Analyzing the disclosure risk of regression coefficients

Felix Ritchie*
* Bristol Centre for Economics and Finance, University of the West of England, Coldharbour Lane, Bristol BS16 1QY. UK
E-mail felix.ritchie@uwe.ac.uk

Received 14 February 2018; received in revised form 4 January 2019; accepted

30 April 2019.

Abstract. A major growth area in social science research this century has been access to highly sensitive confidential microdata, often via restricted-access remote facilities. These allow researchers highly unlimited access to manipulate the data but with checks for disclosure risk before the statistical results can be published. Effective output-based statistical disclosure control (OSDC) is therefore central to effective use of confidential microdata for research.

Multiple regression is a key anaytical tool for researchers, and so knowing whether multiple regression results are 'safe' for release is essential for research facilities. This is a relatively unexplored field; guidelines used by almost all restricted-access facilities reference an informal document from 2006, but more recent work suggests that problems may exist.

This paper demonstrates that linear regression coefficients show no substantive disclosure risks in realistic environments, and so should be considered as 'safe statistics' in the terminology of this field. Conflicting results in the literature reflect institutional perceptions rather than statistical differences, the confusion of statistical quality with disclosure risk, or the failure to identify the source of risk. The result has important implications for those responsible for providing research access to sensitive data.

The paper explores this result on simple linear regression models; more complex models are shown to be 'safer' subsets. Non-linear models pose slightly different problems, but this paper indicates a way such models may be tackled.

Keywords: Statistical disclosure control, output SDC, principles-based, linear regression, safe statistics.

1. Introduction

Since the start of the century, a major development in social science research has been the increased availability of confidential microdata from government and health survey and administrative data. This usually requires giving researchers access to detailed disclosive data in controlled environments such as remote job servers (RJSs) and research data centres (RDCs). Researchers can carry out complex multivariate analysis and data transformations, and the resulting inferences add the most value to the data.

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 is known as 'output-based' statistical disclosure control (output SDC or OSDC).

Until recently, OSDC research meant studying confidentiality risks in tabulations produced by national statistical institutes (NSIs) or other aggregators; for example, the Eurostat-commissioned expert OSDC guidelines in Hundepool et al [1] focus almost exclusively on tabular data protection. Ritchie [2] argued that this was inappropriate for research environments and instead proposed an approach termed 'principles-based' output SDC (PBOSDC) which provided a framework for reviewing any output (see [3] for an extended elaboration). PBOSDC is now common in the 'virtual RDCs' which have dominated recent advances in data access, and it was adopted by Eurostat [4] in training for researchers using distributed data.

A key element of this approach is classifying outputs into 'safe' and 'unsafe' statistics – types of output which can/cannot be published without significant review by the data owner. Such a classification has a substantial effect on both the risk profile and the operational costs of the facility. Of most interest to research facility managers is determining whether linear and non-linear multiple regression outputs are 'safe', as multivariate analysis is the dominant use requiring access to microdata. In particular, is there any need to restrict the release of regression coefficients?

An internal Office for National Statistics (ONS) practice guide [5], arguing that non-linear and linear coefficients are indeed 'safe', has been adopted by RDCs worldwide, and by Eurostat in its expert guidance on OSDC ([6], updated as [7]). The classification also comforts distributors of Scientific Use Files (SUFs) into uncontrolled environments, who worry that researchers do not read guidelines on how to produce non-disclosive statistics when working with sensitive data.

However, both RJS managers and methodologists in statistical agencies worry about this adoption of an informal result. RJS managers fear deliberate falsification of results by researchers manipulating variables to produce 'fake' outputs which appear low-risk but in fact reveal confidential information to the malefactor. Methodologists are concerned that such a simple universal classification must be missing some unspecified but important special cases.

A clear statement on whether multiple regression counts as a 'safe statistic' is therefore of considerable practical value, allowing limited resources to be directed towards research outputs with non-negligible risks. This paper focuses on linear regression, and demonstrates that the classification of estimated coefficients as safe statistics is correct and robust. We analyse the simplest linear regression models, as a worst case. The paper does not consider non-linear modelling, as functional forms vary across estimators; nevertheless, the approach suggested here illustrates how a wider range of statistics may be classified in future.

The next section briefly discusses OSDC and the concept of 'safe' and 'unsafe' statistics. Section 3 describes cases of theoretical exact disclosure; section 4 evaluates whether these are meaningful and useful. Section 5 addresses common concerns raised by methodologists around data and research quality. Section 6 considers approximate disclosure, and proposes simple measures to quantify risk minimum and maximum risk which may have value in RJSs. Section 7 concludes, and proposes how the approach taken here might be usefully extended to non-linear models.

2. Output checking and classification

2.1 Controlled environments, distributed data and the need for output checking

The substantial increase in the availability for research of confidential data this century has been driven by the proliferation of safe 'controlled environments', where the researcher must use the technical facilities provided by the data owner; no data are transferred to the researcher's machine.

The most popular controlled environment is the research data centre (RDC), where the researcher has a relatively free hand to manipulate data but works in an environment fully controlled by the data owner. Virtual RDCs, which provide equivalent security to traditional RDCs but can be accessed remotely, have dominated developments in data access, with governments or universities in over half of EU countries, Canada, the US, Mexico, Australia, NZ and South Africa setting up at least one general-purpose virtual RDC since 2002.

An alternative is the remote job server (RJS) such as LISSY (http://www.lisdatacenter.org/); these allow researchers to submit statistical code and view the results remotely, without direct access to the source data. RJSs are much less common than RDCs. They are popular with some data owners who are reluctant to allow direct access to microdata, but the complex setup and restrictive user environment limits their appeal and only a handful of countries have tried such systems; see [8] for a discussion.

The purpose of such controlled environments is to allow detailed analysis of data which are too sensitive for uncontrolled access. Applying input disclosure control to the microdata (beyond removal of direct identifiers and variables which have no analytical value) defeats this purpose; effort has been invested in ethical processes, training of users, and IT systems to remove the need for input SDC. As researchers are near-enough modelling source data, there is a risk that statistical outputs may inadvertently disclose confidential information: for example, a box plot of the income distribution in a small area might make known the income of a recognisable high-earner. All controlled environments therefore have an element of output checking.

