Catalogue no. 12-001-X

Survey Methodology

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

June 2012

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website at www.statcan.gc.ca, e-mail us at infostats@statcan.gc.ca, or telephone us, Monday to Friday from 8:30 a.m. to 4:30 p.m., at the following numbers:

Statistics Canada's National Contact Centre

Toll-free telephone (Canada and United States):

Inquiries line	1-800-263-1136
National telecommunications device for the hearing impaired	1-800-363-7629
Fax line	1-877-287-4369

Local or international calls:

Inquiries line	1-613-951-8116
Fax line	1-613-951-0581

Depository Services Program

Inquiries line	1-800-635-7943
Fax line	1-800-565-7757

To access and order this product

This product, Catalogue no. 12-000-X, is available free in electronic format. To obtain a single issue, visit our website at www.statcan.gc.ca and browse by "Key resource" > "Publications."

This product is also available as a standard printed publication at a price of CAN$30.00 per issue and CAN$58.00 for a one-year subscription.

The following additional shipping charges apply for delivery outside Canada:

	Single issue	**Annual subscription**
United States	CAN$6.00	CAN$12.00
Other countries	CAN$10.00	CAN$20.00

All prices exclude sales taxes.

The printed version of this publication can be ordered as follows:

• Telephone (Canada and United States)	1-800-267-6677
• Fax (Canada and United States)	1-877-287-4369
• E-mail	infostats@statcan.gc.ca
• Mail	Statistics Canada Finance R.H. Coats Bldg., 6th Floor 150 Tunney's Pasture Driveway Ottawa, Ontario K1A 0T6

- In person from authorized agents and bookstores.

When notifying us of a change in your address, please provide both old and new addresses.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Statistics Canada

Business Survey Methods Division

Survey Methodology

June 2012

Published by authority of the Minister responsible for Statistics Canada

June 2012

Catalogue no. 12-001-XWE
ISSN 1492-0921

Catalogue no. 12-001-XPB
ISSN: 0714-0045

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (n^{o} 12-001-X au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is indexed in The ISI Web of knowledge (Web of science), The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods. It is also covered by SCOPUS in the Elsevier Bibliographic Databases.

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their articles in English or French in electronic form, preferably in Word to the Editor, (smj@statcan.gc.ca, Statistics Canada, 150 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6). For formatting instructions, please see the guidelines provided in the journal and on the web site (www.statcan.gc.ca).

Subscription Rates

The price of printed versions of *Survey Methodology* (Catalogue No. 12-001-XPB) is CDN $58 per year. The price excludes Canadian sales taxes. Additional shipping charges apply for delivery outside Canada: United States, CDN $12 ($6 × 2 issues); Other Countries, CDN $20 ($10 × 2 issues). A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, the Statistical Society of Canada and l'Association des statisticiennes et statisticiens du Québec. Electronic versions are available on Statistics Canada's web site: www.statcan.gc.ca.

Survey Methodology

A Journal Published by Statistics Canada

Volume 38, Number 1, June 2012

Contents

Regular Papers

Short Notes

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 1-10
Statistics Canada, Catalogue No. 12-001-X

A statistical approach to detect interviewer falsification of survey data

Sebastian Bredl, Peter Winker and Kerstin Kötschau [1]

Abstract

Survey data are potentially affected by interviewer falsifications with data fabrication being the most blatant form. Even a small number of fabricated interviews might seriously impair the results of further empirical analysis. Besides reinterviews, some statistical approaches have been proposed for identifying this type of fraudulent behaviour. With the help of a small dataset, this paper demonstrates how cluster analysis, which is not commonly employed in this context, might be used to identify interviewers who falsify their work assignments. Several indicators are combined to classify 'at risk' interviewers based solely on the data collected. This multivariate classification seems superior to the application of a single indicator such as Benford's law.

Key Words: Data fabrication; Falsifier; Benford's law; Cluster analysis.

1. Introduction

Whenever data collection is based on interviews, one has to be concerned about data quality. Data quality can be affected by false or imprecise answers of the respondent or by a poorly designed questionnaire, as well as by the interviewer when he or she deviates from the prescribed interviewing procedure. If the interviewer does so consciously, this is referred to as 'interviewer falsification' (Schreiner, Pennie and Newbrough 1988) or 'cheating' (Schräpler and Wagner 2003).

Interviewer falsification can occur in many ways (*cf.* Guterbock 2008). Rather subtle forms consist of surveying a wrong household member or of conducting the survey by telephone when face-to-face interviews are required. The most severe form of falsifying is the fabrication of entire interviews without ever contacting the respective household. In our analysis, we deal with the latter case.

Fabricated interviews can have serious consequences for statistics based on the survey data. Schnell (1991) and Schräpler and Wagner (2003) provide evidence that the effect on univariate statistics might be less severe, provided the share of falsifiers remains sufficiently small and the 'quality' of the fabricated data is high. But even a small proportion of fabricated interviews can be sufficient to cause heavy biases in multivariate statistics. Schräpler and Wagner (2003) find that the inclusion of fabricated data from the German Socio Economic Panel (GSOEP) in a multivariate regression reduces the effect of training on log gross wages by approximately 80 percent, although the share of fabricated interviews was less than 2.5 percent. This indicates the importance of identifying these interviews.

The most common way to identify falsifying interviewers is the reinterview (Biemer and Stokes 1989). In this case, a supervisor contacts some of the households that should have been surveyed to check whether they were actually visited by the interviewer. However, for reasons of expense, it is impossible to reinterview all households participating in a survey (*cf.* Forsman and Schreiner 1991). Therefore, the question arises of how the reinterview sample can be optimized to best detect falsifiers. Generally, it seems useful to select households for reinterview if the interviews were done by an interviewer − identified by characteristics linked to the answers in his interviews − who is more likely than others to be fabricating data. In this context, Hood and Bushery (1997) uses the term 'at risk' interviewer. If reinterview participants are sampled in a two-stage setting, whereby interviewers are selected in the first stage and participants surveyed by those interviewers in the second stage (as recommended by Forsman and Schreiner (1991)) one might oversample the at risk interviewers in the first stage.

In this paper, we demonstrate a purley statistical approach that relies on the data contained in the questionnaries to define a group of at risk interviewers. This is not a new idea; literature provides several examples for this kind of approach (Hood and Bushery 1997; Diekmann 2002; Turner, Gribbe, Al-Tayyip and Chromy 2002; Schräpler and Wagner 2003; Swanson, Cho and Eltinge 2003; Murphy, Baxter, Eyerman, Cunningham and Kennet 2004; Porras and English 2004; Schäfer, Schräpler, Müller and Wagner 2005; Li, Brick, Tran and Singer 2009). However, with the exception of the work of Li *et al.* (2009), the tests conducted in these studies rely on the examination of single indicators derived from the interviewer's data to detect falsifiers. Some studies calculate several indicators but consider them all separately. We combine multiple indicators in cluster

[1] Sebastian Bredl, Department of Statistics and Econometrics, Justus-Liebig-University, 35394 Gießen, Licher Straße 64, Germany. E-mail: sebastian.bredl@wirtschaft.uni-giessen.de; Peter Winker, Department of Statistics and Econometrics, Justus-Liebig-University, 35394 Gießen, Licher Straße 64, Germany. E-mail: peter.winker@wirtschaft.uni-giessen.de; Kerstin Kötschau, Hanse Parlament, 22587 Hamburg, Blankeneser Landstrasse 7, Germany. E-mail: kkoetschau@hanse-parlament.eu.

analyses, allowing for a better classification of the potential falsifiers compared to previous approaches. To the best of our knowledge, this procedure is an innovation in the context of identifying interviewers who fabricate data, but has already been employed in other fields in order to detect fraudulent behaviour. The basic idea is that characteristics of fraudulent ‘cases’ (what a case is depends on the context) feature striking patterns compared to honest cases that can be revealed if those characteristics are jointly considered in a cluster analysis. Murad and Pinkas (1999) try to detect fraud in the telecommunication industry by means of clustering call profiles of clients. A call is characterized by several indicators like calling time or destination of the call. Thiprungsri (2010) clusters group life claims submitted from clients to life insurance companies based on several characteristics of the claims. Claims that form very small clusters are considered to be suspicious. Donoho (2004) uses cluster analysis, among others, to trace patterns in option markets that might indicate insider trading.

We have a small survey dataset available (see subsection 3.1 for a further description of our dataset), which partially consists of falsified data. With a total of 13 interviewers and 250 questionnaires, the size of the dataset is quite limited and it is not clear to what extent our findings can be generalized to larger datasets. However the dataset enables us to demonstrate our approach. The fact that we know which data was collected honestly and which data was fabricated allows for a first evaluation of our approach. It must be stated that this a priori knowledge is no prerequisite to employ the method.

The problem of identifying at risk interviewers was addressed in the 1980s, however, literature on this issue is still scarce. In 1982, the U.S. Census Bureau implemented the Interviewer Falsification Study. Based on the information collected in the context of this study, Schreiner *et al.* (1988) find that interviewers with a shorter length of service are more likely to fabricate data. Hood and Bushery (1997) use several indicators to find at risk interviewers in the National Health Interview Survey (NHIS). For example, they calculate the rate of households that have been labelled ineligible or the rate of households without telephone number per interviewer and compare the rates to census data from the respective area. When large differences occur, the interviewer is flagged and a reinterview is conducted. Detection rates among the flagged interviewers turn out to be higher than those in random reinterview samples. Turner *et al.* (2002) also find interviewers committing data fabrication to indicate telephone numbers less frequently than honest interviewers when examining the Baltimore STD and Behaviour Survey. For the case of computer assisted interviewing, Bushery, Reichert, Albright and Rossiter (1999) and Murphy *et al.* (2004) propose the use of date and time stamps – the recording of the time and the duration of the interview by the computer – to find suspect interviewers. Those who need a remarkably long or short time to complete the entire questionnaire or certain modules or complete remarkably many questionnaires within a given time period might be flagged as at risk interviewers. Schäfer *et al.* (2005) assume that falsifiers avoid extreme answers when fabricating data. Using data of the GSOEP, the authors calculate the variance of the answers for every question on all questionnaires of an interviewer and sum up all variances. Thanks to other control mechanisms in the GSOEP, falsifiers are known and it turns out that they could be found among the interviewers with the lowest overall variances. Porras and English (2004) use a similar approach and also find falsifiers to produce variances that are smaller to those found in honestly filled questionnaires. Li *et al.* (2009) combine several predictive indicators in a logistic regression model in which the known falsification status of an interview serves as a binary dependent variable. The authors find that reinterview samples that overweight cases with a high probability of being fraudulent according to the logistic regression model identify more cases of actual data fabrication than purely randomly drawn samples. However, it is evident that past reinterview data with known falsification status must be available to conduct the logistic regression.

Further indicators discussed in literature are the number of rare or unlikely response combinations in an interviewer’s questionnaires (Murphy *et al.* 2004; Porras and English 2004) and the comparison of household compositions or descriptive statistics in interviewer’s questionnaires with the entire sample (Turner *et al.* 2002; Murphy *et al.* 2004).

Another means of detecting fabricated data that has gained a lot of popularity in recent years is Benford’s law (Schräpler and Wagner 2003; Swanson *et al.* 2003; Porras and English 2004; Schäfer *et al.* 2005), which will be discussed in section 2, along with its success in detecting fabricated interviews in previous studies. Furthermore, section 2 describes our statistical approach to identify falsifiers. Section 3 presents the data our analysis is based upon as well as our results. The paper concludes with a discussion of our findings.

2. Methods

2.1 Benford’s law

When the physicist Frank Benford noticed that the pages in logarithmic tables containing the logarithms of low numbers (1 and 2) were more used than pages containing logarithms of higher numbers (8 and 9), he started to investigate the distribution of leading digits in a wide range of different types of numbers like numbers on the first page of a newspaper, street addresses or molecular weights (Benford 1938). Benford found that the distribution of the leading non-zero digits could be described by the following formula which has become known as ‘Benford’s law:’

$$\text{Prob(leading digit} = d) = \log_{10}\left(1 + \frac{1}{d}\right). \qquad (1)$$

However, not all series of numbers Benford (1938) investigated seemed to conform to his law. Consequently, the question arose what kind of data can be supposed to produce first digits in line with the law. Discussions of this issue are provided by Hill (1995), Nigrini (1996), Hill (1999) and Scott and Fasli (2001). The detection of financial fraud is a field in which the application of Benford’s law has gained much popularity during the recent decade (Nigrini 1996; 1999; Saville 2006). The results of those studies are not relevant in our context. However, it is interesting to note that there seems to be a consensus in literature that monetary values can be supposed to follow Benford’s law. Swanson *et al.* (2003) show that the distribution of first digits in the American Consumer Expenditure Survey is close to Benford’s distribution.

The basic idea of using Benford’s law to detect fabricated data is that falsifiers are unlikely to know the law or to be able to fabricate data in line with it. Therefore a strong deviation of the leading digits from Benford’s distribution in a dataset indicates that the data might be faked. Of course, one has to be concerned if the nature of the data is such that it can be supposed to follow Benford’s law if it is authentic. Benford’s law cannot be applied if the questionnaires do not contain any or contain only very few metric variables.

Schräpler and Wagner (2003) and Schäfer *et al.* (2005) use Benford’s law to detect data fabrication in the GSOEP. In both studies, all questionnaires delivered by every single interviewer are combined and checked for whether the distribution of the first digits in the respective questionnaires deviates significantly from Benford’s law. This can be done by calculating the χ^2-statistic:

$$\chi_i^2 = n_i \sum_{d=1}^{9} \frac{(h_{id} - h_{bd})^2}{h_{bd}} \qquad (2)$$

where n_i is the number of leading digits in all questionnaires from interviewer i, h_{id} is the observed proportion of leading digit d in all leading digits in interviewer i’s questionnaires and h_{bd} is the proportion of leading digit d in all leading digits under Benford’s distribution. High χ^2-values indicate a deviation from Benford’s distribution and indicate at risk interviewers. Schräpler and Wagner (2003) use different kinds of continuous variables in their analysis, whereas Schäfer *et al.* (2005) restrict theirs to monetary values. In both studies, the critical χ^2-values are assumed to be dependent on the sample size n and are consequently adjusted for this parameter. The results obtained look promising. The fit of the leading distribution of first digits to Benford’s distribution in the questionnaires of falsifiers (which were already known in advance) is, in general, much worse than for honest interviewers. Thus it seems appropriate to use Benford’s law as a means to identify at risk interviewers.

However, when we compared the data of the honest interviewers in our dataset to Benford’s distribution, we observed a large deviation for the digit 5. This might be due to rounding of numbers by the respondents. The same problem is mentioned by Swanson *et al.* (2003) and Porras and English (2004) who opt for applying an alternative approach “in the spirit of Benford” (Porras and English 2004, page 4224). We adopt this approach which consists of comparing the distribution of leading digits in the questionnaires of an interviewer to the distribution of first digits in all questionnaires except their own. The χ^2-value on the interviewer level is calculated as described above but the expected proportion of a digit according to Benford’s law h_{bd} is replaced by the proportion of the digit in all other questionnaires. We then use the resulting χ^2-value as one indicator in the cluster analysis.

With regard to the selection of variables whose first digits are examined, we stick to the approach of Schäfer *et al.* (2005) and include only the first digits of monetary values in the analysis. The survey we are using for demonstration purposes contains monetary values expressed in local currency referring to household expenditures for different items like leasing or buying land, seeds, fertilizer, taxes, and to household income from different sources like agricultural or non agricultural self employment and public or private transfers. Overall we include first digits of 26 different monetary values per interview, ignoring values that were reported to be zero. We then pool first digits of all questionnaires delivered by one interviewer and compare the distribution of first digits to the one for all other interviews according to the method described above. The restriction to monetary values constitutes a clear criterion during the process of selecting data. Furthermore, as mentioned above, financial data is broadly agreed upon to be apt for the analysis with Benford’s law. This is important, although we do not ground our analysis on Benford’s distribution but on an approach based on it.

2.2 Multivariate analyses

Our idea is to combine several indicators, which we derive directly from the questionnaires of each interviewer and which we suppose to be different for falsifiers and honest interviewers. We do this by means of cluster and discriminant analysis. All indicators are derived on the interviewer level. This implies that we pool all questionnaires of one interviewer for the analysis, which increases the amount of data on which every single indicator value is based. This should make the indicator values more reliable and less sensitive to outliers. On the other hand, it is obvious that the discriminatory power of interviewer-level indicators decreases as soon as interviewers only fake parts of their assignments. Looking at indicators on the questionnaire level, therefore, seems to be preferable if the amount of data per questionnaire is sufficiently high.

