Offshore Compliance: Using Regression and Risk Modeling To Select Cases from Large Datasets

Damian Pritchard and Nadeer Khan, Inland Revenue, United Kingdom

---

There has been a growing concern within the UK Inland Revenue that individuals are using offshore tax havens mainly as a tax avoidance measure and that large sums of tax revenue are being lost as a result. This concern is reenforced by independent studies showing vast amounts of deposits held by individuals in offshore jurisdictions. Tax Justice Network, a group of accountants and economists, recently estimated that wealth held in tax havens is costing governments around the world at least $255 billion every year in lost tax revenue[1]. Figures published by Datamonitor,[2] a business analysis company, indicate that approximately $150 billion is held by UK residents in offshore deposits, with just over half in the Channel Islands and the Isle of Man.

At the same time, the UK Inland Revenue has obtained information on hundreds of thousands of UK individuals with offshore bank accounts or trusts. Checking these against our internal systems, we can identify which individuals have declared these accounts for tax purposes and which have not.

There is nothing illegal about placing capital offshore, provided tax has been paid on the source, but UK domiciled individuals should also declare the income generated by these assets for tax consideration. The UK operates a tax system whereby the majority of employees have their incomes taxed at source by the employer (pay as you earn/ PAYE), but, for people with more complex affairs such as high earners and the self-employed, the system is based on self-assessment (SA). It is a requirement of the UK tax system that UK residents with income that has been generated offshore, such as interest on savings or from trusts, report the existence of such on a self-assessment form even if they are normally PAYE tax payers. In many cases, however, there is evidence that the tax liability is not being declared, either on the savings interest or on the source of the capital.

The problem we are presented with is how to best manage the information contained in these large sets of data. We have to consider both the short-term goal of collecting yield and the long-term strategy of enabling taxpayers to comply with their tax obligations. We also want to form a picture of taxpayer behavior and the main characteristics for noncompliance. This al-

lows us to carry out rigorous behavioral analysis of all UK taxpayers with offshore accounts to help us understand their motivations for noncompliance, resulting in a significant reduction in the tax gap.

Data and Sources

The model was developed on data obtained from the Special Compliance Office (SCO) management information systems (MIS). This records all investigations made by SCO and the reason for the investigation. We were able to extract information on individuals who were previously investigated for offshore evasion, though we do not see the exact details of the cases. The data contain information on the taxpayer's name, a Unique Taxpayers Reference (UTR), address, and the amount of yield obtained from the investigation.

These data could then be matched to the self-assessment (SA) tax returns using the taxpayer's Unique Reference Number. This allowed us to obtain information on each of the taxpayers' professions, the nature of their businesses, their incomes, and the sources of those incomes.

We also matched these data to the CACI[3] lifestyle database. This lifestyle database categorizes all UK areas into different ACORN[3] classes reflecting the type of people living in them—for a full description, see Annex 3. The ACORN category is a classification system devised by a commercial marketing company, "CACI Limited,"[3] and it is a means of classifying areas according to various census characteristics. Businesses use this information to improve their understanding of customers and determine where to locate operations. We used it as an approximation of housing 'wealth.'

Developing the model

As our first step, we attempted to develop a risk profile of UK taxpayers who hold offshore accounts or trusts. This involves behavioral analysis of individuals who have previously been investigated for tax avoidance and ascertaining the factors that are linked with noncompliance. The analysis will, in the long run, enable us to create a useful profile of existing offshore account holders and identify high-risk cases.

Through discussions with the 'Special Compliance Office' (SCO) tax inspectors, we identified some factors that they usually perceive to be indicators of high risk, plus some others that we thought would contribute to taxpayers' noncompliance behavior.

The factors that the tax investigators identified were whether the individual is self-employed, a partner in a business, or a Director of a company, as well as whether the industry the individual is involved in has a high element of cash transactions. These factors while not unusual or fraud-related by them-

selves do indicate an increased opportunity to avoid tax.

In addition to the factors suggested by the tax inspectors, we also suggested that age, income, and whether an individual had received offshore dividends or savings income be considered. Further, we also included the individuals' ACORN category into the model.

A list of all variables used and an explanation of their meanings are given in Annex 1. Significant factors included in our final analysis are:

Age: People tend to accumulate wealth over time. This makes age an influential factor in a taxpayer's ability to set up offshore funds. The analysis showed that individuals in the 60-to-70 age range are likely to produce a higher yield from an offshore tax avoidance investigation. Our explanation would be that this links closely with retirement. If an individual has been avoiding tax on the offshore account, then, at this age, he or she will have had the most productive years and economic activity slows down.

Taxpayers' profession/type of business: Being self-employed or a partner in a business represents an increased opportunity to evade tax, as are certain types of businesses where cash transactions are normal.