For RDCs, a typical solution is for researchers to place outputs in a sandpit where a staff member can check them; if there is no reasonable disclosure risk, the staff member will release them. In some RDCs, researchers with a good record are allowed to release their own results. In RJSs, outputs are typically automatically assessed by the system as: “release”, “do not release”, or “get a human to look at it”. The ideal for RJSs is to avoid the last option, but the complexity of statistical analysis means that all extant RJSs keep it in reserve.

Datasets are also distributed to uncontrolled environments, where data owners have little or no say in the outputs produced. For Public Use Files (PUFs) which are deemed to present no disclosure risk, controls are not relevant. However, Scientific Use Files (SUFs), such as those available under licence from Eurostat or the UK Data Archive, are assumed to have some significant residual disclosure risks; these might manifest themselves through poorly-designed outputs. Most distributors of SUFs provide advice on OSDC, but with little follow-up to ensure that the advice is used. Hence data distributors may have difficulties demonstrating that SUFs sent to researchers for marginal analysis do not pose a significant disclosure risk.

2.2 Principles-based OSDC and ‘safe statistics’

Ritchie [2] noted that traditional SDC guidance (based largely on frequency and magnitude tables; exceptions are [4, 9, 10]) is inappropriate for a research environment, and instead proposed ‘principles-based output SDC’ (PBOSDC). This consists of

- replacing hard rules with guidelines, and training for both staff and researchers
- devising models for classes of outputs to focus attention on ‘risky’ outputs
- using evidence-based risk assessment to make decisions on release

A key element of PBOSDC is the division of outputs into ‘safe’ and ‘unsafe’ statistics: respectively, those which do not and do present a substantive disclosure risk in normal use, irrespective of the data used to generate the statistic [11][1]. ‘Substantive’ means that, while there is a theoretical risk, the likelihood of the circumstances occurring which generate that risk is negligible. ‘Normal use’ indicates that PBOSDC protects data from accidental disclosure resulting from

[1] ‘Safe’ and ‘unsafe’ are emotive terms, and hence some NSIs have experimented with “low review” and “high review” statistics, reflecting the operational nature of the division. For this paper, we retain the more common terminology.

research; it is not designed to protect against deliberate falsification of results by a malicious researcher.

A statistic defined as 'safe' (for example, the mode [6]) may be published with minimal checks (for example, checking that there is some variation when reporting the mode). In contrast, an 'unsafe' statistic (such as a frequency table) is assumed to hold a non-negligible disclosure risk: it should be checked before release and only be published once disclosure risk in that specific instance has been assessed as acceptably low. Safe and unsafe can therefore also be characterised as "publish unless…" and "don't publish unless…".

This safe/unsafe distinction is essential for practical management: without it, output checking can become a major cost for the data owner. It can also become a source of frustration for the researcher, a known risk element [2]. In a world of limited resources, the safe/unsafe model reduces overall risk by making those responsible for approving outputs focus on relatively high-risk ones [11, 12].

Whether a statistic is safe or not depends solely on the mathematical properties of the statistic; if it is dependent upon the data, it cannot be safe. There can be qualifications (for example, "the mode is safe unless all observations are identical"; [6]), but these must be specific, few and easily dealt with [11]. Many qualifications, or an unclear definition of a qualification, means that each output must be evaluated; this defeats the purpose of the classification.

All statistics pose some disclosure risk in theory; that is, for a given statistic we cannot prove that some transformation which will uncover a value does not exist. Therefore, the definition of 'safe' requires a subjective judgment: how likely is it that the transformation exists? In practice, the evidence-based ethos of PBOSDC asks: what data are necessary to identify an observation, and how likely is this?

2.3 PBOSDC and analytical results

Prior to the development of PBOSDC in 2003 in the UK, there was almost no literature on disclosure risks in analytical results. If papers on SDC discussed risks from model outputs at all, they took a line such as Reiter [13], for example, whose influential paper proposed perturbing the data as a catch-all protection against unspecified risks. This idea has proved popular and persistent; see, for example, Chipperfield and Yu [14] or O'Keefe and Shlomo [15], who both recommend perturbing results without identifying any specific threat.

Notable exceptions addressing specific threats are Reznek [16] and Reznek and Riggs [17], who analysed conditional explanatory variables; and Corscadden et al [18] who derive expressions for the riskiness of regression results based upon summary statistics. The guide developed in [18] for Statistics New Zealand is the earliest general-purpose guide for researchers which addresses non-tabular outputs.

Ritchie [5] identified the lack of any general statement on the disclosure risk of regressions and derived a set of key results from basic statistical analysis. Initially written in 2003 as part of the internal training documents at the UK Office for National Statistics, it contains drafting notes, unsupported assertions, unresolved queries, and some minor errors. However, the note was widely circulated and the revised 2006 version is cited as evidence for the common assertion that regression results are safe (for example, [4, 6, 10]).

Since 2006, there have been a number of developments. First, several authors have expanded on the capacity of malicious users to produce false results. Second, the literature on remote job servers has stimulated investigations into massively repeated attacks. Third, some authors have looked at particular variable subsets which could make disclosure from coefficients feasible without full information.

The fact that all RDCs take guidelines from an old working paper which conflicts with more recent peer-reviewed papers produces an impression of divided opinion. Moreover, the understanding of relative risk in data access decisions has moved on substantially in the last ten years, albeit in a direction which supports the earlier paper rather than the later ones, with a greater importance placed on acknowledging the subjectivity of decisions and evidence-based reasoning [8, 11, 19, 20]. Hence there is a need for a review and clarification of evidence to produce a new synthesis which reflects more recent thinking in data access and provides unambiguous guidelines for data managers.

