ASSESSING IDENTIFICATION RISK IN SURVEY MICRODATA USING LOG-LINEAR MODELS

CHRIS SKINNER, NATALIE SHLOMO

ABSTRACT

This article considers the assessment of the risk of identification of respondents in survey microdata, in the context of applications at the United Kingdom (UK) Office for National Statistics (ONS). The threat comes from the matching of categorical 'key' variables between microdata records and external data sources and from the use of log-linear models to facilitate matching. While the potential use of such statistical models is well-established in the literature, little consideration has been given to model specification nor to the sensitivity of risk assessment to this specification. In this article we develop new criteria for assessing the specification of a log-linear model in relation to the accuracy of risk estimates. We find that, within a class of 'reasonable' models, risk estimates tend to decrease as the complexity of the model increases. We develop criteria to detect 'underfitting' (associated with overestimation of the risk). The criteria may also reveal 'overfitting' (associated with underestimation) although not so clearly, so we suggest employing a forward model selection approach. We show how our approach may be used for both file-level and record-level measures of risk. We evaluate the proposed procedures using samples drawn from the 2001 UK Census where the true risks can be determined. We also apply our approach to a large survey dataset.

Southampton Statistical Sciences Research Institute
Methodology Working Paper M06/14

Assessing Identification Risk in Survey Microdata using Log-linear Models

CHRIS SKINNER and NATALIE SHLOMO*

ABSTRACT. This article considers the assessment of the risk of identification of respondents in survey microdata, in the context of applications at the United Kingdom (UK) Office for National Statistics (ONS). The threat comes from the matching of categorical 'key' variables between microdata records and external data sources and from the use of log-linear models to facilitate matching. While the potential use of such statistical models is well-established in the literature, little consideration has been given to model specification nor to the sensitivity of risk assessment to this specification. In numerical work not reported here, we have found that standard techniques for selecting log-linear models, such as chi-squared goodness of fit tests, provide little guidance regarding the accuracy of risk estimation for the very sparse tables generated by typical applications at ONS, for example tables with millions of cells formed by cross-classifying six key variables, with sample sizes of 10 or 100 thousand. In this article we develop new criteria for assessing the specification of a log-linear model in relation to the accuracy of risk estimates. We find that, within a class of 'reasonable' models, risk estimates tend to decrease as the complexity of the model increases. We

*Chris Skinner is Professor, Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, United Kingdom. Natalie Shlomo is research student, S3RI, University of Southampton and Department of Statistics, Hebrew University, Jerusalem, Israel.

develop criteria to detect 'underfitting' (associated with overestimation of the risk). The criteria may also reveal 'overfitting' (associated with underestimation) although not so clearly, so we suggest employing a forward model selection approach. Our criteria turn out to be related to established methods of testing for overdispersion in Poisson log-linear models. We show how our approach may be used for both file-level and record-level measures of risk. We evaluate the proposed procedures using samples drawn from the 2001 UK Census where the true risks can be determined. We find the proposed approach is successful in detecting underfitting models which generate overestimates of the risk. The approach also helps to detect overfitting models which lead to underestimation. We employ a forward model selection approach and show how this leads to good risk estimates. There are several 'good' models between which our approach provides little discrimination. The risk estimates are found to be stable across these models, implying a form of robustness. We also apply our approach to a large survey dataset. There is no indication that increasing the sample size necessarily leads to the selection of a more complex model. The risk estimates for this application display more variation but suggest a suitable upper bound.

KEY WORDS: Confidentiality; Disclosure; Key variable; Matching; Model specification.

1 INTRODUCTION

Statistical agencies often wish to provide researchers with access to survey microdata, but must balance this aim against the need to protect the confidentiality of the respondents. In particular, many agencies have policies which require them to control the risk of identification. For example, the key 'confidentiality guarantee' in the United Kingdom (UK) National Statistics Code of Practice (National Statistics, 2004, p.7) is

that 'no statistics will be produced that are likely to identify an individual'.

The developing field of statistical disclosure limitation methodology provides agencies with many methods to protect confidentiality and, in particular, to assess identification risk (Willenborg and de Waal, 2001; Doyle, Lane, Theeuws and Zayatz, 2001). Traditional methods to assess identification risk include the use of rules and check lists based on institutional experience, simple data-based summary measures and re-identification experiments (Federal Committee on Statistical Methodology, 1994). Such methods can be somewhat ad hoc, however, and number of authors (e.g. Paass, 1988; Duncan and Lambert, 1989; Fuller, 1993) have proposed statistical modelling frameworks which permit identification risk to be assessed following clear statistical principles. Identification may be treated as a form of statistical inference by a potential 'intruder', who is assumed to make efficient use of available information to facilitate identification through specified models. There have been some applications of such modelling approaches to assessing risk. Reiter (2005) applied the approach of Duncan and Lambert (1989) to the Current Population Survey. Paass (1988) applied discriminant analysis to two microdata files from the German Federal Statistical Office. Bethlehem , Keller and Pannekoek (1990) applied a Poisson-Gamma model to Dutch data. Nevertheless, more research on issues arising in applications is needed if modelling methods are to become part of the standard risk assessment 'toolkit' of statistical agencies. In particular, more understanding is needed of how to specify models and of how sensitive risk assessment approaches are to specification.

The purpose of this article is to investigate the use of log-linear modelling methods in some risk assessment problems which have arisen at the UK Office for National Statistics (ONS) when releasing microdata from social surveys. In addition to considering here one particular survey application, we draw samples from the 2001 UK Census to mimic social survey data in a setting where population values are avail-

able for validation. In line with the Code of Practice mentioned above, the aim is to protect against identification which could arise from an intruder matching a microdata record to a known population individual using the values of variables which are both available in the microdata and traceable or visible externally. These variables are called key variables (Bethlehem et al., 1990). For the kinds of social survey applications considered by ONS, these key variables are invariably categorical, e.g. sex, age, ethnicity, religion, place of residence or occupation. Previous work has shown that, when multivariate categorical key variables are available, an intruder might be able to use log-linear modelling to improve their chances of identifying records (Skinner and Holmes, 1998; Fienberg and Makov, 1998; Elamir and Skinner, 2006). However, this work has given little attention to the important practical issue of how to specify these models or to the sensitivity of risk assessment to model specification.

The main aim of this paper is to develop and investigate approaches to specifying log-linear models, which are suitable for use in practice by a statistical agency for the very large and sparse cross-classified tables arising in the kinds of application considered here and which directly address the risk assessment objectives. We shall argue that these objectives can be represented as certain prediction problems and thus differ from the standard kinds of objectives of log-linear modelling (e.g. Bishop, Fienberg and Holland, 1975). Our approach will be to develop diagnostic criteria of model adequacy for such prediction purposes.

The kinds of risk measures considered here, based on log linear modelling, may be used to assess the impact of recoding the key variables, which is the primary method of disclosure limitation used at present by ONS in the release of social survey microdata, alongside the use of restrictions on access arrangements, such as via licenses or on-site laboratories. As noted by Fuller (1993), for example, the protection provided by perturbative disclosure limitation methods, such as noise addition, may be better

assessed using other risk measures, such as relating to predictive disclosure. But such perturbative methods are rarely contemplated at present by ONS because of their potential impact on analysis and are not considered further in this article.

The article is organised as follows. The framework for identification risk assessment is set out in Section 2, with the associated log-linear models discussed in Section 3. Section 4 describes possible criteria for assessing the model and Section 5 describes how these might be used to specify a model. Section 6 presents the application to census samples. Section 7 presents the application to a social survey. Finally, Section 8 contains a discussion and areas for future research.

2 IDENTIFICATION RISK ASSESSMENT

Following several authors (e.g. Paass, 1988; Duncan and Lambert; 1989; Bethlehem et al., 1990), we consider a microdata file consisting of records for a sample of individuals from a finite population. We imagine an intruder with access to the file as well as to auxiliary information on the values of the key variables for some known individuals in the population. The intruder matches the two data sources in order to identify one or more records in the microdata. We suppose the intruder assesses whether there is a microdata record and a known individual for which the probability that the former belongs to the latter is high (Paass, 1988; Duncan and Lambert, 1989). Our basic definition of identification risk is the value of this probability when the microdata record does indeed belong to the known individual.

