Australian Bureau of Statistics

1352.0.55.110

Research Paper

Analysis of Micro-Data: Controlling the Risk of Disclosure

New Issue

Research Paper

Analysis of Micro-Data: Controlling the Risk of Disclosure

James Chipperfield and Sebastien Lucie

Analytical Services Branch

Methodology Advisory Committee

18 June 2010, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 18 JULY 2011

ABS Catalogue no. 1352.0.55.110

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Melissa Gare, Analytical Services Branch on Canberra (02) 6252 7147 or email <analytical.services@abs.gov.au>.

ANALYSIS OF MICRO-DATA: CONTROLLING THE RISK OF DISCLOSURE

James Chipperfield and Sebastien Lucie
Analytical Services Branch

QUESTIONS FOR THE COMMITTEE

1. How could the methods discussed here be extended to allow for analysis of mixed continuous and categorical variables?
2. Is perturbing statistical output the best approach to manage disclosure risk? Is there a better approach?
3. What other diagnostics should be supported by the remote server?

CONTENTS

The role of the Methodology Advisory Committee (MAC) is to review and direct research into the collection, estimation, dissemination and analytical methodologies associated with ABS statistics. Papers presented to the MAC are often in the early stages of development, and therefore do not represent the considered views of the Australian Bureau of Statistics or the members of the Committee. Readers interested in the subsequent development of a research topic are encouraged to contact either the author or the Australian Bureau of Statistics.

ANALYSIS OF MICRO-DATA: CONTROLLING THE RISK OF DISCLOSURE

James Chipperfield and Sebastien Lucie
Analytical Services Branch

ABSTRACT

There is very strong demand from analysts, particularly within government and universities, to access micro-data collected by agencies, such as the Australian Bureau of Statistics (ABS), for the purpose of developing and evaluating policy. To help meet this demand, the ABS is planning to develop a remote server which would automatically return the output from remotely submitted statistical programming code. In allowing such access, the ABS is legally obliged to ensure that any information (e.g. analysis output) it releases *is not likely to enable the identification of the particular person or organisation to which it relates.* This paper considers the problem of managing the disclosure risk associated with releasing analysis output, including regression parameters and model diagnostics for generalised linear models, by a remote server. While this paper restricts attention to surveys where all variables are categorical, these variables can be defined without restriction. The disclosure risk is managed by adding noise in two different ways. The first adds noise to the input data prior to analysis and the second adds noise to the counts present in the estimation equation. All inferences using the statistical output released by the server are valid in the presence of adding noise. The methods are evaluated using the 2008 National Health Survey. The results show that perturbing counts in the estimating equation leads to a very small loss in accuracy.

1. INTRODUCTION

1.1 Literature review

There is very strong demand from analysts, particularly within government and universities, to access micro-data collected by agencies, such as the Australian Bureau of Statistics (ABS), for the purpose of developing and evaluating policy. Such data may be from administrative sources, Censuses, or sample surveys. It is part of the ABS' mission to meet this demand or, more broadly, to maximise the utility of the information it collects in order to "… *assist and encourage informed decision making*". However, the ABS is legally obliged to ensure that the information (e.g. population estimates or analysis output) it makes public *is not likely to enable the identification of the particular person or organisation to which it relates.* Managing the balance between utility and the risk of disclosure, particularly when allowing analysts access to micro-data, is a challenging issue faced by many organisations and is an active area of research in the statistical literature. Next we discuss some of the different approaches to managing this balance.

The first approach is to release a synthetic version of the micro-data, say on CD ROM, to the public. As synthetic data are generated from a model, this approach would typically have a low disclosure risk. Such an approach has been extensively studied in the literature (see for example Fienberg and Makov, 1998; Raghunathan, Reiter and Rubin, 2003 and Reiter, 2002). Some disadvantages of this approach are:

- Use of synthetic, rather than real, variables will mean that some information about complex relationships in the micro-data will be lost and may be severely biased. Unlike a remote server, the analyst will have no means to know if their analysis is biased (see Reiter, Oganian and Karr, 2007).
- Generating 'realistic' synthetic micro-data is very time-consuming.
- Analysis with synthetic data can be more complicated, since analysis would need to account for the fact that some variables are generated from a model

A second approach is to confidentialise the micro-data prior to it being released to the public, say via a CD ROM. There are many other ways, discussed at length in the literature, in which micro-data can be confidentialised (see Willenborg and de Waal, 2000). These include reducing detail, sub-sampling, replacing real data with synthetic data, micro-aggregation, swapping variables between records, rounding, and adding noise. Releasing a single set of micro-data amounts to a 'one-size-fits-all' approach. Reducing detail in the variables in the micro-data is a prime example: some analysts may want to retain geographic detail while others are interested in demographic detail. Another possible disadvantage is that analysis can be more complicated, since it would ideally account for the fact that variables may not be equal to their true values.

Another approach, more so defined by its technological rather than statistical properties, is to release analysis output via a remote server (see Gomatam *et al.*, 2008). A simple model for a remote analysis server is:

1. An analyst submits a program, via the Internet, to the analysis server;
2. The analysis server audits the program to confirm that it does not use manipulations associated with a high risk of disclosure;
3. The analysis server processes the analyst's program. The output from the program may be perturbed (i.e. changed in some way) so as to appropriately manage the disclosure risk. Managing the disclosure risk of statistical output is the focus of this paper.
4. The analysis server sends the perturbed output, via the Internet, to the analyst along with information about the how changes to the statistical output impact inference.

The potential benefits of this approach are that:

- the data that are analysed are real so that complex relationships in the micro-data are retained.
- the degree to which a particular statistical output is perturbed (e.g. by adding noise) can depend upon the output itself. For example, on the one hand, estimates at a broad level may only require a small degree of perturbation. On the other hand, diagnostic plots involving small counts may require proportionally more perturbation.
- the impact of perturbation on the output (e.g. regression coefficients) can be broadly indicated to the analyst. If the impact of the perturbation is large the analyst may decide to ignore the results altogether.

Some work in the literature has considered directly protecting the confidentiality of statistical output released by a remote server. Dwork and Smith (2009) discuss how to construct estimators to manage differential privacy. The approach, in principle at least, can be applied to a wide range of analysis. However some implementations have been criticised for adding too much noise to the output to be practically useful. Duncan and Mukherjee (2000) propose a method of adding noise to variables on a data base with continuous variables in order to confidentialise totals. Gomatam *et al.* (2005) provides some high level guidance about how restrictions imposed by a remote server affect disclosure risk and utility. Other work has specifically focused on ensuring the confidentiality of diagnostic plots (see, for example, O'Keefe and Good, 2008, and Reiter and Kohnen, 2005).

Perhaps the most comprehensive work to-date on this approach is by Sparks et al. (2008) in the development of a tool called Privacy Preserving Analytics (PPA). PPA produces a range of exploratory data analysis, caters for both continuous and categorical variables, allows generalised linear, time series, and linear mixed models, and produces useful numerical and graphical model diagnostics. Perhaps the most restrictive feature of PPA as a general purpose tool is that it does not allow the analyst to define new variables. This can be problematic when one analyst is interested in a particular geography, another in a particular demographic, and yet another in specific income ranges relevant to policy development. Relaxing this constraint would require a completely different solution to the problem of managing the risk of disclosure. Other features of PPA that reduce its utility include: it does not report standard errors (it reports p-values in ranges); *all* levels of a categorical variable (e.g. the categorical variable *industry of occupation* may have 20 levels, where '1'= 'Food industry', '2'= 'Retail industry', and so on) are either included or excluded in the model regardless of how many of them are statistically significant. Also, while they formally address disclosure risk for linear regression models, they do not cover the case for generalised linear models.

