Output-based disclosure control for regressions

Felix Ritchie

Abstract

Many recent developments in social science research, especially economics, have arisen from the increased availability of confidential government microdata, often in controlled environments. Output-based statistical disclosure control is increasingly important for making effective use of this resource. A central topic is whether the most common analytical tool, multiple regression, is 'safe' for release. This is a relatively unexplored field: only a handful of papers have been produced over the last decade and the main reference for practitioners is an unreviewed internal document.

This paper analyses the disclosure risks of linear regressions, and demonstrates that, even in the best-case scenario for an intruder, regression results are fundamentally non-disclosive and so come within the class of 'safe statistics'. It shows that conflicting results in papers reflect institutional perceptions, not statistical matters. It notes that simple rules can both guarantee confidentiality and provide measures of the best approximation to confidential data. It discusses a number of statistical concerns that are shown to be misguided. Finally, it summarises these results to produce formal guidelines for data owners managing controlled environments.

Contact:

Felix Ritchie felix.ritchie@uwe.ac.uk

Department of Accountancy, Economics and Finance, University of the West of England, Bristol

JEL classification:

C18, C40

Keywords:

Statistical disclosure control, principles-based, safe statistics, controlled environments, regression

1. Introduction

In the past decade, the one of the major developments in social science research, particularly microeconomics, is the increased availability of confidential government microdata. This usually requires giving researchers access to detailed disclosive data in controlled environments such as remote job servers (RJSs) and virtual research data centres (RDCs). The value of a controlled environment is that researchers have much freedom to work with the data, with confidentiality checks only being applied at the point that the statistical results are prepared for release from the controlled environment.

This 'output-based' statistical disclosure control (SDC) contrasts with the more traditional input-based SDC where the data set was redacted or perturbed before distribution to prevent disclosure risk. There is a large literature on input-based SDC. However, statistical investigation of output-based SDC was focused almost exclusively on the confidentiality of finite, known sets of tables produced by the data owners. Ritchie (2007) argued that this was inappropriate for research environments and instead proposed an approach termed 'principles-based' output SDC (PBOSDC). This has now been formally adopted by the two main UK RDCs, by Statistics Netherlands, and by Eurostat (Brandt et al, 2010); and informally by a number of statistical agencies and RDCs.

A key element of PBOSDC is the division of outputs into 'safe' and 'unsafe' statistics[1]: respectively, those which do not and do present any disclosure risk, irrespective of the data used to generate the statistic (see Ritchie, 2008; Brandt et al, 2010). A 'safe' statistic is one which may be released from the controlled environment without undergoing checks for confidentiality (in some cases, a very small number of automatic checks may be required). This is essential for the practical management of SDC in a controlled environment: without the safe/unsafe distinction, the time needed to check outputs for disclosure increases significantly, leading to increased costs for the data owner, frustration for the researcher, and, potentially, more insecure outputs (Ritchie, 2007).

As a primary reason for controlled environments is to enable multivariate analysis on confidential microdata, a clear statement on the 'safety' or otherwise of outputs is of considerable value. The literature on the most important analytical tool, multiple regression, is spartan, diffuse, and reflects institutional factors as well as statistical ones. However, the key findings are relatively straightforward and robust. The aim of this paper is to provide unambiguous recommendations, and to show how these might be modified in the context of specific organisational arrangements.

The next section briefly discusses the background for this topic, and why there is a need for a revised summary now. Section 3 covers identification in a simple linear regression. Section 4 considers deliberately falsifying estimates to produce disclosive results. Section 5 demonstrates how simple measures of approximate fit can be calculated automatically. Section 6 considers the practical aspects: how realistic are theoretical possibilities, can non-intrusive confidentiality-improving rules be developed, how does statistical quality relate to confidentiality, and how do institutional arrangements matter. Section 7 summarises, and provides a set of revised guidelines for data owners.

[1] Early papers (VML, 2004; Ritchie, 2008; Brandt et al, 2010) referred to 'safe' or 'unsafe' outputs. However, because 'safe output' has a different meaning in the context of the Virtual Microdata Laboratory (VML) Security Model of which PBOSDC is a part, since 2011 the preferred term has been 'safe/unsafe statistic' (see VML-SDS, 2011).

2. Principles-based output SDC and analytical results

In the early days of PBOSDC there was almost no literature on disclosure risks in analytical results. Notable exceptions are Reznek (2004) and Reznek and Riggs (2005), who focused on the problem of conditional explanatory variables; Corscadden et al (2006) whose Statistics New Zealand guidelines derive expressions for the riskiness of regression results based upon summary statistics; and Ritchie (2006) who, identifying the lack of any general statement on the disclosure risk of regressions, derived a number of key results from basic statistical analysis. Although initially written as an internal Office for National Statistics (ONS) practice guideline, this latter paper was widely circulated and is generally the only evidence cited for the popular assertion in RDC literature that analytical results are 'safe'.

In recent years, there have been a number of developments. First, several authors have expanded on the capacity of malicious intruders to produce false results. Second, the revival of interest in remote job servers has stimulated investigations into the use of massively repeated attacks. Third, some authors have looked at particular variable subsets which could make disclosure from coefficients feasible without full information.

