Sample size determination for risk-based tax auditing

TARC DISCUSSION PAPER 026 - 19

PETROS DELLAPORTAS, UNIVERSITY COLLEGE OF LONDON
EVANGELOS IOANNIDIS, ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS
CHRISTOS KOTSOGIANNIS, TARC, UNIVERSITY OF EXETER
Sample size determination for risk-based tax auditing

Petros Dellaportas
University College of London, UK,
Athens University of Economics and Business, Greece
and Alan Turing Institute, UK.

Evangelos Ioannidis
Athens University of Economics and Business, Greece.

Christos Kotsogiannis
Tax Administration Research Centre (TARC), University of Exeter Business School, UK,
and CESifo, Munich, Germany.

Summary. The selection of taxpayers to be audited by tax authorities is an important problem of great economic value. This is a non-standard sample size determination problem because it involves an initial random sample from the population and, based on the statistical information derived from this sample, a consequent risk-based auditing which attempts to select taxpayers with maximal estimated risk in the population. Assuming that the total number of audits is fixed, we provide a methodological approach that estimates the initial random sample size such that the tax authorities maximise their expected tax revenue.

Keywords: bootstrap; Scheffe's multiple comparison; tax gap.

1. Introduction

An efficient management of tax compliance—and the promotion of voluntary compliance—necessitates the development of modern approaches based on risk management. For tax matters, for example, taxpayers are frequently audited in order for the revenue authority (RA) to assess whether taxpayers comply with the tax law. But, in practice, not all tax returns are exhaustively examined. This would not only be, given budgetary constraints, infeasible but it would also be unnecessary to waste scarce enforcement resources on routinely examining low-risk and compliant taxpayers. This discussion is not specific to tax matters but it applies to much broader themes, including, for example, regulation and management. Having said this, and since this paper is motivated by an issue of significant concern of RAs, the discussion will be centered around the compliance issues pertaining tax administrations.

Common practice for tax authorities, notably Her Majesty’s Revenue Customs (HMRC), U.S Internal Revenue Services (IRS) and Canada Revenue Agency, is to use random sample audits. The reason for this is not only to estimate the ‘tax gap’—defined to be equal

Address for correspondence: Petros Dellaportas Department of Statistical Science, University College, Gower Street, London, WC1E 6BT E-mail: p.dellaportas@ucl.ac.uk
the difference between tax voluntarily paid and tax actually owed (the gross compliance gap)—and its component parts, but, also, importantly, to establish the criteria for the selection of accounts for audit in the regular enforcement program which is part of the risk management strategy that identifies areas deemed to be of greatest risk for substantial noncompliance with reporting. The selection of taxpayers to be audited, as well as other types of controls, is based on the assessment of risk and the development of risk-based selection techniques. Taxpayer characteristics are used in risk-scoring systems to identify and assess the risk of noncompliance. This allows appropriate prioritization of audit and taxpayer service workload and enables the allocation of resources to the important high-risk groups; see [OECD (2004)] for a review.

The most extensive analysis based on random audits comes from the IRS studies and the Taxpayer Compliance Measurement Program (TCMP). The studies were conducted to provide the compliance information needed to gauge performance and to inform direction of agency resources through efficient prioritization of audits (see, for example, [Rotz et al. (1994)]) via the construction of Discriminant Inventory Function scores ([Hunter and Nelson, 1996]). The research has showed a considerable compliance gap. The 1992 tax gap of federal individual income tax based on TCMP audits in 1988, the last year of that program, was estimated range from $93 to $95 billion which translates into an individual gross noncompliance rate of about 17 percent. In response to the need for up-to-date measures of taxpayer compliance, and better target of audit resources, the IRS developed the National Research Program (NRP) which is also designed to be less intrusive and burdensome to taxpayers than the TCMP. Under the NRP, 30,000 returns for limited in-person audit and about 2,000 returns for calibration audits when each line is examined and supporting documentation is required from a taxpayer. This is a significant reduction from the 54,000 taxpayers who were required to participate in face-to-face audits in the earlier TCMP program.

In the UK, and HMRC, the Random enquiry programmes (REPs) involve samples of taxpayers being selected at random and their returns subjected to full enquiries by HMRC officers. The results of these programmes show the proportion of taxpayers under-reporting their tax liabilities and the corresponding amount of additional tax due. These results can be used to produce an estimate for the amount of under-declared tax liability for the whole population, as the enquiries are randomly selected and form a representative sample. A proportion of the under-declared liabilities will be recovered as a result of this compliance activity, which is then subtracted from the estimate of under-declared liabilities. Losses through non-payment are also added to obtain final tax gap estimates. The REPs will not identify all incorrect returns or the full scale of tax gaps, especially where independent information from third parties is not available to verify the data supplied by the taxpayer. This means that tax gap estimates produced through random enquiries will under-estimate the full extent of the tax gap. The Internal Revenue Service (IRS) in the United States, and the REP in the UK, has tackled this problem by using a range of ‘multipliers’—supplemented with econometric analysis—to make adjustments for non-detection of under-reported income. ([Andreoni et al., 1998]).