3. Exact identification of values through analysing linear regression results

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

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

or more compactly y=Xβ+u, where y, x, β and u are, respectively, Nx1, NxK, Kx1 and Nx1 matrices. We initially assume

- only genuine regressions are analysed (that is, $N>(K+1)$ and $K>1$); the X matrix may contain both factor and continuous variables, while y is continuous
- the data are the original risky variables containing at least some values which should not be released to the public
- all regression coefficients are published, along with associated statistics (estimated standard errors, goodness-of-fit measures)
- means of the variables used in any regression are published
- the researcher does not transform data specifically for the purposes of re-identifying a value through published results

We term the last the 'outsider' assumption; this will be relaxed later. We will analyse whether disclosure can occur solely as a result of the regression-specific published outputs. That is, we are not interested in other ways an intruder might identify data points from the data set (for example by analysing the means). Initially, we focus on whether an exact value can be determined; section 6 considers 'approximate' disclosure.

No assumptions are made about distributions; the following results depend upon the mathematical qualities of the estimator, not the statistical ones. Hence exact equations are used below, not expectations.

We employ an 'intruder' model of a malicious third party whose simple target is to discover any value or identity that should have been hidden. Although the use of intruder models has been strongly criticised as being overly cautious and largely irrelevant (eg [20]), it meets our purpose here where the aim is to examine worst cases.

One approach to the analysis is to assume that the intruder has a particular information set, and then analyse what information could be uncovered by combining this with the published outputs. We have not taken this approach as it is a methodological dead end: the range of starting assumptions is theoretically unbounded and therefore no definitive conclusions can be reached.

Instead we seek to identify, given the published information, what the intruder would need to know to be able to uncover a specific value. This, as will be shown generates a finite set of requirements for disclosure to occur, which can then be usefully evaluated. Every statistic has a theoretical disclosure risk, and so the definition of whether it is 'safe' or not can be rephrased as 'what information is necessary and/or sufficient to breach confidentiality, and is it likely that such information would be available?' We follow this line of approach.

3.1 Identification in a single regression: the general case

Consider the normal equations used to derive the K coefficients and the estimated standard error:

$$\hat{\beta} = (X'X)^{-1}X'y \quad \hat{\sigma}^2 = \left(y - X\hat{\beta}\right)'\left(y - X\hat{\beta}\right)/(N - K) \tag{2}$$

Direct disclosure from these K equations is, in general, not feasible unless the intruder already knows all but K values out of Nx(K+1) variables in X and y. While this is not impossible in theory, it is difficult to conceive of a real research situation with such an outcome. Disclosure by differencing is not feasible because of the multiplication and inversion.

Knowledge of the coefficients and mean of the explanatory variables does lead to the discovery of the dependent variable mean if it is not already known:

$$\bar{y} = \bar{X}\hat{\beta} \tag{3}$$

While the mean is in general treated as an 'unsafe' statistic because it can be directly informative and differenced [11], a mean reverse-engineered from regression coefficients and published means is lower risk by construction: 3 is usually accepted as the minimum frequency to prevent disclosure [1], and N>3 as K>1 and N>K+1. Even for this outcome to occur requires every mean except the one of interest to have been published. Again, this is not impossible (and it is more feasible than holding exactly (Nx(K-1))-K data points), but it does not reflect practice: researchers are expected to publish means of all continuous variables but not dummy variables, for instance.

Therefore we conclude there is no substantive disclosure risk of disclosing source values, either directly or through the reconstructed mean, in genuine regressions.

There are three exceptions to this rule.

Exception A1: single observation in a single category

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

$$y_1 = \hat{y}_1 = \sum_k x_{1k}\hat{\beta}_k \tag{4}$$

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.

Exception A2: a saturated conditional variable regression

This case was originally described by Reznek [16], who shows that 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 [17] demonstrate that this also holds for weighted regressions. If all the variables are strictly orthogonal (that is, $x_{ij}x_{ik}$=0 for all j≠k), then interactions are irrelevant; the non-interacted model is saturated (and falls into Exception A3, below). A special case would be where the researcher only includes one dummy as the sole regressor.

Ronning [21] argues that analysts 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; as noted above, any reconstructed mean is based on at least K+1 observations. These perspectives can be reconciled by considering that the saturated 'regression' is misclassified: it should be identified as a table - an 'unsafe' statistic in PBOSDC terminology - and assessed as such.

Exception A3: strictly orthogonal variables

Suppose X can be partitioned into two orthogonal variable sets:

$$X \equiv [X_A \; X_B\,] \quad X'_A\, X_B \; = 0 \tag{5}$$

where X_A and X_B are NxK_A and NxK_B. The orthogonality is mathematical, not statistical: that is, $E(X'_A X_B)$=0 is not sufficient. On defining a conformable coefficient vector, this leads to a partitioned estimate:

$$y = [X_A \; X_B] \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix} + u \tag{6}$$

$$\rightarrow \begin{bmatrix} \hat{\beta}_A \\ \hat{\beta}_B \end{bmatrix} = \begin{bmatrix} (X'_A X_A)^{-1} & 0 \\ 0 & (X'_B X_B)^{-1} \end{bmatrix} \begin{bmatrix} (X'_A y) \\ (X'_B y) \end{bmatrix} \tag{7}$$

In this case, the disclosure risk in each section is distinct. This could be the case where, for example, a researcher estimates a wage equation with all the variables interacted with a gender dummy. This is not sufficient to breach confidentiality per se, as noted above, but it is relevant where one orthogonal set consists of a single dummy variable defined as

$$x_A = 1 \rightarrow X_B = 0 \tag{8}$$

$$x_A = 0 \rightarrow X_B \; unrestricted \tag{9}$$

In this case it can be demonstrated [5] that, ignoring the constant term,

$$\hat{\beta}_A = \bar{y}_{x_A=1} \tag{10}$$

If there are multiple variables in an orthogonal set, then the individual means can be determined if the variables are fully interacted, as in Exception A2; and if there are multiple orthogonal sets, each one could present this problem. Again, this is the mean value being uncovered, but in this specific case it is possible to have the mean of a single observation (ie a source value).