ACORN 1 through 5 plus 16: Taxpayers from more affluent areas are more likely to have an offshore account than people who are from less well-off areas. We identified some areas based on this classification and attempted to model the behavior of wealthy individuals living there. Our analysis showed that a disproportionately high number of people with offshore trusts live in ACORN categories 1 through 5 and 16. From our data of individuals who were investigated for offshore evasion using offshore trusts/accounts, 51 percent lived in these areas, whereas the national average accounts for only 17 percent. These areas and the people living in them are defined as "Wealthy achievers." See Annex 3.

Annual income: Annual income is an important element of taxpayers' total wealth. Income comes into the regression in two different forms, the absolute value of the income which was categorized into four income groups based upon UK income tax thresholds and the various sources of income (i.e., dividend/ savings/ other earnings).

Logistic Regression

This study intends to test empirically how taxpayers' noncompliance behaviors are associated with their personal attributes. The logistic regression model

seeks to establish a relationship between actual yield on taxpayers' failure to comply with the regulations and a range of quantitative and qualitative variables attributable to the taxpayers. We used data from the Special Compliance Office management information systems and matched them to the UK self-assessment database and a commercially available lifestyle database to create a behavioral profile. The regression has been able to identify some factors that explain key characteristics of a group of taxpayers who have faltered on compliance. The model has been further extended to assess the extent of risks associated with individuals with similar characteristics and to predict yields.

Logistic regression models the relationship between some independent predictor variables and the categorical response variable. For a simple case, where a response variable is dichotomous, i.e., either an event occurs (Y=1) or the event does not occur (Y=0), ordinary linear regression would not be appropriate to fit the data for numerous reasons. First, the assumption that the residual errors follow normal distribution fails when the response variable is categorical. Second, it is really the probability that each individual in the population responds with 0 or 1 that we are interested in modelling.

Logistic regression has many analogies to Ordinary Least Squares (OLS) regression: logit coefficients correspond to b coefficients in the logistic regression equation, the standardized logit coefficients correspond to beta weights, and a pseudo R^2 statistic is available to summarize the strength of the relationship. Unlike OLS regression, however, logistic regression does not assume linearity of relationship between the independent variables and the dependent, does not require normally distributed variables, does not assume homoschedasticity, and in general has less stringent requirements. The success of the logistic regression can be assessed by looking at the classification table, showing correct and incorrect classifications of the dichotomous or ordinal dependent. However, there is no widely-accepted direct analogy to OLS regression's R^2. This is because an R^2 measure seeks to make a statement about the "percent of variance explained," but the variance of a dichotomous or categorical dependent variable depends on the frequency distribution of that variable.

The response, Y, of a subject can take one of two possible values, denoted by 0 and 1. Let, $X = (x_1, x_2, ..., x_k)'$ be the vector of explanatory variables. The logistic model used to explain the effects of the explanatory variables on the binary response is:

$$\log it\{\Pr(Y=1) = \ln\{\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\} = b_0 + x'b$$

where, b_0 is the intercept and b_1 is the vector of slope parameters.

This equation can be extended beyond the case of dichotomous variables to ordered categories. Let the response Y, of a subject take one of m ordinal values, denoted by 1,2,…..,m. The fitted cumulative logit model (also known as proportional odds model) would be:

$$\log it\{\Pr(Y \le r \mid X)\} = \alpha_r + X'\beta \qquad 1 \le \mathrm{r} \le \mathrm{m}$$

where $\alpha_1, \alpha_2, \ldots \alpha_{\mathrm{m\text{-}1}}$ are $(m-1)$ intercept parameters.

The logistic distribution (the relationship between logit of P and P itself) is sigmoidal (S-shaped) in shape. The estimated coefficients from the logistic regression must be interpreted with care. Instead of the slope coefficients being the rate of change in Y (the response variable) as X changes (as in the OLS regression model), the slope coefficient here is interpreted as the rate of change in the "log odds" as X changes. A more intuitive interpretation of logit coefficient is the "odds ratio"—the ratio of the odds of an event occurring to the odds of the event not occurring, derived by taking the exponential value of the coefficient. For example, if $\exp(B_3)=2$, then a one-unit change in X_3 would make the event twice as likely (0.67/0.33) to occur. We used the proportional odds model (POM) (McCullagh, 1980) of logistic regression for ordered categorical dependent variable.

Interpretation of logistic output

The standard score test is generally used to see if the assumptions of proportional odds model are valid for a given data set. A nonsignificant test is taken as evidence that the proportional odds model holds and that the odds ratios can be interpreted as constant across all possible cut points of the outcome. Standard SAS output of the score test for our regression model is given in Annex 2.