In the spirit of the HMRC and IRS practice, we assume the following RA approach. There is a fixed number of audits that can be conducted due to budgetary constraints which they are split into two stages. In the first stage, a random sample is drawn from the
population which is used to estimate the risk of non-compliance for all population units. In the second stage, the remaining number of audits are chosen to be those with maximal estimated risk in the population. We address the problem of optimally splitting the available number of audits into the two stages. This is an issue of great policy interest for RAs as they strive to improve the efficiency and effectiveness of their operations through sound compliance risk management approaches.

We provide two methodological approaches. The first one is similar in spirit to traditional sample size determination strategies that seek to achieve a given precision for an estimator. However, this goal is too strong in the present context: from the RA’s perspective it may be advantageous to give up some precision in the first stage in order to increase revenue from the second stage. This results in a non-standard sample size determination problem, which is highly non-linear and does not have a straightforward analytical solution. Our second methodological approach tackles this problem with an innovative numerical algorithm based on bootstrap.

The structure of this paper is as follows. Section 2 sets up the background within which the analysis is conducted. Section 3 presents our first methodological approach whereas the second one is discussed in Section 4. Section 5 presents a numerical illustration and Section 6 provides some concluding remarks.

2. Preliminaries

2.1. Notation
Indices $i$ and $j$ are used for sampling units and indices 1 and 2 are used for the two stages of the sampling scheme. When information from previous year(s) audits is used, it is indexed by old. Elements of matrices are presented within square brackets and $[A]_{ij}$ denotes the ($i,j$) element of the matrix $A$. $\Phi$ denotes the cdf of the standard normal distribution, $F_{n_1,n_2}$ denotes the cdf of an F distribution with $a$ and $b$ degrees of freedom, $\chi^2_c$ denotes the chi-squared distribution with $c$ degrees of freedom and $Ga(a,b)$ denoted the Gamma distribution with mean $a/b$. The indicator function is denoted as $I\{\cdot\}$ and understood as $I\{A\} = 1$ if $A$ is true and $I\{A\} = 0$ otherwise. All vectors are understood as column vectors.

2.2. The problem
The RA has resources to select and audit a subset of taxpayer (henceforth ‘units’) of size $n$ from a population of size $N$. The cost of an audit is the same for all units. The tax gap (the difference between tax voluntarily paid and tax actually owed) of the $i$-th unit is denoted by $y_i$. The RA, as it is typically the case, has access to additional information for each unit, summarised in a vector of covariates $x_i \in \mathbb{R}^{p+1}$, which is correlated to $y_i$. More precisely, we assume that the $y_i$’s are independent observations of random variables $Y_i$ with expectation and variance, conditional on the vector of covariates $x_i$, given via a linear model

$$
\mu_i := \mathbb{E}(Y_i|x_i) = x_i^T \beta, \quad \text{Var}(Y_i|x_i) = \sigma^2
$$

where $\beta$ is a vector of unknown coefficients. We call $\mu_i$ the ‘expected tax gap’. Our methodology can be easily extended to cases in which $Y_i$ in (1) is replaced by log $Y_i$. Furthermore,
the assumption that \( Y_i \) are mutually independent is also a simplifying assumption as it would only hold in a finite population if the sample was drawn with replacement. Notice that the methodology developed in Section 4 can be readily modified to accommodate other sampling designs.

The RA seeks an approach to utilize this relation in order to select the subset of \( n \) units in an optimal way. It is common practice, as touched upon in the introductory section, to perform this task in two stages. In a first stage a random sample of \( n_1 \) units is drawn independently from the population. Each one of them is then audited and its tax gap is determined. This sample is then used to estimate \( \beta \) in (1) which yields estimates \( \hat{\beta}, \hat{\sigma}^2 \) and \( \hat{\mu}_i := x_i^T \hat{\beta} \). In a second stage then the RA selects and audits \( n_2 = n - n_1 \) units, those with a maximal estimated tax gap \( \hat{\mu}_i \).

The objective of the RA is to detect as much non-declared income as possible and therefore minimize the tax gap. Given the information obtained from the first stage, the detected tax gap per unit is expected to be higher for the units selected in the second stage, but it is not optimal for the RA to allocate all its resources in this second stage because the second stage sample selection relies on the accuracy of the estimate of the first stage: \( n_1 \) should be sufficiently high so that \( \hat{\beta} \) provides adequate information to select the second stage units. Section 3 provides an intuitive treatment of the problem together with a resulting sampling scheme that provides protection against a worst case scenario by keeping the probability of such an event happening below a certain threshold. In Section 4 we provide an alternative approach that proposes a sample size \( n_1 \) such that the expected tax revenues are maximised.