3.2 Disclosure by repeated estimation

This section considers whether having the coefficients from repeated estimates poses a disclosure risk. We relax one 'outsider' assumption, that the intruder only has access to published information. Feasible scenarios observed in practice include:

- two researchers use the same dataset to estimate the same model but have slightly different exclusion criteria for observations
- a researcher estimates the same model twice with an additional observation, to show the effect of a problematic observation

Using the base regression (1), above, two cases are relevant.

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

Define y_0, and X_0 as Sx1 and SxK matrices of S additional observations, and β_0 as the coefficient vector to be estimated from the combined dataset. Even if X_0 is known, this does not lead to the direct identification of the dependent variables as

$$\hat{y}_o = X_o\hat{\beta}_o \neq X_o\hat{\beta}_o + e_o = y_o \tag{11}$$

where e_0 is the vector of estimation residuals. However,

$$\hat{\beta} - \hat{\beta}_0 = \left(X'X\right)^{-1} X'y - \left(X'X + X_0'X_0\right)^{-1}\left(X'y + X_0'y_0\right) \tag{12}$$

This is a system of K equations, so if there are 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:

$$y_0 = \left(X_0X_0'\right)^{-1} X_0X'X\left(\hat{\beta}_0 - \hat{\beta}\right) + X_0\hat{\beta}_0 \tag{13}$$

This has an exact solution if S=K; a similar case can be presented for known y_0 but S missing explanatory variables.

In general this solution requires full knowledge of the explanatory variables X and X_0, but there are plausible situations for which less knowledge is required.

First, Ritchie [12] notes that the estimated variance-covariance matrix (VCM) allows the unknown cross-product matrix to be recovered (which is why the VCM is an 'unsafe' statistic) as long as the estimated variance is available:

$$(VCM)^{-1}\hat{\sigma}^2 = \{(X'X)^{-1}\hat{\sigma}^2\}^{-1}\hat{\sigma}^2 = X'X \tag{14}$$

If X_0 is known and the VCM and estimated variance made available to the intruder, then there is now sufficient information to calculate y_0 without knowing all the values of X.

Second, as the regression line goes through the mean of the variables, the mean value of the new observations can be identified if the means of the explanatory variables are available:

$$\bar{y}_0 = \frac{N}{S}.\bar{X}(\hat{\beta}_0 - \hat{\beta}) + \bar{X}_0\hat{\beta} \tag{15}$$

If there is only one additional observation, this discloses the value of the additional dependent variable; this result can be derived in several ways (eg [1]). Note that disclosure of the mean y_0 does not require disclosure of the mean y.

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

- S=K, Sx(K-1) other values in the additional (X, y) set are known, and either all the explanatory variables are known, or the estimated VCM and model error from the initial regression are available (case B1a)
- S=1 and the explanatory variable means from the initial and augmented regression are known (case B1b)

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 [22] and Sparks et al [23] note that repeated estimation on subsets provide a potential solution to the normal equations. Define the Nx3 matrix X=[a b c]. Then the normal equations (X'X)β=X'y give:

$$\begin{bmatrix} a'a & a'b & a'c \\ b'a & b'b & b'c \\ b'c & c'b & c'c \end{bmatrix} \hat{\beta} = \begin{bmatrix} a'y \\ b'y \\ c'y \end{bmatrix} \tag{16}$$

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

$$\begin{bmatrix} a'a & a'b \\ b'a & b'b \end{bmatrix} \hat{\beta}_{ab} = \begin{bmatrix} a'y \\ b'y \end{bmatrix} \tag{17}$$

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 X'X and X'y.

This is not necessarily disclosive, but cross-tabulations are 'unsafe' statistics: information could be revealed, for example through interactions with sparse categories. This is a rare example of how an 'unsafe' statistic could, in theory, be generated from repeated estimation of a 'safe' statistic.

3.3 Exact identification using insider information

We now relax the other 'outsider' assumptions, that regressions are not created purely to breach confidentiality. Ritchie [5] explicitly excluded deliberate falsification of results, arguing that there were 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. Several authors (eg [22, 23, and 24]; [11, 25] review these papers using a consistent terminology) 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.

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 direct access to the data, only that coding is unrestricted.

Case C1: Known unique value of one or more 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 that x_1=m and $x_{2..N}$≠m, and wishes to know the value of y_1. Bleninger et al [24] characterise the options as

$$y_i = \alpha + \tilde{x}_i\beta + u_i \ , \ \tilde{x}_i = \frac{1}{|x_i - m| + \varepsilon} \tag{18a}$$

$$y_i = \alpha + x_i\beta + z_i\gamma + u_i \ , \ z_i = 1 \ iff \ x_i = m \qquad (18b)$$

Bleninger et al [24] label these 'artificial outliers' (18a) and 'strategic dummies' (18b), summarise the relative advantages (to the intruder) of these alternative approaches, and test 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' conforms 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-specified model.

Some RJS systems run checks on code to ensure that the frequency of regressors is sufficient. Sparks et al [23] note that matrix transformations can effectively hide the presence of single or sparse observations from simple tests, while still allowing the intruder to interpret directly the regression coefficients. 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. Although the examples in [24] have only a single regressor, disclosure with more variables could be achieved by artificially setting other variables to zero using the same techniques as above (eg multiplying through by 1-z_i).

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 (see below). 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.

Case C2: deliberate falsification based on rank

It would be possible to use rank to generate the disclosive value. Define m=Max(x_i). Then the techniques of case C1 could be applied to find the exact value of the largest observation. A variation could identify the smallest observation or, more generally, y_n where m=x_n. If m is unique, then an exact value of y_n can be determined; otherwise, it is the mean of those observations. This technique does not require any prior knowledge of the explanatory variables other than rank order.