None of these developments change the basic premise and key conclusions of the earlier papers, but they do qualify some results and provide context which is important for designing institutional arrangements. Moreover, Ritchie (2006), as a non-peer-reviewed internal practice document, contains drafting notes, unsupported assertions, unresolved queries, and some minor errors. Hence there is a need for a review of evidence and a restatement of the current consensus.

3. Exact identification of values in a linear regression through outsider knowledge

Consider a linear least-squares regression on N observations and K variables:

$$y_i = x_{i1}\beta_1 \ldots x_{iK}\beta_K + u_i \quad i = 1..N$$

or more compactly y=Xβ+u. Only "genuine" regressions are analysed; that is, where N>(K+1) and K>1. For now, it is assumed that a researcher does not "create" regressions solely for the purpose of disclosure. This papers concentrates on the 'extreme' scenario of a simple OLS regression on untransformed variables; more complex models just reduce the potential for disclosure. Finally, no assumptions are made about the distribution of variables or error terms. The following results depend upon the mathematical qualities of the estimators, not the statistical ones.

This section consider an extreme 'outsider' scenario: that an intruder acquires a set of regression coefficients and standard summary statistics from repeated estimation on the same or a similar sample; that he/she has a large amount of information about the type and means of the variables and the sample; and that his/her only interest is in discovering something that should have been hidden – for example, just to embarrass the data owner.

3.1 Identification in a single regression

Direct disclosure from a single genuine linear regression is, in general, not possible (Ritchie, 2006). Intuitively this may be explained as follows.

A linear regression to determine K parameters implies K independent normal equations; therefore at most K unknowns can be identified. The regression contains K+1 variables (K explanators and one dependent variable), which reflect K+1 variables. An intruder wishing to ascertain specific values must therefore know (N(K+1)-K) values. Conversely, disclosure can be prevented by ensuring that at least K+1 item responses are not known to the intruder.

There are three exceptions to this rule, the first two originally described by Reznek and co-authors.

Case A1: single observation in a single category

Suppose x_{i1}=1 if i=1, 0 in all other cases. Then the estimated coefficient on that category will ensure that the fit is exact ie u_1=0. Therefore

$$y_1 = \hat{y}_1 = \sum_k x_{1k}\hat{\beta}_k$$

In other words, the value of $y_1$is disclosed if the intruder has all the coefficients and the actual values of x_1. This is a smaller information requirement than in the general case, and the result holds irrespective of the type and value of other variables.

Case A2: a saturated conditional variable regression

If the model is fully saturated (that is, only binary variables with all interactions included), then the estimated coefficients reflect the actual means of a conditional magnitude table. Reznek and Riggs (2005) give an example, and demonstrate that this holds for weighted regressions. However, if all the variables are strictly orthogonal (that is, $x_{ij}x_{ik}$=0 for all (i, j, k) except j=k), then interactions are irrelevant; the non-interacted model is saturated.

Ronning (2011) argues that Reznek and others have misinterpreted this case: the fact that regression coefficients have generated conditional means does not necessarily mean that a disclosure has occurred as the means may be non-disclosive. These perspectives can be reconciled by considering that the saturated 'regression' is misclassified: Ritchie (2006) argues that it should be identified as a table (an 'unsafe' statistic in PBOSDC terminology) and assessed as such.

Case A3: strictly orthogonal variables

Suppose X can be partitioned into two variable sets, mathematically orthogonal:

$$X = [X_A\ X_B\ ] \quad X'_A\ X_B\ = 0$$

where X_A and X_B are NxK_A and NxK_B. On defining a conformable coefficient vector, the normal equations lead to a partitioned estimate:

$$y = [X_A\ X_B]\begin{bmatrix}\beta_A \\ \beta_B\end{bmatrix} + u$$

$$\rightarrow \begin{bmatrix}\hat{\beta}_A \\ \hat{\beta}_B\end{bmatrix} = \begin{bmatrix}(X'_A X_A)^{-1} & 0 \\ 0 & (X'_B X_B)^{-1}\end{bmatrix}\begin{bmatrix}(X'_A y) \\ (X'_B y)\end{bmatrix}$$

This only occurs where one of the variable sets consists wholly of dummy variables. In practice, this requires two sets of dummy variables, mutually exclusive between sets although not necessarily within sets. If dummies are also mutually exclusive within sets, then this collapses to case A2.

This is relevant where one set consists of a single dummy variable where

$$x_A = 1 \rightarrow X_B = 0$$

$$x_A = 0 \rightarrow X_B \neq 0$$

In this case it can be demonstrated that

$$\hat{\beta}_A = \bar{y}_{x_A=1}$$

3.2 Disclosure by repeated estimation

Consider the matrix formulation y=Xβ+u where y, x, β and u are, respectively, Nx1, NxK, Kx1 and Nx1 matrices. Two cases are relevant.