We conceive of this probability as conditional on data, which might reasonably be assumed available to the intruder, and defined with respect to a model and assumptions, which are justifiable from analysis of the data and from knowledge of the processes (sample selection, measurement error etc.) generating the data. We treat the key

variables as given by a specified scenario, as in Paass (1988). In the kinds of census and social survey applications of concern here, we may assume that the key variables are categorical. A stronger assumption that we shall make is that the key variables are measured in the same way in the two sources, so there is no measurement error to create discrepancies. Ignoring such discrepancies may be expected to lead to overestimation of risk and the risk estimates reported in this article may therefore be considered to be conservative. The treatment of measurement error would be a key extension of our approach but is beyond the scope of this paper.

To introduce our basic measure of identification risk, let F_k be the population count in cell k of the multi-way contingency table formed by cross-classifying the key variables (with cells labelled $k = 1, \cdots, K$). Under the above assumptions, together with weak exchangeability assumptions about the selection of records and known population individuals, and the assumption that F_k is known to the intruder, the definition of identification risk above, i.e. the probability that a microdata record may be identified, takes the form $1/F_k$, where k is the cell to which the record belongs (Duncan and Lambert, 1989). The risk is maximum when the record is population unique, i.e. $F_k = 1$. In practice, the agency should ensure that key variables are not released where intruders are able to determine small values of F_k using, for example, population lists (Skinner, Marsh, Openshaw and Wymer, 1994). A more realistic measure is therefore given by $E(1/F_k) = \sum_r P(F_k = r)/r$, where $P(F_k = r)$ denotes the probability that $F_k = r$ under the model (r=1, 2, $\cdots$), given data available to the intruder (Skinner and Holmes, 1998). Given the particular concern about population uniqueness (e.g. Bethlehem et al., 1990), a related risk measure of interest is $P(F_k = 1)$, the probability of population uniqueness. This is the first term in the sum $\sum_r P(F_k = r)/r$. Given the models we shall consider later and treating the microdata as the available data, the sufficient statistics will consist of the sample counts f_k in the cells k=1, $\cdots$, K.

Treating the pairs $(F_k,\ f_k)$ as independent, the first risk measure may be expressed more explicitly in terms of the available data as $E(1/F_k \mid f_k)$ and will generally be highest when $f_k = 1$, i.e. in sample unique cells. Moreover, the probability of population uniqueness is only non-zero when $f_k = 1$. Consideration of worst cases thus leads to a focus on the measures $r_{1k} = P(F_k = 1 \mid f_k = 1)$ and $r_{2k} = E[1/F_k \mid f_k = 1]$.

These are referred to as record-level or per record measures (Willenborg and de Waal, 2001, p.52) since they vary between records. More generally, we write $r_k = E[g(F_k) \mid f_k = 1]$, where $g(F) = I(F = 1)$ or $1/F$ in the case of r_{1k} or r_{2k}, respectively. Estimation of such record-level measures may help the agency identify and target 'high risk' records for the application of 'local' disclosure limitation methods. Nevertheless, agencies often also need measures of risk at the file level in their decision making processes, such as in the assessment of recoding options, and this leads to consideration of aggregating such record-level measures (Lambert, 1993; Fienberg and Makov, 1998). Here, we consider simply summing the record-level measures across sample unique records, to give $\tau^* = \sum_{SU} r_k$ and, in particular:

$$\tau_1^* = \sum_{SU} r_{1k} = \sum_{SU} P(F_k = 1 \mid f_k = 1), \tag{1}$$

the expected number of sample uniques that are population unique, and

$$\tau_2^* = \sum_{SU} r_{2k} = \sum_{SU} E(1/F_k \mid f_k = 1), \tag{2}$$

the expected number of correct matches for sample uniques, where $SU = \{k : f_k = 1\}$ denotes sample unique cells. Our focus will be on situations where K is large (and the (F_k, f_k) may be treated as independent) so that a law of large numbers implies that τ^* will closely approximate $\tau = \sum_k I(f_k = 1)g(F_k)$, which takes the particular forms $\tau_1 = \sum_k I(f_k = 1, F_k = 1)$ or $\tau_2 = \sum_k I(f_k = 1)/F_k$. Such measures may be more appealing to some statistical agencies since they have a model-free interpretation.

For any of the measures above, the problem of risk assessment becomes one of statistical inference if the f_k are observed but the F_k are not. In the case of τ, we may view this as a problem of finite population prediction. While there do exist some measures for which design-based survey sampling techniques can provide reliable inference (Skinner and Elliot, 2002), it is mostly necessary to base inference upon models.

3 LOG-LINEAR MODELS

Models are required not only for the explicit definition of most of the risk measures in the previous section, but also for inference about these measures. Following standard methods for contingency tables (e.g. Bishop et al., 1975) and previous work on disclosure control (e.g. Bethlehem et al., 1990), we consider models where the F_k are realisations of independent Poisson random variables with means λ_k ($k = 1, \cdots, K$). We write $F_k \sim P(\lambda_k)$. In order to develop relatively simple procedures, we shall assume that the sample is drawn by Bernoulli sampling with common inclusion probability π so that the sample counts f_k are also independent Poisson random variables: $f_k \sim P(\pi\lambda_k)$. In practice, the sampling schemes employed in surveys are more complex than this and we shall comment on this issue further in section 8. At least in the applications we consider in sections 6 and 7, the inclusion probabilities are equal. It follows from the above assumptions that $F_k \mid f_k \sim P[\lambda_k(1-\pi)] + f_k$ so that the record level measures may be expressed as $r_{1k} = \exp[-(1-\pi)\lambda_k] = h_1(\lambda_k)$, say, and, $r_{2k} = \{1 - \exp[-(1-\pi)\lambda_k]\}/[(1-\pi)\lambda_k] = h_2(\lambda_k)$ say, or, more generally, $r_k = E[g(F_k) \mid f_k = 1] = h(\lambda_k)$, say, where $h(\lambda)$ is a monotonic decreasing function of λ. We write the aggregated risk measures as:

$$\tau^* = \sum_k I(f_k = 1)h(\lambda_k). \tag{3}$$

The modelling assumptions so far are generally insufficient to make precise inference about these risk measures since the measures depend on unknown λ_k values for cells where the observed counts f_k are just one. In order to 'borrow strength' between cells we suppose the λ_k are related via the log linear model:

$$\log \lambda_k = x_k' \beta, \tag{4}$$

where x_k is a qx1 design vector, depending on the values of the key variables in cell k, and β is a qx1 parameter vector. Typically, we shall specify x_k to include main effects and low order interactions of the categorical key variables (Bishop et al., 1975). Since the f_k are the outcomes of independent $P(\pi\lambda_k)$ random variables, the maximum likelihood (ML) estimator $\hat{\beta}$ may be obtained by solving the score equations:

$$\sum_k [f_k - \pi \exp(x_k'\beta)] x_k = 0. \tag{5}$$

using numerical techniques. The risk measures in Section 2 may then be estimated by replacing λ_k by $\hat{\lambda}_k = \exp(x_k'\hat{\beta})$ in the expressions above, for example $\hat{\tau} = \sum_k I(f_k = 1) h(\hat{\lambda}_k)$. Such an approach has been described in Skinner and Holmes (1998) and Elamir and Skinner (2006), who have shown how it may generate useful risk measures. See also Fienberg and Makov (1998). The problem addressed in this paper is that inference may be sensitive to the specification of (4). We propose an approach in the next section to check the adequacy of this specification. We shall assume that, given a specified model of form (4), inference proceeds in the simple manner above, i.e. by plugging $\hat{\lambda}_k$ in for λ_k in the risk measure expressions. Other more sophisticated approaches are possible, for example averaging over alternative models (Fienberg and Makov, 1998), but will not be considered here.

4 CRITERIA FOR MODEL ASSESSMENT

4.1 Rationale

We seek criteria for assessing whether the vector x_k in the log-linear model in (4) may be expected to lead to accurate estimated risk measures. One approach would be to use goodness-of-fit criteria such as Pearson or likelihood-ratio tests. These are not designed for finite population prediction problems, however. Moreover, the usual conditions on the average cell size n/K required for their validity (e.g. at least 1 or 5) do not hold for the large and sparse tables typical of the kinds of applications considered here. For example, the survey that is assessed in Section 7 has 127,200 records in 2,366,000 cells defined by six identifying key variables, and the average cell size is 0.05. Some work on sparse tables (Koehler, 1986) suggests that the Pearson test is preferable to the likelihood ratio test in such circumstances. Nevertheless, our empirical work has suggested that neither of these criteria, nor other standard approaches such as Akaike's Information Criterion, are very successful in deciding whether the disclosure risk measures will be well estimated and we shall not consider them further in this paper.