4. Is exact disclosure feasible?

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

The above examples demonstrate feasible possibilities. Although authors have detailed more variants (eg [14]), to date no exceptions have been uncovered which do not fit those listed above. However, these cases have very specific information requirements, as summarised in Table 1 -

Case	**Regression conditions**	**Intruder knowledge required**	**Consequence**
A1 Unique explanatory dummy	Binary variable with only one non-zero observation	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 set	Single orthogonal binary variable	Orthogonal variable values	Mean of flagged dependent variables
B1 Direct differencing, S additional observations	Smaller sample (N) is exact subset of larger (N+S); same variable set	If S=K, all X values *or* original VCM; original sample means	Identification of S values if S=K and either X values or VCM available; identification of mean of additional Y values if only sample means available

B2 Differencing by repeated estimation	Same sample for all models; all combinations of variables estimated	None	Reconstruction of VCM
C1 Deliberate falsification based on unique values	Single variable regression only (artificial outliers); unique explanatory variable	Unique value of explanatory variable	Identification of one other variable associated with that unique value
C2 Deliberate falsification based on rank	Single variable regression only (artificial outliers)	Position of observation in ranking	Mean of variables at a particular rank

Table 1 Summary of problematic cases

We now consider how feasible these are.

4.1 Disclosure in genuine regressions (A1, A2, A3, B1, B2)

For A2, A3 and B2, there is little empirical support for the disclosure conditions being met in genuine research environments; even if the conditions are meet, the outcome is, at worst, another summary table.

Exception A2 sounds plausible: researchers have been observed running regressions just on categorical variables. However, it is rare to include all interactions unless this is a single set of mutually exclusive categories (eg highest level of education). For this to produce a 'table' outcome, it is essential that all interactions are included; otherwise only partial means are generated. Exception A3 appears less restrictive, by focusing on a subset of orthogonal variables; it is not uncommon for researchers to interact all variables with a dummy, as this simplifies the testing of restrictions. However, this case requires that at least one variable is orthogonal to every other one; and even in this case, only the mean of the dependent variable for the isolated variable is revealed. It is difficult to rationalise why a researcher would include such a variable or set of variables.

Case B2 makes little sense. Researchers include or exclude variables systematically; there is no research value in estimating exhaustive non-hierarchical combinations of variables.

Exceptions A1 and B1 are feasible. A researcher may include a dummy with only one positive value, particularly if this is complex data and the uniqueness of the observation is not spotted; and running the same regression on a subset of the data is an accepted way of analysing the impact of sampling.

Whether this is useful to an intruder is less certain. Both require substantial additional knowledge. In the case of A1, the intruder needs to know that one observation is unique, and all the other explanatory variables for that observation; this is only likely to be the case if all the explanatory variables are in the public domain and the exact sample is known. For B1, all other observations or the VCM (which is unsafe and so not automatically released in controlled environments, and unlikely to be published) must be known.

The one realistic situation is under exception B2 where the intruder knows the means of the two variable sets X and X_0; this is possible as researchers will often publish means in the data description. This enables the means of the additional observations to be identified. A mean is an 'unsafe' statistic: disclosure risk is non-negligible[6], and so specific instances of a mean should be assessed for risk before release. However, it would be unusual for the researcher to describe all except the dependent variable; it is far more likely that a researcher would omit explanatory variables (for example, seasonal, geographical or event dummies) from a list of variable means. If the researcher includes means for all variables, as is not uncommon when few variables are used, then exception B2 is uninformative; the information is already available.

In summary, the likelihood of disclosure to 'outside' intruders 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. These conditions are not fulfilled by genuine research activity.

4.2 Disclosure by 'inside' intruders (C1, C2)

The likelihood of disclosure by 'inside' intruders is higher, because genuine regressions are no longer considered; the regression outputs under consideration have been falsified to produce a specific result. The recent literature on insider attacks reflects an interest in remote job servers (RJSs). Currently all existing RJSs have some human oversight, but the ideal RJS is fully automatic. In this case, the possibility arises of both 'insider' attacks and 'outsider' attacks based on multiple repeated estimation. Hence, O'Keefe and Chipperfield [25] argued regression should not be seen as risk-free. Note however that even these require some very specific conditions. Bleninger et al [24] demonstrate that the ability to fake results is crucially dependent on a single value being both unique and known to be unique. C2 shows that uniqueness can be created such that the means of groups of values are known, but this requires an exact knowledge of rankings.

However, these 'inside' scenarios are not problems of OSDC but of access management.

The widely-used 'five safes' framework [26] shows that safe use of sensitive data is affected by both statistical and non-statistical factors. OSDC assumes that (a) research results are genuine, but that (b) mistakes are made. There are routes, such as C1 and C2 above, for an ill-intentioned researcher to falsify analyses or the presentation of results; but this is not what OSDC is designed to uncover; these are management problems.

Consider an RDC where a researcher, learning that regression is always approved but small table cells are not, chooses to hide small-cell tabular output as regression output using the saturated model noted above. This is a failure of researcher training and oversight. Alternatively, consider an RJS designed to be available to the general public, with 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. Some authors [23, 25] suggest that exceptions C1 and C2 could be avoided in RJSs by banning the creation of new or transformed variables; but although this may have a severe impact on genuine research, it cannot stop sophisticated programmers. On the other hand, a management solution (offering menu-based analysis rather than direct coding) does deal effectively with the inside intruder scenarios.

In the case of regressions, an additional factor is that the results presented here only work on a very specific set of data and conditions. They require the ill-intentioned user to be malicious, extremely well-informed, and willing to waste time on complicated measures when other, simpler, mechanisms will deliver much more information.

In short, the RJS literature suggests that regressions are risky because they assume unrestricted access to unlimited data and outputs. However, it is clear that non-statistical mechanisms for managing confidentiality offer better solutions with low impact on research: if there is a significant chance of a user deliberately falsifying results, the data manager would be wiser to invest effort in better accreditation and access procedures than in trying to restrict the outputs of genuine research. Similarly, if the RJS is genuinely designed for open access by unknown users, then basic IT security would suggest limiting access to a finite sets of commands through the use of menu-based interfaces; again, the solution is to improve the safety of the setting, not to limit outputs.