### 3. Controlling for the worst case scenario

The issue is to determine the sample size \( n_1 \) that guarantees, in some probabilistic sense, an accuracy of the OLS estimate \( \hat{\beta} \) of \( \beta \) in equation (1) which is sufficient for the purposes of the selection in the second stage. Assume that accuracy is specified by considering two sample units \( i \) and \( j \) such that \( \mu_i > \mu_j \) and requiring that the estimated expected tax gaps, which would form the basis for the risk driven auditing in the second stage, should respect this ordering so that \( \hat{\mu}_i > \hat{\mu}_j \). Given, however, that \( \hat{\mu}_i, \hat{\mu}_j \) are subject to random fluctuations it is only possible to guarantee this ‘ordering-respecting-property’, uniformly over all pairs of units, with sufficient high probability. Let this probability be at least \( 1 - \delta \), where \( \delta > 0 \) is a small tolerance-threshold specified by the RA. Even with this restriction two units with identical (or very close to identical) expected tax gap values might not be distinguishable from one another. This implies that further restrictions are necessary which will be discussed and introduced further below. In what follows in this section it will be further assumed that \( \hat{\beta} \) is the OLS estimator and that its distribution is assumed to be normal.

#### 3.1. The case of one continuous covariate

To intuitively introduce the problem consider first the case of only one continuous covariate so \( \mu_i := E(Y_i) = \beta_0 + \beta_1 x_i \) and \( \hat{\mu}_i := \hat{\beta}_0 + \hat{\beta}_1 x_i \). Without loss of generality assume further that \( \hat{\beta}_1 > 0 \) and that \( x_i - x_j > 0 \), which implies \( \mu_i - \mu_j > 0 \). This then implies that the
Sample size determination

A sample size determination is crucial in ensuring that the results of a study are meaningful and reliable. The 'ordering-respecting-property' becomes

$$ P(\hat{\mu}_i - \hat{\mu}_j < 0) < \delta $$

which after some simple manipulation it is shown to be

$$ P[\hat{\mu}_i - \hat{\mu}_j < 0] = P[\hat{\beta}_1 (x_i - x_j) < 0] = P[\hat{\beta}_1 < 0] $$

$$ = P\left((\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1) < -\beta_1/se(\hat{\beta}_1)\right) $$

$$ \approx \Phi\left(-\beta_1/se(\hat{\beta}_1)\right) $$

(3)

where

$$ se(\hat{\beta}_1) = \sqrt{\sigma^2[\Sigma^{-1}]_{2,2}/n_1} , \Sigma := n_1^{-1} \sum_{i=1}^{n_1} x_i x_i^T ; x_i^T = [1 \; x_i]. $$

The 'ordering-respecting-property' in (2) now reads

$$ P[\hat{\mu}_i - \hat{\mu}_j < 0] \approx \Phi\left(-\beta_1 \sqrt{n_1}/\sqrt{\hat{\Sigma}_{old}}\right) < \delta $$

and one may choose $n_1$ sufficiently large for this to hold. Notice that beyond $n_1$ this probability also depends on the true value of $\beta_1$, since as the true absolute impact of the covariate increases so does the probability that it will be estimated with the correct sign.

One might specify a lower bound for the size of impact, denoted by $B$, for the magnitude of the covariate coefficient, $\beta_1 \geq B$, such that a covariate of an impact at least equal to $B$ to be imperatively taken into account in the predicting model in (1). If estimates $\hat{\Sigma}_{old}$ and $\hat{\Sigma}_{old}$ are available from past audits, the sample size $n_1$ should then be chosen sufficiently large for

$$ \Phi\left(-B \sqrt{n_1}/\sqrt{\hat{\Sigma}_{old}}\right) < \delta $$

to hold.

3.2. The case of an arbitrary number of covariates

In turns out that in the general case of $p$ covariates, and in order to obtain a bound for $P[\hat{\mu}_i - \hat{\mu}_j < 0]$ uniformly over all possible pairs of units, the appropriate way to bound $\mu_i - \mu_j$ is to use, instead of $\beta_1 \geq B$, the restriction

$$ \mu_i - \mu_j \geq C \|x_i - x_j\|_{\Sigma} $$

(4)

for some fixed value of $C$ where the distance $\|x_i - x_j\|_{\Sigma}$ is defined in terms of the geometry which orthonormalizes the covariates given by

$$ \|x_i - x_j\|_{\Sigma}^2 := (x_i - x_j)^T \Sigma^{-1} (x_i - x_j). $$

(5)