Case B1: direct differencing by adding observations with known explanatory variables

Define y_0, X_0, and u_0 as Sx1, SxK and Sx1 matrices of additional observations, and β_0 as the corresponding estimate from an OLS regression. Even if X_0 is known, this does not lead to the direct identification of the dependent variables as

$$\hat{y}_o = X_o\hat{\beta}_o \neq X_o\hat{\beta}_o + e_o = y_o$$

where e_0 is the vector is estimation residuals. However Ritchie (2006) notes that

$$\hat{\beta} - \hat{\beta}_0 = \left(X'X\right)^{-1} X'y - \left(X'X + X_0'X_0\right)^{-1}\left(X'y + X_0'y_0\right)$$

This is a system of K equations, so if there are less than K unknowns in (y_0, X_0), then it is possible to solve the model. For example, if X_0 is known then solving for y_0 gives:

$$y_0 = \left(X_0X_0'\right)^{-1} X_0X'X\left(\hat{\beta}_0 - \hat{\beta}\right) + X_0\hat{\beta}_0$$

This has a solution if (X_0X_0') is invertible; that is if S≤K.

In general this solution requires full knowledge of the explanatory variables, but there are plausible situations for which less knowledge is required.

First, Ritchie (2008) notes that the estimated variance-covariance matrix (VCM) from the initial estimate allows the unknown cross-product matrix to be recovered (which is why the VCM is an 'unsafe' statistic):

$$(VCM)^{-1}\hat{\sigma}^2 = ((X'X)^{-1}\hat{\sigma}^2)^{-1}\hat{\sigma}^2 = X'X$$

If X_0 is known, then there is now sufficient information to calculate y_0.

Second, Ritchie (2006) exploits the mathematical property that the regression line goes through the mean of the variables to show that the mean value of the new observations can be identified:

$$\bar{y}_0 = \frac{N}{S}.\bar{X}\left(\hat{\beta}_0 - \hat{\beta}\right) + \bar{X}_0\hat{\beta}$$

If there is only one additional observation, then clearly this discloses the value of the additional dependent variable (this result can be also derived in other ways).

In summary, if a regression is duplicated with S≤K additional observations then it is possible to identify up to S unknown values if

- SxK other values in the additional (X, y) set are known, and either
    - the estimated variance-covariance matrix and model error from the initial regression are available, or
    - S=1 and the explanatory variable means from the initial and augmented regression are known

Ritchie (2006) notes that these results are not affected by the orthogonality of the explanatory variables. In models composed entirely of binary variables the identification issues collapses to a problem of table differencing, as described in case A2 above.

Case B2: identification through repeated estimation of subsets

Gomatam et al (2005) and Sparks et al (2008) note that repeated estimation on subsets provide a potential solution to the normal equations. Assume that the 3xN matrix X=[a b c]. Then the normal equations (X'X)β=X'y give:

$$\begin{bmatrix} a'a & a'b & a'c \\ b'a & b'b & b'c \\ b'c & c'b & c'c \end{bmatrix}\hat{\beta} = \begin{bmatrix} a'y \\ b'y \\ c'y \end{bmatrix}$$

Assuming the estimated coefficient vector is known, this gives a system of three equations with nine unknowns (a'a, a'b, a'c, b'b, b'c, c'c, a'y, b'y, c'y). A regression on the subset of variables (a, b) would produce

$$\begin{bmatrix} a'a & a'b \\ b'a & b'b \end{bmatrix}\hat{\beta}_{ab} = \begin{bmatrix} a'y \\ b'y \end{bmatrix}$$

where the subscript denotes that the coefficient vector is estimated only on (a, b). This generates two additional equations with no new unknowns. Overall, the three variables generate twelve equations, meaning that it is theoretically possible to find solutions for all the values of (X'X) and (X'y). In general, for K>2 (or K>3 if a constant term is included and regressions on a constant are disallowed), there will always be more potential equations than unknowns. Thus by repeated subsetting of the variables it is theoretically possible to reconstruct the VCM X'X.

This itself is not necessarily disclosive. However, as Ritchie (2008) notes, the VCM is an 'unsafe' statistic: it is capable of revealing information, for example through the interactions with conditional variables. It is therefore possible. This is a rare example of how an 'unsafe' statistic could, in theory, be generated from repeated estimation of a 'safe' statistic.

4. Exact identification using insider information

Several authors (eg Gomatam et al, 2005; Sparks et al 2008; Bleninger et al 2011) have noted that it is possible for a researcher having access to the source data to generate regression results which, although apparently innocuous, can conceal disclosive results[2].

Ritchie (2006) explicitly excluded deliberate falsification of results, noting that there are simpler and less traceable ways of generating false output from an RDC than manipulating regressions. However, interest in fully-automated remote job systems, where the outputs are approved by simple rules, has stimulated the consideration of unauthorised transformations by those who have access to the data. For completeness, therefore, this section considers disclosure risk in regressions where a researcher

- is prepared to generate nonsense regressions purely to disclose confidential values
- can apply any transformation to the data

It is not necessary for the researcher to have access to view the code.