5. Popular questions about disclosure risk

Since the 'safe regression' assumption began to be used and disseminated, a number of methodologists have raised concerns about special cases other than those above.

5.1 How important is the release of the full coefficient set?

It is clear from the results presented above that the full set of estimated coefficients is necessary for disclosure; therefore, any linear model which includes incidental (implicitly included but unpublished) parameters poses no disclosure risk. For example, consider the longitudinal model (with a constant term included in the X variables):

$$y_{it} = X_{it}\beta + u_{it} \quad u_{it} \equiv \alpha_i + e_{it} \quad i = 1..N, t = 1..T \tag{19}$$

The N individual-specific elements are of little interest per se, and the model is transformed to avoid estimating N individual intercepts. Hence, for example, it is no longer possible to identify an observation under scenario A1 even if all the X values are known, as

$$\hat{y}_{it} = X_{it}\hat{\beta} + \hat{\alpha}_i \tag{20}$$

and the last term is not published. Nor is it possible to generate all interaction terms, or to carry out repeated estimation. This result can be generalised: any linear regression model where parameters contribute to the line of best fit but are not explicitly generated poses no disclosure risk under the scenarios described above. This usefully includes a number of classes of models, such as multilevel models. Note that it does not matter whether α_i is estimated as a fixed or random effect; this result depends on the completeness of the mathematical expression, not on the expected value of a statistic.

Ritchie [5] used this result to argue that researchers should reserve publication of at least one coefficient; this could be done without much impact because, for example, intercepts or time dummies are often of little direct interest. This suggestion has been followed in some guides (eg [6, 10]), but not others (eg [4, 8]). Given the low risk associated even with full disclosure of coefficients, this condition seems excessive and our advice is that this should be a guideline for good practice, not a requirement. This is more likely to avoid negative engagement from researchers, which is known to have a much higher risk than the risk from model estimates.

5.2 Do data transformations and multistage regressions increase or lower risk?

The linear regression described above is the best case for an intruder; the raw variables are directly of interest. Genuine transformations of the data (that is, not done specifically to deceive) cannot increase the risk and almost certainly reduce it, as they create uncertainty over the data values, reduce the chance of one of the exceptions, or both.

For example, consider a generalised error variance:

$$E(uu') = \sigma^2 \Omega \quad PP' \equiv \Omega^{-1} \tag{21}$$

In generic robust estimators, an estimate for Ω (and hence P) is derived from a simple regression, and then used to transform the data so that

$$\widehat{PP}' = (y - X\hat{\beta})(y - X\hat{\beta})' \tag{22}$$

$$\hat{P}y = \hat{P}X\beta + \hat{P}u \tag{23}$$

This is fundamentally the same as equation (1) but with variables transformed:

$$z = W\beta + e \tag{24}$$

where

$$z \equiv \hat{P}y \quad W \equiv \hat{P}X \quad e \equiv \hat{P}u \quad E(ee') = E\left(\hat{P}uu'\hat{P}'\right) = \sigma^2 I \tag{25}$$

For exceptions relying upon knowledge of the explanatory variables, direct identification is no longer possible. For exceptions which rely upon a single orthogonal variable, these are unaffected.

5.3 Is statistical quality associated with risk?

Some suggestions have been made which relate to the quality of the data, but these confuse statistical properties with disclosure:

- **Outliers** 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 even less disclosive than other observations
- **Influential points** are outliers with a significant impact on the regression line, and so differences between regressions are most likely

to be (a) discernible and (b) published; however, an influential point cannot lead to exact disclosure as there must be an interaction between variables for the regression line to shift (A1-A3 are therefore not relevant), unless all the explanatory variables are known (exception B1)

- **Multicollinearity** and **measurement error** increase estimated errors and make attribution of effects to particular variables more difficult; but neither affects the prospect of disclosure as this is based upon the mathematical projection of the variables, not on statistical qualities.

In all cases, quality issues need to be separated from SDC issues. Data problems leading to a highly skewed distribution theoretically lead to more risk of approximate disclosure (see below), 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 zero residual 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. Eurostat guidelines [6, 7] take the more pragmatic line that any regression must have at least ten residual degrees of freedom. 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 (such as N/K>3), while equally arbitrary, may be more appropriate.

5.4 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. For example, a point frequently made is that residuals may be much more disclosive than regression coefficients [6, 10, 13, 17, 27]. This is a particular problem for RJSs, where researchers might want to see distributions of residuals as part of the diagnostics, as they cannot see the source data. Hence, much of the work around RJSs has considered how to present these residuals safely.

This is not a problem of regression coefficent risk, but tabulation risk. Residuals are microdata, albeit modelled, and so depictions of them are treated as 'unsafe' statistics just as graphs, tabulations or distributions of source microdata would be. As for any tabulation, the data owner would be concerned about whether the distributed statistic shows point values 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 output needs to be given.

Similarly, Ritchie [12] defines the variance-covariance matrix (VCM) as an 'unsafe' statistic. Although formally an estimated value, it was noted above that the VCM can be used to derive the cross-product matrix X'X on the reasonable assumption that the estimated standard error is published. The cross-product matrix is 'unsafe' because interactions with dummy variables can generate group means. However, the decision to publish regression coefficients does not require the publication of the VCM; this is a separate decision to release, in effect, a cross-tab.

6. Evaluating the likelihood of approximate disclosure

Sections 3 described exact identification of values in theory; section 4 argued that actual re-identification is extremely unlikely because it relies upon a large amount of information on the variable distribution and structure. Corscadden et al [18] note that, in practice, this massively overstates the likelihood of making accurate predictions unless one of the exceptions listed also holds, even in the intruder's best case of estimation with all the explanatory variables public and a confidential dependent variable.

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 has been identified in the medical literature, where the possibility has been raised of the outcomes of regression-based prescribing models being used to predict hidden factors such as genomic type [28]. As a result, some researchers have suggested using differental privacy algorithms on prescribing models. This is extremely problematic, because the implication of only creating differentially private regression coefficients directly reduces prescribing efficacy, with potentially life-threatening consqeuences.