The cluster analysis constitutes the real method of identifying at risk interviewers. The interviewers are clustered in two groups with the intention of obtaining one that contains a high share of falsifiers and another one that contains a high share of honest interviewers. Clustering does not require a priori information on who is fabricating data and who is not. In fact, this is what it is supposed to reveal. Since we know from the outset which interviewer belongs to which group, we can discover whether the cluster analysis identifies the 'true falsifiers' to be at risk. Clearly, the assumption that our approach is able to separate both groups perfectly is not realistic. The idea is rather that we obtain an at risk interviewer cluster exhibiting a higher share of falsifiers compared to the other cluster. If a reinterview is feasible, subsequent reinterview efforts might be focused on interviewers in the at risk cluster.

To judge the performance of the cluster analysis, we consider the number of undetected falsifiers as well as the number of 'false alarms.' Both types of 'errors' inccur costs: data of undetected falsifiers is likely to impair the results of further statistical analysis. False alarms inccur costs in the sense that an unnecessary effort to reinterview the respective households might be taken or data is unnecessarily removed from the sample. Furthermore, it might be demoralizing for honest interviewers if they see their work being subject to a reinterview, particuliarly if they are aware of the fact that predominantly the work of at-risk interviewers is picked. How to weight an undetected falsifier compared to a false alarm in a loss function is a highly subjective issue. Generally, it seems reasonable to assign more weight to the former than to the latter.

The discriminant analysis requires knowledge on the falsifiers versus non-falsifiers status of each interviewer before it can be conducted. Therefore, it is not an instrument to detect falsifiers. We use the discriminant analysis to verify our hypotheses on the behaviour of falsifiers, which will be discussed below, and to evaluate how well the employed indicators can separate the two groups.

One of the indicators we use is the χ^2-value, calculated by comparing the distribution of first digits in the questionnaires of each interviewer with the respective distribution in all other questionnaires as described in the previous subsection. Furthermore, we derive three other indicators from hypotheses concerning the behaviour of falsifiers fabricating data. Schäfer *et al.* (2005) assume that falsifiers have a tendency to answer every question, thus producing less missing values. Furthermore, in line with Porras and English (2004), they expect falsifiers to choose less extreme answers to ordinal questions. Hood and Bushery (1997) hypothesize that falsifiers will "try to keep it simple and fabricate a minimum of falsified data" (Hood and Bushery 1997, page 820).

Based on these assumptions, we calculate three proportions, which serve as indicator variables in the multivariate analyses along with the χ^2-value. The three indicator variables are calculated as follows:

- The 'item-non-response-ratio' is the proportion of questions which remain unanswered in all questions. We expect this ratio to be lower for falsifiers than for honest interviewers.
- The 'extreme-answers-ratio' refers to answers which are measured in ordinal scales. The ratio indicates the share of extreme answers (the lowest or highest category on the scale) in all ordinal answers. According to the above-mentioned assumptions, this ratio should also be lower for falsifiers.
- The 'others-ratio' refers to questions which, besides several framed responses offer the item 'others' as a possible answer. The choice of this item requires the explicit declaration of an alternative. If falsifiers tend to keep it simple, we can expect them to prefer the framed responses to the declaration of an alternative. Thus, this ratio too (calculated as the proportion of 'others' answers in all answers where the others item is selectable) should be lower for falsifiers.

Of course, the list of indicator variables, which might be included in the cluster analysis, can be extended. Generally, it is possible to derive many more of those variables from hypotheses on the behaviour of interviewers who fabricate data or to use those which have already been proposed in the literature, albeit not in the context of cluster analysis. For example, based on the assumption that falsifiers try to fabricate a minimum of falsified data, Hood and Bushery (1997) expect them to disproportionately often select the answer 'No' to questions, which either lead to a set of new questions or avoid it (assuming that 'No' is generally the answer that avoids further questions). So one could calculate the ratio of 'No' answers to such questions and use this ratio as a variable in the cluster analysis. We do not use this ratio, as two slightly different versions of the questionnaire were used in our empirical sample. There are only a small number of questions that lead to new questions or avoid them depending on the answers, which are identical in both versions of the questionnaire.

Furthermore, when computer assisted interviewing allows the use of date and time stamps as discussed by Bushery *et al*. (1999), the average time needed to conduct an interview or the number of interviews conducted in one day might serve as indicators. Panel surveys offer some additional information to construct indicators. Stokes and Jones (1989) propose to compare the actual rate of non-matched household members in an interviewer's questionnaires to expected nonmatch rates that are calculated conditional on several household characteristics. The authors employ this procedure in the post-enumeration survey that is conducted as follow-up survey for the U.S. Census. If the actual rate of nonmatches strongly exceeds the expected rate, the authors consider this to be an indicator for fabricated data. Generally, this approach is applicable as soon as one has two or more waves of a panel survey available.

It becomes obvious that the first steps of our approach consist of examining the structure of the questionnaire and other types of data like date or time stamps collected during the survey process. Then one might consider which indicators could be derived from those sources that are likely to differ between falsifiers and honest interviewers. Another approach is the use of data mining techniques to identify patterns that are common in fabricated data or patterns in which fabricated data differs from honestly collected data (Murphy, Eyerman, McCue, Hottinger and Kennet 2005). If those patterns are detected, they might be used as indicators instead of deriving indicators from hypothesis on falsifier behaviour. However, this approach requires a huge dataset with known cases of falsification in order to conduct the data mining process. Such a dataset is not always available.

3. Results

3.1 Data sources

The data used in this study are derived from household surveys conducted in November 2007 and February 2008 in a Commonwealth of Independent States (CIS) (*i.e.*, former Soviet Union) country. The survey was part of an international research project on land reforms and rural poverty. We intended to interview 200 households in four villages in 2007. After identifying that all interviews had been fabricated in the first surveyed village we broke the survey off and started a new round with new interviewers in other villages in February 2008. All villages had been selected by qualitative criteria like the agricultural production structure and the implementation of land reforms. The households within one village had been selected by random sample based on household lists, which were provided by the mayors of the villages. This procedure not only assured that all households had been selected at random, but also provided the basis for reinterviews as all households were exactly defined. However, these reinterviews were not planned in the very beginning. Because the households rarely owned telephones, check-calls were not possible and reinterviews in these households were associated with high costs and expenditure of time for traveling to the village for a face-to-face reinterview. Five interviewers were engaged in the first 2007 survey. Two of them had been the local partners of the research project. They had been involved in the development of the questionnaire and were responsible for the coordination of the surveys in their country. The other three interviewers were students hired by the partners. The questionnaire was composed of different sections with regard to household characteristics, resource endowment as well as income and expenditures. Most of the questions were closed questions. Only a few questions included a scale. Metric variables were collected for household expenditures like leasing or buying land, seeds, fertilizer or taxes and household income from different sources like agricultural or non-agricultural self employment and public or private transfers.

When the interviews of the 2007 survey were conducted, none of the German researchers were present in the villages. The questionnaires were collected right after the survey of the first village. In a first review of the questionnaires, we became suspicious because the paper of the questionnaires looked very clean and white. There was no dirt or dog-ears on the paper. Comparing the answers of different questionnaires of one interviewer we found two questionnaires with identical answers. Considering the fact that we asked for the amount of income from different sources in metric numbers it was very unlikely that the answers of two questionnaires would have been identical. Not getting any explanations from the project partners, we reinterviewed a sub-sample of 10% of the original sample face-to-face. None of the reinterviewed households reported having been surveyed. After detecting the fabrication of the interviews, the partners acknowledged that all interviews had been fabricated. As a matter of course, we stopped working with all interviewers and partners and implemented a new local research group.

In February 2008, the survey was repeated in the same country. As mentioned before, we selected new villages and households according to the above-mentioned criteria. We hired nine students for the interviews and arranged the survey with on-site supervision. In most cases, the interviews took place in a school or the city hall so that we could monitor all interviewers. When the interviews took place in the houses of the surveyed families we attended some of them. Due to this procedure, we presume that the questionnaires from the 2008 survey are not fabricated.

In this paper, we use a total of 250 household interviews by 13 interviewers, of which four were falsifiers from the 2007 survey (the interviews submitted by one falsifier were excluded as he filled in only three questionnaires) who definitely faked the results, referred to as F1-F4, and nine interviewers who are supposed to be honest, labelled H1-H9. Table 1 provides an overview of the number of questionnaires per interviewer, which were included in the analysis.

Table 1
Number of questionnaires per interviewer

Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Number of questionnaires	10	12	10	10	22	23	23	24	23	23	23	23	24

3.2 Cluster analysis

In this subsection, we present the results of the cluster analysis. Based on the results, we evaluate the success of our procedure in identifying interviewers who fabricate data. As already mentioned, we use four indicator variables in the cluster analysis: the item-non-response ratio, the proportion of extreme ordinally scaled answers in all ordinally scaled answers referred to as extreme ratio, the proportion of answers where the others item including an alternative was selected in all answers which offered this item (referred to as others ratio) and the χ^2-value stemming from the comparison of the leading digit distribution in the questionnaires of an interviewer with the respective distribution in all other questionnaires.

Table 2 provides the values of the four indicator variables included in the cluster analysis for all 13 interviewers. It shows that the item-non-response ratio and the others ratio are clearly lower for the four falsifiers than for the honest interviewers. F1 and F4 have not chosen the others item at all. For the extreme ratio, things seem to be less clear. All the values range between 40% and 70% except the value of interviewer F1, which is clearly lower. The χ^2-values are quite high for falsifiers F2 and F4. The values of the other two falsifiers do not differ much from the ones observed for honest interviewers.

The general idea of cluster analysis is to identify subgroups of elements in a space of elements that are all characterized by multivariate measurements (see Härdle and Simar (2007) for an introduction to cluster analysis). In the first step, a measure to determine either distance or similarity between elements has to be chosen. In the second step, elements are assigned to different subgroups or clusters. Elements within one cluster should be similar according to the selected measure whereas elements in different clusters should be distant. There is a large variety of methods according to which elements can be assigned to clusters whereby the number of clusters might either be fixed or determined by the cluster method.

Table 2
Values of the variables included in the cluster analysis for each interviewer (all values except χ^2 -value in percent)

Interviewer	Item-Non-Response	Others	Extreme	χ^2 -value
F1	1.36	0.00	28.33	19.63
F2	0.71	0.65	40.85	29.70
F3	0.68	2.33	56.90	11.34
F4	0.51	0.00	58.62	27.33
H1	3.85	18.01	65.12	14.48
H2	1.99	2.40	59.42	6.91
H3	3.10	9.47	70.07	15.49
H4	4.52	13.04	56.43	16.61
H5	1.18	4.48	70.07	12.16
H6	3.46	1.37	50.75	15.42
H7	2.51	12.72	45.65	9.11
H8	1.77	10.95	69.85	3.63
H9	0.14	1.61	69.44	19.14

We measured distance as squared Euclidian distance and employed several cluster procedures in order to check the robustness of the results. In all cases, the interviewers have been clustered in two groups with the intention to obtain one ‘falsifier group’ and one ‘honest interviewer group.’ The advantage of this approach is that a clear classification is obtained. In contrast, when one of the indicator variables is examined separately, it is not clear where to draw the line separating falsifiers and honest interviewers. Before conducting the cluster analysis, we standardized all variables on a mean of zero and on a variance of unity. This eliminates the scale effect as distances are measured in standard deviations and not in different units.

The first clustering method we use is hierarchical clustering. This is a standard procedure that can also be applied to larger datasets and is implemented in standard statistical software packages. Hierarchical clustering merges clusters step by step, combining the two closest clusters. At the beginning, every element is considered as a separate cluster. We measure distance between two clusters as the average squared Euclidian distance between all possible pairs of elements with the first element of the pair coming from one cluster and the second element from the other cluster. We used the software package STATA with the option ‘average linkage’ to conduct the hierarchical cluster analysis.

In hierarchical cluster analysis, two elements will stay in the same cluster once they are merged together. Thus, the procedure does not necessarily lead to a global optimum with regard to a given distance measure. In our case the relatively low number of interviewers allows us to conduct an alternative analysis by simply examining all possible cluster compositions and select the best one with regard to a certain target function. (The analysis was carried out in MATLAB, the programm code is available upon request.)

This procedure is clearly superior to hierarchical clustering as it ensures that the globally optimal cluster composition is identified. However, we also provide the results of hierarchical clustering as it is rather feasible compared to the computationally intensive approach of trying all possible compositions when the number of interviewers rises. Alternatively, one might resort to heuristic optimization techniques.

When examining all possible cluster compositions we use two target functions. The first one combines the ideas that a large distance between the two cluster centers is eligible as well as a small distance between the elements of a cluster and the cluster center. We look for the cluster composition, which maximizes the following expression:

$$\frac{\sum_{i=1}^{4}(\bar{d}_{1i}-\bar{d}_{2i})^2}{\sum_{j=1}^{n_1}\sum_{i=1}^{4}(d_{ij}-\bar{d}_{1i})^2+\sum_{j=n_1+1}^{13}\sum_{i=1}^{4}(d_{ij}-\bar{d}_{2i})^2}. \qquad (3)$$

The index i represents the four different indicator variables, $\bar{d}_{ai}$ with $a = 1, 2$ is the mean of variable i in cluster a, j symbolizes the different elements (interviewers) in cluster 1 and cluster 2, d_{ij} is the value of variable i for element j, and n_1 is the number of elements in cluster 1. Thus the numerator measures the distance between the two clusters, the denominator the distance within clusters and distance is measured in squared Euclidian form.

Alternatively, it could be interesting to see what optimal cluster composition results if instead of maximizing Equation (3) the average squared Euclidian distance between all possible pairs of elements within one cluster is minimized. In fact, this idea is very similar to the relevant target function in the hierarchical cluster procedures we presented before. Our second distance measure, which this time is to be minimized, is calculated as follows:

$$\frac{\sum_{j=1}^{n_1-1}\sum_{k=j+1}^{n_1}\text{SED}_{jk}+\sum_{j=n_1+1}^{13-1}\sum_{k=j+1}^{13-1}\text{SED}_{jk}}{(n_1(n_1-1))/2+((13-n_1)(13-n_1-1))/2} \qquad (4)$$

SED_{jk} is the squared Euclidian distance between elements j and k, calculated as $\text{SED}_{jk}=\sum_{i=1}^{4}(d_{ij}-d_{ik})^2$. The numerator is the sum of distances between all possible pairs of elements in the same cluster. By dividing this sum by the number of possible pairs, one obtains the average within cluster distance.

Table 3 reveals the results of the three cluster procedures. In the hierarchical analysis with linkage between groups, the three falsifiers F1, F2 and F4 form cluster 1, falsifier F3 and all honest interviewers form cluster 2. Thus, we are able to separate both groups of interviewers, except one falsifier. However, without knowing from the outset which interviewers fabricated data and which were honest, one would have to decide which of the two clusters contains the at risk interviewers. This can be done by comparing the means of the indicator variables for each cluster displayed in Table 4. For the hierarchical procedure, means of the item-non-response ratio and the others ratio are clearly lower in cluster 1. The same is true for the mean of the extreme ratio, albeit the difference between the two clusters is less striking. Finally, a higher mean of the χ^2-value can be observed for cluster 1. Given these results, one would – according to the above mentioned hypotheses on the behaviour of falsifiers – correctly identify cluster 1 to be the cluster containing the at risk interviewers. We also tried to improve the results of the hierarchical clustering procedure using the cluster means displayed in Table 4 as starting point for the K-means analysis. However, the application of K-means clustering did not lead to any changes in the cluster composition.

Table 3
Results of the three employed clustering procedures

Hierarchical clustering													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance between clusters divided by distance within clusters													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	2	1	2	2	2	2	2	2	2	2	2
Distance between elements in one cluster													
Interviewer	F1	F2	F3	F4	H1	H2	H3	H4	H5	H6	H7	H8	H9
Cluster	1	1	1	1	2	2	2	2	2	2	2	2	1

Table 4
Indicator variable means by cluster for the three cluster compositions

	Item-Non-Response		**Others**		**Extreme**		**χ^2-value**	
Hierarchical clustering								
Cluster	1	2	1	2	1	2	1	2
Mean	0.86	2.32	0.22	7.64	42.60	61.37	25.55	12.43
Distance between clusters divided by distance within clusters								
Cluster	1	2	1	2	1	2	1	2
Mean	0.86	2.32	0.22	7.64	42.60	61.37	25.55	12.43
Distance between elements in one cluster								
Cluster	1	2	1	2	1	2	1	2
Mean	0.68	2.80	0.92	9.06	50.83	60.92	21.43	11.73

The cluster composition that maximizes Equation (3) is identical to the one obtained using hierarchical clustering. Consequently, as can be seen from Table 4, the indicator means within the two clusters are identical as well.

The cluster composition minimizing Equation (4) is slightly different. Cluster 1 now contains all falsifiers and one honest interviewer. The means of the indicator variables again clearly indicate cluster 1 to be the cluster containing the at risk interviewers. This is a very satisfying result. All falsifiers are identified and only one false alarm is produced.

However, it should be kept in mind that this does not mean that this particular cluster method works best when applied to another dataset.

To evaluate to what extent a higher number of indicators leads to better results, we repeated our cluster approach based on Equations 3 and 4 with all possible combinations of indicators, including cases that only rely on one indicator. The results (see Table 7 in the appendix) generally indicate that an increasing number of indicators improves the results. However, there are also combinations with a smaller number of indicators that lead to similar results compared to those based on all four indicators. Determining which indicator composition is the best would require the highly subjective fixation of the relative cost caused by non-identified falsifiers compared to the cost caused by a false alarm. But one can determine which indicator compositions are not Pareto dominated in the sense that there is no other composition that exhibits less non-identified falsifiers (false alarms) and at the same time not more false alarms (non-identified falsifiers). The indicator composition including all four indicators is the only one that is not Pareto dominated no matter which equation is used. In contrast, compositions including only one indicator are Pareto dominated in six out of eight cases.

3.3 Discriminant analysis

Finally, we turn to the discriminant analysis to check whether the hypotheses on falsifiers' behaviour our cluster analysis is based upon are valid. Discriminant analysis can be used if the clusters are known in order to assess how well the indicators in the analysis can separate the different groups and whether group membership can be predicted correctly (see Härdle and Simar (2007) for an introduction to discriminant analysis). In a linear discriminant analysis, the coefficients b_0 and b_i of the discriminant function $D = b_0 + \sum_{i=1}^{n} b_i x_i$ are determined in such a way that they maximize a function that increases with the difference of the mean D-values of the two different groups and at the same time decreases with the differences of the D-values of elements within the groups. In our case, the x_i are our four indicator variables and we obtain two groups by separating falsifiers and honest interviewers.

We use prior probabilities corresponding to the relative group size (4/13 and 9/13) in order to predict group membership. Table 5 shows the results. Obviously the four variables allow a good separation of the falsifiers and the honest interviewers, as the group membership is correctly predicted in all cases but one.

As can be seen from Table 5 negative values of the discriminant function are associated with the falsifier group. Consequently, Table 6 indicates that three of the four coefficients' signs are in line with the expected falsifier behaviour. Higher item-non-response and extreme ratios lead to a higher probability to observe an honest interviewer as does a lower χ^2-value. The estimated coefficient for the others ratio is negative. Thus an increase in the others ratio ceteris paribus raises the probability that an interviewer has fabricated data. This might appear as a contradiction to our above-mentioned hypotheses. One possible explanation might be that the effect of the others ratio is already captured by the item-non-response ratio. In fact, the correlation coefficient between the two variables is quite high with a value of 0.71. The Wilks' lambda of the discriminant analysis is statistically significant on the 5%-level.

Table 5
Results of the discriminant analysis by interviewer

Interviewer	Predicted group	Actual group	Discriminant function
F1	1	1	-2.878
F2	1	1	-3.376
F3	2	1	-0.541
F4	1	1	-1.955
H1	2	2	1.828
H2	2	2	1.060
H3	2	2	1.747
H4	2	2	1.616
H5	2	2	0.706
H6	2	2	0.777
H7	2	2	-0.041
H8	2	2	1.765
H9	2	2	-0.710

Table 6
Standardized and non-standardized estimated coefficients (discriminant analysis)

Variable	Coefficient (non-standardized)	Coefficient (standardized)
Item-Non-Response	0.767	0.917
Others	-0.025	-0.129
Extreme	0.075	0.821
χ^2 -value	-0.092	-0.562
Constant	-4.250	–
Wilks' lambda (Prob > F)	0.0254	

4. Conclusion

Survey data are potentially affected by interviewers who fabricate data. Data fabrication is a non-negligible problem as it can cause severe biases. Even a small amount of fraudulent data might seriously impair the results of further empirical analysis. We extend previous approaches to identify at risk interviewers by combining several indicators derived directly from the survey data by means of cluster analysis. To demonstrate our approach, we apply it to a small dataset which was partialy fabricated by falsifiers. The fact that we know the falsifiers from the outset allows us to evaluate the results of the cluster analysis and to furthermore conduct a discriminant analysis to reveal how well the two groups of interviewers can be separated by the indicator variables. Different types of cluster analyses are conducted. All of them lead to the identification of an at risk interviewer cluster, with the item-non-response ratio and the others ratio being the clearest indicators. We are not able to identify falsifiers perfectly. However, in all cases the at risk interviewer contains a much higher share of falsifiers than the second cluster. The advantage of clustering is that one obtains a clear classification of interviewers who are at risk and the other interviewers, something that is not the case when indicators like the χ^2-value are examined separately. Furthermore, it allows us to combine the information of several indicators. By investigating the performance of all possible subsets of indicators we find that generally a larger number of indicators is more apt to identify falsifiers. The fact that different clustering methods lead to different results should not necessarily be considered a shortcoming of our approach. Depending on how one weights the costs of an undetected falsifier relative to a false alarm, one might finally assign only those interviewers to the potential falsifier group that always fall into the at risk cluster, no matter what clustering method is applied (which would imply high costs of false alarms), one might assign all interviewers to the potential falsifier group that fall into the at risk cluster at least once (which would imply high costs of undetected falsifiers) or choose a solution in between.

The application to a small dataset demonstrates another merit of our approach: it was tested and worked well in a situation in which the number of questionnaires per interviewer was quite limited (three of the falsifiers only submitted 10 questionnaires). If a small number of questionnaires per interviewer is sufficient to perform the analysis, one might also think about implementing it during the main field period when interviewers have only submitted a certain percentage of their questionnaires. Falsifiers could then be replaced by other interviewers who survey the units that should have been surveyed by the falsifiers.

Of course, when examining our results one has to keep in mind that we applied our method to a dataset in which a very severe form of data fabrication occurred: on the one hand we have falsifiers that faked all of their questionnaires (nearly) completely, on the other hand we have interviewers that (presumably) did all of their work honestly, which eases the discrimination between honest interviewers dishonest interviewers. Furthermore, with 13 interviewers, the size of our sample is quite limited. It would be interesting to explore the usefulness of our approach applied to larger datasets, given that the share of falsified interviews in large surveys has been found to be smaller than in our case. Additionally, larger datasets might allow the construction of additional indicators for the cluster analysis. If the survey has a reinterview program it would be possible to evaluate the usefulness of our approach by comparing the 'success' of a random reinterview with the success of a reinterview focusing on interviewers that were labeled as being at risk. We also intend to pursue the analysis in an experimental setting. An appropriate setting can ensure that one obtains a dataset which was partly collected by conducting real interviews and partly fabricated by telling some participants in the experiment to fill their questionnaires themselves.

Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft through the project 'SPP 1292: Survey Methodology' is gratefully acknowledged.

We furthermore thank John Bushery and four anonymous referees for providing useful comments on our paper.

Appendix

Table 7
Results of the cluster analyses based on Equations 3 and 4 for all possible cluster combinations

Indicators				Equation 3		Equation 4	
Item-Non-Response	**Others**	**Extreme**	**χ^2-value**	**Undetected falsifiers**	**False Alarms**	**Undetected falsifiers**	**False Alarms**
			X	2	0	1	1
		X		2	1	2	2
		X	X	2	0	1[1]	0
	X			0[1]	4	0	4
	X		X	2	0	0	2
	X	X		3	0	0	3
	X	X	X	1[1]	0	1	1
X				0[1]	4	0	4
X			X	2	1	0	2
X		X		3	0	-[2]	-
X		X	X	1[1]	0	1	1
X	X			0[1]	4	0	4
X	X		X	1	1	0	2
X	X	X		0[1]	4	0	4
X	X	X	X	1[1]	0	0[1]	1

[1] Indicator composition not Pareto dominated.
[2] Mean cluster values did not allow for an identification of the 'at risk' cluster.

References

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(1), 551-572.

Biemer, P., and Stokes, S. (1989). The optimal design quality control sample to detect interviewer cheating. *Journal of Official Statistics*, 5(1), 23-29.

Bushery, J., Reichert, J., Albright, K. and Rossiter, J. (1999). Using date and time stamps to detect interviewer falsification. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 316-320.

Diekmann, A. (2002). Diagnose von Fehlerquellen und methodische Qualität in der sozialwissenschaftlichen Forschung. Technical Report Manuskript 06/2002, Institut für Technikfolgenabschätzung (ITA), Wien.

Donoho, S. (2004). Early detection of insider trading in option markets. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 420-429.

Eyerman, J., Murphy, J., McCue, C., Hottinger, C. and Kennet, J. (2005). Interviewer falsification detection using data mining. In *Proceedings*: *Symposium 2005*, *Methodological Challenges for Future Information Needs*. Statistics Canada.

Forsman, G., and Schreiner, I. (1991). The design and analysis of reinterview: An overview. In *Measurement Errors in Surveys*, (Eds., P.B. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz and S. Sudman), New York: John Wiley & Sons, Inc, 279-301.

Guterbock, T.M. (2008). Falsification. In *Encyclopedia of Survey Research Methods*, (Ed., P.J. Lavrakas), Sage Publications, Thousand Oaks, 1, 267-270.

Härdle, W., and Simar, L. (2007). *Applied Multivariate Statistical Analysis*, 2nd Edition. Springer, Berlin.

Hill, T. (1995). A statistical derivation of the significant digit law. *Statistical Science*, 10(4), 354-363.

Hill, T. (1999). The difficulty of faking data. *Chance*, 26, 8-13.

Hood, C., and Bushery, M. (1997). Getting more bang from the reinterviewer buck: Identifying 'At risk' interviewers. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 820-824.

Li, J., Brick, J., Tran, B. and Singer, P. (2009). Using statistical models for sample design of a reinterview program. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 4681-4695.

Murad, U., and Pinkas, G. (1999). Unsupervised Profiling for Identifying Superimposed Fraud. Lecture Notes in Computer Science, 1704, 251-261.

Murphy, J., Baxter, R., Eyerman, J., Cunningham, D. and Kennet, J. (2004). A system for detecting interviewer falsification. Paper Presented at the American Association for Public Opinion Research 59th Annual Conference.

Nigrini, M. (1996). A taxpayers compliance application of Benford's law. *Journal of the American Taxation Association*, 18, 72-91.

Nigrini, M. (1999). I've got your Number. *Journal of Accountancy*, 187(5), 79-83.

Porras, J., and English, N. (2004). Data-driven approaches to identifying interviewer data falsification: The case of health surveys. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 4223-4228.

Saville, A. (2006). Using Benford's law to predict data error and fraud - An examination of companies listed on the JSE Securities Exchange. *South African Journal of Economic and Management Sciences*, 9(3), 341-354.

Schäfer, C., Schräpler, J., Müller, K. and Wagner, G. (2005). Automatic identification of faked and fraudulent interviews in the German SOEP. *Schmollers Jahrbuch*, 125, 183-193.

Schnell, R. (1991). Der einfluss gefälschter Interviews auf survey ergebnisse. *Zeitschrift für Soziologie*, 20(1), 25-35.

Schräpler, J., and Wagner, G. (2003). Identification, Characteristics and Impact of Faked Interviews in Surveys - An analysis by means of genuine fakes in the raw data of SOEP. IZA Discussion Paper Series, 969.

Schreiner, I., Pennie, K. and Newbrough, J. (1988). Interviewer falsification in census bureau surveys. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 491-496.

Scott, P., and Fasli, M. (2001). Benford's law: An empirical investigation and a novel explanation. CSM technical report, Department of Computer Science, University Essex.

Stokes, L., and Jones, P. (1989). Evaluation of the interviewer quality control procedure for the post-enumeration survey. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 696-698.

Swanson, D., Cho, M. and Eltinge, J. (2003). Detecting possibly fraudulent data or error-prone survey data using Benford's law. In *Proceedings of the Survey Research Method Section*, American Statistical Association, 4172-4177.

Thiprungsri, S. (2010). Cluster Analysis for Anomaly Detection in Accounting Data. Collected Papers of the Nineteenth Annual Strategic and Emerging Technologies Research Workshop San Francisco, California.

Turner, C., Gribbe, J., Al-Tayyip, A. and Chromy, J. (2002). Falsification in Epidemiologic Surveys: Detection and Remediation (Prepublication Draft). Technical Papers on Health and Behavior Measurement. Washington DC: Research Triangle Institute. No. 53.

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 11-21
Statistics Canada, Catalogue No. 12-001-X

The application of graph theory to the development and testing of survey instruments

Steven Elliott [1]

Abstract

This paper focuses on the application of graph theory to the development and testing of survey research instruments. A graph-theoretic approach offers several advantages over conventional approaches in the structure and features of a specifications system for research instruments, especially for large, computer-assisted instruments. One advantage is to verify the connectedness of all components and a second advantage is the ability to simulate an instrument. This approach also allows for the generation of measures to describe an instrument such as the number of routes and paths. The concept of a 'basis' is discussed in the context of software testing. A basis is the smallest set of paths within an instrument which covers all link-and-node pairings. These paths may be used as an economic and comprehensive set of test cases for instrument testing.

Key Words: Graph theory; Computer Assisted Interviewing (CAI); Questionnaire development; Software testing; Basis testing; Test cases.

1. Introduction

Graph theory is a branch of mathematics which deals with collections of nodes and links. A visual representation of a collection of nodes and links is referred to as a 'graph'. Graphs have been used in many areas of study to model real-world phenomena. The earliest examples appear in the analysis of transportation logistics (Berge 1976, page VII). In such analyses, a graph-theoretic approach is useful for determining such things as a maximally efficient set of paths to cover a number of locations. The locations are represented by the nodes of the graph, and the links represent routes from one location to another.

Graph theory has applications also in survey methodology. If the questions in a survey questionnaire are represented as nodes and the routes of flow between questions are represented as links, then a graph may be used to model a questionnaire. As such, many of the theorems and descriptive measures from graph theory pertain to questionnaires. In addition, the processes of documenting and testing survey instruments benefit from a graph-theoretic approach. For example, a documentation system that contains one table for questions and another for response alternatives has the ability to verify the connectedness of all instrument components as well as perform simulations of a working instrument. A testing procedure in which the set of test cases minimally spans the 'basis' of an instrument graph guarantees that all combinations of consecutive links and nodes are tested with the smallest possible number of cases.

A graph-theoretic representation is not necessary for the development, documentation, or testing of most survey instruments. In most cases, survey instruments have relatively few questions and the routing through an instrument does not have many branching points. Examples of this are customer satisfaction surveys and short, paper-and-pencil surveys such as the U.S. Census. For these types of instruments, conventional documentation and testing procedures are adequate. However, large and complex surveys, like many current survey efforts, may benefit from a graph-theoretic approach. For example, the Canadian Financial Capability Survey (CFCS) is a survey that was conducted in 2009 to determine Canadians' knowledge and behavior with respect to financial decision making. It was a computer-assisted telephone interview comprised of 12 sections each of which had approximately 12 questions (Statistics Canada 2010). Another example is the Consumer Expenditure Surveys Quarterly Interview CAPI Survey (2010) conducted by the United States Department of Labor, Bureau of Labor Statistics. This survey has 22 sections most of which have 3 or more subsections, and within each subsection there may be as few as six or as many as 90 questions (US Bureau of Labor Statistics 2010). Either of these examples would be a good candidate for a graph-theoretic approach to documentation and testing.

This paper addresses the application of mathematical graph theory to survey research instruments. The next section of the paper which follows immediately below contains a description of a questionnaire as a graph and a delineation of the special properties that set apart a questionnaire graph from other types of graphs. The third section outlines the implications of a graph-theoretic representation on the structure of databases used for documentation/specifications systems for computer-assisted surveys. In Section 4, the specific features of graph-theoretic data structures are discussed. Sections 5 and 6 pertain to software testing and

[1] Steven Elliott, Westat, Incorporated, 1500 Research Blvd., Rockville, MD 20850-3158, U.S.A. E-mail: sdelliott2@verizon.net.

the implications of graph theory on testing. A rationale is presented for the use of a 'basis' set of test cases which covers all pairs of linked nodes. This set of paths constitutes a comprehensive set of test cases for instrument testing.

2. A questionnaire as a graph

A graph may be represented as follows: $G = (V, E)$, where $V = \{v_1, v_2, v_3, ..., v_n\}$ is a set of nodes or vertices and $E = \{(v_i, v_j), (v_1, v_k), ...\}$ is a set of links or relations between pairs of vertices. Links are referred to as 'edges' in the terminology of graph theory, and hence the common usage of "E" to represent them (Chartrand 1985, page 27). A graph need not have any additional special characteristics. However, graphs which are attributed special characteristics are useful in modeling many phenomena in science and engineering. For example, graphs with undirected edges (*i.e.*, where both of the nodes attached to a link may be a predecessor or successor) may be used to model AC electric circuits, and graphs with directed edges may be used to model problems in traffic-pattern design. Other graphs with special characteristics are utilized to model networks in computer science, communications, sociology, and psychology.

In the case of survey questionnaires, the nodes of the graph represent different components or parts of a survey instrument. Most frequently, these are the substantive questions of a survey or decision points where routing is determined. The edges represent the response alternatives or outcomes associated with a node. Edges also represent the routing from one node to the next, and each edge has a unique predecessor and successor node. The graph depicted in Figure 1 represents a simple, 12-question survey instrument. The black circles (*i.e.*, nodes) represent the components of the instrument, and the lines connecting the black circles represent the edges that join one question to another. For example, the first node could represent a question with two response alternatives such as 'yes' and 'no'. The second node could represent a question with five response alternatives, where the first three alternatives branch to node 3, and the fourth and fifth alternatives branch to node 4.

When a graph is used to represent a questionnaire, there are a number of special properties that are attributed to the graph. These properties define the logical nature of a questionnaire. Bethlehem and Hundepool (2004) pointed out a number of these properties. First, a questionnaire has a starting node and an ending node. Second, all nodes other than the starting and ending nodes are connected. This means that for each node in the graph there is at least one route to it from the starting node, and one route away from it to the ending node. A third property of a questionnaire graph is that each of the edges is directed. This means that the route of flow from one node to another is always in one direction. A fourth characteristic of a questionnaire graph is that it may have multiple edges between a single pair of nodes. Many types of graphs are restricted such that only one edge may join a pair of nodes. This restriction does not apply to a questionnaire graph, because questionnaires commonly have more than one response alternative leading from one question to another. A final characteristic is that looping structures are permitted. This means that a node may appear multiple times on a single route. Looping structures are used frequently in questionnaires to modify responses that are determined to be incorrect. For example, financial or time-usage questions may be checked with edits that loop back if component questions do not sum to the correct total.

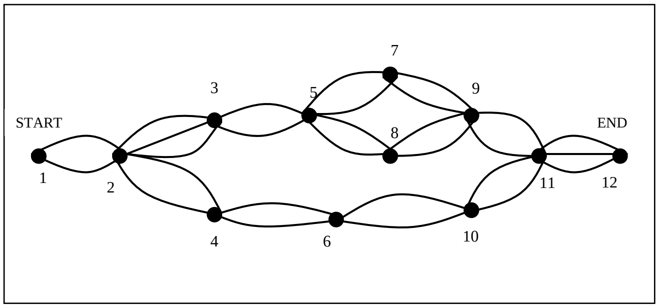


Figure 1 Representation of a Survey Instrument as a Graph

The characteristics of a questionnaire graph may be summarized as follows:

1. a starting node and an ending node,
2. connectedness (*i.e.*, each node is connected to the start and end nodes),
3. all edges are directed,
4. pairs of nodes may have multiple or parallel edges connecting them, and
5. nodes may appear more than once on a route.

Given a set of defining properties, it is possible to determine a number of descriptors including the number of routes and a basis. It is possible also to model a documentation system on the structure of the graph as illustrated in the next section.

3. Documentation and specification systems for survey questionnaires

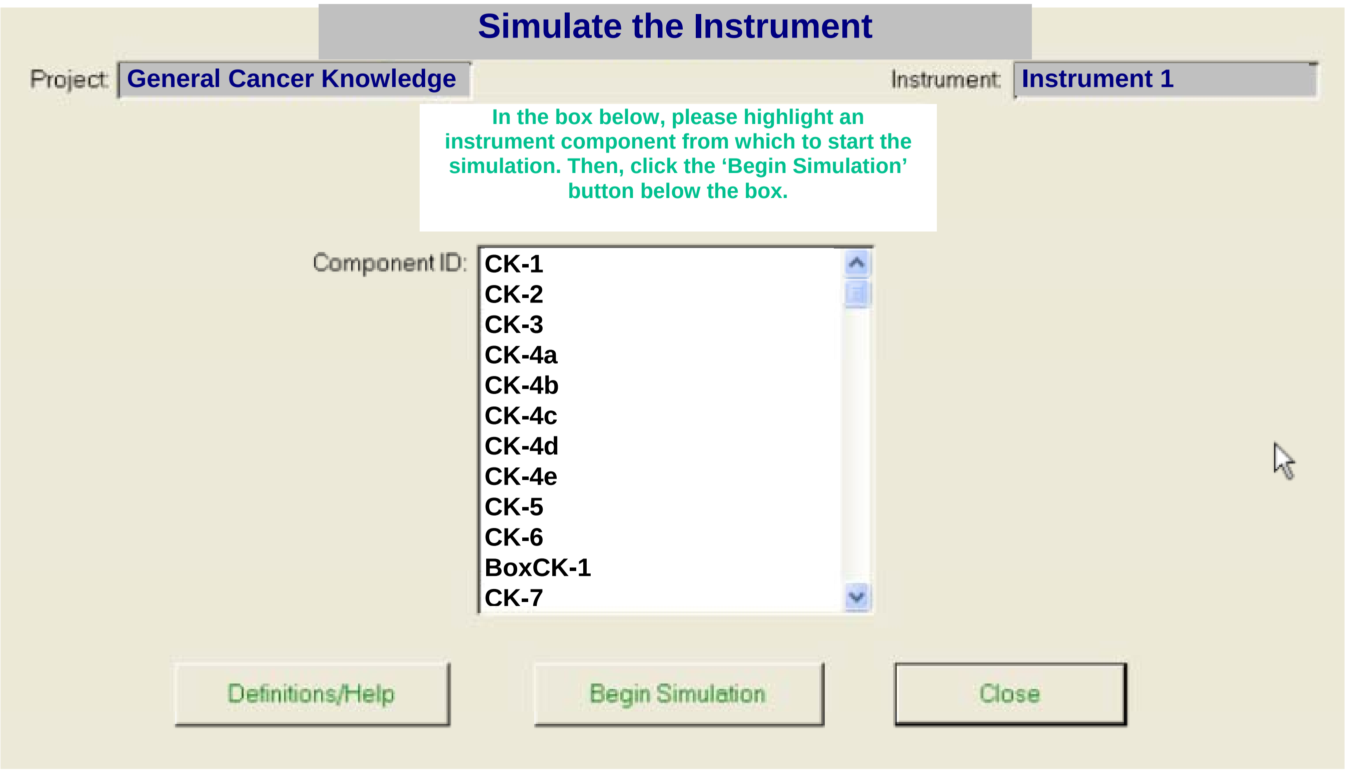
Questionnaire documentation systems are typically one of two types: a text document or a relational database. For text-document systems, the information pertaining to a substantive question or other type of instrument component is most often presented as a section of the document. It consists of the question text, response alternatives, routing, and instructions for programmers. The documentation system itself has no functionality aside from the search and print capabilities available in the word-processing software used to create the documentation. Systems using a relational database, on the other hand, are typically structured as a table where the rows represent the questions of the survey, and the columns represent attributes of the questions. Each record in the table is an n-tuple of question attributes. For example, the attributes of a question might include: name, sequence number, text of the question, response alternatives, routing information, and technical notes. One such specifications system is the *Tool for the Analysis and Documentation of Electronic Questionnaires* (TADEQ) (Bethlehem and Hundepool 2004). Other examples include systems developed at Westat Inc. for the *Medicare Current Beneficiary Survey* (MCBS) sponsored by the US Centers for Medicare and Medicaid Services (Medicare Current Beneficiary Survey: Overview 2010) and the *Medical Expenditure Panel Survey* (MEPS) sponsored by the US Department of Health and Human Services (MEPS: Survey Instruments and Associated Documentation 2010). These database systems have in common a structure of one primary table where each record represents a question.

Despite the advantages afforded by the straightforward nature of conventional systems, a specification system modeled like a graph has capabilities beyond those possible with a conventional structure. Before describing those capabilities and the necessary underlying structure, it should be noted that there are multiple ways in which a graph-theoretic data structure may be constructed (the interested reader is referred to Gibbons (1985, page 73) who described and categorized a number of those structures). The system proposed here is a relational list structure with two primary tables. One table represents the nodes of the graph, and the second table represents the edges. In the table representing nodes, each record or row represents an individual instrument component (*i.e.*, survey question, edit, or routing decision point). The second table represents edges where each record represents an individual edge (*i.e.*, a response alternative or a specific condition existing at a decision point). Each record from either table contains attributes associated with the record. Individual attributes are contained in the columns of the table. In the table of nodes, each column represents a specific attribute such as the component ID and component type. In the table of edges, each column represents an attribute such as the text of a response alternative. Two important distinctions between a documentation system with this structure versus a more conventional documentation system are: 1) the information pertaining to edges is not contained in the table for instrument components and 2) the table of edges (*i.e.*, links) contains identifiers for the predecessor and successor of an edge. As described in the next section, these distinctions allow a documentation system to perform in ways not possible with conventional systems.

4. Features of a graph-based specifications system

The use of separate tables for nodes and links as the building blocks of a specifications systems has several advantages. Most important of these advantages is the ability to simulate an interview. A developer or tester can move through an instrument selecting response alternatives while being routed from one instrument component to another just as if they were administering the instrument to a respondent. Figure 2 is an example of a screen display for simulating an instrument. The component from which simulation begins is selected from this screen. Figure 3 is the actual simulation screen itself. It shows the current component with the question text or conditional in the center of the screen. The lower left is a display of all components from which one may have come in order to arrive at the current component (*i.e.*, predecessors). These are referred to as ‘origination points’ in the screen display. The lower right is a display of destination points or components to which one may go from the current component (*i.e.*, successors). Thus, one may move through an instrument one component at a time in either direction by selecting either an origination point or a destination point. In Figures 2 and 3, the questionnaire used as an example is one on general knowledge about cancer, and the question depicted in Figure 4 has only one predecessor and one successor. This will be the case for most survey questions, however if multiple predecessors or successors did exist, they would be listed in the display.

The ability to simulate the operation of a survey instrument is made possible because a separate table is utilized for links. This table may be queried to find all predecessors and successors for any component in the questionnaire. During the design phase of development, this feature can be used to insure that all sections and questions are properly connected and all routing is correct. In the testing phase of development, this feature may be used to perform side-by-side comparisons of an instrument and the specifications upon which it was built. A tester could have the specifications system simulating the instrument on one monitor while running the actual instrument on a second. Such comparisons can be used to check not only the wording and formatting of questions and response alternatives, but also to verify that the instrument is going to the appropriate question at the appropriate time. Reports of errors or problems may then be entered directly into the specification system as an attribute of an instrument component.



Simulate the Instrument

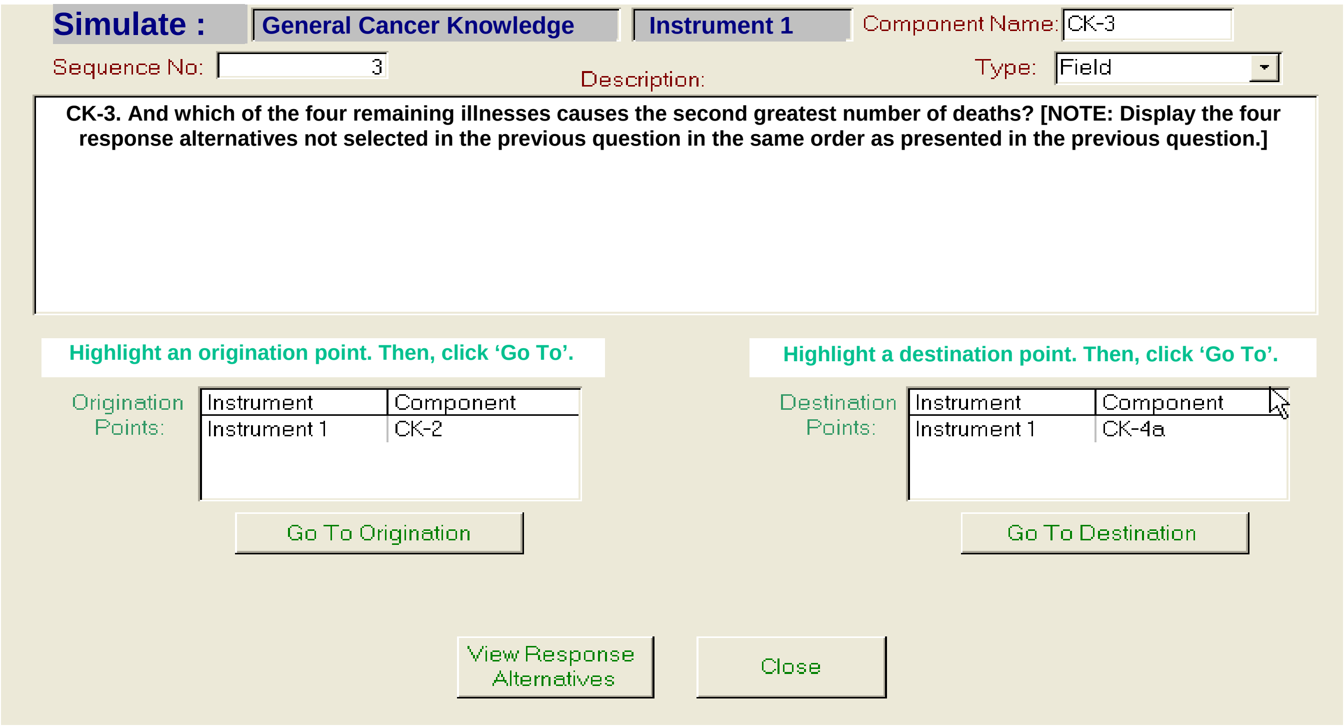
Project: General Cancer Knowledge

Instrument: Instrument 1

In the box below, please highlight an instrument component from which to start the simulation. Then, click the 'Begin Simulation' button below the box.

Component ID:
CK-1
CK-2
CK-3
CK-4a
CK-4b
CK-4c
CK-4d
CK-4e
CK-5
CK-6
BoxCK-1
CK-7

Figure 2 Begin simulation screen



Simulate : General Cancer Knowledge Instrument 1

Component Name: CK-3

Sequence No: 3

Description:

Type: Field

CK-3. And which of the four remaining illnesses causes the second greatest number of deaths? [NOTE: Display the four response alternatives not selected in the previous question in the same order as presented in the previous question.]

Highlight an origination point. Then, click 'Go To'.

Origination Points:

Instrument	Component
Instrument 1	CK-2

Highlight a destination point. Then, click 'Go To'.

Destination Points:

Instrument	Component
Instrument 1	CK-4a

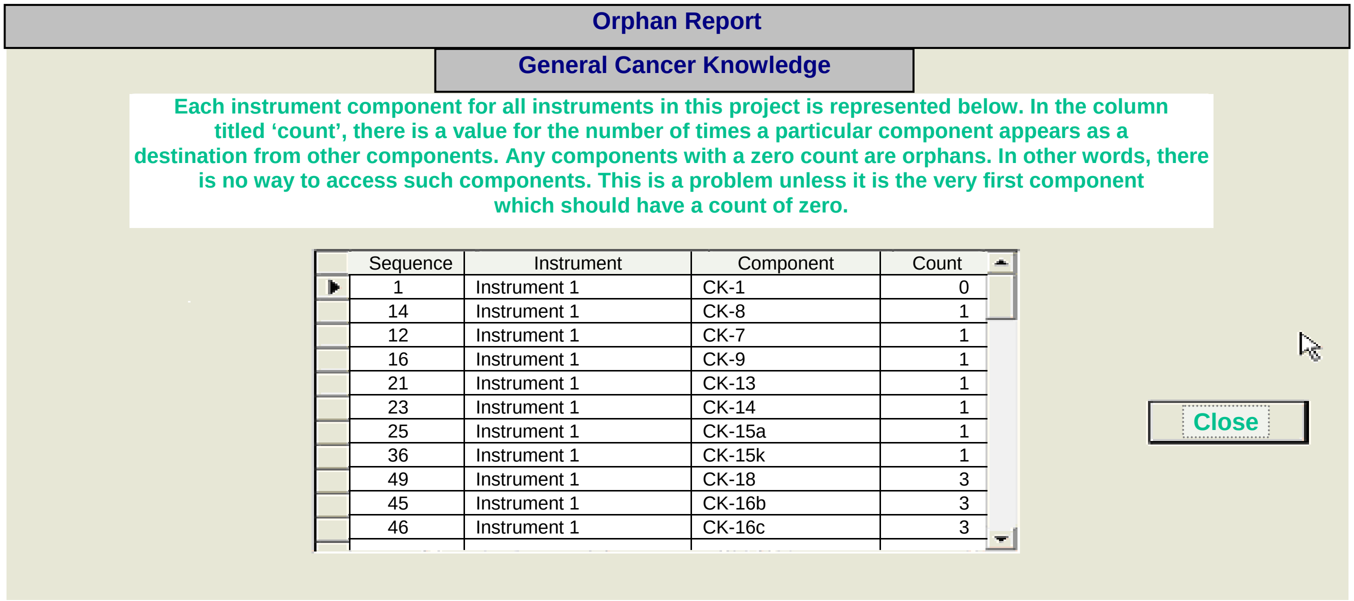
Figure 3 Simulation screen

Another method for evaluating the integrity of a questionnaire is to identify ‘orphan’ instrument components. Sometimes in the course of creating or modifying a questionnaire, an instrument component may become inaccessible. Such components are referred to as ‘orphans’. Since a table exists for links (*i.e.*, response alternatives and conditions), it is possible to run queries on this table to determine if a particular question appears as the successor to any link. If the question does not appear as a successor, then it is an orphan. Figure 4 contains the screen display for a listing of instrument components sorted by the frequency with which each appears as a successor. This is called an ‘Orphan Report’ in the figure. It shows that the first question in the survey has no origination points. This is as it should be since the first component cannot have predecessors. Any other component having zero origination points is an orphan. The orphan report is useful also in characterizing instrument components. For example, a question or component with a large number of originations may be the first question of a section devoted to handling premature terminations. Such a section is accessible from any other section of the interview, and therefore it would have a large number of predecessors.