Case C: Known value of some explanatory variables

In the simplest case, an intruder knows the value of some variable and uses it to weight the regression such that only the observation with that specific value has any explanatory power. Suppose that an intruder knows the value x_1=m, and wishes to know the value of y_1. Two approaches lend themselves

$$(a)\ y_i = \alpha + \tilde{x}_i\beta + u_i\ , \qquad \tilde{x}_i = \frac{1}{|x_i - m| + \varepsilon}$$

$$(b)\ y_i = \alpha + x_i\beta + z_i\gamma + u_i\ , \qquad z_i = 1\ iff\ x_i = m$$

Bleninger et al (2011)'s elegant paper labels these 'artificial outliers' and 'strategic dummies', summarises the relative advantages (to the intruder) of these alternative approaches, and tests the likelihood in the case of the IAB Establishment Panel. The results demonstrate the feasibility of these intruder scenarios, but also highlight the importance of the uniqueness of m.

The 'strategic dummies' confirms to case A1, above; the difference is that here the dummy is being generated specifically to target an observation, rather than the result of a poorly-specific model. Sparks et al (2008) note that applying a known matrix transformation can effectively hide the presence of single or sparse observations from simple tests on the frequency of regressors.

Other transformations, particularly non-linear ones, could be postulated to attenuate the distribution in a covert manner; or observations could simply be dropped to provide the necessary concentration of information in one observation. Finally, it would be possible to generate strategic dummies based on the ranking of the observation, such as the largest company or greatest age.

[2] A reviewer of Ritchie (2006) also noted the possibilities in selecting observations to produce highly-skewed distributions.

These subversive transformations should not be confused with estimation on a skewed distribution; they are designed specifically to target particular observations, so that, in effect, the regression collapses to a single case. Estimation on a skewed distribution *per se* is not disclosive. However, it is clear that if an intruder has accurate information on a specific value and an uncontrolled ability to transform the data, generation of a false regression which appears to be genuine is always feasible.

5. Evaluating the likelihood of approximate disclosure

Sections 3 and 4 described how exact identification of values can occur. In practice this is unlikely because it relies upon being able to difference regressions effectively, which in turn requires detailed information about how the regressions were constructed. However, it may be sufficient for an intruder to have a rough idea of the value of a variable – for example, by taking coefficients and creating fitted values of the dependent variable. This section quantifies this risk, concentrating on created fitted values for dependent variables where the intruder has access to

- the estimated parameters
- the values of the explanatory variables x_i
- common summary statistics on the regression

This section requires evaluating expectations but, as before, no distributional assumptions are made. Note that if the equation is mis-specified (for example, errors are not i.i.d.), then the results here under-estimate confidence intervals and over-estimate the accuracy of approximations.

5.1 Approximating values

Using the same matrix notation as before, suppose an intruder wishes to find the exact value of a dependent variable y_1 and has knowledge of x_1. The residual e_1 has the expected mean 0 and variance (see Ritchie, 2006)

$$V(e_1) = \sigma^2(1 - x'_1(X'X)^{-1}x_1)$$

This is smaller than the standard error of the regression, reflecting the fact that this observation contributed to the estimates. It reaches its minimum value when this observation contributes most to the regression ($X'X \rightarrow x_1x_1'$), and approaches the standard error when the observation has a negligible impact ($x_1 \rightarrow 0$).

When evaluated at the largest vector in X, this enables the minimum predictive error on a dependent variable to be ascertained. In other words, this allows the NSI to automatically determine whether an intruder, working with a set of explanatory variables, the published coefficients and descriptive statistics, would be able to derive a fitted value within a specified level of certainty.

If the published descriptive statistics are available, then an exact confidence interval can be calculated without the need for variable values. Defining TSS and ESS as total and estimated sums of squares, then using

$$\hat{\sigma}^2 = (TSS - ESS)/(N - K)$$

$$R^2 = TSS \,/\, ESS$$

$$ESS = \hat{\beta}'X'X\hat{\beta} \quad \rightarrow \; X'X = \left(\hat{\beta}\hat{\beta}'\right)^{-1}.ESS$$

it can be shown that

$$x'_1(X'X)^{-1}x_1 = \frac{\left(\sum_k x_{1k}^2\hat{\beta}_k^2\right)R^2}{\left(\hat{\sigma}^2(N-K)(1-R^2)\right)}$$

And so

$$\hat{V}(e_1) = \hat{\sigma}^2\left(1 - \frac{\left(\sum_k x_{1k}^2\hat{\beta}_k^2\right)R^2}{\left(\hat{\sigma}^2(N-K)(1-R^2)\right)}\right)$$

Hence, if the intruder knows the value of x_1, then it is possible to calculate a confidence interval for a predicted value from the minimal set of summary statistics.

5.2 Approximation for new observations

If the published coefficients are used for prediction by the application of a new set of observations (y_0, x_0), then a similar limit can be derived. Assuming y_0 and x_o come from the same distribution then (see eg Verbeek , 2004):

$$E(e_0) = 0$$

$$V(e_0) = \sigma^2(1 + x'_0(X'X)^{-1}x_0)$$

The intuition behind this is that the new error is assumed to be uncorrelated with the errors used to generate the coefficients. Therefore, the values of explanatory variables increase uncertainty as they move away from the mean values used in the regression.