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 for a specific observation x_1, and summary statistics on the regression. No distributional assumptions are made, but if the fitted equation is mis-specified the results below under-estimate variances and over-estimate the accuracy of approximations.

Using the same notation as before, suppose an intruder knows x_1 and seeks an approximate value for y_1. The residual e_1 has variance [5]

$$V(e_1) = \sigma^2(1 - x'_1(X'X)^{-1}x_1) \tag{26}$$

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→x_1x_1'), and approaches the standard error when the observation has a negligible impact (x_1→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 data owner 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 coefficients are used for prediction by the application of a new set of observations (y_0, x_0) from the same distribution, then a similar limit can be derived [29]:

$$V(e_0) = \sigma^2(1 + x'_0(X'X)^{-1}x_0) \tag{27}$$

The intuition 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. The predictive error cannot be reduced below this level.

It is not necessary for the intruder to know X'X; the confidence interval can be calculated by using common summary statistics. Defining TSS and ESS as total and estimated sums of squares, and noting that

$$\hat{\sigma}^2 = (TSS - ESS)/(N - K) \tag{28}$$

$$R^2 = TSS \,/\, ESS \tag{29}$$

$$ESS = \hat{\beta}'X'X\hat{\beta} \quad \rightarrow \; X'X = \left(\hat{\beta}\hat{\beta}'\right)^{-1}.ESS \tag{30}$$

it can be shown that

$$x'_1(X'X)^{-1}x_1 = \frac{(\sum_k x_{1k}^2\widehat{\beta}_k^2)R^2}{(\widehat{\sigma}^2(N-K)(1-R^2))} \tag{31}$$

And so

$$\widehat{V}(e_1) = \hat{\sigma}^2\left(1 - \frac{(\sum_k x_{1k}^2\widehat{\beta}_k^2)R^2}{(\widehat{\sigma}^2(N-K)(1-R^2))}\right) \tag{32}$$

Hence, if the intruder knows the value of x_1, then it is possible to calculate minimum confidence intervals for predicted values from the summary statistics.

This is particularly relevant for RJSs; it would be possible to generate an automatic test for the width of confidence intervals on both in-sample and out-of-sample observations, so that particularly accurate regressions could be blocked if so desired. This incidentally would also address the issue of inside intruders creating spurious regression as the exact fits would be identified.

Note however that this result still depends upon a detailed knowledge of the explanatory variables. Hence even approximate disclosure has a notable information requirement, and the data manager's confidence intervals are likely to be much smaller than an intruder's.

Finally, Corscadden et al [18] 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. Although no general relationship between R^2 and predictive uncertainty has been derived, exploratory work by Statistics New Zealand suggested that, in empirical tests, extremely high R^2s (>0.99) were necessary to breach rules on approximate disclosure.

7. Conclusions

This paper has reviewed the opportunities for determining confidential information from regression outputs. This is an important topic, because the efficiency of RDCs, the feasibility of RJSs, and confidence in SUF releases depend upon being able to make quick, reliable and accurate 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 (and the safe-regression rule) by ONS in 2003 cut the target clearance time for results from two weeks to two days. This discussion therefore has a direct impact on researchers and data owners.

Regression coefficients are also used for operational purposes, for example in the US for estimating prescription doses. Some authors (eg [30]) have proposed applying differential privacy adjustments to these, arguing that theoretical risk exists and must be contained, even though the evidence suggests this may be clinically unwise [28]. But as Section 6 showed, there are simple practical tests to assessment approximate risk, rather than automatically adding unnecessary noise to a potentially life-threatening decisions.

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, and the intruder may have access to some of the underlying data. The purpose is to show that, even in an intruder's preferred scenario, the 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.

Such ideal conditions are unlikely; and practical experience in various countries has not shown regression analyses to be problematic. This paper has demonstrated that this is not a happy accident but an expected consequence, and data owners can design access mechanisms with this in mind.

Nevertheless, the widespread adoption of [5] by practitioners and RDC managers caused concerns amongst some SDC specialists: the idea that disclosure risk could be analysed independently of the data being analysed seemed inherently suspicious. Initially, there was substantial resistance, with the hypothetical 'exceptions' in the paper being used to deny the existence of a useful general rule on theoretical grounds. Since 2006 there have been three key developments; none change the basic premise and key conclusions of the earlier paper, but they do provide context for institutional management.

First, the number of exceptions has been expanded by researchers. However, the discussion above shows that the exceptions can be reduced to a small set of extreme cases. It also highlights the extensive information requirements of any intruder, and the improbability of such requirements being met in real situations.

Second, the growth in analysis in controlled facilities has made practical considerations of resource management increasingly important. Although worst-case theoretical models still dominate the SDC literature, operational and strategic decisions are increasingly based upon the evidence-based balance-of-risks models described here.

Third, several authors have investigated the risks in unrestricted access to RJSs. Authors discussing RJSs generally assume that researchers are malicious, even when the same authors consider users of RDCs not to be malicious (see [25, 27] for example). These have undoubtedly shown the potential for deliberate manipulation of regression models, but they also show that malicious intent is a prerequisite for misuse. There remains no regression risk in genuine research use; managerial problems are best dealt with by operational, not statistical, measures.

Thus this paper's affirmation of the classification of linear regression coefficients as 'safe statistics' is both timely and important. This paper has shown that there are exceptions to this statement, but that theoretical possibilities have little practical relevance. This is true for RJSs and SUF outputs, as well as for the RDCs. The paper has also demonstrated the need to separately consider non-statistical elements when considering the likelihood of a solution.

This judgement is explicitly subjective. We have argued that certain outcomes are 'likely' or 'infeasible'. Such values are not quantified, but since 2003, many thousands of regression outputs have been manually checked in the various RDCs worldwide that operate this rule, and RJSs such as LISSY which give automatic approval; there is no evidence to date of regression coefficients leading to a breach of confidentiality. This does not constitute proof that disclosive outputs cannot happen; only that all the accumulated evidence supports the contention that these pose no meaningful risk. For data managers required to demonstrate reasonable precautions against data breaches, this is more relevant than theoretical possibilities.