This paper proposes a method that imposes none of the above restrictions but, instead, perturbs the analysis output to manage disclosure risk. This high degree of flexibility comes at a cost of a small reduction in accuracy. Nevertheless, the validity of inferences will not be compromised. Statistical output from analysis of categorical variables is essentially a function of frequency counts. The proposed method perturbs the output by adding noise in two different ways. The first adds noise to the micro-data prior to analysis and the second adds noise to the counts present in the estimation equations. The latter requires adding noise to a relatively small number of counts and so suffers from only a small loss in efficiency.

Importantly, the method can be used to jointly manage the disclosure risk when releasing estimates of population counts *and* regression parameters. If the risk associated with releasing these estimates is not jointly managed, it may be possible to affect disclosure by combining them. This additional feature is particularly relevant for agencies such as the Australian Bureau of Statistics, as we discuss below.

Agencies naturally measure disclosure risk in different ways. This is because agencies collect different types of data and are guided by different legislation. This would naturally lead agencies to prefer different methods of adding noise to counts. The framework in this paper is flexible in that it allows agencies to adds noise to counts in their own preferred way.

1.2 Statistical agencies

It is hard to over-state the international importance of managing the disclosure risk associated with releasing analysis output, whether via a remote server or not. A range of statistical agencies attended the inaugural meeting of the OECD/ABS Paris Microdata Access Group in May 2009. Papers were presented by Statistics Canada, Australian Bureau of Statistics, Office for National Statistics, Statistics New Zealand, U.S. Census Bureau, Eurostat, German Federal Statistics Agency and others. A typical example is the German Federal Statistics Agency (GSFA) which executes analysts' code on their micro-data and manually checks output for disclosure risk before being released. This is a time-consuming process for the analyst and for the GSFA, taking up to five days to review and return output to analysts. Many agencies see a remote server as a critical development. Some agencies, such as ONS, Eurostat and Statistics New Zealand are keeping abreast of developments. Other agencies, such as Statistics Canada and the U.S. Census Bureau are in the process of developing analysis servers (e.g. the USBC is developing a prototype analysis server called Microdata Analysis System (MAS)).

We now discuss how the ABS currently manages access to its micro-data. Publication estimates for ABS household surveys are based on micro-data, referred to as the Main Unit Record File (MURF). The MURF contains a high level of geographic and demographic detail which could be used to identify individuals with relatively uncommon characteristics. As such, the release of the MURF, say on a CD ROM, for public use would be a breach of the ABS' legal obligations with regard to disclosure risk. The ABS has sought to meet the demand for access to its micro-data by allowing access to:

- CD ROMs containing micro-data called Basic Confidentialised Unit Record Files (CURFs). The analyst can view records on the CURF on their own personal computer. The disclosure risk associated with CURFs is managed by reducing the level of detail (e.g. geographic and demographic). For example, a MURF may contain single year age categories while the CURF may only contain five-year age categories. Feedback from analysts is that the level of detail on the CURF can seriously reduce its utility.
- a Remote Access Data Laboratory (RADL). RADL allows users to submit code (SAS, SPSS, STATA) via the Internet, which is run within the ABS's secure environment, with the output returned to the user. In the literature, this is often called a *remote analysis server*. The submitted code and the output are subject to a range of automated checks, with certain input commands not allowed (including graphical displays of data) and some output not released if it does not meet confidentiality requirements. Queries flagged as higher risk are held over to be manually cleared, and a sample of queries are also audited post-release. RADL allows access to Expanded CURFs and Basic CURFs. An Expanded CURF has more detailed variables than a Basic CURF, though still less than a MURF.

- a Data Laboratory (DL). DLs allows interactive access to unit level Specialised CURFs, Expanded CURFs or Basic CURFs. Specialist CURFs have more detail than an Expanded CURF. There are no limitations on what unit record or summary information a user can view within a DL, but users are supervised at all times by an ABS staff member. All outputs produced by users in a DL are manually cleared for release by ABS confidentiality methodologists. No unit level output is allowed to be released and aggregate output must be unlikely to enable disclosure.

In addition, the ABS is in an advanced stage of developing a web-based tool called *Survey TableBuilder* that allows analysts to remotely submit requests for survey estimates of user-defined population counts. Population counts will be estimated from a MURF, automatically confidentialised, and sent to the analyst without the need for manual intervention. However, an unmet demand for researchers is the ability to run analytical models such as generalised linear models (GLMs) on original MURF data. Assessing model accuracy and model assumptions using standard diagnostic tools is a key part of this modelling process.

Section 2 outlines the statistical requirements of the remote analysis server. Sections 3 and 4 describe two methods of managing the disclosure risk associated with releasing statistical output from analysis of survey data, where the data contains only categorical variables. In particular, Section 3 confidentialises the micro-data prior to analysis and Section 4 confidentialises the counts present in the estimating equation. Section 5 evaluates the methodology on the ABS' *2008 National Health Survey*. Section 6 makes concluding remarks.

2. STATISTICAL REQUIREMENTS OF THE REMOTE ANALYSIS SERVER

The fundamental objective of this paper is to find a good balance between the risk of disclosure associated with releasing output from statistical analysis and its utility, or fitness-for-purpose. The specific *utility* requirements of the remote analysis server are to allow analysts to access the MURF, to create user-defined variables, and to fit generalised linear models (GLMs). The specific *risk* requirement is that statistical output released by the analysis server does not increase the risk of disclosure over and above the risk associated with TableBuilder.

In the remainder of this section we discuss disclosure risk and utility, describe how TableBuilder manages disclosure risk, and define the GLMs considered by this paper.

2.1 Balancing disclosure risk and utility

Disclosure risk

When allowing analysts to view individual records in a set of micro-data there are at least two ways in which disclosure can occur (see deWaal and Willenborg, 2000):

- *Spontaneous recognition* occurs when an analyst recognises that a record corresponds to someone they know.
- *Matching* records from micro-data to other data sets (e.g. administrative data bases with names and addresses) using a set of variables in common to both. Such a set of variables may include age, sex, geographic location and occupation.

Another type of disclosure, which can occur without viewing individual records, is *inferential disclosure*. Inferential disclosure occurs when relationships in micro-data are used to accurately make inferences about a person. For example, consider a model which very accurately predicts personal income from a set of freely available variables (e.g. level within an organisation). Clearly, the model could be used to disclose a person's income.

When access to micro-data is via a remote server an analyst is not able to directly view individual records. Nevertheless, an analyst can attempt to view, through indirect means, individual records and thereby affect disclosure. This is referred to as a *data attack*. A data attack occurs when an analyst submits repeated queries in order to circumvent confidentiality protections. Common forms of attack include differencing, transformations and leverage (see Gomatam *et al.*, 2008). In a leverage or transformation attack, the analyst gives a small number of records a high degree of influence on estimates. Differencing attacks involves repeating a statistical procedure after dropping a small number of records and then taking the difference between the two estimates to disclose information about the records that were dropped.

Considerable work in the literature has focused on measuring and managing the risk of disclosure. Feinberg (1994) considers the relative risk of disclosure before and after the micro-data have been released. Differential Privacy (Dwork and Smith, 2009) asserts that information should be released in such a way that the inclusion of a record in a dataset results in only marginally stronger inferences about its presence and characteristics, compared to the population inferences derivable from the same dataset excluding that record. Duncan and Mukherjee (2000) measure disclosure risk by the variance of the predicted value of a sensitive characteristic for a record on the micro-data.

Fienberg and Makov (1998) consider disclosure risk associated with counts in a contingency table after they have been perturbed, say by adding or subtracting small integers that have been randomly generated. They measure disclosure risk by the probability that a perturbed count of 1 corresponds to a true count of 1. Counts of 1s are clearly a disclosure risk. It is a simple matter to specify a perturbation distribution that will ensure a true count of 1 will be perturbed to a value other than 1, thereby ensuring the disclosure risk is zero. The authors suggest that this process can be repeated for cells with true counts of 2. The Fienberg and Makov (1998) measure of disclosure risk was used to inform the method of confidentiality behind TableBuilder (see Section 2.2).

This paper develops two methods for managing the risk of disclosure associated with releasing analysis output. The aim is to ensure that this risk is no greater than the risk associated with the estimates of population counts available from TableBuilder.

Utility

The concept of utility, though hard to measure, has been broadly discussed in the literature (see Gomatam *et al.*, 2008). Utility is a measure of how well the remote server meets the needs of the analyst. Managing disclosure risk via a remote server necessarily requires imposing restrictions that will reduce utility. In general, high utility and low disclosure risk are conflicting goals. Clearly utility and disclosure risk would be at their greatest if analysts had access to a MURF on their personal computer (i.e. the complete absence of a remote server).

There are at least five ways in which the ABS' planned remote analysis server might reduce utility. First, if either method of managing disclosure risk (see Sections 3 and 4) is implemented in the remote analysis server, all variables on the micro-data are restricted to be categorical. Continuous variables could, of course, be transformed to categorical variables. For analysts interested in distributions of continuous variables, the utility of the server could be low. On the other hand, since most variables on MURFs are categorical, most analysts may not be concerned by such a restriction.

Second, the remote analysis server may restrict analysts to specific statistical packages. For example, at the moment the ABS RADL supports only STATA, SAS, SPSS. Third, analysis through a remote server will take longer than analysis of micro-data that is available on the analysts' personal computer. From a recent survey of key clients of RADL, some reported that analysis could take more than three times longer. Fourth, the remote analysis server may restrict the range of statistical techniques in order to manage the risk of disclosure (e.g. disallow use of influence statistics). Finally, managing disclosure risk requires adjusting the statistical output in some way. Such changes will necessarily make the analysis output less reliable.

2.2 Tabular confidentiality

There is a significant amount of work in the literature on confidentiality for count estimates (for a review see Willenborg and de Waal, 2000). Following the framework of Fienberg and Makov (1998), mentioned previously, TableBuilder manages disclosure risk by perturbing cell counts by adding or subtracting a randomly generated integer. Next we describe the method behind TableBuilder, for two main reasons. Firstly, both proposed Analysis Server methods (Sections 3 and 4) use the TableBuilder algorithm as a mechanism to manage the disclosure risk of statistical output. Secondly, we need to ensure that the outputs generated from the Analysis Server and TableBuilder are compatible: that is, they cannot be combined to effect disclosure. For example, estimates of subpopulation prevalences are available from TableBuilder but could also potentially be derived from Analysis Server outputs (by requesting a categorical logistic regression with appropriate indicator variables).

Denote the i-th unweighted sample count in a contingency table by

$$n_i = \sum\nolimits_{j=1}^{n} \delta_{ij};\ i = 1,\ldots,C$$

where $\delta_{ij} = 1$ if the j-th record on the micro-data belongs to the i-th cell and $\delta_{ij} = 0$ otherwise, $j = 1,2,\ldots,n$ and $n = \sum\nolimits_{i=1}^{n} n_i$.

For each non-zero count, n_i, the corresponding perturbed count is $n_i^* = n_i + e_i^*$. TableBuilder releases n_i^*, instead of n_i, to analysts. The term e_i^* is a random variable restricted to be an integer with a distribution that satisfies the following criteria:

a) $n_i^* \geq 0$

b) $E_\xi\left(n_i^*\right) = n_i$

c) $Var_\xi\left(n_i^*\right) > 0$ and $Var_\xi\left(n_i^*\right)$ is a function of only n_i so that $Var_\xi\left(n_i^*\right) = Var_\xi\left(n_j^*\right)$ if $n_i = n_j$.

d) $Cov\left(n_i^*, n_j^*\right) = 0$ if $i \neq j$. See Fraser and Wooton (2005) for two ways of ensuring this condition holds in practice.

e) $\left|e_i^*\right| \leq L$ for some small positive integer L.

f) whenever the same set of records contribute to a cell count, the value for e_i will always be the same (see Fraser and Wooton, 2005).

Define S to be the perturbation function described above, so that we can say n_i^* is a random variable generated from $S\left(n_i\right)$ or, more concisely, that $n_i^* \sim S\left(n_i\right)$.

Criterion a) ensures that no negative numbers are created as a result of perturbation; criterion b) ensures the estimators of the regression coefficients proposed in this paper are approximately unbiased, under mild conditions (see Sections 3 and 4 for more details); criterion c) ensures that any cell derived by differencing two perturbed cells has a fixed variance, criterion d) ensures that differencing two cells counts does not remove the effect of perturbation, criterion e) is applied to ensure that no perturbation is ever greater than *L* in magnitude, and f) is designed to protect against differencing attacks.

Within the framework of Fienberg and Makov (1998), disclosure risk is measured by the probability that a perturbed cell count of 1 equates to a true cell count of 1 – i.e. $\Pr\left(n_i^* = n_i = 1\right)$. Thus, managing disclosure risk simply requires adjusting the distribution of e_i. However, a true count of 1 could be obtained by taking the difference between two perturbed counts. The distribution of e_i would then need to be such that this occurs with an acceptably small probability.

Denote the i-th weighted sample count in a contingency table by $n_{i(d)} = \sum_j d_j \delta_{ij}$, where d_j is the survey weight for the j-th record.

The corresponding perturbed count, denoted by $n_{i(d)}^*$, is a random variable generated from $\tilde{d}_i S\left(n_i\right)$, where $\tilde{d}_i = n_i^{-1} n_{i(d)}$ is the average weight for records belonging to the i-th cell and n_i is the unweighted sample count. More concisely we may write $n_{i(d)}^* \sim S\left(n_{i(d)}\right) = \tilde{d}_i S\left(n_i\right)$.

Instead of releasing $n_{i(d)}$, TableBuilder releases the perturbed count $n_{i(d)}^*$.

As survey weights are determined at a relatively high geographic level, they are not considered a disclosure risk. The TableBuilder algorithm can be applied to perturb either unweighted or weighted counts. For survey data, current ABS practice is to only allow weighted counts to be produced, using the original weight provided on the MURF (i.e. user-defined weights are not allowed).

2.3 Generalised linear models and diagnostics

Define x to be a P-vector of dichotomous variables which is observed for each record on the micro-data, and u to be a dichotomous outcome variable.

Let $\mathrm{x}_i = \left(x_{i1},\ldots,x_{ip},\ldots,x_{iP}\right)$ denote the i-th different value that x may take.

Let n_i be the number of observations with $\mathrm{x} = \mathrm{x}_i$, y_i be the number of records where $\mathrm{x} = \mathrm{x}_i$ and the outcome variable u takes the value of 1 (e.g. '0'= 'not over-weight' and '1'= 'over-weight'), and m_i be the corresponding number of records with outcome equal to 0 (so that $m_i = n_i - y_i$).

We use the standard formulation for generalised linear models, in which the outcome variable $u_{(k)}$ for observation k is modelled as being drawn from an exponential family:

$$f_{U_{(k)}}\left(u_{(k)};\theta_{(k)},\tau\right) = \exp\left(\frac{u_{(k)}\theta_{(k)} - b\left(\theta_{(k)}\right)}{\tau^2} + c\left(u_{(k)},\tau\right)\right)$$

whose expectation $\pi_{(k)} = E\left(U_{(k)}\right) = b'\left(\theta_{(k)}\right)$ is related to a linear predictor $\eta_{(k)} = x_{(k)}\beta$ via the link equation $\eta_{(k)} = g\left(\pi_{(k)}\right)$.

Maximum likelihood may be used to estimate $\boldsymbol{\beta}$, resulting in the score function:

$$\begin{aligned}
Sco\left(\beta;Y,\tau\right) &= \frac{d}{d\beta}\sum\nolimits_k \log f_U\left(u_{(k)};\theta_{(k)},\tau\right) \\
&= \frac{d}{d\beta}\sum\nolimits_k \left(\frac{u_{(k)}\theta_{(k)} - b\left(\theta_{(k)}\right)}{\tau^2}\right) \\
&= \frac{d}{d\beta}\sum\nolimits_i \left(\frac{y_i\theta_i - n_i b\left(\theta_i\right)}{\tau^2}\right) \\
&= \tau^{-2}\frac{d}{d\theta}\left(Y^{\mathrm{T}}\theta - n^{\mathrm{T}}b\left(\theta\right)\right).\frac{d\theta}{d\eta}.\frac{d\eta}{d\beta} \\
&= \tau^{-2}\left(Y - n_{diag}b'\left(\theta\right)\right)^{\mathrm{T}}.\frac{d\theta}{d\eta}.X \\
&= \tau^{-2}\left(Y - n_{diag}\Pi\right)^{\mathrm{T}}.\frac{d\theta}{d\eta}.X
\end{aligned}$$

where

$$X = \left(x_1^{\mathrm{T}}, \ldots, x_i^{\mathrm{T}}, \ldots, x_C^{\mathrm{T}}\right)^{\mathrm{T}}$$

$$\Pi = \left(\pi_1, \ldots, \pi_i, \ldots, \pi_C\right)^{\mathrm{T}}$$

$$n = \left(n_1, \ldots, n_i, \ldots, n_C\right)^{\mathrm{T}}$$

$$Y = \left(y_1, \ldots, y_i, \ldots, y_C\right)^{\mathrm{T}}$$

$$n_{diag} = diag\left(n\right)$$

Equating this to zero and transposing produces the estimating equation

$$X^{\mathrm{T}}.\frac{d\theta}{d\eta}.\left(Y - n_{diag}\Pi\right) = 0$$

In this paper we will restrict ourselves to considering models with a canonical link function $\eta = \theta$, giving rise to estimating equations of the form

$$X^{\mathrm{T}}Y - X^{\mathrm{T}}n_{diag}\Pi = 0 \qquad (1)$$

This assumption holds for a large number of useful models, including the standard logistic regression, multinomial, log-linear and Poisson models for discrete data.

To account for unequal weights, the pseudo-likelihood estimator (see Chambers and Skinner, 2003) of $\boldsymbol{\beta}$ may be obtained by replacing the unweighted counts Y and n in (1) by their corresponding weighted counts. Analysts can conduct weighted or unweighted analyses through the remote server. In the former case only the weight provided on the micro-data may be used. This is a protection against a leverage attack, which would involve allocating a set of records with specific characteristics a very high weight and hence influence on the analysis results.

The variance of $\hat{\boldsymbol{\beta}}$ can also be estimated using the delete-a-group Jackknife (Rao and Wu, 1988). The Jackknife variance estimator, $\widehat{Var}\left(\boldsymbol{\beta}\right)$, is unbiased when the micro-data have been collected from a sample with a complex design (e.g. clustered sampling and unequal probabilities of selection), as is the case for MURFs. The Jackknife method involves allocating all selection units, in this case geographic clusters of dwellings, to one and only one replicate group in the same way that the sample was selected from the population. Indexing the replicate groups by $r = 1, \ldots, R$, the Jackknife estimator is:

$$\widehat{Var_{JK}}\left(\hat{\boldsymbol{\beta}}\right) = \frac{\left(R-1\right)}{R}\sum\nolimits_r \left(\hat{\boldsymbol{\beta}}\left(r\right) - \hat{\boldsymbol{\beta}}\right)^2$$

where $\hat{\boldsymbol{\beta}}\left(r\right)$ has the same form as $\hat{\boldsymbol{\beta}}$ except that it is calculated after excluding the sample from the r-th replicate group.

Model fitting commonly often involves forward and backward selection, which is a simple and quick way of identifying explanatory variables that are important to the model. The forward-backward selection algorithm could be applied to the true micro-data to obtain the covariates to include in $\mathbf{x}$. Only the perturbed output for the set of covariates selected by the algorithm would need to be released to the analyst – no information about the number of iterations or the order in which covariates were dropped or added from/to the model would be provided to the analyst.

Also, diagnostics are commonly used to assess a fitted model's accuracy and whether its underlying assumptions are reasonable. For illustrative purposes, the diagnostics considered in this paper are the Pearson statistic, R-squared and a plot of observed against predicted probabilities.

3. PERTURBING THE MICRO-DATA PRIOR TO ANALYSIS

The approach described below involves Perturbing the Micro-data (PM), using TableBuilder, prior to analysis.

3.1 Estimation

Estimation of $\boldsymbol{\beta}$ for models of the form (1) requires

$$\mathrm{d} = \left\{ \left(y_i, m_i \right) \middle| \mathrm{x}_i : i = 1, \ldots, C \right\}$$

where d is a table of counts with $2C$ cells.

After applying TableBuilder to d, the corresponding perturbed table of counts is

$$\mathrm{d}^* = \left\{ \left(y_i^*, m_i^* \right) \middle| \mathrm{x}_i : i = 1, \ldots, C \right\}$$

where $y_i^* \sim S\left(y_i\right)$ and $m_i^* \sim S\left(m_i\right)$.

Instead of applying the EM algorithm to d to obtain $\hat{\boldsymbol{\beta}}$, the remote server will apply the EM algorithm to d^* and instead output $\hat{\boldsymbol{\beta}}^*$. Since the disclosure risk associated with d^* is acceptable (d^*could be obtained directly from TableBuilder) it follows that $\hat{\boldsymbol{\beta}}^*$ has an acceptable disclosure risk. The case of unequal weights simply requires replacing unweighted counts in d^* with weighted counts.

3.2 Inference

To make valid inference involving $\hat{\boldsymbol{\beta}}^*$ we need to correctly take into account the uncertainty due to perturbation and the model itself. The estimate of $\boldsymbol{\beta}$ is a function of d so we may write $\hat{\boldsymbol{\beta}}$ by $\hat{\boldsymbol{\beta}}\left(\mathrm{d}\right)$.

Denote the b-th independent random outcome from perturbing d by $\mathrm{d}_{(b)}^*$, where $b = 1, \ldots, B$, and $\hat{\boldsymbol{\beta}}_{(b)}^* = \hat{\boldsymbol{\beta}}^*\left(\mathrm{d}_{(b)}^*\right)$ to have the same form as $\hat{\boldsymbol{\beta}}\left(\mathrm{d}\right)$ except that d is replaced by $\mathrm{d}_{(b)}^*$.