In this case, the standard error of the regression is the minimum level of uncertainty, achieved when the new explanatory variables equal the mean of the variables used to calculate the coefficients. The predictive error cannot be reduced below this level.

5.3 Using R^2 directly as an estimate of riskiness

Corscadden et al (2006), using a similar analytical approach to functional form, develop an alternative measure where a direct relationship between R^2 and the required level of uncertainty in a regression can be quantified. This is a measure of the average riskiness, not the maximum, and, as in the above example, could be relatively easily coded to be a standard output from regressions.[3]

6. Guidelines for practical application

For data owners, the question of whether regressions are 'safe' in the PBOSDC terminology is critical, as it greatly affects the effectiveness of any controlled facility. Regressions are currently seen as 'safe' in formal guidelines (eg Brandt et al, 2010).

[3] Although no general relationship between R^2 and predictive uncertainty has been derived, conversations between the author and Statistics New Zealand staff suggested that, in practical cases, very high R^2s (>0.99) were necessary to breach SNZ rules on approximate disclosure (within 10%-15% of the true value) for any particular observation.

The above discussion showed that, in theory, it is possible for a regression to generate disclosive outputs; and that it is possible for intruders with access to the raw data to falsify regression outputs. This section considers whether the current standard should be adjusted in the light of these findings, and then considers the proposed solution from Ritchie (2006) that claims to guarantee confidentiality.

6.1 Regression risk in practice

No statistic can be guaranteed non-disclosive in that no combination of variable, transformations and repeated calculation would ever produce a single value. Therefore, 'disclosiveness' is a matter of judgement.

The above examples demonstrate feasible possibilities. However they have very specific information requirements:

Case	*Restrictions on regression*	*Intruder knowledge required*	*Consequence*
A1 Unique explanatory dummy	Binary variable with only one observation No interaction terms	That one unique observation exists All other x values for the unique observation	Dependent variable for unique target identified
A2 Saturated regression	Only conditional variables All interactions included	None	Table of conditional means generated
A3 Orthogonal variable	Single orthogonal binary variable	The nature of the orthogonal variable	Mean of flagged dependent variables
B1 Direct differencing, S additional observations	Smaller sample (N) is exact subset of larger (N+S) Same variable set	SxK known values amongst additional observations, and Original VCM or (S=1) and original sample means	Identification of S values
B2 Differencing by repeated estimation	Same sample for all models	None	Reconstruction of VCM
C Deliberate falsification	Single variable regression only (artificial outliers) Unique explanatory variable	Known value of explanatory variable	Identification of any other variable associated with that unique value

It is clear that the likelihood of disclosure depends upon some very specific models and, usually, some stringent information requirements on the intruder. B1 and B2 also require repeated estimation under controlled circumstances. In general, these conditions are not fulfilled by genuine research activity.

Of the 'outside intruder' cases A1-3, B1-2, the first two are most likely to occur by accident. A1 is potentially the most serious. However, even knowing of the existence of a unique marker is only relevant if the intruder knows the other explanatory variable values. A2 is rare because of the need to include all interaction terms on non-orthogonal variables – a regression merely on conditional

variables is in effect a correlation matrix. It is difficult to see how any genuine statistical analysis would generate the case A3.

B1 and B2 do reflect the actions of researchers, who will repeat estimates on different sample sizes and variable sets. Note however, that both cannot be varied, and in applied analysis these will typically both be varying.

The discussion has also assumed that the intruder has a wealth of information about the analyses. In practice, the level of detail described here is not available to those merely looking at journal articles or output from statistical programs. Finally, much analysis is multi-stage (for example to deal with autocorrelation). This does not mean that the above analysis cannot be applied, but that the intruder is identifying variables which have been subject to an unknown transformation.

In summary, the need for a specific (and often statistically unhelpful) form of the variables and sample, and the information requirement on the intruder, mean that the probability of disclosure from genuine research activities is negligible. Only the 'inside intruder' scenario provides a realistic possibility of disclosure.

6.2 Guaranteeing non-disclosiveness

Ritchie (2006) suggested that non-disclosiveness could be guaranteed to a very high degree by not allowing all the coefficients from regression to be published. This

- Prevents the generation of predicted values (A1, A2, C)
- Nullifies disclosure through repeated estimation by introducing new unknowns at every estimate (B1, B2)
- Prevents the calculation of confidence intervals for a specific observation

Ritchie (2006) notes that this does not need to be a significant variable, as the disclosure possibilities noted above depend upon the mathematical properties of regression (statistical requirements to construct useful confidence intervals are additional). However, there are a number of obvious candidates for omission from published output which are both statistically significant and of limited interest to researchers, particularly constant terms, time dummies and other conditioning variables.

The advantages of removing coefficients include simplicity in implementation and effectiveness against a range of potential problems, even deliberate falsification of outputs. Importantly, experience in the UK since 2004 (where 'final' outputs such as submissions to journals are required to above the remove-coefficient rule) has demonstrated that this is acceptable to researchers. One reviewer also argued that this benefits the research community more generally, by encouraging researchers to include only key results in publications.