To see more clearly how (4) relates to (3), observe that under (4) in the case of one continuous covariate we obtain

$$ P[\hat{\mu}_i - \hat{\mu}_j < 0] = P[\hat{\beta}_1 (x_i - x_j) < 0] = P[\hat{\beta}_1 - \beta_1 (x_i - x_j) < -\beta_1 (x_i - x_j)] $$

$$ = P\left((\hat{\beta}_1 - \beta_1)/se(\hat{\beta}_1) < -\beta_1 (x_i - x_j)/se\right) $$

$$ \leq \Phi\left(-C \|x_i - x_j\|_{\Sigma}/se\right) = \Phi\left(-C \sqrt{n_1}/\sigma\right) $$
where
\[ se := \sqrt{\text{var} \left( (\hat{\beta}_1 - \beta_1) (x_i - x_j) \right)} = \sqrt{\sigma^2 \Sigma^{-1}_{2,2} |x_i - x_j|^2 / n_1} \]
and the last inequality holds because
\[ \|x_i - x_j\|_\Sigma^2 = \left[ 0 (x_i - x_j) \right] \Sigma^{-1} \left[ 0 (x_i - x_j) \right]^T = \left[ \Sigma^{-1}_{2,2} \right] |x_i - x_j|^2 = se^2 / \sigma^2. \]

Thus, the term \( \|x_i - x_j\|_\Sigma \) in the bound in (4) serves as a factor which standardizes the quantity of interest. In the general case of \( p \) covariates the following Proposition holds and gives us a way to specify the sample size \( n_1 \) (the proof of this is relegated to the Appendix). This We thus arrive at:

**Proposition 1.** Uniformly, over all pairs satisfying (4), the probability that the ‘ordering-respecting-property’ is violated is less than

\[ \frac{1}{2} \left[ 1 - F_{p,n_1-p} \left( \frac{C^2 n_1}{p \hat{\sigma}^2} \right) \right]. \]

This latter quantity can be made smaller than any \( \delta > 0 \) for sufficiently large \( n_1 \). Note that the choice of \( n_1 \) depends only on \( \delta, C, p, \hat{\sigma}^2 \) but not on the covariate values of the specific pair. The RA would need to specify \( \delta \) and \( C \) and choose \( n_1 \) so that

\[ F_{p,n_1-p} \left( \frac{C^2 n_1}{p \hat{\sigma}^2_{old}} \right) > 1 - 2\delta. \]

The above proposition guarantees that the ‘ordering-respecting-property’ would hold with probability at least \( 1 - \delta \), uniformly over all pairs of units for which the difference in \( \mu_i - \mu_j \) exceeds \( C \) when the orthonormalized covariates differ in norm by a unity: \( \|x_i - x_j\|_\Sigma = 1 \).

In order to specify \( C \) the RA can proceed as follows: First observe that when two units indexed by \( i \) and \( j \) differ only in the value of the \( k \)-th covariate, then (4) reads

\[ \mu_i - \mu_j \geq C \|x_i - x_j\|_\Sigma \quad \iff \quad \beta_k (x_{ik} - x_{jk}) \geq C \sqrt{(\Sigma^{-1})_{kk}} |x_{ik} - x_{jk}| \quad \iff \quad \hat{\beta}_k / \sqrt{(\Sigma^{-1})_{kk}} \geq C. \]

Since the t-statistic for an estimated coefficient \( \hat{\beta}_k \) is given by

\[ \left[ \hat{\beta}_k / \sqrt{(\Sigma^{-1})_{kk}} \right] \left[ \sqrt{n_1} / \hat{\sigma} \right] \]

\( C \) may be interpreted as a lower bound of the (theoretical) t-statistic multiplied by \( \sigma / \sqrt{n_1} \). This induces the following suggestion to specify \( C \): consider the t-statistics of all covariates in the random sample of the previous period, multiplied by \( \hat{\sigma}_{old} / \sqrt{n_{1,old}} \). Take \( C_{old} \) to be the minimum of these values across all covariates which should be included in the model of the current period. Then, choose \( n_1 \) large enough so that

\[ F_{p,n_1-p} \left( \frac{C_{old}^2 n_1}{(p \hat{\sigma}_{old}^2)} \right) > 1 - 2\delta. \]
3.3. Illustration

The tolerated error probability $\delta$ is a parameter which has to be chosen by the RA and clearly if it is chosen to be too small the sample size required to guarantee a smaller error will be very high. To illustrate this, we performed the following exercise. By assuming a standardised design and unit error variance, we calculated the sample sizes required to guarantee the order respecting property for two sampling units differing by unity in only one covariate the coefficient of which equals to $C$. Table 1 shows the corresponding sample sizes so that the estimated ordering is the correct one with error probability less than $\delta$, provided model coefficients of a model with p covariates have been estimated on the basis of a sample with the given sample size.

Note the sample size required in the first stage strongly depends on the number of covariates in the model. This is expected since a higher sample size is necessary to compensate the increased variability of $\hat{\beta}$ induced by large $p$.