Under the conditions b) and d) of TableBuilder (see Section 2.2) it is easy to show that an unbiased estimate of the Mean Squared Error of $\hat{\boldsymbol{\beta}}^*$, that correctly accounts for uncertainty due to perturbation and the model, is

$$\widehat{MSE}_{M\xi}\left(\hat{\boldsymbol{\beta}}^*\right) \approx \widehat{Var}_{JK}\left(\hat{\boldsymbol{\beta}}\right) + \widehat{MSE}_{\xi}\left(\hat{\boldsymbol{\beta}}^*\right) \tag{2}$$

where $$\widehat{MSE}_{\xi}\left(\hat{\boldsymbol{\beta}}^*\right) = B^{-1} \sum_{b=1}^{B} \left(\hat{\boldsymbol{\beta}}_{(b)}^* - \hat{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}_{(b)}^* - \hat{\boldsymbol{\beta}}\right)'$$

The first term in (2) reflects the error due to the model itself and is based on the true micro-data. The second terms reflects the error due to perturbation. An analyst would make correct inference involving $\hat{\beta}^*$ with $MSE_{M\xi}\left(\beta^*\right)$. The second term in (2) is a function of the perturbation distribution, which is not provided to the analyst, and so is not a disclosure risk.

We now consider the disclosure risk of the first term. The Jackknife estimate has a level of uncertainty due to the process, denoted by v, of allocating selection units to replicate groups. The unbiased property of the Jackknife estimator means, in the present context that,

$$E_v\left[\widehat{Var_{JK}}\left(\hat{\beta}\right)\right]=\left(\mathbf{X}'\mathbf{W}\mathbf{X}\right)^{-1}$$

where $\mathbf{W}$ is a $C\times C$ diagonal matrix with elements that depend on the link function.

For example, for the logistic regression model, $\mathbf{W}$ has i-th diagonal element $n_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)$, where $\hat{\pi}_i=\text{logit}^{-1}\left(\mathbf{x}_i'\beta\right)$.

Clearly, in expectation the Jackknife variance estimator is a function of sample counts which we do not want to disclose. However as long as R is not too large, the uncertainty in the Jackknife variance estimates will provide sufficient protection against disclosure. The important property of the Jackknife here is that, conditional on the sample,

$$CV_v\left[\widehat{Var_{JK}}\left(\hat{\beta}_p\right)\right]\approx\frac{2}{R-1} \tag{3}$$

where
$$CV\left[\hat{\theta}\right]=\frac{Var\left[\hat{\theta}\right]}{\hat{\theta}^2}$$

(see Shao and Wu, 1995, p. 196).

This means that the 95% confidence interval for $\widehat{Var_{JK}}\left(\hat{\beta}_k\right)$ is

$$\widehat{Var_{JK}}\left(\hat{\beta}_p\right)\left(1-2\sqrt{\frac{2}{R-1}},1+2\sqrt{\frac{2}{R-1}}\right)$$

For example if $R=60$ (a standard for ABS MURFs) then the confidence interval is

$$\widehat{Var_{JK}}\left(\hat{\beta}_p\right)\left(0.63,1.37\right)$$

which has a range spanning more than 50% of the magnitude of $\widehat{Var_{JK}}\left(\hat{\beta}_k\right)$. This degree of uncertainty is sufficient protection against a data attacker's attempt to isolate the sample counts contained in $\widehat{Var_M}\left(\hat{\beta}\right)$.

In small samples, there is no simple expression for $CV_V\left[\widehat{Var_{JK}}\left(\hat{\beta}_p\right)\right]$.

As a result, we suggest only releasing $\widehat{Var_{JK}}\left(\hat{\beta}_p\right)$ when the sample size is reasonably large, say when $n>60$. This is not an onerous restriction since $\widehat{Var_{JK}}\left(\hat{\beta}_p\right)$ would be quite variable when $n<60$.

3.3 Diagnostics

A range of diagnostics that evaluate the model's accuracy and assumptions (see Hosmer and Lemeshow, 2000) are available for models of the form (1). In providing analysts with diagnostics, we need to ensure they could not be used to affect disclosure.

Some diagnostics, such as the likelihood ratio test or the Chi-squared test, require p-values. Reporting p-values in ranges is a simple and effective way to manage disclosure risk. In fact, if an analyst specifies a value for the significance level (e.g. 0.05) a remote server need only indicate whether the significance of the test statistic is greater or less than this value. We propose indicating whether the significance of the test statistic falls within four ranges (e.g. less than 0.01, between 0.01 and 0.05, between 0.05 and 0.1, or greater than 0.1) as long as $C>10$. The Chi-squared statistic is a function of $2C$ unknowns – given it is only provided in ranges, it is impossible for a data attack to solve for d.

Other diagnostics such as the Area under the ROC Curve, are often reported in ranges (see Hosmer and Lemeshow, 2000, p. 164). The well known Pearson R-squared can be reported in conservatively wide ranges. We propose reporting the R-squared in ranges of 0.05 $(0.00-0.05, 0.05-0.10, \ldots, 0.95-1.00)$ when $C>10$, the R-squared statistic is effectively suppressed.

Diagnostic plots are useful ways to check the model's assumptions. One example is to plot the predicted probabilities, $\hat{\pi}_i^*$, against the true observed probabilities, y_i/n_i. To manage disclosure risk, we propose plotting $\hat{\pi}_i^*$ against the perturbed probability $p_i^* = y_i^*/n_i^*$, both of which have an acceptable disclosure risk. The uncertainty introduced by perturbing the observed probabilities can be expressed as

$$Var_\xi\left(p_i^*\right) \le \frac{2\sigma_i^2}{n_i^2}$$

if we make the reasonable simplifying assumption $Var_\xi\left(y_i^*\right) = Var_\xi\left(n_i^*\right) = \sigma_i^2$.

This impact on the diagnostic plot will become increasingly small as n increases. For example, if $\sigma_i^2 = 2$ and $n_i = 100$, then true observed probability will fall within the confidence interval $p_i^* \pm 0.02$ 95% of the time.

Extending this approach to multinomial regression is straight-forward, in which case there is a separate residual plot for each outcome indicator variable. Reiter and Kohnen (2005) consider a similar approach to that described above, but where they impose the additional restriction that the plotted counts must have a minimum number of contributing records. They showed their approach worked well at identifying model mis-specification.

Appendix A.2 gives the Leverage and Cooks distance diagnostics, commonly used to assess the fit of a logistic regression model. It also suggests a corresponding confidentialised version of the diagnostics that may be released to analysts. This is an active area of research. Appendix A.3 discusses other approaches to confidentialising diagnostic plots that have been proposed in the literature.

4. PERTURBING THE COUNTS IN THE ESTIMATING EQUATION

What information could be obtained from giving $\hat{\boldsymbol{\beta}}$ to analysts which could be used in a data attack? The answer is that, given $\hat{\boldsymbol{\beta}}$, an analyst could solve (1) for $\mathbf{Y}$ and $\mathbf{n}$, the true counts we do not want to disclose. Such a solution could easily be found by conducting a grid search centred around $\mathbf{n}^* = \left(n_1^*, \ldots, n_i^*, \ldots, n_C^*\right)$ and $\mathbf{Y}^* = \left(y_1^*, \ldots, y_i^*, \ldots, y_C^*\right)$, both of which could be obtained from TableBuilder.

Consider the example when $\mathrm{x} = 1$ (implying $C = 1$ so we may drop the i subscript) and $\hat{\pi}$, obtained directly from $\hat{\boldsymbol{\beta}}$, is provided to the analyst. Equation (1) simplifies to $y - n\hat{\pi} = 0$, where the solution for y and n could easily be found from a grid search centred around y^* and n^*. For example if $\hat{\pi} = 1/13$, the possible solutions are $(y, n) = (k, 13k)$, where k is an integer. If $\left(y^*, n^*\right) = (3, 11)$ then an analyst can be very certain that $k = 1$, which effectively amounts to disclosure: a value of $k = 2$ would imply that n=26 but was perturbed to a value of 11, an unlikely scenario.

(Also see Appendix A.1 for an example of how output from TableBuilder and an analysis server could be combined to affect disclosure.)