Instead, we consider an approach motivated more directly by our aim to estimate the risk measures accurately. Specifically, we seek a criterion for choosing a specification of model (4) which minimises the error (in a sense to be defined) of $\hat{\tau} = \sum_k I(f_k = 1)h(\hat{\lambda}_k)$ as an estimator of $\tau^* = \sum_k I(f_k = 1)h(\lambda_k)$ or as a predictor of $\tau = \sum_k I(f_k = 1)g(F_k)$. See Rao and Wu (2001) for a general discussion of the use of prediction criteria in model selection. Empirical work suggests that, within a neighbourhood of 'reasonable' models, $\hat{\tau}$ tends to decline the more complex the model. To provide some heuristic

theoretical reasoning for this phenomenon, let $\tilde{\beta}$ be the solution of

$$\sum_k [\lambda_k - \exp(x_k'\beta)]x_k = 0, \tag{6}$$

interpreted as an 'average' value of $\hat{\beta}$ across its sampling distribution and let $\tilde{\lambda}_k = \exp(x_k'\tilde{\beta})$ be a corresponding 'average' value of $\hat{\lambda}_k$. We can think of the estimation error $\hat{\lambda}_k - \lambda_k$ as composed of the sum of a 'sampling error' $\hat{\lambda}_k - \tilde{\lambda}_k$ and a 'misspecification error' $\tilde{\lambda}_k - \lambda_k$ and, via these components, consider two problems.

Overfitting: this is the case where the model is 'too complex' in the sense that the sampling error is positively associated with f_k (in the extreme case of a saturated model $\hat{\lambda}_k = f_k/\pi$) and where this sampling error is the dominant component of estimation error. We consider applications where the expected sample size per cell is less than one so that $I(f_k = 1)$ tends to be positively associated with f_k. Since h is a monotonic decreasing function, we may expect that, in the presence of overfitting, $I(f_k = 1)$ tends to be positively associated with $\hat{\lambda}_k - \lambda_k$ and negatively associated with $h(\hat{\lambda}_k) - h(\lambda_k)$ and thus for $\hat{\tau}$ to underestimate τ^*. Another reason to expect this outcome is that overfitting may produce too many fitted marginal zero counts where sample marginal counts are random zeros, leading to fitted cell counts being too high for the non-zero cells of the table and risk measures being underestimated.

Underfitting: this is the case where $\tilde{\lambda}_k$ is 'oversmoothed', so that there is negative association between $\tilde{\lambda}_k - \lambda_k$ and λ_k, and misspecification error is the dominant component of estimation error. It follows that $\hat{\lambda}_k - \lambda_k$ is also negatively associated with λ_k. Now, we expect f_k to be positively associated with λ_k and thus (when the expected sample size per cell is less than one) for $I(f_k = 1)$ to be negatively associated with $\hat{\lambda}_k - \lambda_k$ and positively associated with $h(\hat{\lambda}_k) - h(\lambda_k)$ and thus for $\hat{\tau}$ to overestimate τ^*. Another reason to expect this outcome is that structural

zero counts in tables (which often cannot be identified easily in advance) may fail to be fitted correctly in the presence of underfitting, leading to expected cell counts tending to be too low for the non-zero cells of the table and risk measures being overestimated.

Our empirical experience (as will be illustrated in sections 6 and 7) is that it is harder to detect the impact of overfitting than underfitting. Our development of a data-based criterion for minimising estimation error is therefore led by consideration of the impact of the latter.

4.2 Development of Criterion

We represent the impact of underfitting by the component of the bias of $\hat{\tau}$ as an estimator of τ^* or predictor of τ arising from misspecification of the model, that is from the difference between $\tilde{\lambda}_k$ and λ_k, i.e:

$$B = \sum_k E[I(f_k = 1)][h(\tilde{\lambda}_k) - h(\lambda_k)] = \sum_k \pi\lambda_k \exp(-\pi\lambda_k)[h(\tilde{\lambda}_k) - h(\lambda_k)]. \quad (7)$$

We approximate the term $h(\tilde{\lambda}_k)$ in this expression by

$$h(\tilde{\lambda}_k) \doteq h(\lambda_k) + h'(\lambda_k)(\tilde{\lambda}_k - \lambda_k) + h''(\lambda_k)(\tilde{\lambda}_k - \lambda_k)^2/2, \quad (8)$$

using a quadratic expansion of $h(\tilde{\lambda}_k)$ around λ_k. For example, when $h(\lambda) = h_1(\lambda)$, we obtain $h'(\lambda_k) = -(1-\pi)h_1(\lambda_k)$ and $h''(\lambda_k) = (1-\pi)^2 h_1(\lambda_k)$. To illustrate the quality of the approximation, consider the value $\lambda_k = 1$ which might be taken to be the value of most concern, being the value when $F_k = 1$ is most likely. Figure 1 plots $h(\tilde{\lambda})$ and its approximation in (8) against $\tilde{\lambda}$ for π=0.05 and the two choices of h function considered above equation (3). The approximation works well for the range of $\tilde{\lambda}$ values plotted

and potential problems with the approximation at the extremes are mitigated by the lower bound $\lambda_k > 0$ and the damping effect of $\exp(-\pi\lambda_k)$ in (7) for large values of λ_k.

Substituting approximation (8) into (7) gives:

$$B \doteq \sum_k \pi\lambda_k \exp(-\pi\lambda_k)[h'(\lambda_k)(\tilde{\lambda}_k - \lambda_k) + h''(\lambda_k)(\tilde{\lambda}_k - \lambda_k)^2/2]. \tag{9}$$

Since $E(f_k) = \mu_k = \pi\lambda_k$ and $E[(f_k - \pi\tilde{\lambda}_k)^2 - f_k] = \pi^2(\lambda_k - \tilde{\lambda}_k)^2$, it follows that, for a large number of cells, expression (9) may be approximated by

$$\tilde{B} = \sum_k \lambda_k \exp(-\mu_k)\{-h'(\lambda_k)(f_k - \pi\tilde{\lambda}_k) + h''(\lambda_k)[(f_k - \pi\tilde{\lambda}_k)^2 - f_k]/(2\pi)\}. \tag{10}$$

In the case of underfitting, when $f_k - \pi\tilde{\lambda}_k$ may be reasonably approximated by $f_k - \pi\hat{\lambda}_k$, a natural estimator of $\tilde{B}$ and hence of B is

$$\hat{B} = \sum_k \hat{\lambda}_k \exp(-\hat{\mu}_k)\{-h'(\hat{\lambda}_k)(f_k - \hat{\mu}_k) + h''(\hat{\lambda}_k)[(f_k - \hat{\mu}_k)^2 - f_k]/(2\pi)\}. \tag{11}$$

We write $\hat{B}$ as $\hat{B}_1$ or $\hat{B}_2$ when $h(\lambda) = h_1(\lambda)$ or $h(\lambda) = h_2(\lambda)$ respectively, for example

$$\hat{B}_1 = \sum_k \hat{\lambda}_k \exp(-\hat{\lambda}_k)(1-\pi)\{(f_k - \hat{\mu}_k) + (1-\pi)[(f_k - \hat{\mu}_k)^2 - f_k]/(2\pi)\}. \tag{12}$$

We have argued that $\hat{B}$ may be viewed as an estimator of the bias of $\hat{\tau}$ in the presence of underfitting, when this bias may be expected to be positive. The properties of $\hat{B}$ in the case of overfitting are more difficult to assess. As will be discussed further below, we expect the first part of expression (11) involving $(f_k - \hat{\mu}_k)$ to contribute less than the second component involving $[(f_k - \hat{\mu}_k)^2 - f_k]$. In the second component, we expect that overfitting will lead to $(f_k - \hat{\mu}_k)^2$ tending to be less than $(f_k - \mu_k)^2$ and thus, since $E[(f_k - \mu_k)^2] = E(f_k)$, we may expect the second component to tend to be negative

and hence for $\hat{B}$ to be negative. We thus conclude that $\hat{B}$ will tend to be negative in the presence of overfitting, although we do not suggest that it will estimate the bias of $\hat{\tau}$ in this case. We refer to $\hat{B}$ as a *minimum error criterion*, since it is constructed with the aim of minimising the error of $\hat{\tau}$ as an estimator of τ^* or predictor of τ.