One recommendation of [5] has not stood the test of time: that one or more coefficients should be supressed in publications to guarantee privacy. With no meaningful exceptions to the rule coming to light, this now seems overly restrictive. We therefore propose that, while researchers be encouraged to suppress unnecessary coefficients (such as incidental dummy variables), this is done for reasons of clarity rather than SDC.

We have focused on linear regression coefficients. Ritchie [5] briefly discussed non-linear models; he argued that intuitively the indirect interpretation of coefficients made them inherently more non-disclosive but then gave a simple example where a non-linear model is theoretically more disclosive than its linear counterpart. We have avoided non-linear models because the range of functional forms is so large (whereas all linear models must be of the form y=a+bx). However, we believe that the approach adopted in this paper would allow many more classes of models to be reviewed and classed as 'safe' or not, and this would be a productive further area of research.

Acknowledgements

I am grateful to the referees (one of whom suggested the much improved title), Richard Chandler and Don Webber for comments to improve the structure of the paper, clarify terms and identify jargon. Christine O'Keefe's views on disclosure risk in RJSs were influential in rephrasing the debate around institutional factors. I am particularly indebted to the various methodologists who have argued the toss on this subject with me, particularly those who were more helpful than just 'something must be wrong'. All remaining errors are mine.

References

[1] Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nord-holt, E., Seri, G. and De Wolf, P. (2010). Handbook on Statistical Disclosure Control. ESSNet SDC. http://neon.vb.cbs.nl/casc/.\SDC_Handbook.pdf

[2] Ritchie F. (2007) Statistical disclosure control in a research environment. Mimeo, Office for National Statistics. Edited and reprinted as WISERD Data and Methods Working Paper no. 6 (2011).

[3] Ritchie F. and Elliot M. (2015). Principles- versus rules-based output statistical disclosure control in remote access environments. IASSIST Quarterly 39:5-13

[4] Eurostat (2016) Self-study material for the users of Eurostat microdata sets. http://ec.europa.eu/eurostat/web/microdata/overview/self-study-material-for-microdata-users

[5] Ritchie F. (2006) Disclosure control of analytical outputs. Mimeo: Office for National Statistics. Edited and reprinted as WISERD Data and Methods Working Paper no. 5 (2011).

[6] 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

[7] Bond S., Brandt M., de Wolf P-P (2015) Guidelines for Output Checking. Eurostat. https://ec.europa.eu/eurostat/cros/system/files/dwb_standalone-document_output-checking-guidelines.pdf

[8] Department of Social Services (2016) Data Access Strategy: final report. Australian Department of Social Services, June.

[9] Statistics New Zealand (2015). Microdata Output Guide (Third edition). Available from www.stats.govt.nz.

[10] O'Keefe C., Westcott M., Ickowicz A., O'Sullivan M. and Churches T. (2015) Guidelines for Confidentiality Protection in Public Health Research Results. CSIRO.

[11] Ritchie F. (2014) Operationalising safe statistics: the case of linear regression. Working papers in Economics no. 1410, University of the West of England, Bristol. September

[12] Ritchie F. (2008) Disclosure detection in research environments in practice. In: Work session on statistical data confidentiality 2007, Eurostat; pp399-406

[13] Reiter, J. (2003). Model diagnostics for remote-access regression servers. Statistics and Computing. 13:371–380

[14] Chipperfield J. and Yu F. (2012) Protecting confidentiality in a remote analysis server for tabulation and analysis of data. In: Work session on statistical data confidentiality 2011, Eurostat

[15] O'Keefe C., and Shlomo N. (2012) Comparison of remote analysis with statistical disclosure control for protecting the confidentiality of business data. Transactions on Data Privacy 5:403–432

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

[17] 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

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

[19] Skinner C. (2012) Statistical disclosure risk: separating potential and harm. International Statistical Review. 80(3):349–368

[20] Hafner H-P., Lenz R., Ritchie F., and Welpton R. (2015) Evidence-based, context-sensitive, user-centred, risk-managed SDC planning: designing data access solutions for scientific use. In: Worksession on Statistical Data Confidentiality 2015, Eurostat.

[21] Ronning G. (2011) Disclosure risk from interactions and saturated models in remote access. IAW Discussion Papers No. 72, June.

[22] 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. Statistical Science 20(2):163-177

[23] Sparks R., Carter C., Donnelly J., O'Keefe C., Duncan J., Keighley T., and McAullay D. (2008) Remote access methods for exploratory data analysis and statistical modelling: privacy-preserving analytics. Computer Methods and Programs in Biomedicine 91(3):208–222

[24] Bleninger P., Drechsler J., and Ronning G. (2011) Remote data access and the risk of disclosure from linear regression. Statistics and Operational Research Transactions, Special Issue: Privacy in Statistical Databases. 35:7-24

[25] O'Keefe C. and Chipperfield J. (2013) A summary of attack methods and confidentiality protection measures for fully automated remote analysis systems. International Statistical Review 81(3)426–455

[26] Desai T., Ritchie F., and Welpton R. (2016) The Five Safes: designing data access for research. Working papers in Economics no. 1601, University of the West of England, Bristol. January

[27] O'Keefe C., Westcott M., Ickowicz A., O'Sullivan M. and Churches T. (2014) Protecting confidentiality in statistical analysis outputs from a virtual data centre. In: Work session on statistical data confidentiality 2013, Eurostat.

[28] Fredrikson M., Lantz E., Jha S., Lin S., Page D., and Ristenpart T. (2014) Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Conference on Security Symposium, pp 17–32.

[29] Verbeek M. (2017) A Guide to Modern Econometrics, 5e. Wiley.

[30] Wang Y., Si C., and Wu X. (2015) Regression model fitting under differential privacy and model inversion attack. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence.