University of the West of England
BRISTOL

Faculty of Business and Law

Operationalising 'safe statistics': the case of linear regression

Felix Ritchie

Department of Accounting, Economics and Finance, University of the West of England, Bristol, UK

Economics Working Paper Series
1410

Operationalising 'safe statistics': the case of linear regression

Felix Ritchie

Department of Economics, University of the West of England, Bristol

Abstract The recent growth in research access to confidential government microdata has prompted the development of more general 'output-based statistical disclosure control' (OSDC) methods which go beyond tabular protection. Central to OSDC is the concept of 'safe/unsafe statistics', allowing researchers and facility owners to make informed judgments about the types of research output that pose a disclosure risk. While increasingly accepted in specialist environments, in the wider community this novel approach causes some concern: how can 'safe' be unconditional? This paper therefore demonstrates the new approach using linear regression, a key research output, as an example. In doing so, the paper reconsiders the objectivity of SDC decision-making, arguing that 'safety' be explicitly acknowledged as a relative concept.

Key words: statistical disclosure control, confidentiality, safe statistics, regression.

JEL Classification: C18, C20, C81, C89

Address for correspondence: Felix Ritchie, Bristol Business School, University of the West of England Bristol, Frenchay Campus, Coldharbour Lane, Bristol BS16 1QY. Email: felix.ritchie@uwe.ac.uk.

1. Introduction

In the last decade or so, a major growth area in the social science research has been driven by the increasing availability of confidential data from government sources for research through controlled environments. The growth in controlled environments such as research data centres (RDCs) enables researchers to access data at previously unreleased levels of detail, as well as linking datasets. One particular factor has been the growth of remote access systems: since the creation in 2002 in Denmark of the first remote RDC designed for general access to confidential government data, remote RDCs have been set up in many European countries, North and South America, and Oceania; the latter has also been investing in remote job systems. The UK alone has two national remote RDC networks for general purpose research (that is, not tied to any particular type of data or research) and is in the process of building a third to improve research access to administrative government data.

This expansion in research access to confidential data causes problems for facility owners. Traditionally, statistical disclosure control (SDC) has focused on protection of tabular data and on creating safe microdata; these are the core outputs of National Statistical Institutes (NSIs) which drive much, although not all, of the research in this field. However, the statistical outputs produced by researchers cover a much wider variety of statistics: regressions, odds ratios, factor models, kernel densities, and so on. Even the magnitude and frequency tables that researchers produce are likely to be different: they are more often used to illustrate characteristics of the data selected for analysis, and hence reflect subjective preferences for particular observations.

Whilst some statisticians considered confidentiality of non-tabular outputs (eg Gomatam et al, 2005; Corscadden et al, 2006), it is clear that there was no systematic approach to checking research outputs for disclosure. This led to proposals for a field of 'output SDC' (OSDC) to counterpoint the 'input SDC' of data anonymisation, with the aim of generating a framework within which all types of output could be analysed and classified; Ritchie (2007) provides an early summary of the problem.

Central to OSDC is the concept of 'safe/unsafe statistics' (Ritchie, 2007 and 2008). A 'safe statistic' is an inherently non-disclosive statistical output, whereas an 'unsafe statistic' means that the output in question can only be asserted as non-disclosive in a specific context. The successful classification of outputs into unsafe and safe statistics has significant resource implications for the researcher and the facility owner, and has proved its effectiveness in a number of settings; since 2010 it has been Eurostat recommended practice for OSDC (see Brandt et al, 2010).

However, to date only a relatively small number of outputs have been classified, largely because classification of a core set of outputs was done by those behind the original concepts. Those coming later to this approach have tended to take these classifications for granted. Moreover, the novelty of the approach has caused some concern to those familiar only with more traditional approaches: how can a definition of 'safe' be unconditional? As Ritchie (2014) notes, uncertainty tends to lead to the rejection of novelty.

This paper therefore demonstrates how the classification is determined, taking the example of a linear regression, a key output for researchers. As part of this process, the paper demonstrates the essentially subjective nature of all SDC and argues for a more realistic perspective costs and benefits of data protection.

The next section defines the mean of ‘safe/unsafe statistics’, and the classification process in outline. Section 3 applies this approach to linear regression: it explains the role of functional form as an underlying principle, it shows how exceptional cases are identified and reviewed, and it describes the role of evidence in deciding whether theoretical exceptions should be considered. As a counter-example, Section 4 briefly demonstrates the 'unsafeness' of magnitude tables. Section 5 discusses what 'safe' means in the wider context of data access, and the implications for the risk management strategies of facility owners. Section 6 concludes.