The basic idea here is to instead solve the following 'adjusted estimating equation' for $\boldsymbol{\beta}$,

$$S(\mathbf{X}'\mathbf{Y}) - S\left(\mathbf{X}'\mathbf{n}_{diag}\boldsymbol{\Pi}\right) = 0 \tag{4}$$

where $S(\mathbf{X}'\mathbf{Y})$ is the table of counts $\mathbf{X}'\mathbf{Y}$ after it has been perturbed by TableBuilder and $S\left(\mathbf{X}'\mathbf{n}_{diag}\boldsymbol{\Pi}\right)$ is the table of counts $\mathbf{X}'\mathbf{n}_{diag}\boldsymbol{\Pi}$ with weight matrix $\boldsymbol{\Pi}$ after it has been perturbed by TableBuilder.

Denote the solution for $\boldsymbol{\beta}$ in (4) by $\hat{\boldsymbol{\beta}}^{**}$.

Now, given $\hat{\boldsymbol{\beta}}^{**}$, the solution for the counts present in (4) are in fact perturbed counts obtained from TableBuilder.

We now rewrite (4) in a way that clearly shows how it differs from (1).

Define $$N_{(p)} = \sum_i x_{ip} n_i \text{ and } N_{(yp)} = \sum_i x_{ip} y_i$$

and their corresponding perturbed values by $N_{(p)}^* = S\left(N_{(p)}\right)$ and $N_{(yp)}^* = S\left(N_{(yp)}\right)$.

Also define $$\Lambda_{(p)}^* = \frac{N_{(p)}^*}{N_{(p)}} \text{ and } \mathrm{T}_{(yp)}^* = \frac{N_{(yp)}^*}{N_{(yp)}}$$

to be the multiplicative impact of perturbation on $N_{(p)}$ and $N_{(yp)}$, respectively.

Defining

$$\Lambda^* = diag\left\{\left(\Lambda^*_{(1)}, \ldots, \Lambda^*_{(p)}, \ldots, \Lambda^*_{(P)}\right)'\right\}$$

and $$\mathrm{T}^* = diag\left\{\left(\mathrm{T}^*_{(y1)}, \ldots, \mathrm{T}^*_{(yp)}, \ldots, \mathrm{T}^*_{(yP)}\right)'\right\}$$

we may now rewrite (4) by

$$\mathrm{T}^*\mathrm{X}'\mathrm{Y} - \Lambda^*\mathrm{X}'\mathrm{n}_{diag}\Pi = 0 \qquad (5)$$

Notice that (5) is the same as (1) except that the two terms are pre-multiplied by T^* and Λ^*. Notice that there only $2P$ counts that are perturbed (e.g. if there are 20 covariates in the model there will be 40 counts that are perturbed). In contrast, the method discussed in Section 3 perturbs up to $2C$ counts. As the number of cells to be perturbed increases, so does the uncertainty introduced into the estimates.

Also, since $E_\xi\left(\mathrm{T}^*\right) = E_\xi\left(\Lambda^*\right) = \mathrm{I}_P$, it follows that $E_\xi\left(\hat{\beta}^{**}\right) = \hat{\beta}$.

An algorithm for solving (5) is

1. Initialise the estimate of $\hat{\beta}^{**}_{(0)}$.
2. Update $\hat{\beta}^{**}_{(t+1)} = \hat{\beta}^{**}_{(t)} + \left(\mathrm{X}'\mathrm{W}_{(t)}\mathrm{X}\right)^{-1}\left(\mathrm{T}^*\mathrm{X}'\mathrm{Y} - \Lambda^*\mathrm{X}'\mathrm{n}_d\Pi_{(t)}\right)$,

   where for the logistic regression model,
   $\mathrm{W}_{(t)} = diag\left\{n_t\hat{\pi}^{**}_{i(t)}\left(1-\hat{\pi}^{**}_{i(t)}\right)\right\}$ and $\hat{\pi}^{**}_{i(t)} = \mathrm{logit}^{-1}\left(\mathrm{x}_i'\hat{\beta}^{**}_{(t)}\right)$.
3. Repeat 1 and 2 until convergence.

It is interesting to note that, for the simple model with $\mathrm{x}_i = 1$, $\hat{\beta}^{**}$ is obtained by solving the equation $y^* - n^*\pi = 0$ for β. It is easy to see that the method in Section 3 solves the same estimating equation meaning that, for this simple model, $\hat{\beta}^{**} = \hat{\beta}^*$. Since this equation is a function of only y^* and n^*, $\hat{\beta}^{**}$ must have an acceptable disclosure risk in this simple model.

5. EVALUATION ON THE NATIONAL HEALTH SURVEY

In this section we evaluate the method of perturbing the micro-data (PM) prior to analysis (see Section 3) and the method of Perturbing the counts that are present in the estimating equations (PEE) (see Section 4) when fitting a logistic model using the 2008 National Health Survey (NHS). The NHS is a multi-stage sample of dwellings and collects information from 20,788 people about their weight, health related aspects of life-style, and use of health services. The results below account for the complex design and unequal weighting of the NHS.

5.1 Models

All logistic models have 'over-weight' as the outcome variable (i.e. Body Mass Index (BMI) ≥ 25). Here we consider three different sets of explanatory variables, denoted by BMI1, BMI2 and BMI3.

The explanatory variables for BMI1 included sex and 10 year age groups, giving $P = 8$ dichotomous explanatory variables, including the intercept, and $C = 14$. The Perturbing Micro-data (PM) method perturbs $2 \times 14 = 28$ cells counts, whereas PEE perturbed $2 \times 8 = 16$ cells counts.

The explanatory variables for BMI2 included cross-classifying state with sex, 10 year age groups and five different exercise levels. Due to small sample sizes, some cross-classified variables were collapsed together. The BMI2 model had P = 77 parameters, which meant that EE perturbs 154 cell counts. In contrast, PM perturbs 753 cell counts – while C= 392 and the table underlying the logistic regression therefore had 784 cells, only 753 were non-zero and were therefore perturbed.

The explanatory variables for BMI3 included sex, 10 year age groups, asthma condition, type 2 diabetes status, smoker status, socio-economic index, and state cross classified by metropolitan/ex-metropolitan. The BMI3 model had $P = 37$ parameters. The table underlying the logistic regression model had 12,103 non-zero cell counts. This meant that PM perturbs 12,103 nonzero cells, most of which had very small unweighted cell counts and received a high amount of perturbation. In contrast, PEE perturbs only $2 \times 37 = 74$ cell counts, considerably less than the input method.

5.2 Evaluation

We now define terms that allow the impact of perturbation on the analysis output to be appreciated. The Increase in Mean Squared Error (IMSE) and the Standardised measure of Bias (SB) due to using $\hat{\boldsymbol{\beta}}^*$ (or $\hat{\boldsymbol{\beta}}^{**}$), rather than $\hat{\boldsymbol{\beta}}$, to estimate β_p are measured by

$$\widehat{IMSE_{M\xi}}\left(\hat{\beta}_p^*\right) = \frac{\widehat{MSE_{M\xi}}\left(\hat{\beta}_p^*\right)}{\widehat{Var_M}\left(\hat{\beta}_p\right)} - 1$$

and

$$\widehat{SB}_{\xi}\left(\hat{\beta}_p^*\right)=\left|\frac{\frac{1}{100}\sum_{b=1}^{100}\hat{\beta}_{(b)}^{*}-\hat{\beta}_p}{\sqrt{\widehat{Var}\left(\hat{\beta}_p\right)}}\right|$$

where $\hat{\beta}_{(b)}^{*}=\left(\hat{\beta}_{1(b)}^{*},\ldots,\hat{\beta}_{\nu(b)}^{*},\ldots,\hat{\beta}_{P(b)}^{*}\right)'$ and is defined in Section 3.

