information available to the agency is unlikely to definitively show an income tax filing requirement; moreover, any tax balance due is likely to be rather modest. Perhaps for this reason, the results indicate that such individuals are about 7 percentage points less likely to file than other taxpayers with similar characteristics.

The theoretical framework indicated that a high filing burden can discourage filing compliance. Overall, it appears that burden does, in fact, serve as a deterrent to compliance. The filing burden for a

typical taxpayer in tax year 2010 is estimated to have been about $450. Compared to such a taxpayer, an individual facing a filing burden of $900 is estimated to be about 8 percentage points less likely comply with his filing obligation. An exception to this finding applies to individuals with gross income near the filing threshold. For such individuals, the deterrent effect of the filing burden is largely mitigated. Many individuals with relatively low earnings are eligible for refundable tax credits, such as the Earned Income Credit. Although it is somewhat burdensome to claim such credits, the benefits in most cases will outweigh the additional filing burden, which helps to explain the differential impact of filing burden for those near the filing threshold.

Eligible taxpayers with three or more children receive a larger Earned Income Credit than those with comparable earnings but fewer children. Consistent with this incentive, the results indicate that households with three or more children are relatively more likely to file a required tax return.

Somewhat surprisingly, the results do not indicate any significant relationship between filing compliance and having self-employment earnings or living in a state that does not impose an income tax.

[Insert Table 5 about here]

6. Directions for Future Research

This study has confirmed the existence of ghosts in the US income tax system. At the time that federal individual income tax returns were due for each of tax years 2008 through 2010, about 10 to 12 percent of the required returns had not been filed. While about 40 percent of this initial ghost population eventually emerged to file their returns, around 8 to 10 million ghosts never did so.

The taxes owed by ghosts are significant. Returns not filed by their due date were associated with an estimated net tax balance due of some $30 billion per year over the tax year 2008 through 2010

period. Late filings (which were sometimes prompted by IRS enforcement activity) ultimately remitted a portion of this balance. However, $17 to $20 billion per year continued to go unpaid by ghosts who never filed.

Although ghosts as a group have a substantial tax balance due on net, many of them are actually entitled to tax refunds, owing to their excessive prepayment of taxes and their eligibility for refundable credits. The econometric analysis indicates that burden can be an important deterrent to filing a return.

In future research, we plan to explore filing compliance behavior over a large number of tax years. We are also pursuing an extension of our estimation framework to account for past filing behavior. This is important, because individuals who have filed in the past are more visible to the tax agency. It is commonly believed that once a filer enters the income tax system, he is likely to remain in the system. However, this hypothesis would benefit from empirical confirmation. We are also exploring an extension of our framework to multinomial outcomes, so that we can examine not only the factors that drive timely filing behavior, but also those that drive late filing behavior.

Funding

The work of the first author was supported by the Internal Revenue Service [contracts TIRNO-10-D-00021, TIRNO-14-P-00157 and TIRNO-15-P-00172]. The opinions expressed herein are those of the authors and do not necessarily reflect the views of the IRS.

Acknowledgements

We thank Christos Kotsogiannis, the other participants at the CESifo Economic Studies Conference on New Perspectives on Tax Administration Research, and an anonymous referee for their helpful comments.

Table 1: Estimated breakdown of federal individual income tax filing behavior, tax years 2008-2010

Tax Year	Millions of Required Returns				Voluntary Filing Rate
	Filed on Time	Filed Late	Never Filed	Total Required	
2008	116.3	5.0	8.3	129.5	89.7%
2009	113.4	5.1	9.5	128.0	88.6%
2010	115.4	5.9	9.8	131.0	88.1%

Source: Estimates developed by the IRS

Table 2: Nonfiler tax gap estimates ($billions) based on the Administrative Data and Census Methods, three-year average (TY2008-TY2010)

Tax Concept	Administrative Data Method	Census Method
Total Income	**591.8**	**604.7**
…minus adjustments	-13.6	-13.1
…minus exemptions	-67.3	-63.1
…minus itemized deductions	-145.5	-130.6
Taxable income	**365.4**	**397.9**
Tentative tax	**68.8**	**73.9**
…plus self-employment tax	10.9	10.1
…minus tax offset by refundable credits	3.0	3.4
Net tax due	**77.3**	**80.8**
…minus tax offset by prepayments	-43.6	-45.1
…minus tax offset by refundable credits	-4.8	-5.0
Overall nonfiler tax gap	**28.9**	**30.8**
Never-filer share of overall gap	**17.6**	**19.9**

Table 3: Nonfiler tax gap by decile, TY2010

Decile	Balance Due ($Billions)	Percent Share
1	0.0	0.1%
2	0.2	0.5%
3	0.2	0.8%
4	0.4	1.4%
5	0.7	2.2%
6	1.0	3.4%
7	1.5	5.3%
8	2.4	8.2%
9	4.3	14.4%
10	18.7	63.7%

Table 4: Average values of explanatory variables by data source

Variable	Data Source	
	IRTF	CPS-ASEC
AGE 65	0.1420	0.1417
MARRIED	0.4339	0.4563
CHILD3UP	0.0685	0.0624
LN(GROSSINC)	10.7046	10.7501
NO TAX STATE	0.2062	0.2027
SEFILREQ	0.1303	0.1361
NEARTHRESH	0.0364	0.0416
LN(BURDEN)	6.0604	6.1013
LN(BURDEN)*NEARTHRESH	0.1994	0.2265
MIDATL	0.1376	0.1359
EASTNC	0.1535	0.1515
WESTNC	0.0691	0.0679
SOUTHATL	0.1934	0.1922
EASTSC	0.0565	0.0568
WESTSC	0.1111	0.1111
MOUNTAIN	0.0691	0.0717
PACIFIC	0.1597	0.1628