2. Safe/unsafe statistics

The concept of ‘safe statistics’ was formally defined in Ritchie (2008)[1]; see Brandt et al (2010) for the formal Eurostat definition and SDS (2011) for a practical definition. The concept is straightforward: researchers produce an infinite range of outputs, but these fall within a much more limited set of ‘types’ of output: frequency or magnitude tables; actual and estimated probability; linear regression model; residuals; simple and weighted indexes; correlation matrices, and so on. More complex outputs can be built from these: simultaneous equation systems, odds ratios, etcetera.

The key to the classification system is the focus on the functional form of the statistic. This is what enables it to break away from questions about whether this or that data provides a disclosure risk. If the mathematical characteristics of the statistic mean that identification of separate data points is not feasible, then it is ‘safe’. ‘Safety’ is defined irrespective of the data used.

If the type of output can be shown to be ‘safe’ then no further analysis of specific outputs is required. For example, if an odds ratio is declared ‘safe’, then researchers can be confident that all the odds ratios produced in their research will be acceptable for general release. At the same time, the owner of the controlled environment need expend no resources checking these outputs, as they have been pre-determined as acceptable for release. In contrast, an output defined as an ‘unsafe’ statistic cannot be released unless it can be shown, in the specific context, that the output presents no disclosure risk. This involves resource expenditure by both researchers and facility owners.

This classification has two positive outcomes:

- Researchers are encouraged to produce outputs with no disclosure risk, as the certainty and speed of approval is better for ‘safe statistics’
- Owners of controlled facilities expend resources checking only those outputs which present a notable disclosure risk; this makes both manual and automatic checking simpler

As a result, both facility owners and researchers spend the most resources on the most risky outputs.

Note that the classification can be challenged: for example, an instance of an ‘unsafe’ frequency table can be shown to be safe, perhaps by having many observations. However, while it is common for notionally ‘unsafe’ statistics to be declared safe in particular circumstances, it is extremely

[1] The concept was developed at the UK Office for National Statistics (ONS) from 2003, but initially only used in internal training documents. Ritchie (2008) refers to ‘safe outputs’; the term was changed later to avoid confusion with use of the former phrase in the ‘five safes’ data access model.

unlikely that a ‘safe’ statistic need be assessed for context-sensitive disclosiveness; otherwise, the statistics should not have been declared ‘safe’ (see Section 5 below).

3. Classification of linear regression coefficients

There are four stages to classification:

1. Define the functional form
2. Identify the disclosure potential in the functional form by considering
    a. Can an element of the statistic relate to a single data point?
    b. Can the statistic be differenced to reveal values?
    c. Anything else, specific to the statistic
3. If the provisional definition is ‘safe’, identify and evaluate ‘special cases’
4. Draft guidelines

In this section we use the example of simple linear regression coefficients. Ritchie (2012) provides a more detailed discussion of the issues and derivation of results.

3.1 Defining the functional form

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

$$y = X\beta + u$$

where y, X, β and u are, respectively, Nx1, NxK, Kx1 and Nx1 matrices with N>(K+1) and K>1. The estimated coefficient vector

$$\hat{\beta} = (X'X)^{-1}X'y$$

is the functional form to be analysed.

3.2 Identifying the disclosure potential in the functional form

In general, a coefficient cannot be related to a single data point. Even in the case of a single explanatory variable (X= x and a constant),

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

From the single estimated coefficient, it is not possible to identify any values. It is also not possible to difference estimated coefficients. Consider adding an additional set of observations so that the variable matrices y_a and X_a are now (N+1) x 1 and (N+1) x K, respectively. Then

$$\widehat{\beta_a} = ({X_a}'X_a)^{-1}{X_a}'y_a$$

But this does not help to identify additional observations:

$$\hat{\beta} - \widehat{\beta_a} = (X'X)^{-1}X'y - ({X_a}'X_a)^{-1}{X_a}'y_a$$

The ‘anything else’ test allows transformations to be applied to the statistics where it is thought that these might enable identification. For example, in the case of regression coefficients, the inverse matrix prevents differencing. However, if the variance-covariance matrix $\hat{V} = (X'X)^{-1}\widehat{\sigma^2}$ was available, then

$$\widehat{V\beta} = X'y\widehat{\sigma^2} \qquad \text{and} \qquad \widehat{V\beta} - \widehat{V_a\beta_a} = X'y\widehat{\sigma^2} - X_a{}'y_a\widehat{\sigma_a^2}$$

seems to linearise the difference. However, the estimated standard deviations depend on the interaction between variables, so the linearised difference is still the result of a weighted sum. Hence this transformation has no meaningful impact, and the preliminary classification of regression coefficients is ‘safe’.