4. Maximizing the expected tax revenue

We now present an alternative approach that aims at directly maximising the expected tax revenue and does not necessitate a choice of $\delta$. Let us fix for a moment the values of $n_1$ and $n_2$. Moreover, for some $\beta$ let $F_\beta (m)$ denote the cdf of $\mu_i(\beta) = x^T \beta$ given as

$$F_\beta (m) := N^{-1} \sum_{i=1}^N I\{\mu_i(\beta) \leq m\}$$

and let $F(m) := F_\beta(m)$ for the true value of the parameter $\beta$. The sum of the tax gaps detected in the first stage is on average $n_1$ times the mean expected tax gap

$$R_1 (n_1) = n_1 \int m \, dF(m).$$

Let us assume at the moment that the true parameter $\beta$ is known. In the second stage all units with a tax gap $\mu_i := \mu_i(\beta)$ exceeding the $(1 - n_2/N)$th quantile $F_\beta^{-1} (1 - n_2/N)$ are audited and the tax revenue equals the expectation of the right tail of the tax gaps. Thus, denoting by

$$J_2 (\beta) := \{ i \leq N | \mu_i(\beta) \geq F_\beta^{-1} (1 - n_2/N) \}$$

<table>
<thead>
<tr>
<th>$C$</th>
<th>$p=4$</th>
<th>$p=50$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1172</td>
<td>7274</td>
</tr>
<tr>
<td>0.2</td>
<td>297</td>
<td>1828</td>
</tr>
<tr>
<td>0.3</td>
<td>135</td>
<td>820</td>
</tr>
<tr>
<td>0.4</td>
<td>75</td>
<td>467</td>
</tr>
<tr>
<td>0.5</td>
<td>52</td>
<td>304</td>
</tr>
</tbody>
</table>
the set of indices with the \( n_2 \) largest values of \( \mu_i \), we obtain that the revenue from the second stage equals \( n_2 \) times the expectation of the right tail of the tax gaps:

\[
R_2(\beta, n_2) = n_2 \int_{\{m \geq F_\beta^{-1}(1-n_2/N)\}} m \, dF(m) = \sum_{i \in J_2(\beta)} \mu_i.
\]

If \( \beta \) is not known, but estimated in the first stage by \( \hat{\beta} \), then the expected revenue from the first stage \( R_1(n_1) \), remains unaffected, but the revenue from the second stage is a random variable \( R_2(\hat{\beta}, n_2) \) and depends on the distribution of \( \hat{\beta} \).

We propose to choose \( n_1 \) so as to maximize the expected revenue \( R(\hat{\beta}, n_2) := R_1(n_1) + R_2(\hat{\beta}, n_2) \):

\[
E[R(\hat{\beta}, n_2)] = n_1 \int m \, dF(m) + \int f_{\hat{\beta}, n_1}(b) R_2(b, n_2) \, db,
\]

where \( f_{\hat{\beta}, n_1}(b) \) is the density of \( \hat{\beta} \), which depends on the value of the true parameter \( \beta \) and \( n_1 \).

For example, in the case of only one continuous covariate, \( \mu_i := \beta_0 + \beta_1 x_i \) and under \( \beta_1 > 0 \) we would get, denoting by \( \mu(r) \) the \( r \)-th order statistic of \( \{\mu_i\}_{i=1,...,N}^r \):

\[
R_2(\beta, n_2) = \sum_{i \in J_2(\beta)} \mu_i = \sum_{r=N-n_2}^N \mu(r)
\]

\[
E[R(\hat{\beta}, n_2)] = \sum_{i \in J_2(\beta)} \mu_i = \mathbb{P}_{\hat{\beta}, n_1}(\hat{\beta}_1 > 0) \sum_{r=N-n_2}^N \mu(r) + \mathbb{P}_{\hat{\beta}, n_1}(\hat{\beta}_1 < 0) \sum_{r=1}^{n_2} \mu(r).
\]

The reason for the last equality is that, when \( \hat{\beta}_1 > 0 \) the \( n_2 \) units with largest \( \mu_i(\hat{\beta}) \) are those with largest \( \mu_i \), while when \( \hat{\beta}_1 < 0 \) the \( n_2 \) units with largest \( \mu_i(\hat{\beta}) \) are those with lowest \( \mu_i \). Thus, with probability \( \mathbb{P}_{\hat{\beta}, n_1}(\hat{\beta}_1 < 0) \) the RA selects in the second stage the units with lowest tax gap, instead of those with highest tax gap.

When the number of parameters is greater than one the expressions involved are getting even more complicated and so is an analytical solution to the maximization of (6).