5. Testing

Testing a computer-assisted survey instrument is the process of verifying that the behavior of the instrument is consistent with the design specifications. Several approaches have been utilized to accomplish this. One is to test first the building block components of a system, and then move to increasingly larger and more integrated assemblages of components (*i.e.*, ‘bottom-up’ testing). Testing the building block components is referred to as ‘unit testing’ (Beizer 1995, page 5). After each of the building blocks has been tested separately, the blocks are assembled, and testing is concentrated on how the components interact. This is referred to as ‘integration testing’ (Hetzel 1984, page 11). The final stage of integration testing is ‘system testing’ where the entire system as a whole just as it would be used in a true production environment (Myers 1979, page 110).

Other approaches and terminology have also been applied to testing procedures. These include ‘black-box’, ‘white-box’, and ‘regression’ testing. In black-box testing, a program is treated as if it were in a black box where the inner workings not visible. Inputs and outputs are the only observable aspects of program function (Beizer 1995, page 8). White-box testing utilizes knowledge of the program code to decide how to conduct the tests and which cases are used in testing (Patton 2006, page 55). For example, a programmer might conduct a series of white-box tests such that every line of code is ‘exercised’ (*i.e.*, ‘code coverage’) or such that every branching point is exercised (*i.e.*, ‘branch coverage’). Regression testing is used to insure code integrity after changes or additions have been introduced to an operational program (Beizer 1995, page 235). Regression tests utilize a set of test cases. This set is selected such that each of the major branches of the program is exercised. Other types of testing (*e.g.*, alpha, beta, usability) are also used in software development, and there are many sources for a more comprehensive description of testing procedures (see Kaner, Falk and Nguyen 1999, page 277).



Orphan Report

General Cancer Knowledge

Each instrument component for all instruments in this project is represented below. In the column titled ‘count’, there is a value for the number of times a particular component appears as a destination from other components. Any components with a zero count are orphans. In other words, there is no way to access such components. This is a problem unless it is the very first component which should have a count of zero.

Sequence	Instrument	Component	Count
1	Instrument 1	CK-1	0
14	Instrument 1	CK-8	1
12	Instrument 1	CK-7	1
16	Instrument 1	CK-9	1
21	Instrument 1	CK-13	1
23	Instrument 1	CK-14	1
25	Instrument 1	CK-15a	1
36	Instrument 1	CK-15k	1
49	Instrument 1	CK-18	3
45	Instrument 1	CK-16b	3
46	Instrument 1	CK-16c	3

Figure 4 Orphan report

In any testing procedure, a major concern is testing bias. This results when some components or functionality of an instrument are excluded from testing. For example, questions which appear toward the end of a survey or in an obscure section may be more likely to be excluded. Testing bias is eliminated completely if a set of test cases is selected such that all instrument components, links between components, and aspects of functionality are included. However, given the length and complexity of some surveys, comprehensive testing is not a practical option. Consider, for example, the questionnaire represented in Figure 1. This questionnaire has only 12 questions and 28 response alternatives, and yet, there are 672 possible routes through the instrument. In large surveys such as those mentioned above, the number of routes could be well over 10,000. Thus, if comprehensive coverage is not a viable approach for large surveys, it is possible to avoid testing bias by taking a probability sample of potential test cases. A graph-theoretic approach can be useful in both the specification of the universe of test cases and in the determination of a rational approach to sampling test cases.

6. A graph-theoretic approach to testing

A universe of test elements can be defined in several different ways. One could use the elements already discussed - test cases, where each case is a mock interview. Alternatively, a universe of test elements could be survey questions, response alternatives, or any of a variety of combinations of questions and response alternatives. The discussion here is limited to test cases, and therefore, it will be helpful to provide precise definitions of a test case and two closely related terms, ‘path’ and ‘route’.

A path is a unique, ordered set of nodes, which traverses an instrument from beginning to end. Each node in a given path, provided that it is not a starting or ending node, is linked to a predecessor and a successor (this definition is consistent with Bethlehem and Hundepool 2004). A unique path results whenever a component has more than one successor component. In Figure 1, multiple successors appear for components 2 and 4. These two branching nodes result in three paths:

Path 1 - 1, 2, 3, 5, 7, 9, 11, 12
Path 2 - 1, 2, 3, 5, 8, 9, 11, 12
Path 3 - 1, 2, 4, 6, 10, 11, 12

A ‘route’, on the other hand, is a unique, alternating series of nodes and links beginning with the starting node and terminating with the ending node. Like a path, a route must satisfy the properties of connectedness and direction. ‘Route’ is the graph-theoretic term which is synonymous with what is commonly called a ‘test case’ in software testing. Since a route takes into account which link connects a pair of nodes, the number of routes in a graph is greater than or equal to the number of paths. The number of routes contained within a particular a path is equal to the product of the number of links between each pair of nodes along the path. Thus for the example in Figure 1, the number of routes for each path is:

Path 1 - 2 x 3 x 2 x 2 x 2 x 2 x 3 = 288
Path 2 - 2 x 3 x 2 x 2 x 2 x 2 x 3 = 288
Path 3 - 2 x 2 x 2 x 2 x 2 x 3 = 96

The total number of routes is the sum of routes over all paths (*i.e.*, 288 + 288 + 96 = 672). A formula for computing the number of routes is:

$$\text{Routes} = \sum_{i}^{P} \prod_{j}^{NP_i} \text{links}_{ij}$$

where $\boldsymbol{i}$ represents the i^{th} path, $\mathbf{P}$ represents the total number of paths, $\boldsymbol{j}$ represents the j^{th} set of links on a given path, $\mathbf{NP}_{\boldsymbol{i}}$ represents the number of pairs of connected nodes on a given path, and *links* represents the number of links connecting a pair of nodes.

If a testing protocol is based on a sample of routes, then a minimum and comprehensive suite or universe of test cases is contained in the ‘basis’ of a graph. The term, ‘basis’, in this context is analogous to a ‘basis’ in geometry. The basis of a geometric space is a set of vectors which is sufficient to span the space, or in other words, a basis is a set of vectors sufficient to locate any point in the space. Likewise, the basis of a graph is a set of paths sufficient to include all predecessor-successor pairings of nodes. This implies that all nodes and at least one of the links between any connected pair of nodes are included. A basis is a subset of all possible paths. All questionnaires have a set of paths (**P**) in which each member satisfies the definition of a path as stated above (*i.e.*, a unique sequence of nodes). Within this set is a subset which has the special characteristic that each member path contains at least one pair of connected nodes that is not contained in any other path within the subset. This subset will be referred to as ‘basis paths’ (**BP**).

In order to gain a better understanding of the difference between the paths in **BP** and those in the complement of **BP** (*i.e.*, **P** – **BP**), consider the graph presented in Figure 5. The set of all paths (**P**) for the graph in Figure 5 is:

Path 1 - 1, 2, 4, 5, 7
Path 2 - 1, 2, 4, 6, 7
Path 3 - 1, 3, 4, 5, 7
Path 4 - 1, 3, 4, 6, 7

Any one of the four paths could be eliminated and the remaining three would include each pair of connected nodes, and therefore any three constitutes a set of basis paths (**BP**). For example if Path 1 were eliminated, each of the node pairings would still be contained in Paths 2, 3, and 4. However, if both Paths 1 and 2 were eliminated, then node pairings 1 - 2 and 2 - 4 would be excluded. Thus, the set of two paths would be insufficient to span all of the independent sequences of nodes in the graph.

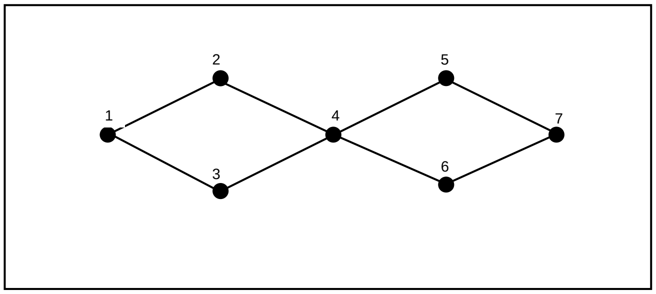


Figure 5 Representation of paths and basis paths

As illustrated above in Figure 1, many questionnaires encountered in practice have so many routes that testing all routes is not practical. Further, a typical route within an instrument has one or more similar routes which involve the same set of nodes, and these routes may be so similar that they differ by only a single, parallel link. Therefore, testing all routes would be not only impractical due to the large number of routes, but also redundant due to the similarity of many routes. The task for a test designer is to select a subset of routes that maximizes coverage and minimizes redundancy. This may be accomplished by using **BP** as a first step in sampling from the universe of routes. The utilization of **BP** in this manner is equivalent to beginning the sampling process with a purposive sample (Cochran 1977, page 10). Another way to think of this first step is as a redefinition of the universe of elements for the purpose of eliminating redundancy. This universe is comprehensive in its coverage, and it contains the smallest set of cases necessary to include all connected node pairs. A second stage of sampling could then be to select one or more routes from each of the paths contained in **BP**. This could be accomplished in several ways. One way would be to consider each path as a cluster of test cases and then take a probability sample from each cluster. Another way would be to select one route from each cluster by randomly selecting one parallel link at each node.

If one accepts the notion of basis testing, then it must be determined how much of the basis should be tested. If all paths in **BP** are tested, then the only elements of an instrument excluded from testing are redundant links. While redundant links may contain spelling or formatting errors, they are unlikely to contain routing errors. This stems from the nature of the programming task involved in creating CAI instruments. Response alternatives are typically ‘bundled’ in the sense that alternatives which lead to the same next question are likely to be either all misdirected or none misdirected. For this reason, comprehensive testing of a basis is an effective method for minimizing errors of a type most likely to lead to loss of data.

On the other hand, non-comprehensive testing may be the only reasonable strategy if constraints due to time or level of effort exist and the number of paths in a basis is large. Despite the fact that any part of an instrument not tested may contain an error, any fraction of the paths in a basis may constitute an unbiased test. Thus, the percentage of paths to be included in a test should probably depend on factors specific to a particular development situation. For example, an instrument may contain modules which have been used previously or modules that have had only minor modification since previous use. These modules need not be tested as thoroughly as newer ones. As a general rule, a minimum sample of test cases should include each distinct section of an instrument in one or more paths, and paths should be included to cover all inter-sectional connections.

7. Discussion and conclusions

A graph-theoretic approach to software development has two major advantages over conventional approaches. First, it allows for a documentation system that can simulate the behavior of a computer-assisted interview. This is useful in verification of routing and as an aid to testers in side-by-side comparisons of instrument behavior versus design specifications. The second major advantage is in selecting cases for testing. The use of the basis of a questionnaire allows for the specification of a universe of test cases which covers all node pairings with a minimum number of paths. Probability sampling from this universe insures that no bias is incorporated into the testing procedures.

In practice, the first advantage can be achieved by structuring the database behind a specifications system such that it contains a table for nodes and a table for links. If the links table specifies a predecessor and a successor node, then queries of the tables will provide the functionality for verification of routing and simulation. The second advantage can be achieved with an algorithm for the identification of a basis. As pointed out by Poole (1995), one of the most important things to do when setting out to test software is to determine which test cases to use. He presented an algorithm for doing this that is based on the flowgraph of a program. Using a flowgraph for this purpose is useful as long as the program is not too large. With large and complicated programs, flow diagrams become unwieldy. The same is true of large and complicated questionnaires (Bethlehem and Hundepool 2004). The appendix contains output from an algorithm which generates a basis, counts routes, and specifies basis paths for an example questionnaire graph (the algorithm used to generate the output appearing in the appendix is available from the author (sdelliott2@verizon.net). This algorithm does not handle looping structures as would be inherent in edits or 'go back' features. These structures may be tested as separate from the questionnaire graph. An algorithm which handles looping is under development).

A graph-theoretic approach is valuable also in that it allows for the use of a number descriptive measures of questionnaires such as the number of routes, the number of paths, cyclomatic complexity (cyclomatic complexity is a measure of complexity in software code (see Hetzel 1984; McCabe 1976; and Watson and McCabe 1996). It is equal also to the number of paths in the basis of a graph. For directed graphs where parallel links are not permitted, cyclomatic complexity $(\mathbf{CC}) = \mathbf{L} - \mathbf{N} + 2,$ where $\mathbf{L}$ is the number of links and $\mathbf{N}$ is the number of nodes), and several types of descriptive matrices (see appendix). Future enhancements to a graph-theoretic approach will likely involve such things as: 1) taxonomies for components, links, and errors; 2) secondary tables in the specification database containing attributes specific to different types of nodes and links; 3) sophisticated sampling plans for selecting test cases; and 4) purposive route sampling.

Taxonomies will promote the specification of special types of instrument components and the incorporation of secondary tables in the documentation system. An example of a special type of instrument component is one with a randomization feature. Such a component would be used in multi-phase respondent selection where a respondent reporting a particular disease, for example, has an increased probability of being routed to a follow-up section pertaining to that disease. In this case, the initial question pertaining to the disease may be a special type called 'respondent selection'. A secondary table in the documentation system for 'respondent selection' questions may have attributes pertaining to a random number generator such as generator seed and selection threshold.

Enhancements to sampling may include stratified sampling (Cochran 1977, page 89) and sampling with probability proportional to size (*i.e.*, PPS). Stratified sampling could be used to insure that all sections within a questionnaire are included with certainty. Paths would be stratified according to the sections they traverse. With PPS sampling, size might be a measure of path length, and the probability of selection for a particular path would be dependent on the number of nodes included in the path. Thus, longer paths could be included with greater frequency. Purposive route sampling may be utilized for testing instrument characteristics other than programming errors. For example, later phases of questionnaire development might target specific sequences of questions for tests of the cognitive characteristics of an instrument.

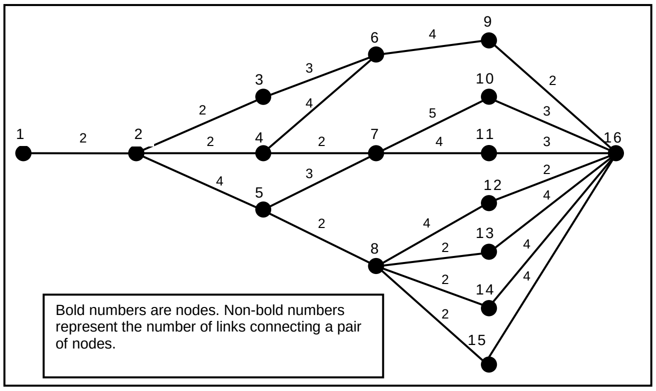
Other researchers in this area likely will provide further enhancements to the application of graph theory to questionnaire development. It does seem clear that graph theory lends itself well to the description, development, and testing of complex CAI instruments. The current trends in CAI usage seem to be in the direction of more sophisticated and larger instruments. For this reason, tools which help to document instrument components and identify errors are valuable to development efforts.

Acknowledgements

A significant portion of the work that went into this paper was supported by Westat, Incorporated Rockville Maryland. Substantial improvements and contributions to this paper were made by the reviewers and editors of *Survey Methodology*.

Appendix

Example of Basis Generation



Bold numbers are nodes. Non-bold numbers represent the number of links connecting a pair of nodes.

Links (*i.e.*, excluding redundant links) = 23
Nodes = 16

Figure 6 Questionnaire graph

Table 1
Branches count for each node

Node Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Number of Branches	1	3	1	2	2	1	2	4	1	1	1	1	1	1	1	0

Table 2
Link matrix

Node	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
1		2														
2			2	2	4											
3						3										
4						4	2									
5							3	2								
6									4							
7										5	4					
8												4	2	2	2	
9																2
10																3
11																3
12																2
13																4
14																4
15																4
16																

Each cell contains a value for the number of links between the row and column nodes.

Table 3
Path matrix

	1st node	**2nd node**	**3rd node**	**4th node**	**5th node**	**6th node**	**7th node**	**8th node**	**9th node**	**10th node**
Path 1	1	2	3	6	9	16				
Path 2	1	2	4	6	9	16				
Path 3	1	2	5	7	10	16				
Path 4	1	2	4	7	10	16				
Path 5	1	2	5	8	12	16				
Path 6	1	2	5	7	11	16				
Path 7	1	2	4	7	11	16				
Path 8	1	2	5	8	13	16				
Path 9	1	2	5	8	14	16				
Path 10	1	2	5	8	15	16				

Cell values represent nodes. Each row represents a path.
[*Note*: *The paths in this example all have 6 nodes. However in general, all paths will not have the same number of nodes.*]

Table 4
Link counts and number of routes for each path

	Node Pairings					**Routes**
	1st to 2nd	**2nd to 3rd**	**3rd to 4th**	**4th to 5th**	**5th to 6th**	
Path 1	2	2	3	4	2	96
Path 2	2	2	4	4	2	128
Path 3	2	4	3	5	3	360
Path 4	2	2	2	5	3	120
Path 5	2	4	2	4	2	128
Path 6	2	4	3	4	3	288
Path 7	2	2	2	4	3	96
Path 8	2	4	2	2	4	128
Path 9	2	4	2	2	4	128
Path 10	2	4	2	2	4	128

Paths = 10 Total Routes = 1,600
Cells represent the number of links between successive nodes in a path.

Table 5
Basis path matrix

	1st node	2nd node	3rd node	4th node	5th node	6th node	7th node	8th node	9th node	10th node
Basis Path 1	1	2	3	6	9	16				
Basis Path 2	1	2	4	6	9	16				
Basis Path 3	1	2	5	7	10	16				
Basis Path 4	1	2	4	7	10	16				
Basis Path 5	1	2	5	8	12	16				
Basis Path 6	1	2	5	7	11	16				
Basis Path 7	1	2	5	8	13	16				
Basis Path 8	1	2	5	8	14	16				
Basis Path 9	1	2	5	8	15	16				

Cell values represent nodes. Each row represents a basis path.

Table 6
Link counts and number of routes for each basis path

	Node Pairings					Routes
	1st to 2nd	2nd to 3rd	3rd to 4th	4th to 5th	5th to 6th	
Basis Path 1	2	2	3	4	2	96
Basis Path 2	2	2	4	4	2	128
Basis Path 3	2	4	3	5	3	360
Basis Path 4	2	2	2	5	3	120
Basis Path 5	2	4	2	4	2	128
Basis Path 6	2	4	3	4	3	288
Basis Path 7	2	4	2	2	4	128
Basis Path 8	2	4	2	2	4	128
Basis Path 9	2	4	2	2	4	128

Basis Paths = 9 Total Routes in Basis = 1,504
Cells represent the number of links between successive nodes from the Basis Paths Matrix above.

References

Balakrishnan, V.K. (1997). *Graph theory*. New York: McGraw Hill, Inc.

Beizer. B. (1995). *Black-box testing*. New York: John Wiley & Sons, Inc.

Berge, C. (1976). *Graphs and hypergraphs*. New York: North-Holland Publishing Company - Amsterdam, London and American Elsevier Publishing Company, Inc.

Bethlehem, J., and Hundepool, A. (2004). TADEQ: A tool for the documentation and analysis of electronic questionnaires. *Journal of Official Statistics*, 20, 233-264.

Centers for Medicare and Medicaid Services (2010). *Medicare Current Beneficiary Survey*: *Overview*. August 2002. Internet address: https://www.cms.gov/LimitedDataSets/11_MCBS.asp.

Chartrand, G. (1985). *Introductory Graph Theory*. New York: Dover Publications, Inc., Mineola.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York: John Wiley & Sons, Inc.

Cohen, J. (1997). *Design and methods of the Medical Expenditure Panel Survey Household Component*. Rockville (MD): Agency for Health Care Policy and Research. MEPS Methodology Report No. 1. AHCPR Pub. No. 97-0026.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. and O'Reilly, J.M. (1988). *Computer Assisted Information Collection*. New York: John Wiley & Sons, Inc.

Gibbons, A. (1985). *Algorithmic Graph Theory*. Cambridge University Press.

Harary, F., and Palmer, E. (1973). *Graphical Enumeration*. New York: Academic Press.

Hetzel, W. (1983). *The Complete Guide to Software Testing, QED*. Massachusetts: Information Sciences, Inc., Wellesley.

Kaner, C., Falk, J. and Nguyen, H. (1999). *Testing Computer Software*, 2nd Edition. New York: John Wiley & Sons, Inc.

Medical Expenditure Panel Survey (MEPS) (2010). *Survey Instruments and Associated Documentation*. http://www.meps.ahrq.gov/mepsweb/.

Myers, G.J. (1979). *The Art of Software Testing*. New York: John Wiley & Sons, Inc.

Patton, R. (2006). *Software Testing*, 2nd Edition. Sams Publishing, Inc.

Poole, J. (1995). NISTIR 5737 – A Method to Determine a Basis Set of Paths to Perform Program Testing, National Institute of Standards and Technology, Gaithersburg, MD, November, 1995.

Statistics Canada (2010). *Canadian Financial Capability Survey* (*CFCS*): *Questionnaire 2009*. http://www.statcan.gc.ca.

Statistics Netherlands (2002). Blaise Developer's Guide. Department of Statistical Informatics, Statistics Netherlands, Heerlen.

US Bureau of Labor Statistics (2010). *Consumer Expenditure Surveys Quarterly Interview CAPI Survey 2010.* United States Department of Labor, http://www.bls.gov/cex/capi/2010/cecapihome.htm.

Watson, A., and McCabe, T. (1996). NIST Special Publication 500-235 Structured Testing: A Testing Methodology Using the Cyclomatic Complexity Metric. National Institute of Standards and Technology, Gaithersburg, MD, September, 1996.

On sample allocation for efficient domain estimation

G. Hussain Choudhry, J.N.K. Rao and Michael A. Hidiroglou [1]

Abstract

Sample allocation issues are studied in the context of estimating sub-population (stratum or domain) means as well as the aggregate population mean under stratified simple random sampling. A non-linear programming method is used to obtain "optimal" sample allocation to strata that minimizes the total sample size subject to specified tolerances on the coefficient of variation of the estimators of strata means and the population mean. The resulting total sample size is then used to determine sample allocations for the methods of Costa, Satorra and Ventura (2004) based on compromise allocation and Longford (2006) based on specified "inferential priorities". In addition, we study sample allocation to strata when reliability requirements for domains, cutting across strata, are also specified. Performance of the three methods is studied using data from Statistics Canada's Monthly Retail Trade Survey (MRTS) of single establishments.

Key Words: Composite estimators; Compromise allocation; Direct estimators; Domain means; Non-linear programming.

1. Introduction

Stratified simple random sampling is widely used in business surveys and other establishment surveys employing list frames. The population mean $\bar{Y} = \sum_h W_h \bar{Y}_h$ is estimated by the weighted sample mean $\bar{y}_{st} = \sum_h W_h \bar{y}_h$, where $W_h = N_h / N$ is the relative size of stratum h (= 1,..., L) and $\bar{Y}_h$ and $\bar{y}_h$ are the stratum population mean and sample mean respectively. The well-known Neyman sample allocation to strata is optimal for estimating the population mean in the sense of minimizing the variance of $\bar{y}_{st}$ subject to $\sum_h n_h = n$ where n is fixed or minimizing $\sum_h n_h$ subject to fixed variance of $\bar{y}_{st}$, where n_h is the stratum sample size. But the Neyman allocation may cause some strata to have large coefficients of variation (CV) of the means $\bar{y}_h$. On the other hand, equal sample allocation, $n_h = n/L$, is efficient for estimating strata means, but it may lead to a much larger CV of the estimator $\bar{y}_{st}$ compared to that of Neyman allocation.

Bankier (1988) proposed a "power allocation" as a compromise between Neyman allocation and equal allocation. Letting $C_h = S_h / \bar{Y}_h$ be the stratum CV, the power allocation is

$$n_h^B = n \frac{C_h X_h^q}{\sum_h C_h X_h^q}, \; h = 1, ..., L \qquad (1.1)$$

where X_h is some measure of size or importance of stratum h and q is a tuning constant. Power allocation (1.1) is obtained by minimizing $\sum_h \{X_h^q \, \mathrm{CV}(\bar{y}_h)\}^2$ subject to $\sum_h n_h = n$, where $\mathrm{CV}(\bar{y}_h)$ is the CV of the stratum sample mean $\bar{y}_h$. The choice $q = 1$ and $X_h = N_h \bar{Y}_h$ in (1.1) leads to Neyman allocation

$$n_h^N = n \frac{N_h S_h}{\sum_h N_h S_h}, \; h = 1, ..., L \qquad (1.2)$$

and $q = 0$ gives equal allocation if $C_h = C$ for all h, where S_h^2 is the stratum variance. Bankier (1988) viewed values of q between 0 and 1 as providing compromise allocations. He gave a numerical example to illustrate how q may be chosen in practice. The choice $X_h = N_h$ and $q = 1/2$ in (1.1) gives "square root allocation" $n_h = n\sqrt{N_h} / \sum_h \sqrt{N_h}$ if $C_h = C$. Power allocation (1.1) and some other allocations generally depend on the variable of interest y and hence in practice a proxy variable with known population values is used in place of y.

Costa *et al.* (2004) proposed a compromise allocation based on a convex combination of proportional allocation, $n_h = nW_h$, and equal allocation $n_h = n/L$, see section 2.1. Longford (2006) made a systematic study of allocation in stratified simple random sampling by introducing "inferential priorities" P_h for the strata h and G for the population. In particular, he assumed that $P_h = N_h^q$ for a specified q $(0 \le q \le 2)$, see section 2.4. He also studied the case of small strata sample sizes n_h in which case composite estimators of strata means $\bar{Y}_h$ may be used.

The main purpose of our paper is to propose an "optimal" allocation method, based on non-linear programming (NLP), see section 2.3. It minimizes the total sample size $\sum_h n_h$ subject to specified tolerances on the CVs of the strata sample means $\bar{y}_h$ and the estimated population mean $\bar{y}_{st}$. The case of indirect (composite) estimators of strata means is studied in Section 3. In Section 4, we study optimal sample allocation to strata when reliability requirements for domains, cutting across strata, are also specified.

[1] G.H. Choudhry, Statistical Research and Innovation Division, Statistics Canada. E-mail: ghchoudhry@gmail.com; J.N.K. Rao, School of Mathematics and Statistics, Carleton University. E-mail: jrao@math.carleton.ca; M.A. Hidiroglou, Statistical Research and Innovation Division, Statistics Canada. E-mail: mike.hidiroglou@statcan.gc.ca.

The proposed method readily extends to multiple variables, but for simplicity we omit details. Using the optimal total sample size obtained from NLP, we make a numerical study of the performances of Costa *et al.* and Longford methods in terms of satisfying reliability requirements, Section 5.

2. Allocation for direct estimators

In this section, we consider direct estimators, $\bar{y}_h$, of strata population means, assuming stratified simple random sampling. The case of indirect estimators of strata means is studied in Section 3. Indirect strata estimators are used in the case of strata with small sample sizes n_h.

2.1 Costa *et al.* allocation

The sample allocation of Costa *et al.* (2004) is

$$n_h^C = k(nW_h) + (1-k)(n/L) \qquad (2.1)$$

for a specified constant $k(0 \leq k \leq 1)$. This allocation reduces to equal allocation when $k = 0$ and to proportional allocation when $k = 1$. Formula (2.1) needs to be modified when $n / L > N_h$ for some h in a set of strata A. The modified allocation is

$$\tilde{n}_h^C = k(nW_h) + (1-k)\, n_h^0, \qquad (2.2)$$

where $n_h^0 = N_h$ if $h \in A$ and $n_h^0 = (n - \sum_{h \in A} N_h) / (L - m)$ otherwise, where m is the number of strata in the set A. Note that when $k = 0$, (2.2) gives modified equal allocation. We study different choices of the constant k in the numerical study of Section 5, based on data from Statistics Canada's Monthly Retail Trade Survey (MRTS).

2.2 Longford allocation

Longford's (2006) method attempts to simultaneously control the reliability of the strata means $\bar{y}_h$ and the estimated population mean $\bar{y}_{st}$ by minimizing the objective function

$$\sum_{h=1}^{L} P_h V(\bar{y}_h) + (GP_+) V(\bar{y}_{st}) \qquad (2.3)$$

with respect to the strata sample sizes n_h subject to $\sum_h n_h = n$, where $P_+ = \sum_h P_h$. The first component in (2.3) specifies relative importance, P_h, of each stratum h while the second component attaches relative importance to $\bar{y}_{st}$ through the weight G. Longford (2006) assumed that $P_h = N_h^q$ for some constant $q(0 \leq q \leq 2)$. The term P_+ in (2.3) offsets the effect of the sizes P_h and the number of strata on the weight G.

Under stratified simple random sampling, the sample allocation minimizing (2.3) is

$$n_h^L = n \frac{S_h \sqrt{P_h'}}{\sum_h S_h \sqrt{P_h'}}, \; h = 1, \ldots, L \qquad (2.4)$$

where $P_h' = P_h + GP_+ W_h^2$. If $q = 2$, then (2.4) does not depend on the value of G and it reduces to Neyman allocation, n_h^N, given by (1.2)

2.3 Nonlinear programming (NLP) allocation

We now turn to the NLP method of determining the strata sample sizes n_h subject to specified reliability requirements on both the strata sample means and the estimated population mean. Letting $\mathbf{f} = (f_1, \ldots, f_L)^T$ with $f_h = n_h / N_h$, we minimize the total sample size

$$g(\mathbf{f}) = \sum_{h=1}^{L} f_h N_h \qquad (2.5)$$

with respect to f subject to

$$\text{CV}(\bar{y}_h) \leq \text{CV}_{0h}, \; h = 1, \ldots, L \qquad (2.6)$$

$$\text{CV}(\bar{y}_{st}) \leq \text{CV}_o \qquad (2.7)$$

$$0 < f_h \leq 1, \; h = 1, \ldots, L \qquad (2.8)$$

where CV_{0h} and CV_0 are specified tolerances on the CV of the stratum sample mean $\bar{y}_h$ and the estimated population mean $\bar{y}_{st}$, respectively. Inequality signs are used in (2.6) and (2.7) because the resulting CVs for some strata h and/or for the aggregate may be smaller than the specified tolerances (Cochran 1977, page 122).

Letting $k_h = f_h^{-1}$, (2.5) becomes a separable convex function of the variables k_h,

$$\tilde{g}(\mathbf{k}) = \sum_{h=1}^{L} N_h k_h^{-1}. \qquad (2.9)$$

We re-specify the constraints (2.6) and (2.7) in terms of relative variances so that the constraints are linear in the variables k_h. The relative variance (RV) of $\bar{y}_h$ is the square of its CV,

$$\text{RV}(\bar{y}_h) = \frac{k_h - 1}{N_h} C_h^2. \qquad (2.10)$$

Similarly, the relative variance of $\bar{y}_{st}$ is the square of its CV,

$$\text{RV}(\bar{y}_{st}) = \bar{Y}^{-2} \sum_{h=1}^{L} W_h^2 \frac{k_h - 1}{N_h} S_h^2. \qquad (2.11)$$

We used the SAS procedure NLP with the Newton-Raphson option to find the optimal k_h that would minimize (2.9) subject to

$$\text{RV}(\bar{y}_h) \leq \text{RV}_{oh}, \; h = 1, \ldots, L, \qquad (2.12)$$

$$\mathrm{RV}(\bar{y}_{st}) \le \mathrm{RV}_0, \qquad (2.13)$$

$$k_h \ge 1,\ h = 1, \ldots, L. \qquad (2.14)$$

$\mathrm{RV}(\bar{y}_h)$ and $\mathrm{RV}(\bar{y}_{st})$ are given by (2.10) and (2.11) where $\mathrm{RV}_{0h} = \mathrm{CV}_{0h}^2$ and $\mathrm{RV}_0 = \mathrm{CV}_0^2$. By expressing the constraints as linear constraints and the objective function as a separable convex function, we achieve faster convergence of the re-formulated NLP. Denoting the solution to NLP as $\mathbf{k}^0 = (k_1^0, \ldots, k_L^0)^T$, the corresponding vector of optimal strata sample sizes is given by $\mathbf{n}^0 = (n_1^0, \ldots, n_L^0)^T$, where $n_h^0 = N_h / k_h^0$. We can modify (2.14) to ensure that $n_h^0 \ge 2$ for all h which permits unbiased variance estimation.

The NLP method can be readily extended to multiple variables $y_1, \ldots, y_P$ by specifying tolerances on the CVs of strata means and the estimated population mean for each variable $(p = 1, \ldots, P)$. If the number of variables P is not small, then the resulting optimal total sample size $n^0 = \sum_h n_h^0$ may increase significantly relative to n^0 for a single variable. Huddleston, Claypool and Hocking (1970), Bethel (1989) and others studied NLP for optimal sample allocation in the case of estimating population means of multiple variables under stratified random sampling.

3. Allocation for composite estimators

Longford (2006) studied composite estimators of strata means of the form

$$\hat{\theta}_h = \alpha_h \bar{y}_h^S + (1 - \alpha_h) \bar{y}_h \qquad (3.1)$$

where $\bar{y}_h^S$ is a synthetic estimator; here we take $\bar{y}_h^S = \bar{y}_{st}$. The MSE of $\hat{\theta}_h$ is

$$\begin{aligned}\mathrm{MSE}(\hat{\theta}_h) &= V(\hat{\theta}_h) + [B(\hat{\theta}_h)]^2 \\ &= \alpha_h^2 \sum_{h=1}^{L} W_h^2 \frac{S_h^2}{n_h} + (1-\alpha_h)^2 W_h^2 \frac{S_h^2}{n_h} \\ &\quad + 2\alpha_h (1-\alpha_h) W_h \frac{S_h^2}{n_h} + \alpha_h^2 (\bar{Y}_h - \bar{Y})^2 \\ &\quad + \text{terms not depending on the } n_h. \qquad (3.2)\end{aligned}$$

Longford (2006) showed that the optimal coefficient α_h in (3.1) minimizing (3.2) is approximately equal to $\alpha_h^* = S_h^2 (S_h^2 + n_h \Delta_h^2)^{-1}$, where $\Delta_h = \bar{Y}_h - \bar{Y}$. He then replaced Δ_h^2 in α_h^* by its average over the strata, denoted by $\sigma_B^2 = L^{-1} \sum_h (\bar{Y}_h - \bar{Y})^2$, leading to $\alpha_h^* \approx (1 + n_h \omega_h)^2$, where $\omega_h^2 = \sigma_B^2 / S_h^2$. The resulting MSE of $\hat{\theta}_h$ is approximated as

$$\mathrm{MSE}(\hat{\theta}_h) \approx \frac{\sigma_B^2}{1 + n_h \omega_h}. \qquad (3.3)$$

Longford's allocation is obtained by minimizing the objective function

$$\sum_{h=1}^{L} P_h \,\mathrm{MSE}(\hat{\theta}_h) + (\mathrm{GP}_+) V(\bar{y}_h) \qquad (3.4)$$

with respect to the n_h. The resulting solution satisfies

$$\frac{P_h \sigma_B^2 \omega_h}{(1 + n_h \omega_h)^2} + (\mathrm{GP}_+) W_h^2 \frac{S_h^2}{n_h} = \text{const.},\ h = 1, \ldots, L. \qquad (3.5)$$

Longford used an iterative method to obtain the solution to (3.5) since it does not have a closed-form solution.

Our NLP procedure minimizes $g(\mathbf{f})$ given by (2.5) subject to

$$\mathrm{RMSE}(\hat{\theta}_h) \le \mathrm{RMSE}_{0h},\ h = 1, \ldots, L;\ \mathrm{RV}(\bar{y}_{st}) \le \mathrm{RV}_o \qquad (3.6)$$

and (2.8), where $\mathrm{RMSE}(\hat{\theta}_h) = \mathrm{MSE}(\hat{\theta}_h) / \bar{Y}_h^2$ and RMSE_{0h} is a specified tolerance. The approximation (3) to $\mathrm{MSE}(\hat{\theta}_h)$ is used in (3.6).

4. Allocation for domain estimation

Suppose that the population U is partitioned into domains ${}_dU$ $(d = 1, \ldots, D)$ that cut across the strata. Also, suppose that the estimators of domain means need to satisfy specified relative variance tolerances, ${}_d\mathrm{RV}_0$, $d = 1, \ldots, D$. We find the optimal additional strata sample sizes that are needed to satisfy the domain tolerances, using the NLP method.

An estimator of domain mean ${}_d\bar{Y} = {}_dN^{-1} \sum_{k \in U_d} y_k$ is the ratio estimator

$${}_d\hat{\bar{Y}} = \frac{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k \in s_h} {}_d\delta_k\, y_k}{\sum_{h=1}^{L} N_h n_h^{-1} \sum_{k \in s_h} {}_d\delta_k}, \qquad (4.1)$$

where ${}_d\delta_k = 1$ if $k \in {}_dU$ and ${}_d\delta_k = 0$ otherwise, s_h is the sample from stratum h and ${}_dN$ is the size of domain d. The relative variance of the ratio estimator (4.1) is $\mathrm{RV}({}_d\hat{\bar{Y}}) = V({}_d\hat{\bar{Y}}) / {}_d\bar{Y}^2$, where the variance $V({}_d\hat{\bar{Y}})$ is obtained by the usual linearization formula for a ratio estimator.

Let $\tilde{n}_h$ denote the revised total sample size from stratum h so that the sample increase from stratum h is $\tilde{n}_h - n_h^0$. Let $\tilde{f}_h = \tilde{n}_h / N_h$ be the corresponding sampling fraction. We obtain the optimal $\tilde{\mathbf{n}} = (\tilde{n}_1, \ldots, \tilde{n}_L)^T$ by minimizing the sample increase

$$g(\tilde{\mathbf{f}}) - \sum_{h=1}^{L} n_h^0 N_h = \sum_{h=1}^{L} (\tilde{f}_h - f_h^0) N_h \qquad (4.2)$$

with respect to $\tilde{\mathbf{f}} = (\tilde{f}_1, \ldots, \tilde{f}_L)^T$ subject to

$$f_h^0 \le \tilde{f}_h \le 1, \;\; h = 1, \ldots, L \qquad (4.3)$$

$$\mathrm{RV}({}_d\hat{\bar{Y}}) \le {}_d\mathrm{RV}_0, \;\; d = 1, \ldots, D. \qquad (4.4)$$

As before, we reformulate the problem by expressing (4.2), (4.3) and (4.4) in terms of $\tilde{\mathbf{k}} = (\tilde{k}_1, \ldots, \tilde{k}_L)^T$, where $\tilde{k}_h = \tilde{f}_h^{-1}$. This leads to minimization of the separable convex function

$$g^*(\tilde{\mathbf{k}}) = \sum_{h=1}^{L} N_h \tilde{k}_h^{-1} \qquad (4.5)$$

with respect to $\tilde{\mathbf{k}}$ subject to the linear constraints

$$1 \le \tilde{k}_h \le k_h^0, \;\; h = 1, \ldots, L \qquad (4.6)$$

and

$$\mathrm{RV}({}_d\hat{\bar{Y}}) = {}_d\bar{Y}^{-2} \sum_{h=1}^{L} \left(\frac{N_h}{{}_dN} \right)^2 \frac{\tilde{k}_h - 1}{N_h} \, {}_dS_{eh}^2 \le {}_d\mathrm{RV}_0, \;\; d = 1, \ldots, D \qquad (4.7)$$

where ${}_d\mathrm{RV}_0$ is the specified tolerance, ${}_dS_{eh}^2$ denotes the stratum variance of the residuals ${}_de_k = {}_d\delta_k (y_k - {}_d\bar{Y})$ for $k \in U_h$ and U_h denotes the stratum population. Denote the resulting optimal $\tilde{k}_h$ and $\tilde{n}_h$ as $\tilde{k}_h^0$ and $\tilde{n}_h^0$ respectively, so that the optimal sample increase in stratum h is $\tilde{n}_h^0 - n_h^0$.

It can be shown that the minimization of total sample size subject to all the constraints $\mathrm{RV}(\bar{y}_h) \le \mathrm{RV}_{0h}$, $h = 1, \ldots, L$, $\mathrm{RV}({}_d\hat{\bar{Y}}) \le {}_d\mathrm{RV}_0$, $d = 1, \ldots, D$, $\mathrm{RV}(\bar{y}_{st}) \le \mathrm{RV}_0$, and $0 < f_h \le 1$, $h = 1, \ldots, L$ will lead to the same optimal solution, $\tilde{n}^0 = (\tilde{n}_1^0, \ldots, \tilde{n}_L^0)^T$. However, domain reliability requirements may often be specified after determining n^0.

5. Empirical results

In this section, we study the relative performance of different sample allocation methods, using data from the MRTS. Section 5.1 and 5.2 report our results for direct estimators and composite estimators of strata means, respectively. Results for the domain means are given in section 5.3.

5.1 Strata means: Direct estimators

For the empirical study, we used a subset of the MRTS population values restricted to single establishments. Strata sizes, N_h, strata population means, $\bar{Y}_h$, strata standard deviations, S_h, and strata CVs, $C_h = S_h / \bar{Y}_h$, are given in Table 1 for the ten provinces in Canada (treated as strata). For the NLP allocation, we have taken the CV tolerances as $\mathrm{CV}_{0h} = 15\%$ for the strata means $\bar{y}_h$ and $\mathrm{CV}_0 = 6\%$ for the weighted sample mean $\bar{y}_{st}$, denoted Canada (CA).

The NLP allocation satisfying the specified CV tolerances resulted in a minimum overall sample size $n^0 =$ 3,446. Table 2 reports the sample allocation n_h^0 and the associated $\mathrm{CV}(\bar{y}_h)$ and $\mathrm{CV}(\bar{y}_{st})$ for the NLP allocation. It shows that the NLP allocation respects the specified tolerance $\mathrm{CV}_0 = 6\%$, gives CVs smaller than the specified tolerance $\mathrm{CV}_{0h} = 15\%$ for two of the larger provinces (QC: 11.4% and ON: 11.0%) and attains a 15% CV for the remaining provinces.

Table 1
Population values for the MRTS

Provinces	N_h	$\bar{Y}_h$	S_h	C_h
Newfoundland (NL)	909	963	1,943	2.02
Price-Edward-Island (PE)	280	712	1,375	1.93
New-Brunswick (NB)	1,333	1,368	3,200	2.34
Nova-Scotia (NS)	1,153	1,568	4,302	2.74
Quebec (QC)	11,135	2,006	4,729	2.36
Ontario (ON)	21,531	1,722	6,297	3.66
Manitoba (MN)	1,700	1,295	2,973	2.30
Saskatchewan (SK)	1,743	1,212	3,019	2.49
Alberta (AL)	5,292	1,698	5,358	3.16
British Columbia (BC)	7,803	1,291	4,013	3.11
Canada (CA)	52,879	1,654	-	-

Table 2
Equal, proportional, square root and NLP allocations and associated CVs (%)

Province	Equal		Proportional		Square-Root		NLP	
	n_h	CV_h	n_h	CV_h	n_h	CV_h	n_h	CV_h
NL	352	8.4	59	25.4	169	14.0	151	15.0
PE	280	0.0	18	44.1	94	16.2	104	15.0
NB	352	10.7	87	24.2	205	15.0	206	15.0
NS	352	12.2	75	30.6	191	18.1	259	15.0
QC	352	12.4	726	8.5	593	9.4	410	11.4
ON	352	19.3	1,403	9.4	824	12.5	1,056	11.0
MN	352	10.9	111	21.1	232	14.0	206	15.0
SK	352	11.9	114	22.6	234	15.2	238	15.0
AL	352	16.3	345	16.4	408	15.0	409	15.0
BC	352	16.2	508	13.3	496	13.5	407	15.0
CA	3,446	9.1	3,446	5.2	3,446	6.3	3,446	6.0

Using the optimal overall sample size 3,446, we calculated the sample allocations n_h and the associated $\mathrm{CV}(\bar{y}_h)$ and $\mathrm{CV}(\bar{y}_{st})$ for the modified equal allocation, proportional allocation and square-root allocation, reported in Table 2. It is clear from Table 2 that the modified equal allocation is not suitable in terms of satisfying specified CV tolerances because it leads to $\mathrm{CV}(\bar{y}_{st}) = 9.1\%$ which is significantly larger than the specified $\mathrm{CV}_0 = 6\%$. Also, under the modified equal allocation, $\mathrm{CV}(\bar{y}_h)$ equals 19.3%, 16.3% and 16.2% for the larger provinces ON, AL and BC respecttively. Note that for the smallest province PE Table 2 gives $\mathrm{CV}(\bar{y}_h) = 0\%$ for the modified equal allocation because for PE it gives $n_h = N_h$.

Turning to proportional allocation, Table 2 reports $\text{CV}(\bar{y}_{st}) = 5.2\%$ but it leads to considerably larger strata CVs relative to the specified 15% for seven of the provinces, ranging from 16.4% to 44.1%. On the other hand, Table 2 shows that square-root allocation offers a reasonable compromise in terms of desired CV tolerances. We have $\text{CV}(\bar{y}_{st}) = 6.3\%$ and $\text{CV}(\bar{y}_h) \le 15\%$ for seven of the provinces and the three provinces with CVs greater than 15% are SK with 15.2%, PE with 16.2% and NS with 18.1%.

Table 3 reports the results for the Costa *et al.* allocation (2.1) with $k = 0.25$, 0.50 and 0.75, using $n = 3{,}446$ obtained from NLP. We observe from Table 2 that the choice $k = 0.25$, which assigns more weight to equal allocation, is not satisfactory for the estimation of the population (Canada) mean: $\text{CV}(\bar{y}_{st}) = 7.2\%$, but performs well for strata means, except AL with $\text{CV}(\bar{y}_h) = 16.3\%$. On the other hand, the choice $k = 0.75$, which assigns more weight to proportional allocation, performs poorly in estimating provincial means, with $\text{CV}(\bar{y}_h)$ ranging from 16.2% to 21.4% for seven of the provinces, although $\text{CV}(\bar{y}_{st})$ is smaller than the desired tolerance, 6%. The compromise choice $k = 0.50$ performs quite well, leading to $\text{CV}(\bar{y}_{st}) =$ 6.4% and $\text{CV}(\bar{y}_h)$ around 15% or less except for two provinces (NS and AL) with CVs of 17.0% and 16.5% respectively. Performance of the Costa *et al.* method with $k = 0.50$ and square-root allocation are somewhat similar, and both allocations do not depend on the variable of interest, y, unlike the Longford and NLP allocations.

We next turn to Longford's allocation (2.4) which depends on q and G. Table 4 provides results for $q = 0$, 0.5, 1.0, 1.5 and $G = 0, 10, 100$, using $n = 3{,}446$ obtained from NLP. For $q = 2.0$, Longford's allocation does not depend on G and in fact it reduces to the Neyman allocation (1.2) which minimizes $\text{CV}(\bar{y}_{st})$ for fixed n but leads to highly inflated $\text{CV}(\bar{y}_h)$, ranging from 16% to 85% for seven provinces. We see from this table that $\text{CV}(\bar{y}_h)$, for a given q, increases with G rapidly while $\text{CV}(\bar{y}_{st})$ decreases slowly as G increases and in fact is virtually a constant ($\approx$5.1%) for $G > 100$ (values not reported here). Also, $\text{CV}(\bar{y}_h)$ for a given G, increases rapidly as q increases while $\text{CV}(\bar{y}_{st})$ decreases. Langford's allocation, for $q \ge 0.5$ and/or $G \ge 10$, leads to significantly larger $\text{CV}(\bar{y}_h)$ than the specified tolerance $\text{CV}_{0h} = 15\%$ for several provinces, even though $\text{CV}(\bar{y}_{st})$ respects the specified tolerance of 6%. On the other hand, for $q = 0$ and $G = 0$, $\text{CV}(\bar{y}_h)$ is below the specified tolerance except for BC with 15.7%, but $\text{CV}(\bar{y}_{st}) = 7.3\%$ significantly exceeds the specified tolerance. For $q = 1.0$ and $q = 1.5$, $\text{CV}(\bar{y}_{st})$ stays below 6% when $G = 0$, but $\text{CV}(\bar{y}_h)$ exceeds 15% for six provinces, ranging from 17.7% to 34.0% for $q = 1.0$ and 22.0% to 54.6% for $q = 1.5$. On the whole, Table 4 suggests that no suitable combination of q and G can be found that ensures that all the specified reliability requirements are satisfied even approximately.