3.3 Identifying and evaluating special cases

Ritchie (2012) identifies a number of potential special cases, set as sets of orthogonal variables, single observations on a category, and regression on dummy variables. Ritchie (2012) demonstrates that although these are all theoretically feasible, the information set required by an intruder trying to extract information from regression coefficients is infeasibly high.

Take an example: one observation x_b, is the only positive observation for a particular category in a set of dummy variables. The error term in this case is identically zero: $y_b = x_b\hat{\beta}$. If an intruder knew that there was only one observation in a category, and the value of all the other regressors for that observation, then the exact value of the dependent variable is revealed. Is this a likely occurrence for someone reading statistical output? It is difficult to conceive; there is no indication in the t-statistic that this is a unique value. A possibility is that all the regressors are public variables, and it known with certainty that only one observation, included in the data has one value. This seems an unlikely scenario, particularly as it implies that this observation has no explanatory value and should be omitted.

Ritchie (2012) also examines the possibility of approximate (rather than exact) identification, where R^2 can provide an upper bound for the accuracy of estimates. This is harder to evaluate as a very wide range of possible outcomes could be considered (for example, one exceptional value amongst many small values). A difficulty with specifying these alternatives is that they make little statistical sense. For example, for an outlier to be identifiable with some certainty it needs to be orders of magnitude larger than all other records on all variables. Ritchie (2012) cites studies by Statistics New Zealand which concludes that R^2s of 0.99 are needed before approximate identification becomes a realistic probability.

In summary, while there are theoretical cases where regression coefficients could be disclosive (note, even then disclosiveness has to be assessed in context), these cases are not meaningful in a practical research environment.

3.4 Draft guidelines

Given this analysis, Ritchie (2012) concludes that linear regression coefficients should be identified as ‘safe’; exceptions are not deemed sufficiently relevant in practice.

Ritchie (2012) highlights that even the exceptions disappear if some of the coefficients are not released. This has important implications for models which estimate 'incidental' parameters; for example, fixed-effect panel data models are non-disclosive per se, as the fixed effects are never published. Brandt et al (2010) and SDS (2011) use this to argue that an extra-safe regression rule should be used, whereby researchers are required to hide at least one published coefficient (for example, the constant or time dummies).

Note that there were qualifications: that (N+1)>K and K>1; that is, this is a feasible regression and not an equation set. Brandt et al (2012)suggest that a more sensible (if entirely arbitrary) rule is 'at least ten degrees of freedom'. While this has no strict statistical basis (and should have no impact on research), it serves to highlight the assumption that these are 'genuine' regressions being estimated.

Having disposed of the simplest linear regression, it is then straightforward to show that more complex regressions (for example, weighted, multi-stage, robust or simultaneous models) have even less scope for disclosure risk. All add additional non-linearity to the simpler result and increase the convolution of variables. From this Ritchie (2012) generalises that linear regression coefficients are 'safe' irrespective of the estimation method, even though not all methods have been assessed. Hence a relatively specific question (is the simplest linear model disclosive?) has been used to generate a general which covers a vast amount of research outputs.

4. Classification of magnitude tables

As a counter-point to the case of linear regression, consider a simple magnitude table of sums of observations.

4.1 Define the functional form

For any one table cell, the value is

$$t_N = \sum\nolimits_N x_i$$

4.2 Identify the disclosure potential

This is potentially disclosive if the cell size is less than three, but not necessarily; it depends upon the data. Hence a rule that N>2 might guard against direct disclosure but prevents legitimate totals being published. Moreover, even in tables with N>2 it is quite likely that one observation dominates the total, and so that observation can be approximately disclosed; this is a particular problem with business data. The total therefore fails the first test, preventing primary disclosure.

Applying the second test, add an extra observation to the cell:

$$t_{N+1} = \sum\nolimits_{N+1} x_i$$

Therefore differencing allows exact disclosure: $t_{N+1} - t_N = x_{N+1}$, and the totals in the table fail the second test too. Thus each table cell presents a significant risk of disclosure; one must evaluate disclosure risk by examining the data and the context of disclosure; this is therefore an 'unsafe' statistic.