Table 5.1 gives the SB and IMSE averaged over the *P* regression coefficients. For PM, the perturbation variance increases with the number of cell counts in the underlying table that are perturbed. For example, table 5.1 shows that the IMSE is only 1% for BMI1, but it increases dramatically to 185% for BMI3. This means PM manages disclosure risk at a cost of effectively decreasing the sample size by almost two thirds. As a result, PM appears to be unsuitable as a general way of managing disclosure risk for a remote analysis server.

In contrast, for PEE, the loss of accuracy for BMI3 equates to a sample size reduction of only 1.7%. This loss of accuracy is insignificant. Also, the standardized bias supports the theoretical result that the regression coefficients from PEE are unbiased.

5.1 Effect of perturbation on estimates of regression coefficients

Method	*Model*	*SB(%)*	*IMSE (%)*
Perturbing Micro-data (PM)	BMI1	0.2	1.2
	BMI2	2.2	32.7
	BMI3	6.1	184.6
Perturbing Estimating Equation (PEE)	BMI1	0.1	1.4
	BMI2	0.7	20.0
	BMI3	0.3	1.7

Table 5.2 gives the average value and the Monte-Carlo standard error over 100 independent perturbations for the Pearson Chi-squared and R-squared diagnostics. The results show that these diagnostics are essentially the same as those based on the true data.

5.2 Impact of perturbation on PEE select diagnostics

	Pearson Chi-squared			*Pearson R-squared*		
		PEE			*PEE*	
Model	*True Micro-data*	*Average*	*Standard error*	*True Micro-data*	*Average*	*Standard error*
BMI1	15,749	15,749	5.0	0.05	0.05	8×10^{-5}
BMI2	15,671	15,683	19.0	0.06	0.06	4×10^{-4}
BMI3	15,447	15,449	8.4	0.06	0.06	1.2×10^{-4}

6. DISCUSSION

Historically agencies disseminated estimates, from surveys or from administrative micro-data, to the public in a publication format. There was little or no scope for analysts to access the micro-data itself. With the increasing sophistication and demands of analysts this situation is rapidly changing. Internationally there is very strong demand from analysts, particularly within government and universities, for flexible access to micro-data for the purpose of developing and evaluating policy. The utility or public benefit of allowing such access is hard to over-state, especially given the range of micro-data that is collected by government agencies. In allowing access, the confidentiality of the individuals or businesses about which the data relate must be protected.

One way of facilitating access to micro-data is through a remote server. A remote server automatically returns the output from remotely submitted statistical programming code. The purpose of the remote server is to ensure that any output that is returned to the analyst meets certain confidentiality criteria. On the other hand, if the server is to be a useful tool for analysts, any restrictions imposed by the server will need to be minimal.

This paper evaluates methods of managing the risk of disclosure when releasing both population counts *and* statistical output from generalised linear models fitted to categorical variables. In the latter case, the preferred method perturbs the counts present in the model's estimation equation which, the empirical evaluation shows, has an insignificant impact on the output. The method allows analysts to define variables without restriction, releases variance estimates on the regression coefficients, and select diagnostics.

REFERENCES

Binder, D.A. (1983) "On the Variances of Asymptotically Normal Estimators from Complex Surveys", *International Statistical Review*, 51(3), pp. 279–292.

Chambers, R.L. and Skinner, C.J. (eds.) (2003) *Analysis of Survey Data*, John Wiley and Sons, Chichester.

Duncan, G.T. and Mukherjee, S. (2000) "Optimal Disclosure Limitation Strategy in Statistical Databases: Deterring Tracker Attacks Through Additive Noise", *Journal of the American Statistical Association*, 95(451), pp. 720–729.

Dwork, C. and Smith, A. (2009) "Differential Privacy for Statistics: What We Know and What We Want to Learn", *Journal of Privacy and Confidentiality*, 1(2), pp. 135–154.

Fienberg, S.E. (1994) "Conflicts between the Needs for Access to Statistical Information and Demands for Confidentiality", *Journal of Official Statistics*, 10(2), pp. 115–132.

Fienberg, S.E. and Makov, U.E. (1998) "Confidentiality, Uniqueness and Disclosure Limitation for Categorical Data", *Journal of Official Statistics*, 14(4), pp. 385–397.

Fraser, B. and Wooton, J. (2005) "A Proposed Method for Confidentialising Tabular Output to Protect against Differencing", *UNECE Work Session on Statistical Data Confidentiality*.

Gomatam, S.;, Karr, A.F.; Reiter, J.P. and Sanil, A.P. (2008) "Data Dissemination and Disclosure Limitation in a World Without Microdata: A Risk–Utility Framework for Remote Access Analysis Servers", *Statistical Science*, 20(2), pp. 163–177.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*, John Wiley and Sons, New York.

Lucero, J. and Zayatz, L. (2010) "The Microdata Analysis System at the U.S. Census Bureau. Privacy in Statistical Databases", in Domingo-Ferrer, J. and Magkos, E. (eds.) *PSD 2010: Proceedings of the 2010 International Conference on Privacy in Statistical Databases*, Springer-Verlag, Berlin, Heidelberg, pp. 234–248.

McCulloch, C.E. and Searle, S.R. (2001) *Generalized, Linear and Mixed Models*, John Wiley & Sons, New York.

O'Keefe, C.M. and Good, N. (2008) "A Remote Analysis Server – What Does Regression Output Look Like?", in *PSD 2008: Proceedings of the 2008 International Conference on Privacy in Statistical Databases*, Springer-Verlag, Berlin, Heidelberg.

Raghunathan, T.E.; Reiter, J.P. and Rubin, D.B. (2003) "Multiple Imputation for Statistical Disclosure Control", *Journal of Official Statistics*, 19(1), pp. 1–16.

Rao, J.N.K. and Wu, C.F.J. (1988) "Resampling Inference with Complex Survey Data", *Journal of the American Statistical Association*, 83(401), pp. 231–241.

Reiter, J.P. (2002) "Satisfying Disclosure Restrictions with Synthetic Data Sets", *Journal of Official Statistics*, 18(4), pp. 531–543.

Reiter, J.P. (2003) "Model Diagnostics for Remote-Access Regression Servers", *Statistics and Computing*, 13(4), pp. 371–380.

Reiter, J.P. and Kohnen, C.N. (2005) "Categorical Data Regression Diagnostics for Remote Servers", *Journal of Statistical Computation and Simulation*, 75, pp. 889–903.

Reiter, J.P.; Oganian, A. and Karr, A.F. (2007) "Evaluating the Disclosure Risks of Reporting Quality Measures to the Public", in *Work Session on Statistical Data Confidentiality*, UNECE/Eurostat, pp. 54–65.

Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*, Springer-Verlag, New York.

Sparks, R.; Carter, C.; Donnelly, J.; O'Keefe, C.M.; Duncan, J.; Keighley, T. and McAullay, D. (2008) "Remote Access Methods for Exploratory Data Analysis and Statistical Modelling: Privacy-Preserving Analytics^{TM}", *Computer Methods and Programs in Biomedicine,* 91(3), pp. 208–222.

Willenborg, L. and de Waal, T. (2001) *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, 155 , Springer-Verlag, New York.

APPENDIX

A.1 Using output from TableBuilder and the Analysis Server to affect disclosure

By way of a simple example, we show that we need to defend against attacks that use the output from TableBuilder and the analysis server to affect disclosure. Consider the simple linear regression model with $\mathrm{x}_i = 1$ and constant variance (so we may drop the i subscript) given by

$$y = \beta + e$$

The estimate of β is $\beta = y/n$. If an analyst is given β from the analysis server, they can estimate y by βn, where $n = S(n)$ is obtained from TableBuilder.