The score test result may raise a question to the extent to which data may be fit using the proportional odds model. However, there is some debate about the adequacy of the proportional odds test, e.g., it may be too sensitive to the sample size. Therefore, the test may be misleading. There remains a scope for further investigation into this matter. SAS produces both maximum likelihood estimates as well as odd ratio estimates for a logistic regression.

For the proportional odds model, the odds ratios for a predictor can be interpreted as the summary of the odds ratios obtained from separate binary logistic regressions using all possible cut points of the ordinal outcome (Scott et al., 1997). Whereas a binary logistic regression models a single logit, the proportional odds model models several cumulative logits. Therefore, if the

ordinal outcome has four levels (1,2,3, and 4), three logits will be modeled, one for each of the following cut points: 1 versus 2,3,4.

The maximum likelihood estimates table reports the intercept of each independent of regression coefficients. The intercept parameters quantify the shift in location between the five cumulative logits (risk groups) being modeled. It also reports the chi-squared test value indicating if an explanatory variable is significant in the regression. Considering a 10-percent significance level, we would expect the p-value to be less than .10 for a coefficient to be considered significant. Apart from the ACORN variables, all other in this regression, especially age (between 60 and 70), income (over 100,000), self-employment status, declaration of dividend in the tax return, and company directorship status appear to be quite significant. This result is somewhat expected. However, distortionary effects, such as covariance between the variables and inherent bias in reporting by taxpayers, may cause unexpected noise in our analysis. Hence, this result is to be treated with caution.

It is rather more intuitive and conventional to interpret odds ratios in a logistic regression. The odds ratios are derived taking the exponential value of the intercept of a coefficient. In the odds ratio table, the odds ratio for the age group 60 to 70 (age60270) is 2.36, i.e., a person aged between 60 and 70 is approximately 2½ times more likely to be in a risk group than someone who is not. Thus, the model enables us to understand taxpayers' behaviors and model how noncompliance behaviors can be predicted with a given set of available information.

Explanatory power

The first thing to look at in the output is the reported standard errors and z statistics, looking to be sure that no standard error was reported to be absurdly small (no z absurdly large because z=coef/se). A good indication that the model has explanatory power is to compare the log likelihood and the number of observations. In this model, the reported log likelihood value is -370.72 for 314 observations. If the category to which an observation belonged was unknown, we could use a probability of 1/6 = 0.167 for each of the 6 outcome categories. This model says that, on average, for an observation, the probability of being observed (what was observed conditional on the estimates) is exp(-370.72/314)=0.307. That number is larger than 0.167, and that is good news because it means that our model has explanatory power.

Limitations

Meaningful coding. Logistic coefficients will be difficult to interpret if not coded meaningfully. The convention for binomial logistic regression is to code

the dependent class of greatest interest as 1 and the other class as 0. For multinomial logistic regression, the class of greatest interest should be the last class. Logistic regression is predicting the log odds of being in the class of greatest interest.

Inclusion of all relevant variables in the regression model. If relevant variables are omitted, the common variance they share with included variables may be wrongly attributed to those variables, or the error term may be inflated. We have had discussions with the SCO investigators to ensure that all relevant variables have been considered.

Exclusion of all irrelevant variables. If irrelevant variables are included in the model, the common variance they share with included variables may be wrongly attributed to the irrelevant variables. The more the correlation of the irrelevant variable(s) with other independents, the greater the standard errors of the regression coefficients for these independents will be. This is something we will look into in more detail with our future developments.

Error terms are assumed to be independent (independent sampling). Violations of this assumption can have serious effects. Violations are apt to occur, for instance, in correlated samples and repeated measures designs. That is, subjects cannot provide multiple observations at different time points. The nature of the investigations and the method of collecting our data ensure that this assumption is met.

Linearity. Logistic regression does not require linear relationships between the independents and the dependent, as does Ordinary Least Squares regression, but it does assume a linear relationship between the logit of the independents and the dependent. When the assumption of linearity in the logits is violated, then logistic regression will underestimate the degree of relationship of the independents to the dependent and will lack power (generating Type II errors, thinking there is no relationship when there actually is). We were aware of this issue and took steps at the design stage to ensure that linearity holds.

Outliers. As in Ordinary Least Squares regression (OLS), outliers can affect results significantly. We analyzed the data for outliers to consider removing any.

Large samples. Also, unlike OLS regression, logistic regression uses maximum likelihood estimation (MLE) rather than ordinary least squares to derive parameters. MLE relies on large-sample asymptotic normality, which means that reliability of estimates decline when there are few cases for each ob-

served combination of independent variables. That is, in small samples, one may get high standard errors. Our sample size of 314 while not exactly large we consider sufficient to meet this assumption.