It could be argued that the summation is too simple: it is too sweeping to damn magnitude tables because of the unsafeness of the worst-case scenario. Consider instead a mean, which is just as likely to be published as a sum:

$$m_N = \frac{1}{N}\sum\nolimits_N x_i$$

Then differencing, for example leads to:

$$m_{N+1} - m_N = \frac{1}{N+1}\sum\nolimits_{N+1} x_i - \frac{1}{N}\sum\nolimits_N x_i = \frac{1}{(N+1)}x_{N+1} - \frac{1}{(N+1)}m_N$$

And so

$$x_{N+1} = (N+1)m_{N+1} - Nm_N = t_{N+1} - t_N$$

Thus publication of means is a special case of totals; it is non-disclosive only if frequencies are not published, which in general seems a heroic assumption on which to base confidentiality. A weighted average is less likely to be disclosive, as it requires knowledge of the data; but if weights are known and not unique to each observation (as in, for example, the stratified sampling of business surveys), the potential arises again. Thus, for magnitude tables, it is important to demonstrate that a particular output cannot be reduced to a simple summation; again, knowledge of the specific data is required to determine whether the output is safe or not.

4.3 Identify special cases

There is no need for this as the provisional classification of the base case is 'unsafe'.

4.4 Draft guidelines

As this is an 'unsafe' output, guidelines should be drawn up to help output checkers make decide*, in specific contexts*, what makes a table releasable.

Seen in this context, it is clear why the bulk of SDC research focuses on frequency and magnitude tables. These outputs, the primary product of NSIs, are inherently 'unsafe' and need context-sensitive checking. Note that the methods used to ensure tables are safe (minimum numbers of observations, recoding, perturbative procedures etc) do <u>not</u> change the fundamental 'unsafeness' of the statistics; they are solutions which reduce disclosure risk but that risk still needs to be evaluated *once the method has been applied*. Thus controlled rounding is not enough to guarantee non-disclosure; the effectiveness of the rounding still needs to be assessed.

5. What does 'safe' mean?

This paper has used 'safe' and 'unsafe' to describe types of statistics. It should be clear that 'safe' in this context is a relative and subjective term.

The relativity arises because the process acknowledges that there are cases where the 'safety' of the output does not hold. A statistic is 'safe' because these cases are exceptional and unlikely, not because they are theoretically impossible. This is an explicit acknowledgement that one cannot prove a statistic has no theoretical disclosure risk, as conceptually a disclosive transformation of the data could always be devised. The more useful point is that such a transformation, or the exceptional cases identified, imply that the intruder already has more information than would be gained. In contrast, the circumstances under which 'unsafe' statistics could be disclosive are not so unusual and the information requirement relatively low.

Subjectivity arises from the importance of judgment calls on whether a particular circumstance is 'unlikely'. These are grounded in practical experience of research activity rather than the conceptual concerns. In contrast, more traditional approaches to SDC tend to focus on 'worst-case' scenarios, claiming these provide a degree of objectivity. However, as Skinner (2012) notes, these perceptions of objectivity are misplaced: the parameters of even the most 'objective' risk measure are based on the subjective risk preferences of the facility owner.

What does this mean for the facility owner? Put simply, the 'safe/unsafe statistic' model requires the facility owner to explicitly acknowledge that risk management depends upon effective resource management: a 'safe' statistic does not mean no possible risk, rather that the risk is so small that limited resources are better concentrated on the more pertinent statistics. A positive side-effect is that the safe/unsafe statistic model makes much more sense to researchers; and as Desai and Ritchie (2010) point out, a contented researcher presents a much lower disclosure risk than an irritated one.

It should be noted that the safe/unsafe classification makes one essential assumption: that researchers are genuinely producing research outputs. As a number of authors (Reznek and Riggs, 2005; Gomatam et al, 2005; Belninger et al, 2011, for example) have pointed out, it is possible to manipulate regression models such that a specific value is disclosed in the coefficients. This is a potential concern for remote job systems set to automatically release regressions.

On the face of it, this is a serious issue. The counter-response is that there is no evidence from any country of a researcher carrying out such activity; and in the circumstances which would enable a researcher to carry out such an attack, more plausible attack mechanisms (leaving a less obvious audit trail) present themselves.

Moreover, while it is clear that no SDC checking mechanism can prevent deliberate falsification of results, it is equally clear that this is not the domain of SDC. Deliberate misuse is a failure of the accreditation process and is more appropriately addressed there.

In summary, whilst there are theoretical circumstances under which the 'safe statistics' classification could be subverted, experience and the analysis of incentives suggests that this is not a practical problem. Again, judgment is necessary.