We therefore propose the use of a numerical solution based on bootstrap (Efron and Tibshirani, 1994). More precisely the idea is to approximate \( f_{\hat{\beta}, n_1}(b) \) in (6) by a bootstrap analogue, based on past sample(s) of the first stage. The details are given in Algorithm 1.

Note that in Algorithm 1 we do not make any specific assumptions on the estimate \( \hat{\beta} \), beyond those necessary for bootstrap to work, as long as in line 8 we use the same estimation method as the one which will be used when we obtain the new sample of the current tax year.

A further plausible assumption that accelerates Algorithm 1 is that \( \hat{\beta} \sim N(\beta, n_1^{-1} \Sigma) \). We can then generate \( \hat{\beta}_{k, b}^* \) from \( N(\hat{\beta}_{old, n_1, k}^{-1} \Sigma_{old}) \) in line 8 and ignore lines 5-8. Moreover, one may further accelerate the procedure to an important extent by using a simple sequential Monte Carlo algorithm; see, for example, Doucet and Johansen (2009). Algorithm 2 provides the details.
Algorithm 1 Calculation of optimum sample size for tax auditing

1: Compute the residuals from the first stage of past audit(s):
\[ \hat{r}_i := y_{\text{old},i} - x_{\text{old},i}^T \hat{\beta}_{\text{old}}, i = 1, \ldots, n_{\text{1,old}} \]

2: Compute the empirical cdf of \( \hat{r}_i \):
\[ \hat{G}_{n_{\text{1,old}}} (s) := n_{\text{1,old}}^{-1} \sum_{i=1}^{n_{\text{1,old}}} 1\{ \hat{r}_i \leq s \} \]

3: for all \( k \) of a grid of values \( n_{1,k}, k = 1, \ldots, K \) do:
4:   for all \( b = 1, \ldots, B \) do:
5:     Draw a sample \( x_{b,i}^* \), \( i = 1, \ldots, n_{1,k} \) with replacement from the population \( \hat{G}_{n_{\text{1,old}}} \).
6:     Draw a sample \( \hat{r}_{b,i}^* \), \( i = 1, \ldots, n_{1,k} \) with replacement from \( \hat{G}_{n_{\text{1,old}}} \).
7:     Compute \( \hat{y}_{b,i}^* := x_{b,i}^* \hat{\beta}_{\text{old}} + \hat{r}_{b,i}^*, i = 1, \ldots, n_{1,k} \).
8:     From \( \hat{y}_{b,i}^* \), \( x_{b,i}^* \), \( i = 1, \ldots, n_{1,k} \) compute \( \hat{\beta}_{k,b}^* \).
9:     for all \( j = 1, \ldots, N \) do:
10:        Compute \( \mu_{k,j}^* := x_j^T \hat{\beta}_{k,b}^* \).
11:     end for
12:     Determine \( J_{2,k} \left( \hat{\beta}_{k,b}^* \right) \), the index set of the \( n_{2,k} = n - n_{1,k} \) largest values of \( \mu_{b,j}^* \).
13:     Compute \( R_{k,b}^* := n_{1,k} \left( n_{\text{1,old}}^{-1} \sum_{i=1}^{n_{\text{1,old}}} x_{i,\text{old}} \right)^T \hat{\beta}_{\text{old}} + \sum_{j \in J_{2,k} \left( \hat{\beta}_{k,b}^* \right)} \mu_j \left( \hat{\beta}_{\text{old}} \right) \).
14: end for
15: Compute \( \hat{E}R_k := B^{-1} \sum_{b=1}^{B} R_{k,b}^* \).
16: end for
17: Set \( n_{1,k'} \) as the optimal sample size where \( k' = \arg \min_{k=1,\ldots,K} (\hat{E}R_k) \).
Algorithm 2 Calculation of optimum sample size for tax auditing using sequential Monte Carlo

1: Let $n_{1,1}$ be the smallest from a grid of values $n_{1,k}$, $k = 1, \ldots, K$
2: For $n_{1,1}$ generate $\hat{\beta}_{1,1}^*, \ldots, \hat{\beta}_{1,B}^*$ from $N(\hat{\beta}_{old}, n_{1,1}^{\frac{1}{2}} \Sigma_{old})$
3: for all $b = 1, \ldots, B$ do
4:   Compute $\mu_{b,j}^* := x^T_j \hat{\beta}_{1,b}^*$ for $j = 1, \ldots, N$
5: for all $k = 1, \ldots, K$ do
6:   Determine $J_{2,k}^* \left( \hat{\beta}_{1,b}^* \right)$, the index set of the $n_{2,k} = n - n_{1,k}$ largest values of $\mu_{b,j}^*$.
7:   Compute $R_{k,b}^* := n_{1,k} \left( n_{1,old}^{\frac{1}{2}} \sum_{i=1}^{n_{1,old}} x_{i,old} \right)^T \hat{\beta}_{old} + \sum_{j \in J_{2,k}^* (\hat{\beta}_{1,b}^*)} \mu_j (\hat{\beta}_{old})$
8: $w_{k,b} := \left( \frac{n_{1,k}}{n_{1,1}} \right)^{\frac{p+1}{2}} \left( \exp \left[ -\frac{1}{2} (\hat{\beta}_{1,b}^* - \hat{\beta}_{old})^T \Sigma_{old}^{-1} (\hat{\beta}_{1,b}^* - \hat{\beta}_{old}) \right] \right)^{n_{1,k} - n_{1,1}}$
9: end for
10: end for
11: for all $k = 1, \ldots, K$ do
12:   $\hat{ER}_k := B^{-1} \sum_{b=1}^{B} w_{k,b} R_{k,b}^* / \sum_{c=1}^{B} w_{k,c}$
13: end for
14: Set $n_{1,k'}$ as the optimal sample size where $k' = \arg\min_k (\hat{ER}_k)$
5. Simulation example