4.3 Test Statistics

We propose to use the closeness of $\hat{B}$ to zero as evidence of an absence of underfitting. We emphasise that this criterion is designed to assess the quality of the estimates arising from the model, not whether the model is correct, i.e. the purpose is estimation not testing. Nevertheless, we need to quantify 'closeness' to zero since $\hat{B}$ will differ from zero because of sampling error, even in the absence of underfitting, and thus we consider estimating the variance of $\hat{B}$. We assume that it is reasonable to approximate the distribution of $\hat{B}$ by the distribution of $\tilde{B}$. This approximation may be justified by standard asymptotic theory for contingency tables where the cells (and K) are fixed and the population and sample sizes per cell increase. Alternatively, it may be justified in an asymptotic framework (Haberman, 1977) in which K increases alongside the population and sample sizes and where the contribution of the sampling error in $\hat{\beta}$ via the $\hat{\lambda}_k$ to the variance of $\hat{B}$ becomes negligible relative to the contribution of the terms involving f_k in (11). This framework seems more realistic for our applications, where K is large and the individual cell sizes may be small, but the two-way and three-way marginal counts upon which $\hat{\beta}$ is based tend to increase with sample size.

If the model is correctly specified, so that $\tilde{\lambda}_k = \lambda_k$ and $f_k \sim P(\mu_k)$, then $\tilde{B}$ has zero expectation and, using standard results for the first four moments of a Poisson random variable, $\mathrm{var}(\tilde{B}) = \sum_k a_k^2 \mu_k + 2b_k^2 \mu_k^2$, where $a_k = -\lambda_k \exp(-\pi\lambda_k) h'(\lambda_k)$ and $b_k = \lambda_k \exp(-\pi\lambda_k) h''(\lambda_k)/(2\pi)$. For $h(\lambda) = h_1(\lambda)$, we have $a_k = (1-\pi)\lambda_k \exp(-\lambda_k)$ and $b_k =$

$(1-\pi)^2\lambda_k \exp(-\lambda_k)/(2\pi)$ and for $h(\lambda) = h_2(\lambda)$, we have $a_k = \exp(-\pi\lambda_k)r_{2k} - \exp(-\lambda_k)$ and $b_k = \{\exp(-\pi\lambda_k)r_{2k} - \exp(-\lambda_k)[1 + (1-\pi)\lambda_k/2]\}/[\pi\lambda_k]$, where r_{2k} is given above (3).

A natural estimator of $\text{var}(\tilde{B})$ is given by

$$\nu = \sum_k \hat{a}_k^2 \hat{\mu}_k + 2\hat{b}_k^2 \hat{\mu}_k^2, \tag{13}$$

where $\hat{\mu}_k = \pi\hat{\lambda}_k$, and

$$\hat{a}_k = -\hat{\lambda}_k \exp(-\hat{\mu}_k) h'(\hat{\lambda}_k), \tag{14}$$

and

$$\hat{b}_k = \hat{\lambda}_k \exp(-\hat{\mu}_k) h''(\hat{\lambda}_k)/(2\pi). \tag{15}$$

An alternative variance estimator is obtained by assuming just that $\tilde{\lambda}_k = \lambda_k$ and the f_k are independent with mean and variance equal to μ_k but without assuming that the third and fourth moments follow those of a Poisson distribution. In this case, we obtain $\text{var}(\tilde{B}) = \sum_k E\{a_k(f_k - \mu_k) + b_k[(f_k - \mu_k)^2 - f_k]\}^2$ and an alternative estimator of $\text{var}(\tilde{B})$ is given by

$$\nu_R = \sum_k \{\hat{a}_k(f_k - \hat{\mu}_k) + \hat{b}_k[(f_k - \hat{\mu}_k)^2 - f_k]\}^2, \tag{16}$$

where the subscript R denotes robust.

Given our assumptions above, $\hat{B}/\sqrt{\nu}$ or $\hat{B}/\sqrt{\nu_R}$ have an approximate standard normal distribution under the hypothesis that the expected value of $\hat{B}$ is zero. We shall refer to the associated tests as *minimum error tests*. They are diagnostic tests, designed to assess whether a model displays evidence of underfitting or overfitting for estimation purposes and not to test whether a given model is correct.

4.4 Relation to Existing Tests of Overdispersion

The expression for $\hat{B}$ in (11) or (12) may be considered as the sums of two components $\hat{B} = \hat{B}_a + \hat{B}_b$. The first component, $\hat{B}_a = \sum_k \hat{a}_k(f_k - \hat{\mu}_k)$, is of the same form as the estimating function appearing in (5) so that if β is estimated using ML and the vector of weights $\hat{a}_k$ is in the linear space spanned by x_k then this component will be zero. In general, this argument suggests that the first component may be less important than the second component, $\hat{B}_b = \sum_k b_k[(f_k - \hat{\mu}_k)^2 - f_k]$. We shall consider this empirically in Section 6. The component $\hat{B}_b$ may be interpreted as an estimator of the degree of overdispersion or underdispersion, since f_k and $(f_k - \hat{\mu}_k)^2$ are unbiased estimators of the conditional mean and variance of f_k respectively, again ignoring differences between $\hat{\beta}$ and β and assuming $\mu_k = \exp(x_k'\beta)$. Hence, an average of $[(f_k - \hat{\mu}_k)^2 - f_k]$ is a measure of overdispersion or underdispersion. This reveals a close connection between the proposed test procedure above and existing tests of overdispersion. In particular, Cameron and Trivedi (1998, p.78), construct $z_k = [(f_k - \hat{\mu}_k)^2 - f_k]/\hat{\mu}_k$ and test whether it has zero expectation by referring the test statistic $\hat{\kappa}/\sqrt{\nu_\kappa}$ in the usual way to a standard normal distribution, where $\hat{\kappa} = K^{-1}\sum_{k=1}^{K} z_k$, and $\nu_k = \sum_{k=1}^{K}(z_k - \hat{\kappa})^2/[K(K-1)]$. This is a score test of $H_0 : \kappa = 0$ for a model with a conditional variance of the form $(1 + \kappa)\mu_k$. It can also test for underdispersion.

5 USE OF MODEL ASSESSMENT CRITERIA

We propose to use the criteria developed in the previous section to select a specification of the log-linear model in (4) via a search algorithm. The criteria might also be used as a diagnostic approach to assess whether a given specified model may be expected to provide adequate risk measures.

Since the criterion $\hat{B}$ in (11) and the associated minimum error tests were derived primarily as a means to detect underfitting (and numerical work we have undertaken suggests that indeed they are more effective for this purpose than for detecting overfitting) we suggest a forward search algorithm, starting from simpler models and adding terms until the specification is judged to be adequate.

In many empirical experiments that we have undertaken, we have found that the independence log-linear model tends to underfit and lead to overestimation of the disclosure risk measures. At the other extreme, the all 3-way interactions model tends to overfit and lead to under-estimation of the risk measures. Thus we expect a reasonable solution to lie between these extremes and indeed the all 2-way interactions log-linear model often leads to good estimates of the risk measures for the types of datasets and size of keys that are used in practice. As a practical approach, we suggest first computing the criteria of Section 4 for the independence model and the all 2-way interactions model. If the latter model shows no sign of underfitting then we propose starting with the independence model and adding the 2-way interaction terms for different pairs of key variables, chosen sequentially in order to reduce $\hat{B}$, until a model is identified which is judged to show no evidence of underfitting. On the other hand, if the all 2-way interactions model is found to exhibit underfitting, then we propose to start a similar forward model search algorithm from this model as the initial model, adding 3-way interaction terms for different triples of key variables. As in any model search algorithm for a hierarchical log-linear model, the inclusion of a higher order term containing an interaction implies that all subsidiary lower order effects should also be included.

Given the alternative choices of test procedures, as well as the alternative measures of overdispersion mentioned in section 4.4, there are alternative possible stopping rules for the search algorithm. We shall discuss these in the context of the real applications

in the next sections. There will, of course, be no single 'correct' model and there are likely to be a number of models between which the criteria will not discriminate. We suggest that in the disclosure risk assessment context, it is sensible to produce risk estimates for each of a number of such 'reasonable models' and to use the differences between the estimates as a diagnostic to check the sensitivity of the measures to the specification of the model.