Table 5: Estimation Results: Drivers of Filing Compliance in TY2010*

Variable	Coefficient Estimate	t-statistic	Marginal Effect	t-statistic
CONST	1.9188	9.70		
AGE 65	0.6112	4.12	0.0812	6.36
MARRIED	-0.3404	-5.71	-0.0615	-5.65
CHILD3UP	0.4235	4.89	0.0585	6.42
LN(GROSSINC)	0.1980	8.01	0.0350	7.15
NO TAX STATE	0.0590	1.05	0.0102	1.07
SEFILREQ	-0.0438	-0.92	-0.0079	-0.90
NEARTHRESH	-2.7349	-11.92	-0.1421	-7.77
LN(BURDEN)	-0.4563	-10.37	-0.0806	-8.89
LN(BURDEN)*NEARTHRESH	0.3935	9.63	0.0695	9.28
MIDATL	0.2931	2.76	0.0472	2.76
EASTNC	0.1134	1.27	0.0205	1.25
WESTNC	0.1137	1.18	0.0205	1.17
SOUTHATL	0.0760	0.89	0.0141	0.88
EASTSC	0.0876	0.78	0.0161	0.79
WESTSC	0.1079	1.06	0.0196	1.06
MOUNTAIN	-0.1812	-2.10	-0.0390	-2.16
PACIFIC	-0.0573	-0.71	-0.0115	-0.72

*The estimated coefficients and marginal effects are based on a calibrated probit specification of the likelihood of filing a timely return. The omitted Census division is New England.

References

Cosslett, S.R. (1981), ``Efficient Estimation of Discrete Choice Models'', in C. Manski and D. McFadden, eds., *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, pp. 51—111.

Erard, B. (2018), "Modeling Qualitative Outcomes by Supplementing Participant Data with General Population Data: A New Approach", Working paper.

Erard, B., P. Langetieg, M. Payne, and A. Plumley (2014) ,"Missing Returns vs. Missing Income: Estimating the Extent of Individual Income Tax Filing Compliance from IRS and Census Data", *Proceedings of the 107th Annual Conference of the National Tax Association*.

Erard, B. and C.-C. Ho (2001) ,"Searching for Ghosts: Who are the Nonfilers and How Much Tax Do They Owe?", *Journal of Public Economics* **81**, 25—50.

Lancaster, T. and G. Imbens, (1996) "Case Controlled Studies with Contaminated Controls", *Journal of Econometrics* **71**, 145-160.

Langetieg, P., M. Payne and A. Plumley (2016), " Individual Income Tax and Self-Employment Tax Nonfiling Gaps for Tax Years 2008-2010", Paper presented at the 2016 IRS-TPC Research Conference, Washington, D.C., June 23.

Manski, C.F. and D. McFadden (1981), "Alternative Estimators and Sample Designs for Discrete Choice Analysis", in *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. Manski and D. McFadden, Cambridge: MIT Press, 2-49.

Plumley, A. (1996), "The Determinants of Individual Income Tax Compliance: Estimating the Impacts of Tax Policy, Enforcement, and IRS Responsiveness", Internal Revenue Service, Publication 1916 (Rev 11-96), Washington, DC.

Footnotes

1 There are some exceptions to this rule, such as when the individual has to file to report self-employment payroll taxes or the receipt of advance payments of certain credits.

2 These estimates are computed as the ratio of the number of required returns (i.e., federal individual income tax returns for which taxpayers had a legal filing obligation) that actually were filed on time to an estimate of the total number of required returns. The latter includes timely filed required returns, late-filed required returns, and required returns that were never filed. Tax returns records were used to identify timely filed and late-filed returns, while other administrative data, such as third-party information reports and income tax returns filed for other tax years, were used to identify required returns that were never filed. A computer algorithm was employed to distinguish required from non-required returns. The estimated filing rates for tax years 2008-2010 are broadly consistent with IRS estimates for more recent tax years.

3 For a detailed discussion of these approaches, refer to Langetieg, Payne, and Plumley (2016).

4 Imputations of filing status relied on characteristics such as age and gender. The imputation methodology was calibrated to ensure that the overall age and filing status distribution of timely filers, late filers, and nonfilers was consistent with Census tabulations.

5 It was not possible to assign a PIK to certain CPS-ASEC respondents, because a match to IRS records could not be assigned with a reasonable degree of certainty. Consequently, we have re-weighted the PIK subsample to account for potential nonfilers among those CPS-ASEC respondents without a PIK.

6 For simplicity, the minimum required federal tax prepayment amount is assumed to be fixed. In practice, it depends on one's overall tax liability, so the tax authority may not detect that insufficient tax was withheld if the taxpayer underreported income and no audit was undertaken.

7 In future work, we hope to perform a more traditional qualitative data analysis using a sample of filers and nonfilers in the matched CPS-ASEC-administrative sample discussed previously. To employ this method, we will first need to overcome some sample selection issues involving returns without a PIK as well as some data quality issues.

8 We rely on the GMM option in the SAS MODEL procedure to evaluate the standard errors of our estimates. The estimated standard errors for the marginal effects are then obtained via the delta method.