The recognition that dropping coefficients makes regression output practically non-disclosive highlights that some regression models are fundamentally non-disclosive; specifically, those which estimate unpublished incidental slope or intercept parameters, such as nested or multilevel models. For example, the simple linear panel model

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}$$

is non-disclosive unless all unit means are published.

This also guards against the possibility raised by Gomatam et al (2005), that certain kinds of regression relationship may be undesirable. Gomatam et al note that repeated estimation on different variable subsets could allow the conditional estimation of models which the data owner wishes to suppress for some reason. Again, without a full set of coefficients this is in general not possible.

Some authors have suggested alternatives, such as randomly dropping a small proportion of observations, limiting the number of combinations of variables on the same sample, blocking combinations of variables, and limiting the publication of sample means. The difficulty with all of these is that they restrict what may be valid statistical analyses. The drop-coefficient rule does not restrict analysis itself, only the presentation of results. In the UK RDCs, where researchers themselves choose which coefficients to hide, there is effectively no restriction on research activity.

Ritchie (2006) notes that for this to be wholly successful, the coefficient that is dropped 'could not reasonably be determined from published information'. If it is assumed that variable means are available (which is the assumption in the UK), then more than one coefficient should be dropped to guarantee the non-disclosiveness of released results.

Note that case A3, as practised by an 'inside intruder' is only prevented if the coefficient associated with the single orthogonal variable is dropped.

6.3 The role of the security model

As noted above, much of the recent research has focused on deliberate attempts to unpick regression results. This reflects an interest in remote job servers (RJSs) in some countries. Currently all existing RDCs and RJSs have some human oversight, but ideally, all of the RJS processes are automatic. If however the RJS is completely unmonitored than the possibility arises of both 'insider' attacks and 'outsider' attacks based on multiple repeated estimation.

The 'VML Security Model' (Ritchie, 2010) uses five dimensions to consider how access to data can be made safely and efficiency: projects, data, people, setting and outputs. The latter three are of concern here. As VML (2004), VML-SDS (2011) note, 'safe setting' (the physical or technical environment) and 'safe outputs' (the SDC model) are designed to guard against accidental disclosure of information. In contrast, 'safe people' (the vetting and training of researchers) and, again, 'safe settings' are primarily to lower the chance of deliberate misuse.

In this light, the role of PBOSDC can be put in context. PBOSDC assumes that (a) research results are genuine, but that (b) mistakes are made. In a genuine research setting with well-intentioned researchers, the disclosure risk from regression is negligible for both the direct risk posed by a single output, and the chance of disclosure by repeated analysis on similar subsets, even in the light of mistaken analyses. There are routes for an ill-intentioned researcher to falsify analyses or the presentation of results; this is not what PBOSDC is designed to uncover. If a researcher *chooses* to mislead, this is a failure of the 'safe people' strategy. If a researcher *is able* to mislead, this is a failure of the 'safe settings' strategy.

Consider an RDC where researchers are trained, so that both the 'safe people' and 'safe setting' dimensions are in play. Suppose a researcher, learning that regression is always approved, chooses to hide small-cell tabular output ((which might be blocked) as regression output, using the saturated model noted above. This is a failure of the training, and hence the 'safe people' policy.

Alternatively, consider an RJS designed to be available to the general public; no individual training is possible, and so the security is vested wholly in the 'safe setting'. If there are no restrictions on the number of regressions that may be run on the same subset, a malicious user could exploit the repeated-attack scenarios above. In this case, the 'safe setting' is inappropriate.

The VML Security Model is a system model: discussion about risk and the 'person strategy', for example, only have meaning in the context of project, data, setting and output strategies. Hence, the guidelines developed in this paper are robust in the context of the PBOSDC model. They do not provide a definitive defence against disclosure in all institutional scenarios, because they make assumptions about the motivation and ability of researchers which may not be appropriate.

6.4 Non-issues

6.4.1 Statistical quality

One point of genesis for this work was the dispute between users and methodologists over the impact of quality issues. Ritchie (2006) summarises:

- **Outliers** are observations which deviate strongly from the regression line but in themselves are not significant in determining the relationship; as an outlier has large variance and poor fitted value, it is less disclosive than other observations and reduces the overall disclosiveness of the regression

- **Influential points** are similar to outliers but have a significant impact on the regression line. This is the situation in which differences between regressions are most likely to be (a) discernible and (b) published. However, the interaction between terms means that exploiting this to uncover information still requires detailed variable knowledge

- **Multicollinearity** and **measurement error** increase estimated errors and make attribution of effects to particular variables more difficult; both lower the prospect of disclosure.

- **Estimation on public explanatory variables** is, in theory, the case outlined in B1 above; however, Corscadden et al (2006) seem to show that in practice this overstates the likelihood of making accurate predictions.

In all cases, quality issues need to be separated from SDC issues. Data problems leading to a very skewed distribution theoretically lead to more risk of disclosure, but generally low quality data and poor models reduce disclosure risk.