Unobserved data. An attempt to measure noncompliance and modelling taxpayers' behaviors is generally fraught with an unobserved element. Given that taxpayers do not voluntarily declare their noncompliant activities, any available information should be treated with caution. We often came across this problem of unobserved data, which limited our ability to draw optimal conclusion on behavioral aspects of noncompliant taxpayers.

Missing values. The problem of missing values and omitted variables may severely affect the power on statistical analysis. While we endeavored to minimize the number of missing values and reduce the effect of omitted variables, their existence is inevitable. These may arise from the preceding issue of unobserved data, as well as bad quality of data or data collection method. Though it may sometimes be possible to estimate or interpolate missing values under the assumption of linearity, the logistic regression model does not allow for it. Hence, unlike a simple linear regression model, missing values more often significantly influence the output of a logistic regression.

Multicollinearity. This can be a problem when high in logistic regressions because the standard errors of the coefficients will be high, rendering the interpretation of relative importance of the independent variable unreliable. Given the nature of our analysis, an extent of covariance between the explanatory variables may be expected. We were aware of this issue and made an effort to minimize the effect of multicollinearity. In some cases, there may exist some relationship between the variables, such as age and income of individuals, or people in a similar income group may possess some distinct characteristics. We closely examined the data to avoid any obvious covariance between the variables, and also excluded one subcategory of each explanatory variable type from our analysis. Given the nature and the scope of noncompliance activities, we would not expect any significant covariance existing with the dependent variable. However, it is to be appreciated that time and resource constraint limit our ability to carry out any formal parametric test on multicollinearity, and the extent of this has not been quantified.

However, multicollinearity does not change the estimates of the coefficients, only their reliability. High standard errors flag possible multicollinearity, but multicollinearity is not generally considered a problem if the variance inflation factor (VIF) $1/(1\text{-}r^2) <= 4$, a level which corresponds to doubling the

standard error of the b coefficient is considered. In this model, the VIF = 1.04 and is therefore in an acceptable range.

Additivity. Like OLS regression, logistic regression does not account for interaction effects except when interaction terms (usually products of standardized independents) are created as additional variables in the analysis.

Data quality. The quality of data available for our work has been a major concern for us and to some extent limits our ability to draw any inference with significant confidence. Our analysis involved working with a number of large data sets from different sources. While we strove to collect and process the highest quality of data, matching a number of large data sets on taxpayers' behaviors may yet raise the issue of the quality of the data analyzed.

Application

The purpose of developing this model on existing evidence from historical cases was to apply the model to new and untried data sets. The assumption has been made that all individuals in the data set are high-risk. This assumption is made because these individuals are known to have offshore trusts or accounts but no declarations of these have been made to the tax authority. Given that each individual is considered high-risk, the model attempts to predict a yield from the given characteristics. These individuals are then sorted by expected yield, with the individuals at the top considered "more risky" than those at the bottom.

It is expected that, even from the highest "risk" group 30 percent of investigations will not result in yield. However, the average yield from investigations in this group is expected to be twice that expected from a random sample. This gives us the advantage of targeting high-yield cases.

However, to test the true effectiveness of the model, we need to look at investigating cases from across the range of yield estimates.

Future Developments

This has been the first attempt, within the UK Revenue, to model offshore taxpayers' behaviors, and to create a risk profile. We will continue to look at this area of research and try to overcome some of the limitations we encountered. We are planning a programme of enquiries into a large number of individuals in our data set. Following from this investigation exercise we will have much more results on which to judge the effectiveness of our model.

We want to develop the model further so that its predictive power is better and we will be looking at the possibility of an additive model to determine any second-order interaction effects. One such interaction that we are interested in is income and ACORN, the possible interpretation being that any individual with a particularly low income to high ACORN ratio could be more at risk.

So, the next steps will be:

i. look at the limitations of our analysis
ii. determine how the model can be improved further/explanatory power enhanced
iii. test for the assumptions of the model more rigorously
iv. look into multicollinearity and covariance
v. determine how the model can be extended to evaluate a range of data sets that will be available in future.

Endnotes

[1] http://www.taxjustice.net/e/press/tax_losts_costs.pdf

[2] Datamonitor's Offshore Financial Services Databook

[3] http://www.caci.co.uk/acorn/whatis.asp

Annexes

The following briefly describes the three annexes referred to in the text of this paper:

- **Annex 1—**This annex includes a list of 33 variables used in the model and an explanation of their meanings.
- **Annex 2—**This annex provides a set of three tables:
    - o Score Test for the Proportional Odds Assumption
    - o Analysis of Maximum Likelihood Estimates
    - o Odds Ratio Estimates
- **Annex 3—**This annex describes the lifestyle categories into which all people living in different ACORN classes have been assigned.

Each of these annexes may be accessed on the Tax Stats portion of the IRS Web site at: http://www.irs.gov/taxstats/productsandpubs/article/0,,id=130103,00.html .