6. Conclusion

In a world where potentially all output poses a disclosure risk, effort is best concentrated on the outputs most likely to provide a risk. The 'safe/unsafe statistic' model has proved increasingly popular for OSDC as it provides clear and useful guidelines on how that effort could be expended most effectively.

The conceptual world embodied in this approach presents a different perspective to more traditional analyses of disclosure risk. In this view, all risks are possible but empirical evidence is used to identify which are the ones of most concern; there is no 'worst-case' planning. This explicitly creates a hierarchy of risks, places judgment at the forefront of decision-making, and openly acknowledges the potential residual risk in unchecked outputs.

This subjectivity and emphasis on empirical evidence causes concerns for facility owners unused to this perspective; but the 'safe/unsafe' approach has not created that subjectivity, it has only revealed it. Some facility owners also feel uncomfortable about the focus on resources, taking the view that as many resources as necessary should be thrown at a significant risk. However, the message of this paper is that that is counter-productive: it is the recognition that some risks can be effectively ignored that allows more time to be spend on genuinely risky output. Hence a risk management strategy which does not reflect the ideas in this paper is likely to be inefficient and less safe.

References

Bleninger P., Drechsler J., and Ronning G. (2011) Remote data access and the risk of disclosure from linear regression, *Stat. Op. Res. Trans. Special Issue: Privacy in statistical databases* pp 7-24

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; Eurostat

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

Desai T. and Ritchie F. (2010) Effective researcher management, *Work session on statistical data confidentiality 2009*; Eurostat

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

Reznek A. and Riggs T. (2005) Disclosure Risks in Releasing Output Based on Regression Residuals, *ASA 2004 Proceedings* (Govt. Stat. and Social Stat.) pp1397-1404

Ritchie F. (2007) *Statistical disclosure control in a research environment*, mimeo, Office for National Statistics; available as WISERD Data Resources Paper No. 6

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

Ritchie F. (2012) *Output-based disclosure control for regressions.* University of the West of England Discussion Papers in Economics no. 1209

Ritchie F. (2014) Access to sensitive data: satisfying objectives rather than constraints, *J. Official Stat.*, October

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

Skinner C. (2012) Statistical Disclosure Risk: Separating Potential and Harm, *Int. Stat. Rev.* v80:3 pp349–368

Recent UWE Economics Papers

See http://www1.uwe.ac.uk/bl/research/bristoleconomics/research for a full list

1308 **Risk taking and monetary policy before the crisis: the case of Germany**
Iris Biefang-Frisancho Mariscal

1307 **What determines students' choices of elective modules?**
Mary R Hedges, Gail A Pacheco and Don J Webber

1306 **How should economics curricula be evaluated?**
Andrew Mearman

1305 **Temporary employment and wellbeing: Selection or causal?**
Chris Dawson, Don J Webber and Ben Hopkins

1304 **Trade unions and unpaid overtime in Britain**
Michail Veliziotis

1303 **Why do students study economics?**
Andrew Mearman, Aspasia Papa and Don J. Webber

1302 **Estimating regional input coefficients and multipliers: The use of the FLQ is not a gamble**
Anthony T. Flegg and Timo Tohmo

1301 **Liquidity and credit risks in the UK's financial crisis: How QE changed the relationship**
Woon Wong, Iris Biefang-Frisancho Mariscal, Wanru Yao and Peter Howells

2012

1221 **The impact of the quality of the work environment on employees' intention to quit**
Ray Markey, Katherine Ravenswood and Don J. Webber

1220 **The changing influence of culture on job satisfaction across Europe: 1981-2008**
Gail Pacheco, De Wet van der Westhuizen and Don J. Webber

1219 **Understanding student attendance in Business Schools: an exploratory study**
Andrew Mearman, Don J. Webber, Artjoms Ivļevs, Tanzila Rahman & Gail Pacheco

1218 **What is a manufacturing job?**
Felix Ritchie, Andrew D. Thomas and Richard Welpton

1217 **Rethinking economics: Logical gaps – empirical to the real world**
Stuart Birks

1216 **Rethinking economics: Logical gaps – theory to empirical**
Stuart Birks

1215 **Rethinking economics: Economics as a toolkit**
Stuart Birks

1214 **Rethinking economics: Downs with traction**
Stuart Birks

1213 **Rethinking economics: theory as rhetoric**
Stuart Birks

1212 **An economics angle on the law**
Stuart Birks

1211 **Temporary versus permanent employment: Does health matter?**
Gail Pacheco, Dominic Page and Don J. Webber