6 APPLICATION TO CENSUS SAMPLES

We now apply the proposed methods to samples drawn from the 2001 UK population census. Treating one region of N=944,793 individuals as the population, we compute the true aggregated risk measures and compare them to the estimated risk measures for simple random samples from this population and thus examine the performance of the model choice criteria.

We consider two keys defined by six traceable and visible key variables. The first key is defined by (number of categories in parenthesis): area (2), sex (2), age (101), marital status (6), ethnicity (17) and economic activity (10), giving K=412,080 cells. The second key has 73,440 cells and is defined as the first key except that age is grouped into 18 bands. Our choice of key variables follows considerations at ONS and in Dale and Elliot (2001). To fit the log-linear models, we used iterative proportional fitting (Bishop et al., 1975) which is simple to program and directly generates the fitted values $\hat{\mu}_k$ required for the risk estimates. Log-linear model fitting procedures in standard statistical software will often not cope with the large numbers of variables and cells that we have. We experienced no problems of convergence despite the presence of many cells with $f_k = 0$. Our estimation method dealt 'automatically' with zero marginal counts corresponding to a given model, for example because of impossible

combinations of key variable values (structural zeros), by setting the fitted values for cells falling in these margins to zero.

Table 1 presents true and estimated values of τ_1 and and τ_2 for three samples with 0.5%, 1% and 2% sampling fractions and for three log-linear models: the independence model, the all 2-way interactions model and the all 3-way interactions model. We see a consistent pattern of estimates decreasing with increasing model complexity, with the independence model always leading to overestimation and the all 3-way interactions model always leading to underestimation. The all 2-way interactions model performs rather better, mostly generating underestimates but twice generating overestimates. The errors of estimation of $\hat{\tau_1}$ and $\hat{\tau_2}$ always share the same sign and suggest that a fitting criterion which 'works' for one measure should also work for the other measure. The five test statistics also tend to have the same signs. The serious overestimation (and underfitting) of the independence model is consistently predicted by the large positive values of all five test statistics. The signs of the five test statistics are also always the same for the all 2-way interactions model and all consistently predict whether $\hat{\tau_1}$ and $\hat{\tau_2}$ will overestimate or underestimate τ_1 and τ_2 respectively. The underestimation (and overfitting) of the all 3-way interactions model is consistently predicted by the negative signs of the test statistics $\hat{\kappa}/\sqrt{\nu_\kappa}$, $\hat{B}_2/\sqrt{\nu}$ and $\hat{B}_2/\sqrt{\nu_R}$. There are inconsistencies, however, in the behaviour of $\hat{B}_1/\sqrt{\nu}$ and $\hat{B}_1/\sqrt{\nu_R}$, especially for the smaller sample sizes, and this suggests that these tests should be used primarily to detect underfitting.

Although the test statistics have similar signs, their magnitudes vary. The two test statistics, using a variance estimator based upon the Poisson assumption, seem most sensitive (i.e. have the largest values) to underfitting, but least sensitive to overfitting. In contrast, the test statistics based upon the variance estimator ν_R (or the Cameron-Trivedi test) are more sensitive to overfitting and less sensitive to underfitting.

Table 2 presents some values of the underlying statistics $\hat{B}_1$ and $\hat{B}_2$ for the large key.

For the all 2-way interactions model, there is some similarity between these values and those of the estimation errors $\hat{\tau}_1 - \tau_1$ and $\hat{\tau}_2 - \tau_2$, respectively, as might be expected as the former are intended to estimate the expectation of the latter. For example, for the 1% sample and the large key, we have $\hat{B}_1$ = -59.3, $\hat{\tau}_1 - \tau_1$ = -54.1 and $\hat{B}_2$=-72.9, $\hat{\tau}_2 - \tau_2$ = -75.8. Nevertheless, the statistics $\hat{B}_1$ and $\hat{B}_2$ were derived using approximations around the true model and when the assumed model provides a poor fit, as for the independence and all three-way interactions models, we observe that $\hat{B}_1$ and $\hat{B}_2$ bear little relation to the estimation errors. Moreover, there will be no reliable interpretation of the values of $\hat{B}_1$ or $\hat{B}_2$ when they are of a similar magnitude to their standard errors, the case that will be of most interest in our approach to model selection. Henceforth, we shall therefore only consider the values of the test statistics associated with $\hat{B}_1$ and $\hat{B}_2$, not the unstandardized values. Table 2 also includes breakdowns of the $\hat{B}_1$ and $\hat{B}_2$ statistics according to the $\hat{B} = \hat{B}_a + \hat{B}_b$ decomposition in section 4.4. As discussed there, we observe that the second component $\hat{B}_b$ dominates for the independence and all 2-way interactions models, i.e. except for the case of serious overfitting. Thus, as discussed in section 4.4., the tests based on $\hat{B}$ are similar to tests of overdispersion when the model underfits.

We now undertake a forward model search, as discussed in Section 5, for the data defined by the large key and the 1% sample (n=9,448). Table 1 suggests that the independence model underfits and the all 2-way interactions model overfits. We therefore start from the independence model and consider adding 2-way interaction terms until we find a model for which there is no evidence of lack of fit. Table 3 presents results of the best fitted models obtained for each round of a forward search, starting with the independence model, labelled as Model I. Note that the 1-way (main effects) terms become obsolete when adding in 2-way interaction terms that contain them. The first four rounds are clear-cut in the sense that, at each round, there is a clear choice of

the set of 2-way interactions which best reduces all of the test criteria. The set of interaction terms between age and economic activity, denoted {a*ec}, is included in round 1 (leading to the model denoted 1). Three further rounds leads to the addition of the sets {a*et},{a*m} and {s*ec} to give Model 4. This model provides a good fit in the sense that the values of all the test statistics based upon $\hat{B}_1$ or $\hat{B}_2$ are less than 2 (although the Cameron-Trivedi test still suggests some underfitting). It is less clear how to proceed beyond Model 4. A simple approach in practice might be a forward search using only one criterion (we suggest $\hat{B}_2/\sqrt{\nu}$ in section 8) stopping at the round prior to which the criterion becomes negative for every added term. Here, we adopt a more informal approach, selecting more than one model at a round if they are nearly indistinguishable with respect to the multiple criteria and permitting very slight negative values of one or two criteria. Thus, at round 5, we select two models, 5a and 5b, which each provide improvements over model 4 but neither appears to be uniformly better than the other in terms of all the criteria. We fail to find any terms to add to Model 5a without one of the criteria becoming strongly negative and thus treat Model 5a as one candidate 'terminal' model. There are, however, three models, 6b, 6c, and 6d, which may be obtained from Model 5b and which appear reasonable. Model 6b is again a candidate terminal model since we cannot add any terms without one of the criteria becoming strongly negative. Finally we obtain an additional two candidate terminal models, 7c and 7d from Models 6c and 6d. We thus have four potential 'terminal' models, 5a, 6b, 7c and 7d. In fact each of these models gives very similar estimates $\hat{\tau}_1$ and $\hat{\tau}_2$ of around 148 and 336 respectively, implying a robustness of the search procedure to the choice of criterion. Moreover, similar estimates are obtained from models 4, 5b, 6c and 6d, implying a robustness to the precise form of the stopping rule.

The model search is represented graphically in Figure 2. The points $(\hat{\tau}_2, \hat{B}_2/\sqrt{\nu})$

in the scatterplot correspond to all the models in Table 3 as well as all the models which were considered in the forward search but not selected. The points are scattered around a line with a postive slope which, as desired, is around zero when $\hat{\tau}_2$ is equal to the true value of τ_2, although the search jumps across the true value $\hat{\tau}_2 = \tau_2$ when the term {a*m} is included (the change from Model 2 to 3). The plot tends to display some curvature (convexity) implying that the interval of values of τ_2 for well-fitting models is shorter above its true value than below, i.e. underfitting is easier to detect than overfitting.

We next examine the record-level risk measure $\hat{r}_{2k}$ for the different models. Figure 3 presents a scatterplot of $1/F_k$ against $\hat{r}_{2k}$ for 2,304 sample uniques under Model 5a in Table 3 of the 1% Census sample with the large key. Table 4 provides a corresponding cross-classification of these values within bands. We observe a strong positive relationship with a Spearman rank correlation of 0.80, i.e. the model is effective in using the key variable information to predict $1/F_k$. Nevertheless, it is good news from the point of view of disclosure protection that the prediction is far from perfect with, for example, many population unique cells not being picked up by high $\hat{r}_{2k}$ values. The values of $1/F_k$ range above and below the diagonal line in Figure 3, as anticipated if $\hat{r}_{2k}$ is to be interpreted as an expected value of $1/F_k$. There is no strong evidence of the $\hat{r}_{2k}$ being smoothed to have smaller dispersion than the $1/F_k$ with similar marginal distributions observed in Table 4.