In this section we conduct a simulation experiment in order to check the performance of the bootstrap method proposed in section 3.1. More precisely, we investigate (i) to which extend the estimation of the true expected revenue $E_{\hat{\beta}} \left[ R \left( \hat{\beta}, n_2 \right) \right]$ by the estimated expected revenue $E_{\hat{\beta},old} \left[ R \left( \hat{\beta}, n_2 \right) \right]$ works well, (ii) how the estimated optimal $\hat{n}_1$ which minimises $E_{\hat{\beta},old} \left[ R \left( \hat{\beta}, n - n_1 \right) \right]$ compares to the true optimal $\bar{n}_1$ which maximises $E_{\beta} \left[ R \left( \hat{\beta}, n - n_1 \right) \right]$ obtained by the RA when using the estimated optimal $\hat{n}_1$ compares to the maximal true expected revenue $E_{\beta} \left[ R \left( \hat{\beta}, n - \bar{n}_1 \right) \right]$ which would have hypothetically been obtained by the RA when the (unknown) truly optimal $\bar{n}_1$ that maximises the true expected revenue had been used. We will further explore how they both compare to the simple random sampling (SRS) expected revenue achieved by setting $n_1 = n$, and thus selecting all units for audit by simple random sampling. Finally, we will also investigate how they compare to the Oracle expected revenue achieved by setting $n_2 = n$ and assuming that the true $\beta$ is known and does not need to be estimated.

We first generated independent vectors of covariates $x_i$ of dimension $p$ for $i = 1, 2, \ldots, N$ with each element of $x_i$ drawn from a $\chi^2_4$ distribution re-centered and re-scaled to have expectation equal to zero and variance equal to 1. Then, we fixed parameter values $\beta_1, \beta_2, \ldots, \beta_p$ and a true population of tax sample units $Y_i$ was generated by adding to $\mu_i := x_i^T \beta$ a e-scaled and re-centred) draw from a $Ga(1/2, 20)$ distribution. This sample has been considered to be the ‘true’ population so it has been kept fixed throughout the simulation exercise. This implies that the simulation variability emanated from the random fluctuations of the estimators $\hat{\beta}_{old}$.

Our results are based on two simulated scenarios, both based on sample size $N = 50,000$. Both scenarios assume that the true population model has four covariates given by $\beta_2 = 1, \beta_3 = 2, \beta_4 = 0.5$ and $\beta_5 = 0.3$, but in scenario A we estimated a model with $p = 4$ whereas in scenario B we estimated a model with $p = 50$. Note that the large number of covariates in the estimation of the model in scenario B makes the problem particularly difficult as it inflates the variance of the tax gap estimator. We further assumed that the budget constraints restricted the total number of audits to $n = 2000$ and that the stage 1 sample of the previous period had a sample size of the true optimal $n_1$, that is $n_{1,old} = 275$ and $n_{1,old} = 675$ under the two scenarios respectively (these numbers are obtained by a preliminary Monte Carlo simulation since the true $\beta$ is known). The grid for optimizing across $n_1$ which we used varied from 50 to 600 with a step of 25 for scenario A and from 200 to 900 with a step of 25 for scenario B. The number of bootstrap replications was set to $B = 400$ and the number of simulations was set to $M = 1000$.

Figure 1 depicts a typical realisation of the estimated expected revenue along with the true expected revenue as functions of $n_1$. The discrepancy between the two is due to the difference between and $\hat{\beta}_{old}$ and $\beta$. In both scenarios, it is interesting to note the similarity of the two shapes which implies that the estimation of the optimal $n_1$ still makes sense.

Indeed in Figures 2 and 3 it is shown that the estimated optimal $\hat{n}_1$ has a negative bias as an estimator of $n_1$. This bias increases considerably when the number of covariates is large, resulting to an inflation of variance in the estimation of $\beta$. Similarly, for the revenue,
note that the maximal true expected revenue indicated by the dashed vertical line, is by definition an upper bound for the achieved true expected revenue, the distribution of which is indicated by the black line. The densities were estimated by kernel estimators, the bandwidth of which were chosen empirically so as to smooth out the effects of evaluating the revenue at a discrete grid of possible values.

On the other hand, Figure 4 reveals that the true and maximal revenues seem relatively close when compared with the two extremes, namely the SRS expected revenue and the oracle revenue. Note that it is natural that the SRS expected revenue will be lower than the revenue achieved by the risk based method, as it aims at detecting the average tax gap and not the upper tail of the tax gap distribution, as the risk based method does. Also note that it is also natural that the oracle revenue will be higher than the revenue generated by the method which tries to estimate the individual’s expected tax gap: the oracle “knows” exactly which units have the highest expected tax gap, so it does not need to devote resources in spotting them \((n_2 = n)\), and moreover, has still a zero error probability in detecting them.

6. Concluding remarks

Revenue authorities in a number of countries have given considerable attention to the development of risk management practices. This paper has contributed to this issue by investigating optimal random sample selection in audits, providing statistical methodologies that view the problem with two perspectives: controlling for the worst case scenario
and maximizing the expected tax revenue. We believe that the resulting algorithms will offer a valuable probabilistic risk management tool to RAs.

There is a plethora of issues that may require extension of our methodology. An RA may have empirical evidence that the cost of an audit is related to the probability of detecting under-reporting income and this information may be included in our optimisation
exercise. This will require relaxing the assumption of equal cost per sampling unit and optimising across costs per audit.

The assumption that random auditing identifies the tax gap amongst the class of taxpayers who are audited is, of course, quite strong. In reality, auditors do make errors in their assessment. Indeed, existing estimates show that there is a considerable heterogeneity in detection rates across examiners for some income items; see, for example, Erard et al. (2010).

A deep, and very interesting, statistical problem arises when one combines the issue of model choice and maximisation of expected tax revenue. Throughout our paper we have assumed that the ‘best’ model is fixed and previously estimated by the RA. However, it is clear that a proper model choice procedure may not be based on standard statistical techniques but, rather, be based on the ultimate purpose of the RA which is maximisation of the tax revenue.

Last but not least, our methodology should be incorporated in a larger audit framework in which audits take place in a stratified sampling fashion with a series of practical constraints. We have provided only the first step of such a large, probabilistically sound approach to deal with this problem.
Acknowledgements

We thank Anthony Burke and the Tax Gap Team of HMRC for useful discussions on sampling for tax auditing. ESRC funding under ES/S00713X/1 and EP/N510129/1 through Alan Turing Institute are gratefully acknowledged.

References


Appendix

Proof of Proposition 1.

Here is why the proposition holds: for a pair such that $\mu_i - \mu_j > 0$ the probability that the ‘ordering-respecting-property’ is violated equals

$$P[\inf_{i,j} (\mu_i - \mu_j) < 0] = P[\inf_{i,j} \beta^T (\mathbf{x}_i - \mathbf{x}_j) < 0]$$

$$= P[\inf_{i,j} (\beta - \beta)^T (\mathbf{x}_i - \mathbf{x}_j) + \beta^T (\mathbf{x}_i - \mathbf{x}_j) < 0]$$

$$= P[\inf_{i,j} (\beta - \beta)^T (\mathbf{x}_i - \mathbf{x}_j)/\|\mathbf{x}_i - \mathbf{x}_j\|_{\Sigma} + \beta^T (\mathbf{x}_i - \mathbf{x}_j)/\|\mathbf{x}_i - \mathbf{x}_j\|_{\Sigma} < 0]$$

$$\leq P[\inf_{i,j} (\beta - \beta)^T (\mathbf{x}_i - \mathbf{x}_j)/\|\mathbf{x}_i - \mathbf{x}_j\|_{\Sigma} + C < 0]$$

$$= (1/2) P[\sup_{i,j} [(\beta - \beta)^T (\mathbf{x}_i - \mathbf{x}_j)]^2/\|\mathbf{x}_i - \mathbf{x}_j\|_{\Sigma}^2 > C^2]$$
By Scheffe’s multiple comparison method (Scheffe, 1999) this probability equals

\[
(1/2) \mathbb{P} \left[ (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta) > C^2 \right]
\]

\[
= (1/2) \mathbb{P} \left[ \frac{n_1 (\hat{\beta} - \beta)^T \Sigma (\hat{\beta} - \beta)}{p\hat{\sigma}^2} > \frac{n_1 C^2}{p\hat{\sigma}^2} \right]
\]

\[
= (1/2) \left[ 1 - F_{p,n_1-p} \left( \frac{C^2 n_1}{p\hat{\sigma}^2} \right) \right]
\]