An exception to the "bad is good" rule is where there are few observations. A model with no degrees of freedom clearly leads to a set of equations allowing identification of variables. It could be argued that this is not a 'regression' as such, and so from a philosophical point of view the above considerations do not apply. Brandt et al (2010) and VML-SDS (2011) take the more pragmatic line

that any regression must have at least ten degrees of freedom[4]. This is an arbitrary rule, and ignores the fact that, for example a model with fifty dummy variables and 60 or so observations is likely to have many single-case dummies. A proportionate rule, while equally arbitrary, may be more appropriate.

6.4.2 Transforming variables and relationships

Transforming equations does not change the above conclusions about whether the form of estimated relationship is itself disclosive. Whether the variables themselves are useful is another issue.

Clearly, however, the discussion above has been taking place in an idealised world for intruders. In practice, data transformations, sample selection, treatment of missing values, simultaneous equations, solution algorithms, method of estimation etc will all make the reproduction of the regression environment by intruders extremely difficult.

6.5 Regressions on a single unit

The above discussions relate to regressions on several units, and assumes that an intruder is trying to get information on one unit. However, it is possible that a regression could be run on a single unit – for example, quarterly data on the performance of a company could provide sufficient observations to run regressions solely on that company.

In this case, all coefficients are directly informative. Hiding some does not reduce the disclosiveness of the others, and nor does the uncertainty surrounding actual values: knowing that a relationship is positive and significant may breach confidentiality standards. Eurostat and the Dutch and UK RDCs deal with this again by an arbitrary standard, banning regressions on a single unit[5].

6.6 Releasing additional information

Several authors have raised concerns about the confidentiality of regressions arising from the release of other, related information. These include the release of residuals, the VCM, minimal and maximal values, and quantiles. But in each case, the problem is that the statistic under consideration is a table – an 'unsafe' statistic in PBOSDC terminology, which means that the particular statistic should not be released unless it can be shown to be non-disclosive in that a specific instance.

For example, releasing residuals amounts to tabulating a distribution. As for any other tabulation, the data owner would be concerned about whether that distribution contains outliers which could be associated with unit responses. The fact that the distribution is based on generated rather than observed values may increase the data owner's willingness to release it, but the default assumption is that this is 'unsafe' and the case for releasing this specific instance needs to be given.

[4] This rule has not been invoked in the UK RDC's eight years of existence.

[5] In the context of PBOSDC, 'banning' something does not mean it is never possible, but that the researcher needs to demonstrate to the data owner that the particular output in question is (a) non-disclosive and (b) worth arguing about; see Ritchie (2008a)

6.7 Releasing coefficients for prediction

One aim of modelling is to release a set of coefficients that can be used to predict values in other datasets (for example, using earnings information in one dataset to construct a model which can then be used to generate a predicted income variable in a second dataset). In this case, holding back coefficients is not a valid operation. However, as shown above and in Corscadden et al (2006), it is perfectly possible to assess the prediction risk for a full set of coefficients so that the risk of re-identification in the original dataset can be quantified. This is a maximum risk estimate, and could be adjusted to take account of, for example, the unavailability of the original explanatory variables.

7. Conclusion and recommendations

This paper has reviewed the opportunities for determining confidential information from regression outputs. This is an important topic, because the efficiency of RDCs and the feasibility of RJSs depend upon being able to make quick and reliable decisions about the main analytical tools of researchers. For researchers, waiting for cleared results to be released from a controlled environment can be frustrating and unproductive. The adoption of PBOSDC by ONS cut the target clearance time for results from two weeks to two days, with the actual median clearance time less than one day. This discussion therefore has a direct impact on researchers and data owners.

Ritchie (2006) made some initial proposals, and these have been widely adopted amongst RDCs. However, this was an unreviewed internal practice paper which contained a number of minor errors; moreover, recent research has suggested other potential problems. This paper has argued that, although the conclusions of Ritchie (2006) generally stand, they need to be modified in the light of recent research and a better understanding of the way the separate parts of the security model interact.

This paper recommends a revised set of guidelines:

1. Evaluation of regressions in normal situations

    a. Regressions coefficients and some summary statistics (R2, estimated variance, F-tests etc) should be treated as ‘safe statistics’ in PBOSDC terminology; that is, they can be released from the controlled environment with no further checking

    b. Regressions containing only fully-interacted conditional variables should be assessed as tables

    c. As it might be difficult to identify whether a regression is fully-interacted, for simplicity data owners may wish to consider any model with only conditional variables as a table

    d. Models which generated unreported incidental parameters are inherently ‘safe’

    e. Regressions on multiple observations of a single unit are ‘unsafe’, as are regressions with a single observation on one dummy variable, and ones with no degrees of freedom; although of limited research value and so unlikely to occur, in the interests of giving a clear signal to researchers data owners may wish to consider banning the first two and setting a minimum (actual and proportional) number of degrees of freedom

2. Provision of additional information

    a. Provision of additional statistics associated with the regression (means, residuals, VCMs et cetera) should be evaluated based on their functional form, not on the fact that they are associated with a regression