7 APPLICATION TO SOCIAL SURVEY DATA

We now descibe an application to a social survey with a sample size of $n = 127,200$ individuals drawn with equal probability sampling from the adult population of the UK. Although the true values of $\tau_1$and τ_2 are no longer available for validation, we

can still compare the behaviour of the alternative criteria and the stability of risk estimates. The microdata first underwent disclosure control based on initial recoding or suppression of key variables. The visible and traceable key variables that were used for the evaluations were: area (20), sex (2), age in years (top-coded at 90) (91), marital status (5), ethnicity (13) and economic activity (10) resulting in a key of $K = 2,336,000$ cells. There were 13,954 sample uniques. Some results are presented in Table 5. There is clear underfitting of the independence model and clear overfitting of the all 3-way interactions model. The all 2-way interactions model, however, appears to provide a reasonable fit. It is interesting that this model 'fits' despite the sample size being much larger than in the census samples. The all 2-way interactions model cannot be exactly true. Experience with the increasing power of conventional goodness-of-fit tests with sample size might lead us to expect that this model would be rejected for a sample as large as this. This is not what we see. Table 1 provides further evidence that increasing the sample size does not necessarily result in the selection of a more complex model. We see no tendency in this table for the test statistics for the all 2-way interactions model to deviate more significantly from zero the larger the sample size. Such evidence lends further support for the practical feasibility of using our criteria across a range of survey settings.

Returning to Table 5, since the values of some of the test statistics for the all 2-way interactions model approach 2, we consider adding in 3-way interactions. Among the twenty possible combinations of 3 from 6 key variables, we present results for the eight models (1a-1h) which reduced the values of all the minimum error test criteria (without making any negative). Selecting the two of these models (1c and 1d) with the smallest values of $\hat{B}_2/\sqrt{\nu}$ we also present results for nine further models which lead to a reduction of all the minimum error test criteria by adding in 3-way interaction terms.

We observe that the value of the Cameron-Trivedi test now differs clearly from the

minimum error tests. We have found such discrepancies with other survey examples, both in positive and negative directions. Table 3 provides examples of relatively minor discrepancies in the opposite direction for Models 4 and 5a for the census data, where the Cameron-Trivedi test indicates significant underfitting, unlike the other test criteria. Exploration of these discrepancies indicates a number of sources, mainly related to the fact that the Cameron-Trivedi statistic is not designed with a focus on sample uniques. In particular, cells with higher expected frequencies $\hat{\mu}_k$ may make a more important contribution to the Cameron-Trivedi statistic than the minimum error criteria, because the contributions of these cells are downweighted less severely by $1/\hat{\mu}_k$ than by $exp(-\hat{\mu}_k)$. Moreover, we have found a number of survey examples where the $\hat{B}_b$ term no longer dominated $\hat{B} = \hat{B}_a + \hat{B}_b$ (see section 4.4.). Our broad conclusion is that it is inappropriate to use the Cameron-Trivedi statistic as a general diagnostic criterion for the risk measures considered here, since it is not designed for this purpose.

The values of $\hat{\tau}_1$ and $\hat{\tau}_2$ are spread across the intervals (157.6, 266.9) and (681.5, 845.3) respectively for the well-fitting models in Table 5, exhibiting rather greater variation than in Table 3. We observe that the impact of adding in extra terms is either to reduce the risk measures (e.g. adding terms to Model II) or to have little effect (e.g. adding terms to Model 1d). The values 264.9 and 844.5 of $\hat{\tau}_1$and $\hat{\tau}_2$ for the all 2-way interactions model act as reasonable upper bounds. A clear lower bound is less easy to obtain and this appears to reflect the greater difficulty in detecting overfitting than underfitting. Fortunately, for risk assessment purposes, an upper bound is usually considered to be of greater importance. The variation of values of $\hat{\tau}_1$ and $\hat{\tau}_2$ provides some guidance to the sensitivity of the risk estimate provided by this upper bound.

8 DISCUSSION

We have examined the use of Poisson log-linear models to estimate disclosure risk measures for microdata, with applications to census and survey samples. As in Skinner and Holmes (1998) and Elamir and Skinner (2006), we have found that an all 2-way interactions model often leads to reasonable estimates. We have sought to improve on the use of this model as a default, by developing diagnostic criteria for model choice, suitable for risk assessment with the kinds of large and sparse contingency tables spanned by key variables that are typical in practical applications in official statistics. We have shown that our criteria do help to select models that show appreciable improvements in risk estimation relative to the all 2-way interactions model, especially by enabling us to detect overestimation arising from underfitting models. Since our criteria are more effective at detecting underfitting than overfitting, we have proposed a forward selection approach to model selection. There will invariably be several models which are effectively indistinguishable in terms of our criteria. We have found empirically that the risk estimates tend to be rather stable across the simplest models which show no evidence of underfitting. We have found that there may be additional more complex models, obtained by adding terms to the simplest models without leading to significant overfitting (or underfitting), and they may display somewhat more variable risk estimates, but these estimates always tend to be lower than those for the simpler models. Thus the risk estimates for the simplest well-fitting models tend to provide a good upper bound and a conservative approach to risk assessment.

We considered four different criteria, depending on the choice of risk measure ($\hat{B}_1$ vs. $\hat{B}_2$) and the choice of variance estimator (ν vs. ν_R). We found that models which 'work' for one risk measure (τ_1 or τ_2) tend to work also for the other risk measure. However, our results suggest a slight preference for $\hat{B}_2$ compared to $\hat{B}_1$ since the for-

mer did not generate misleading results for the all 3-way interactions model in Table 1. There may also be a slight preference for ν rather than ν_R if a forward selection approach is to be used since it appears to lead to a test statistic $\hat{B}_2/\sqrt{\nu}$ with more power for rejecting underfitting models.

We have suggested that differences between risk estimates for alternative well-fitting models may be used to represent uncertainty in a form of sensitivity analysis. Further research would be needed to assess the impact of sampling error in the parameter estimates and the construction of confidence intervals, although we suspect such sampling error effects are somewhat less important than the impact of model choice. One critical assumption in this paper is that there are no discrepancies in the values of the key variables between the microdata and the intruder's other data source; we plan to extend our approach to handle such discrepancies. Another assumption is that a Bernoulli sampling scheme is employed. There are at least two departures from this assumption that merit attention. First, even if equal probability sampling is employed, it is possible that a complex sampling scheme could invalidate the conclusion in section 3 that the f_k are Poisson distributed, e.g. if cluster sampling took place with cells k. Although the sample individuals in the survey in section 7 were clustered within households and although some of the key variables, e.g. ethnicity, may be expected to display strong household-level clustering, we anticipate that our risk assessment approach will be fairly robust to such complex sampling, since we anticipate generally negligible dependence between the selection of different individuals within each cell k. This expectation would, nevertheless, benefit from further research. A second possible departure from the Bernouilli sampling assumption would be unequal probability sampling. This would change the actual risk measure and not just the estimation problem. Skinner and Carter (2003) provide some ideas for this case, but more research is needed. Most of these areas for further research involve greater complexity. There is

also a need to consider more simplicity, in particular since our approach can generate significant computational demands when there are many cells. In particular, it would be useful to research ways of splitting the risk assessment by subpopulations (defined by key variables) in order to simplify computation.

References

Bethlehem, J., Keller, W., and Pannekoek, J. (1990), "Disclosure Control of Microdata," *Journal of the American Statistical Association*, 85, 38-45.

Bishop, Y., Fienberg, S., and Holland, P. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA.: MIT Press.

Cameron, A. C. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press.

Dale, A. and Elliot, M. (2001), "Proposals for 2001 samples of anonymized records: an assessment of disclosure risk," *Journal of the Royal Statistical Society, Series A*, 164, 427-447.

Doyle, P., Lane, J. I., Theeuws, J. J. M. and Zayatz, L. V. (eds.) (2001), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, Amsterdam, The Netherlands: North Holland.

Duncan, G. and Lambert, D. (1989), "The Risk of Disclosure for Microdata," *Journal of Business and Economic Statistics*, 7, 207-217.