Consider what this means by way of a simple example. Let $n = 100$ and $y = 1$, neither of which are disclosed to the analyst. If an analyst obtains $\hat{\beta} = 0.01$ from the analysis server and $n^* = 103$ (the perturbed value of $n = 100$) from TableBuilder, their estimate of y is $103 \times 0.01 = 1.03$. The estimate 1.03 is very close to the true value of 1 and effectively constitutes disclosure. It is easy to see that this kind of attack will always be successful when n is much larger than y.

A.2 Influence diagnostics

Leverage

The leverage, h_i is often used to determine the degree of influence of a particular set of covariates, x_i. Plotting h_i against predicted probabilities, π_i, for example, helps to identify systematic patterns in the leverage that may require adjustments to the model's specification.

The standard leverage for the logistic regression model with coefficient $\hat{\boldsymbol{\beta}}^*$ is

$$h_i = n_i \hat{\pi}_i^* \left(1 - \hat{\pi}_i^*\right) \mathrm{x}'_j \left(\mathrm{X}'\mathrm{W}^*\mathrm{X}\right)^{-1} \mathrm{x}_j$$

where W^* has i-th diagonal element $n_i \hat{\pi}_i^* \left(1 - \hat{\pi}_i^*\right)$. The corresponding confidentialised version of the leverage, that can be released to the analyst, is

$$h_i^* = n_i^* \hat{\pi}_i^* \left(1 - \hat{\pi}_i^*\right) \mathrm{x}'_j \widehat{Var_{JK}} \left(\hat{\boldsymbol{\beta}}\right)^{-1} \mathrm{x}_j$$

The difference between h_i and h_i^* will tend to be dominated by the uncertainty in $\widehat{Var_{JK}}\left(\hat{\boldsymbol{\beta}}\right)$, which decreases with R.

Cook's Distance

Cooks distance, adapted to the logistic regression model (see Hosmer and Lemeshow, 2000, p. 173), measures the influence of a covariate pattern on $\hat{\beta}^*$. This distance measure is given by

$$\Delta\hat{\beta}_j^* = \left(\hat{\beta}^* - \hat{\beta}_j^*\right)' \left(X'W^*X\right)\left(\hat{\beta}^* - \hat{\beta}_j^*\right)$$

where $\hat{\beta}_j^*$ has the same form as $\hat{\beta}^*$ except that all records with the i-th covariate pattern are excluded.

The corresponding distance measure with an acceptable disclosure risk is

$$\Delta^*\hat{\beta}_j^* = \left(\hat{\beta}^* - \hat{\beta}_j^*\right)' \widehat{Var_{JK}}\left(\hat{\beta}\right)^{-1} \left(\hat{\beta}^* - \hat{\beta}_j^*\right)$$

The difference between $\Delta^*\hat{\beta}_j^*$ and $\Delta\hat{\beta}_j^*$ is only due to the uncertainty in $\widehat{Var_{JK}}\left(\hat{\beta}\right)$, which decreases with R.

Both h_i^* and $\Delta^*\hat{\beta}_j^*$ require $\widehat{Var_{JK}}\left(\hat{\beta}_k\right)$. Section 3 shows that this term has a sufficient degree of uncertainty in reasonably large samples. In small samples, the properties of $\widehat{Var_{JK}}\left(\hat{\beta}_k\right)$ are not clear.

As a result, we suggest only releasing h_i^* and $\Delta^*\hat{\beta}_j^*$ when the sample size is reasonably large (i.e when $n > 60$). This is not an onerous restriction since $\widehat{Var_{JK}}\left(\hat{\beta}_k\right)$ (and therefore h_i^* and $\Delta^*\hat{\beta}_j^*$) would be poorly estimated when $n < 60$.

A.3 Alternative methods for confidentialising diagnostic plots

Confidentiality for a graph is equivalent to ensuring that every (x-axis, y-axis) data point plotted has low confidentiality risk – in other words, each data point only discloses a limited amount of information about each of the unit records that it summarises over.

In general, confidentialisation is generally achieved through aggregation, perturbation, or synthesisation of data values. Aggregation diminishes risk by averaging values across several units and suppressing finer-grained information, making it difficult to reverse to determine information about any particular record. If the aggregates created are large enough, additional information about them can be released, such as quantiles or standard deviations.

In contrast, perturbation and synthesisation are techniques for lowering the risk of specific variables, and may be applied to each axis independently. Perturbation is generally applied in combination with aggregation as this diminishes the influence of any single unit on the outputs, enabling less noise to be added to the data for effective confidentialisation.

The variables used for diagnostic plots cover a wide range of risk levels. Plots of observed values or residuals for units are of most concern, as these directly disclose information about specific unit characteristics.

Predicted values for units are somewhat less sensitive, being a function $\pi_{(k)} = f\left(x_{(k)}\beta\right)$ of the regression parameter values $\boldsymbol{\beta}$ and units' dependent variables $x_{(k)}$.

Assuming that the $\boldsymbol{\beta}$ values used to calculate these are not sensitive or have been confidentialised, predicted values provide strictly less information than would be gained by releasing the set of dependent variables $x_{(k)}$ for each unit. In the presence of continuous x-variables, it is also difficult to infer information about individual members of $x_{(k)}$ from $x_{(k)}\beta$.

The majority of graphs containing categorical variable x-axes are of low risk, as they naturally aggregate data by x-categories. Provided all x-categories contain a minimum threshold number of records, this makes it difficult to relate plotted values back to individual records.

The Reiter and Kohnen (2005) proposal used as a basis for this paper's approach achieves confidentiality for observed, residual, and partial residual scatterplots mainly through aggregating identical x-data values and replacing individual y-data values with the aggregate means. It also advises using a small amount of perturbation to avoid potential confidentiality breaches when these means are close to the minimum or maximum possible y-values. We additionally suggest that predicted values displayed in the plot should be computed based on perturbed parameter estimates, to be consistent with other analysis server modules.

Privacy-Preserving Analytics (O'Keefe and Good, 2008) uses aggregation alone to confidentialise these scatterplots, grouping units with similar x-axis values together to create a display of parallel boxplots. The y-axis structure of the data is shown through outputting medians and quartiles. For discrete x-variables with a sufficient number of observations in each category, grouping may not be required. They also recommend using smoothing to confidentialise qq-plots (which plot the actual distribution of residuals against the theoretical GLM distribution in order to check model distributional assumptions).

Both of the above papers advise that it is likely impossible to release confidentialised influence plots that still allow individual outliers to be identified. PPA instead suggests using either robust regression or automatic outliering procedures to mitigate outliers' effect on analyses.

Reiter (2003) gives a practical method of constructing synthetically confidentialised diagnostic plots. His approach is to preconstruct a dataset of synthetic x-values using simple techniques such as resampling (for categorical variables) or kernel density estimation (for continuous variables), then generate synthetic y-residuals for these on-the-fly by drawing samples from a generalised additive model, fitted on the actual x-values and residuals. The USBC Microdata Analysis System (MAS) uses similar techniques to confidentialise diagnostic plots (Lucero and Zayatz, 2010).

Future ABS investigations will include an evaluation of these various options to determine which approach is best able to provide useful diagnostics to users while maintaining data confidentiality.

1352.0.55.110 • RESEARCH PAPER: ANALYSIS OF MICRO-DATA: CONTROLLING THE RISK OF DISCLOSURE (METHODOLOGY ADVISORY COMMITTEE) • June 2010

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS website is the best place for data from our publications and information about the ABS.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS	www.abs.gov.au

Produced by the Australian Bureau of Statistics