3. Risk assessment

    a. There is no statistic such that a particular combination of variables, transformations and partial knowledge cannot produce disclosure

    b. Withholding two or more coefficients guarantees non-disclosure in practice, and so this should be encouraged where the impact on research and researchers is not significant

    c. Summary statistics can automatically generate adequate measures of disclosure risk

4. Researcher behaviour

    a. The 'inside intruder' scenario provides the only realistic opportunity to generate disclosive outputs, and so data owners should consider the feasibility of this

    b. Output-only controls cannot reliably distinguish between genuine and false outputs

    c. Inappropriate behaviour should be addressed by considering the 'deliberate misuse' dimensions of the security model (particularly people and settings), rather than restricting output

    d. If, however, there is thought to be a significant and permanent risk of misuse, then rule 3b (remove coefficients) should be enforced

This paper has presented the intruder with a near-ideal environment – the data is inherently interesting, has not been transformed or sampled in some way that would make it difficult to identify the included observations, values of additional explanatory variables may be known, or the intruder may have access to the internal variables. The purpose is to show that, even in an intruder's preferred scenario, that chances of being able to uncover information are negligible; and so, in realistic applications, data owners can feel confident about the application of the results here.

In practice, these conditions are unlikely to hold. Experience in various countries suggests that regression analyses are not problematic; this paper has demonstrated that this result is not a happy accident but an expected consequence, and data owners can design their access mechanisms with this in mind.

References

Bleninger P., Drechsler J., and Ronning G. (2011) Remote data access and the risk of disclosure from linear regression", *Stat. and Op. Res. Trans. Special Issue: Privacy in statistical databases*, pp 7-24 http://www.idescat.cat/sort/sortspecial2011/DataPrivacy.1.bleninger-etal.pdf

Brandt M., Franconi L., Guerke C., Hundepool A., Lucarelli M., Mol J., Ritchie F., Seri G. and Welpton R. (2010), *Guidelines for the checking of output based on microdata research*, Final report of ESSnet sub-group on output SDC http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

Corscadden, L., Enright J., Khoo J., Krsnich F., McDonald S., and Zeng I. (2006) *Disclosure assessment of analytical outputs*, mimeo, Statistics New Zealand, Wellington

Gomatam S., Karr A., Reiter P., and Sanil A.(2005) "Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk–Utility Framework for Remote Access Analysis Servers" Stat. Science v20:2 pp163-177 http://projecteuclid.org/DPubS?verb=Display&version=1.0&service=UI&handle=euclid.ss/1121347638&page=record

Reznek, A. (2004) *Disclosure risks in cross-section regression models*, mimeo, Center for Economic Studies, US Bureau of the Census, Washington

Reznek A. and Riggs T. (2005) "Disclosure Risks in Releasing Output Based on Regression Residuals" ASA 2004 Proceedings of the Section on Government Statistics and Section on Social Statistics pp1397-1404 http://www.amstat.org/sections/srms/proceedings/y2005/Files/JSM2005-000448.pdf

Ritchie F. (2006) *Disclosure control of analytical outputs*, mimeo, Office for National Statistics; available as WISERD Data Resources Paper No. 5 http://www.wiserd.ac.uk/wp-content/uploads/2011/12/WISERD_WDR_005.pdf

Ritchie F. (2007) *Statistical disclosure control in a research environment*, mimeo, Office for National Statistics; available as WISERD Data Resources Paper No. 6 http://www.wiserd.ac.uk/wp-content/uploads/2011/12/WISERD_WDR_006.pdf

Ritchie F. (2008) “Disclosure detection in research environments in practice”, in *Work session on statistical data confidentiality 2007*; Eurostat; pp399-406 http://epp.eurostat.ec.europa.eu/portal/page/portal/conferences/documents/unece_es_work_session_statistical_data_conf/TOPIC%203-WP.37%20SP%20RITCHIE.PDF

Ritchie F. (2010) “UK Release Practices for Official Microdata”, *Stat. J. of the Int. Ass. for Official Statistics*, v26:3-4, pp103-111 http://neon.vb.cbs.nl/casc/ESSnet/guidelines_on_outputchecking.pdf

Ronning G. (2011) *Disclosure Risk from Interactions and Saturated Models in Remote Access*, IAW Discussion Papers No. 72, June http://www.iaw.edu/RePEc/iaw/pdf/iaw_dp_72.pdf

Sparks R., Carter C., Donnelly J., O’Keefe2 C., Duncan J., Keighley T., McAullay D. (2008) “Remote access methods for exploratory data analysis and statistical modelling: Privacy-Preserving Analytics”, *Computer Methods and Programs in Biomedicine* v91:3, September, pp208–222 http://www.sciencedirect.com/science/article/pii/S016926070800093X

VML (2004) *Virtual Microdata Laboratory Researcher Training*, presentations 2004-2010; mimeo, Office for National Statistics

VML-SDS (2011) *Safe Researcher Certification*, training presentation, 2011 onwards; mimeo, UK Secure Data Service