Elamir, E. A. H. and Skinner, C. (2006), "Record Level Measures of Disclosure Risk for Survey Microdata," *Journal of Official Statistics*, to appear

Federal Committee on Statistical Methodology (1994), "Statistical Policy Working Pa-

per 2: Report on Statistical Disclosure Limitation Methodology", Subcommittee on Disclosure Limitation Methodology, Office for Management and Budget, Washington, D.C.

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

Fienberg, S. E., Makov, U. E. and Sanil A. E. (1997), "A Bayesian Approach to Data Disclosure: Optimal Intruder Behaviour for Continuous Data," *Journal of Official Statistics*, 13, 75-89.

Fuller, W. (1993), "Masking Procedures for Microdata Disclosure Limitation," *Journal of Official Statistics*, 9, 383-406.

Haberman, S.J. (1997), "Log-linear models and frequency tables with small expected cell counts," *The Annals of Statistics*, 5, 1148-1169.

Koehler, K. J. (1986), "Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables," *Journal of the American Statistical Association*, 81, 483-493.

Lambert, D. (1993), "Measures of Disclosure Risk and Harm," *Journal of Official Statistics*, 9, 313-331.

National Statistics (2004), *Code of Practice: Protocol on Data Access and Confidentiality*, Norwich, United Kingdom: Her Majesty's Stationary Office.

Paass, G. (1988), "Disclosure Risk and Disclosure Avoidance for Microdata," *Journal of Business and Economic Statistics*, 6, 487-500.

Rao, C. R. and Wu, Y. (2001), "On Model Selection," in *Model Selection*, ed. P. Lahiri, Beechwood, Ohio: Institute of Mathematical Statistics Lecture Notes- Monograph Series 38, 1-57.

Reiter, J. (2005), "Estimating Risks of Identification Disclosure in Microdata," *Journal of the American Statistical Association*, 100, 1103-1112.

Skinner, C.J. and Carter, R.G. (2003) "Estimation of a Measure of Disclosure Risk for Survery Microdata under Unequal Probability Sampling," *Survey Methodology*, 29, 177-180.

Skinner, C. J., and Elliot, M. J. (2002), "A Measure of Disclosure Risk for Microdata", *Journal of the Royal Statistical Society*, Ser. B, 64, 855-867.

Skinner, C. J. and Holmes, D. (1998), "Estimating the Re-identification Risk Per Record in Microdata", *Journal of Official Statistics*, 14, 361-372.

Skinner, C. J. , Marsh, C., Openshaw, S., and Wymer, C. (1994), "Disclosure Control for Census Microdata," *Journal of Official Statistics*, 10, 31-51.

Willenborg, L., and De Waal, T. (2000), *Elements of Statistical Disclosure Control*, Vol. 155, Lecture Notes in Statistics, New York: Springer.

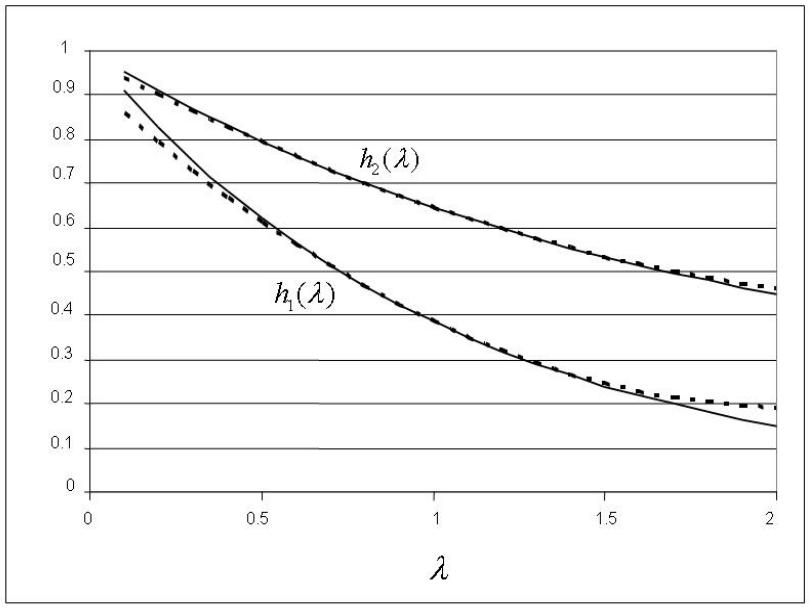

Figure 1: *Quadratic approximations of $h(\lambda)$ functions for $\pi = 0.05$. Solid lower line is $h_1(\lambda)$. Solid upper line is $h_2(\lambda)$. Dotted lines are approximations from equation (8).*

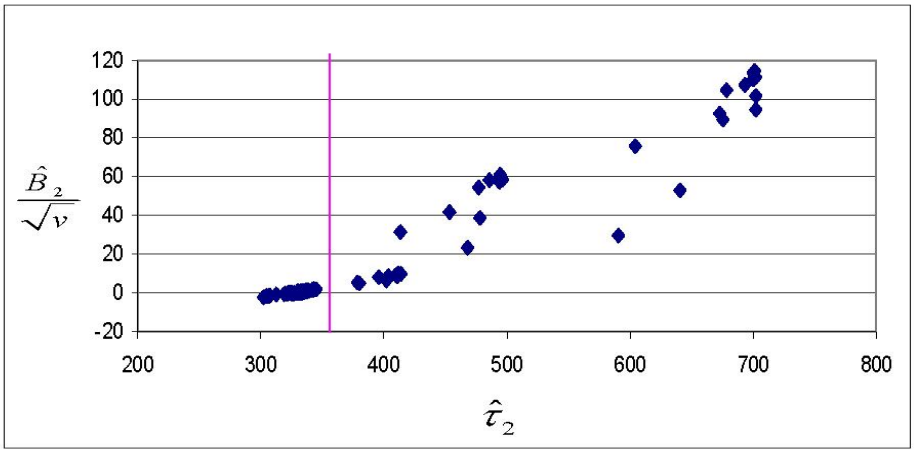

Figure 2: *Scatterplot of* $\frac{\hat{B}_2}{\sqrt{v}}$ *against* $\hat{\tau}_2$ *for all models considered in forward search, summarised in Table 3.*

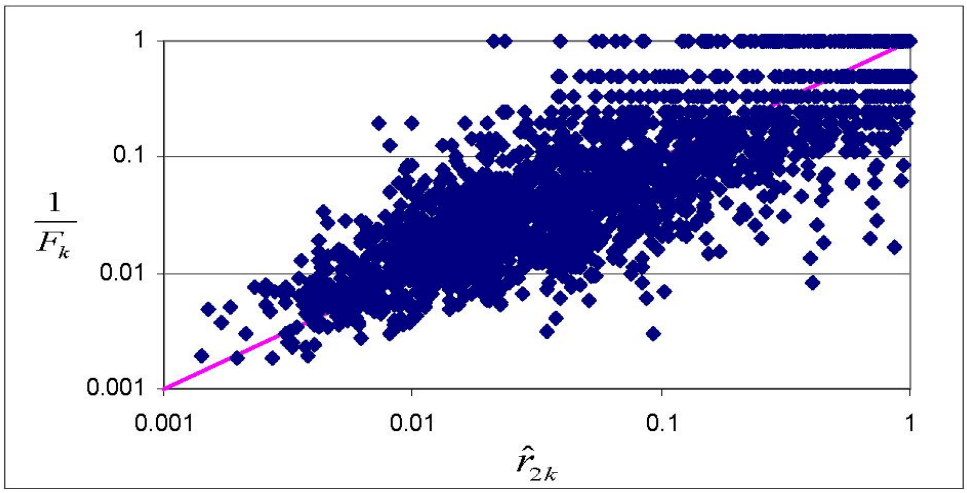

Figure 3: *Scatterplot (on logarithmic scales) of* $1/F_k$ *against* $\hat{r}_{2k}$ *for 2,304 sample uniques for model 5a in Table 3 with 1% census sample and large key.*

Table 1: *Aggregated Risk Measures and Test Statistics for Samples Drawn from the 2001 UK Census.*

n	Model	τ_1	τ_2	$\hat{\tau}_1$	$\hat{\tau}_2$	Test Statistics				
						$\hat{\kappa}/\sqrt{\nu_\kappa}$	$\hat{B}_1/\sqrt{\nu}$	$\hat{B}_1/\sqrt{\nu_R}$	$\hat{B}_2/\sqrt{\nu}$	$\hat{B}_2/\sqrt{\nu_R}$
Small Key K = 73440										
4724	I	23	68.2	54.2	126.9	8.6	12.5	3.3	30.4	7.2
	II			16.0	52.2	-3.6	-0.5	-6.4	-0.8	-2.9
	III			0.0	7.1	-26.4	0.0	2.2	-1.0	-13.1
9448	I	39	127.1	99.3	230.2	8.6	32.1	4.2	60.6	6.8
	II			37.8	117.9	-3.9	-1.3	-9.0	-1.6	-4.2
	III			0.5	24.7	-28.8	-0.2	-2.8	-2.3	-14.3
18896	I	75	215.3	174.3	355.7	9.6	70.7	6.1	125.5	9.1
	II			85.5	222.0	2.0	0.7	0.5	0.7	0.6
	III			11.0	82.1	-28.6	-1.2	-7.4	-4.1	-20.8
Large Key K = 412080										
4724	I	80	183.9	197.4	385.1	10.6	16.8	4.8	53.1	7.4
	II			35.9	112.3	-8.0	-0.5	-1.6	-1.0	-1.4
	III			0.0	11.0	-40.7	0.0	1.1	-1.3	-19.3
9448	I	159	355.9	386.6	701.2	14.4	48.5	8.0	114.2	8.8
	II			104.9	280.1	-10.3	-1.6	-11.1	-2.7	-4.9
	III			1.1	42.2	-45.1	-0.3	-3.0	-3.1	-22.1
18896	I	263	628.9	672.0	1170.5	16.8	105.2	10.3	226.1	10.4
	II			252.0	591.3	-5.7	-1.1	-1.5	-1.5	-1.8
	III			11.3	150.2	-51.9	-1.3	-8.5	-7.0	-37.0

Model I = independence model, Model II = all 2-way interactions model, Model III = all 3-way interactions model.

Table 2: *Aggregated Risks Measures and Components of Model Choice Criteria for Samples Drawn from the 2001 UK Census with a Large Key.*

n	Model	τ_1	τ_2	$\hat{\tau}_1$	$\hat{\tau}_2$	Components of Test Criteria					
						$\hat{B}_1$	$\hat{B}_{1a}$	$\hat{B}_{1b}$	$\hat{B}_2$	$\hat{B}_{2a}$	$\hat{B}_{2b}$
4724	I	80	183.9	197.4	385.1	117.9	-11.8	1190.7	2555.4	11.2	2544.2
	II			35.9	112.3	-16.8	4.2	-21.0	-23.7	1.7	-25.4
	III			0.0	11.0	0.1	-0.6	0.7	-6.1	-3.0	-3.1
9448	I	159	355.9	386.6	701.2	3400.8	-12.1	3412.8	5463.2	25.2	5437.9
	II			104.9	280.1	-59.3	6.6	-65.9	-72.9	2.4	-75.2
	III			1.1	42.2	-2.1	-1.6	-0.6	-24.1	-5.9	-18.3
18896	I	263	628.9	672.0	1170.5	7269.9	-32.1	7302.0	10618.0	55.7	10562.0
	II			252.0	591.3	-43.6	3.9	-47.5	-43.0	2.5	-45.5
	III			11.3	150.2	-17.0	-5.1	-11.9	-84.7	-9.3	-75.4

Table 3: *Models Selected by a Forward Search for 1% Census Sample with Large Key*

Model	$\hat{\tau}_1$	$\hat{\tau}_2$	Test Statistics				
			$\hat{\kappa}/\sqrt{\nu_\kappa}$	$\hat{B}_1/\sqrt{\nu}$	$\hat{B}_1/\sqrt{\nu_R}$	$\hat{B}_2/\sqrt{\nu}$	$\hat{B}_2/\sqrt{\nu_R}$
I	386.6	701.2	14.4	48.5	8.0	114.2	8.8
II	104.9	280.1	-10.3	-1.6	-11.1	-2.7	-4.9
1: I + {a*ec}	243.4	494.3	6.5	54.8	3.3	59.2	3.5
2: 1 + {a*et}	180.1	411.6	13.3	3.1	1.4	9.8	4.5
3: 2 + {a*m}	152.3	343.3	5.2	0.9	0.6	1.7	1.1
4: 3 + {s*ec}	149.2	337.5	2.7	0.3	0.2	0.9	0.6
5a: 4 +{ar*a}	148.5	337.1	2.3	0.0	0.0	0.8	0.6
5b: 4 +{s*m}	147.7	335.3	2.2	0.0	0.0	0.7	0.4
6b: 5b + {ar*a}	147.0	335.0	1.8	-0.2	-0.2	0.6	0.4
6c: 5b + {ar*m}	148.9	337.1	2.1	0.0	0.0	0.7	0.5
6d: 5b + {m*ec}	146.3	331.4	1.1	-0.2	-0.2	0.0	0.0
7c: 6c + {m*ec}	147.5	333.2	1.0	-0.3	-0.3	0.1	0.0
7d: 6d + {ar*a}	145.6	331.0	0.7	-0.4	-0.4	0.0	0.0

Area-ar, Sex-s, Age-a, Marital Status-m, Ethnicity-et, and Economic Activity-ec; true values are $\tau_1 = 159$, $\tau_2 = 355.9$

Table 4: *Cross-classification of* $1/F_k$ *against* $\hat{r}_{2k}$ *for Sample Uniques within Bands for Model 5a of 1% Census Sample with Large Key.*

$1/F_k$	$\hat{r}_{2k}$			
	0 - 0.1	0.1 - 0.5	0.5 - 1	Total
0 - 0.1	1391	150	11	1552
0.1 - 0.5	162	253	76	491
0.5 - 1	26	91	144	261
Total	1579	494	231	2304

Table 5: *Models Selected by a Forward Search for a Social Survey.*

Model	$\hat{\tau}_1$	$\hat{\tau}_2$	Test Statistics				
			$\hat{\kappa}/\sqrt{\nu_\kappa}$	$\hat{B}_1/\sqrt{\nu}$	$\hat{B}_1/\sqrt{\nu_R}$	$\hat{B}_2/\sqrt{\nu}$	$\hat{B}_2/\sqrt{\nu_R}$
I	879.5	2301.6	15.51	561.4	9.77	1206.7	9.19
II	264.9	844.5	0.68	1.80	0.99	1.93	1.41
III	10.5	211.4	-82.74	-0.48	-9.12	-3.54	-43.15
1a: II+{ar*s*et}	263.5	840.9	-0.02	0.96	0.66	1.59	1.23
1b: II+{ar*s*ec}	263.4	843.0	0.51	1.35	0.98	1.83	1.35
1c: II+{ar*a*m}	232.1	787.6	-3.01	1.61	0.88	0.94	0.70
1d: II+{ar*a*ec}	217.9	748.3	-3.65	1.46	0.76	0.36	0.30
1e: II+{ar*et*ec}	191.2	739.2	-0.94	0.98	0.69	1.27	0.99
1f: II+{s*m*et}	266.9	845.3	0.58	1.73	0.95	1.83	1.35
1g: II+{a*m*et}	188.5	727.8	-0.96	1.50	0.88	1.24	0.90
1h: II+{m*et*ec}	244.3	813.0	0.16	1.59	0.89	1.35	1.03
2c1: 1c+{ar*s*et}	230.5	784.1	-5.38	0.53	0.43	0.49	0.41
2c2: 1c+{ar*s*ec}	231.2	786.4	-3.22	1.52	0.83	0.84	0.63
2c3: 1c+{ar*et*ec}	157.6	681.5	-6.99	0.32	0.28	0.04	0.03
2c4: 1c+{s*a*m}	232.5	785.0	-3.54	1.61	0.88	0.88	0.65
2c5: 1c+{s*a*et}	226.7	772.7	-4.41	1.39	0.81	0.78	0.59
2c6: 1c+{s*m*et}	234.2	788.7	-3.21	1.55	0.85	0.90	0.67
2d1: 1d+{ar*s*et}	216.0	745.2	-6.77	0.91	0.56	0.16	0.13
2d2: 1d+{ar*s*ec}	217.8	747.8	-3.76	1.45	0.76	0.28	0.23
2d3: 1d+{s*m*ec}	216.6	743.8	-3.86	1.43	0.75	0.32	0.26

Area-ar, Sex-s, Age-a, Marital Status-m, Ethnicity-et, and Economic Activity-ec