Table 3
Costa *et al.*'s allocation and associated CVs (%) for *k* = 0.25, 0.50 and 0.75

Province	*k* = 0.25		*k* = 0.50		*k* = 0.75	
	n_d	CV_d	n_d	CV_d	n_d	CV_d
NL	278	10.1	205	12.4	132	16.2
PE	214	6.4	149	10.8	83	17.8
NB	286	12.3	219	14.5	153	17.8
NS	282	14.2	213	17.0	144	21.4
QC	446	10.9	539	9.9	633	9.1
ON	615	14.5	878	12.1	1,140	10.5
MN	292	12.2	231	14.0	171	16.6
SK	292	13.3	733	15.2	174	17.9
AL	350	16.3	349	16.3	347	16.4
BC	391	15.3	430	14.6	469	13.9
CA	3,446	7.2	3,446	6.2	3,446	5.6

Table 4
CVs (%) for Longford's allocation with *q* = 0, 0.5, 1.0, and 1.5

Province	*q* = 0			*q* = 0.5			*q* = 1.0			*q* = 1.5		
	G = 0	*G* = 10	*G* = 100	*G* = 0	*G* = 10	*G* = 100	*G* = 0	*G* = 10	*G* = 100	*G* = 0	*G* = 10	*G* = 100
NL	13.5	19.3	29.7	17.2	23.0	33.4	22.7	29.0	38.3	30.4	36.2	40.6
PE	12.7	20.4	34.6	21.4	29.6	48.5	34.0	45.4	67.3	54.6	67.3	85.6
NB	12.0	17.1	25.0	14.5	19.4	26.8	18.3	23.1	29.0	23.5	27.6	30.3
NS	11.1	16.7	25.5	14.2	19.5	27.9	18.7	24.1	30.9	24.9	29.4	32.8
QC	11.0	9.8	9.1	9.9	9.4	9.0	9.2	9.0	8.9	8.9	8.9	8.8
ON	14.9	9.8	8.7	12.3	9.5	8.6	10.5	9.1	8.5	9.3	8.7	8.5
MN	12.7	17.6	24.3	14.7	19.1	25.2	17.7	21.9	26.5	22.0	25.4	27.5
SK	13.6	18.9	25.9	15.7	20.5	26.9	19.0	23.5	28.3	23.5	27.0	29.4
AL	13.5	15.7	16.1	13.3	15.2	15.9	13.6	15.2	15.9	14.6	15.5	15.9
BC	15.7	16.1	15.4	14.7	15.4	15.3	14.3	15.0	15.1	14.5	15.0	15.1
CA	7.3	5.5	5.1	6.2	5.3	5.1	5.5	5.2	5.1	5.2	5.1	5.1

5.2 Strata means: Composite estimators

We now report some results for the composite estimators, $\hat{\theta}_h$, of strata means. We obtained the optimal total sample size as $n = 3{,}368$ using NLP and the reliability requirements (3.6). This value is slightly smaller than the optimal $n^0 =$ 3,446 for the direct estimators. For the Longford allocation, we used $n = 3{,}368$ and calculated the sample allocation and associated CVs of the composite estimators $\hat{\theta}_h$ and the weighted mean $\bar{y}_{st}$ for specified q and G, constraining n_h to be at least two. For the MRTS data we have used, the first term of (3.5) is small relative to the second term. As a result, the sample allocation is flat across G - values for a given q which means that the CVs for the Longford allocation did not vary significantly with G.

Therefore, we have reported results in Table 5 only for $G = 0$ and $q = 0, 0.5, 1.0$ and 1.5. We note from Table 5 that $\mathrm{CV}(\hat{\theta}_h)$ decreases with q for the two largest provinces (QC and ON) because the sample shifts from the smaller provinces to these two provinces as q increases. Also, $\mathrm{CV}(\hat{\theta}_h)$ initially decreases for AL and BC but it starts increasing when q is large because the sample starts shifting to QC and ON from these provinces as well. Further, $\mathrm{CV}(\hat{\theta}_h)$ increases for all other provinces with q except for NS for which it starts decreasing for large q because of larger synthetic component and very negligible bias. In particular, $\mathrm{CV}(\hat{\theta}_h)$ increases rapidly for NL and PE because of very large bias.

Table 5
CVs (%) for the composite estimators using Longford's allocation: G = 0 and q = 0, 0.5, 1.0 and 1.5

Province	q = 0	q = 0.5	q = 1.0	q = 1.5
NL	12.7	17.0	24.2	37.3
PE	12.4	23.8	46.0	112.2
NB	10.4	12.8	16.1	20.4
NS	9.4	11.9	14.5	11.7
QC	10.3	9.0	8.3	8.0
ON	13.9	11.1	9.3	8.2
MN	11.2	13.1	16.0	20.3
SK	12.4	14.6	17.9	23.2
AL	11.4	11.2	11.5	12.2
BC	14.4	13.3	12.9	13.1
CA	8.0	6.3	5.4	5.6

On the other hand, $\mathrm{CV}(\bar{y}_{st})$ decreases initially with q but starts increasing when q is large because most of the sample gets allocated to QC and ON and very little sample is assigned to the smaller provinces. It appears from Table 5 that the Longford allocation performs reasonably well only for $q = 0$ and $G = 0$, giving $\mathrm{CV}(\hat{\theta}_h)$ less than 15% for all provinces at the expense of $\mathrm{CV}(\bar{y}_{st}) = 8.0\%$.

5.3 Domain means

Establishments on the Canadian Business Register are classified by industry using the North American Industry Classification System (NAICS). NAICS is principally a classification system for establishments and for the compilation of production statistics. The industry associated with each establishment on the Canadian Business Register is coded to six digits using the North American Industry Classification System. There are 67 six digit codes associated with the Retail sector. These six digit codes are regrouped into 19 trade groups (TG) for publication purposes.

We took the trade groups as domains that cut across provinces (strata). The trade group with the smallest number of establishments is TG 110 (beer, wine and liquor stores) with 307 establishments and the TG with the largest number of establishments is TG 100 (convenience and specialty food stores) with 7,752 establishments. Establishments were coded to all the 19 trade groups in all but one province: in PE, establishments were coded to only 16 trade groups.

We applied NLP based on (4.5), (4.6) and (4.7), and obtained the optimal total sample size increase to meet specified reliability requirements on the domain estimators ${}_d\hat{\bar{Y}}$. We found that no increase in the total sample size is needed if the tolerance on $\mathrm{CV}({}_d\hat{\bar{Y}})$ is less than or equal to 30% for each domain. If the tolerance is reduced to 25%, then the optimal total sample size increase is 622 and the total sample size after the increase is 4,068. If the tolerance is further reduced to 20%, then the optimal total sample size increase is 2,100 and the total sample size after the increase is 5,546, which is considerably larger than the original 3,446. Note that as the total sample size is increased, CVs of strata means $\bar{y}_h$ and the weighted sample mean $\bar{y}_{st}$ decrease.

6. Summary and concluding remarks

We have proposed a non-linear programming (NLP) method of sample allocation to strata under stratified random sampling. It minimizes the total sample size subject to specified tolerances on the coefficient of variation of estimators of strata means and the population mean. We considered both direct estimators of strata means and composite estimators of strata means. The case of domains cutting across strata is also studied. Difficulties with alternative methods in satisfying specified reliability requirements are demonstrated using data from the Statistics Canada Monthly Retail Trade Survey of single establishments. We also noted that NLP can be readily extended to handle reliability requirements for multiple variables. Compromise allocations that perform reasonably well in terms of reliability requirements are also noted.

Acknowledgements

The authors thank two referees and an associate editor for constructive comments and suggestions.

References

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, 15, 47-57.

Bankier, M. (1988). Power allocation: Determining sample sizes for sub-national areas. *The American Statistician*, 42, 174-177.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd Edition. New York : John Wiley & Sons, Inc.

Costa, A., Satorra, A. and Ventura, E. (2004). Using composite estimators to improve both domain and total area estimation. *SORT*, 28, 69-86.

Huddleston, H.F., Claypool, P.L. and Hocking, R.R. (1970). Optimum allocation to strata using convex programming. *Applied Statistics*, 19, 273-278.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 87-96.

North American Industry Classification System, Version 1.4 (2008). Catalogue 12F0074XCB. Statistics Canada.

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 31-41
Statistics Canada, Catalogue No. 12-001-X

Calibration alternatives to poststratification for doubly classified data

Ted Chang [1]

Abstract

We consider alternatives to poststratification for doubly classified data in which at least one of the two-way cells is too small to allow the poststratification based upon this double classification. In our study data set, the expected count in the smallest cell is 0.36. One approach is simply to collapse cells. This is likely, however, to destroy the double classification structure. Our alternative approaches allows one to maintain the original double classification of the data. The approaches are based upon the calibration study by Chang and Kott (2008). We choose weight adjustments dependent upon the marginal classifications (but not full cross classification) to minimize an objective function of the differences between the population counts of the two way cells and their sample estimates. In the terminology of Chang and Kott (2008), if the row and column classifications have *I* and *J* cells respectively, this results in *IJ benchmark* variables and $I + J$ - 1 *model* variables. We study the performance of these estimators by constructing simulation simple random samples from the 2005 Quarterly Census of Employment and Wages which is maintained by the Bureau of Labor Statistics. We use the double classification of state and industry group. In our study, the calibration approaches introduced an asymptotically trivial bias, but reduced the MSE, compared to the unbiased estimator, by as much as 20% for a small sample.

Key Words: Calibration; Poststratification; Prediction model.

1. Introduction

Suppose we have a population $\mathcal{U}$ which is doubly stratified by two categorical variables whose indices are denoted $(i, j), i = 1, \ldots, I, j = 1, \ldots, J$ and write $\mathcal{U}_{ij}$ for the (i, j)-stratum. If a simple random sample $\mathcal{S}$ of size n is taken and if y denotes the variable of interest a natural estimator for the total $T_y = \sum_{k \in \mathcal{U}} y_k$ is the poststratified estimator

$$\hat{t}_{y\mathrm{PS}} = \sum_{i,\, j} N_{ij}\, \bar{y}_{ij} \qquad (1)$$

where N_{ij} is the size of $\mathcal{U}_{ij}$ and $\bar{y}_{ij}$ is the sample mean of y over $\mathcal{S} \cap \mathcal{U}_{ij}$. This estimator is widely used as long as all the sample sizes n_{ij} of $\mathcal{S} \cap \mathcal{U}_{ij}$ are reasonably large.

What to do if some of the n_{ij} are small, or even zero?

The standard approach would be to collapse some of the cells until all the n_{ij} are big enough. However such a collapsing might not be possible in a way that maintains the double classification scheme: that is the indices j might depend upon i.

The poststratified estimator $\hat{t}_{y\mathrm{PS}}$ is a special case of a calibration estimator. Define for each $k \in \mathcal{U}$ the $I \times J$ vector variable $\mathbf{x}_k = (x_{11k}, \ldots, x_{IJk})^T$ where $x_{ijk} = 1$ if $k \in \mathcal{U}_{ij}$ and $x_{ijk} = 0$ otherwise. The population total $T_{\mathbf{x}}$ of $\mathbf{x}$ is $(N_{11}, \ldots, N_{IJ})^T$ and letting $d_k = N / n$ be the sampling weight and $\beta = (N^{-1} n_{11}^{-1} N_{11} n, \ldots, N^{-1} n_{IJ}^{-1} N_{IJ} n)^T$

$$\hat{t}_{y\mathrm{PS}} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta)\, y_k$$

$$T_{\mathbf{x}} = \sum_{k \in \mathcal{S}} d_k (\mathbf{x}_k^T \beta)\, \mathbf{x}_k.$$

These equations establish that if the *benchmark* variables $\mathbf{x}$ are used, then $\hat{t}_{y\mathrm{PS}}$ is the resulting calibrated estimator of T_y.

Chang and Kott (2008) derived the asymptotic properties of a calibrated estimate of the form

$$\hat{t}_{y,\, \mathbf{z} f \mathbf{V}} = \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \hat{\beta})\, y_k \qquad (2)$$

where $\hat{\beta}$ minimizes an objective function of the form

$$Q(\beta) = \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\, \mathbf{x}_k \right)^T \mathbf{V}^{-1} \left(T_{\mathbf{x}} - \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\, \mathbf{x}_k \right). \qquad (3)$$

In equations (2) and (3), $\mathbf{z}$ is a vector of *model* variables whose length Q is at most the length P of the *benchmark* variables $\mathbf{x}$, f is a positive real valued function which Chang and Kott (2008) calls the *back link* function, and $\mathbf{V}$ is some positive definite symmetric $P \times P$ matrix. $\mathbf{V}$ is allowed to depend upon β as would occur if $\mathbf{V}(\beta)$ is some measurement of the variability of $\sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\, \mathbf{x}_k$.

In Chang and Kott (2008), the realized sample $\mathcal{S}$ is the respondents from an original sample with sampling weights d_k. The respondent sample $\mathcal{S}$ is assumed to be a Poisson subsample of the original sample with Poisson probabilities $f(\mathbf{z}_k^T \beta_0)^{-1}$, for some β_0. The asymptotic formulas derived there were under an asymptotic framework for this quasi-randomization (design based) model. We use the term *quasi-randomization* to remind ourselves that the assumed Poisson response mechanism is actually model based.

It should be noted that the use of calibration to correct for nonresponse goes back to Fuller, Loughin and Baker (1994), at least when $\mathbf{z} = \mathbf{x}$ and $f(\eta) = 1 + \eta$.

[1] Ted Chang, University of Virginia, Department of Statistics, Charlottesville, VA, U.S.A. E-mail: tcc8v@virginia.edu.

We propose to use the Chang and Kott (2008) methodology with $\mathbf{x}$ remaining as indicator variables for the complete $I \times J$ cross classification but letting $\mathbf{z}$ be a vector of $I + J - 1$ indicator variables for the marginal classifications. In other words, we propose to rebalance the sample to come as close as possible, in the sense of minimizing (3), to the correct cell proportions in the complete cross classification, but requiring the rebalancing weights to depend only upon the marginal classifications.

The Chang and Kott (2008) framework applies in the presence of nonresponse (and/or noncoverage) if $f(\mathbf{z}_k^T \beta_0)^{-1}$ is the response (or combined response and coverage) probability. We note that poststratification, a special case of calibration, is often used for the purpose of nonresponse/noncoverage correction. In our test example below, there is no nonresponse or noncoverage to correct for, and hence, the Chang and Kott (2008) framework applies with $\beta_0 = \mathbf{0}$ for any f with $f(0) = 1.$ In other words, if the calibration is used solely for the purposes of sample rebalancing, we can use Chang and Kott (2008) with almost any $f.$ But if we are trying to correct for nonresponse and/or noncoverage, stronger assumptions are required.

It should be noted that raking is simply the calibrated estimate using the $I + J - 1$ indicator variables of the marginal classifications as both benchmark and model variables and using $f(\eta) = e^{\eta}.$ Thus we will also explore the use of this back link function.

Section 2 gives the precise formulas for the estimators we will use in this study. Chang and Kott (2008) can be applied to derive sample based variance estimators and these derivations are given in the Appendix.

In Section 3, we give the results on an empirical study using the 2005 first quarter Quarterly Census of Employment and Wages, collected by the Bureau of Labor Statistics. We will restrict ourselves to the five states which we will denote by A, B, C, D, E and to five industry groupings denoted by 1, 2, 3, 4, 5. We will not further identify either the states or the industry groupings to prevent identification of the outlier in the discussion below. This population has 283,725 firms. From this population we will take Monte Carlo simple random samples of size $n =$ 200, 1,000, 5,000 and use the double classification of state and industry group.

It should be noted that 0.18% of the population has the double classification of state E and industry grouping 5. Thus when $n = 200,$ the expected sample size in this cell is 0.36 and poststratification using the double classification is out of the question.

Kott and Chang (2010) derives the properties of $\hat{t}_{y,\mathbf{z} f \mathbf{V}}$ using a model based framework. The models considered there do not apply with our selection of $\mathbf{x}$ and $\mathbf{z}$ variables. However, motivated by their approach, we examine in Section 4 the behavior of the estimator $\hat{t}_{y,\mathbf{z} f \mathbf{V}}$ defined by equation (2), under highly simplified assumptions, including that $f(\eta) = 1 + \eta.$ This leads in Section 5 to the choice of a new weight matrix $\mathbf{V}^{-1}$ for use in (3). We then continue with our empirical exploration using this new estimator.

2. Mathematical formulas

In this section we list the formulas used in this study. They are all special cases of formulas in Chang and Kott (2008). We assume that a simple random sample of size n is taken from a population of size N and we use $\mathcal{S}$ and r to denote the respondents from that sample and the size of $\mathcal{S}.$ We assume that the calibration weight function has a β_0 such that $f(\mathbf{z}^T \beta_0)^{-1}$ is the response probability for an element with model variables $\mathbf{z}.$ In particular, and without loss of generality, if there is no nonresponse problem, we assume $f(0) = 1.$

The same formulas work with noncoverage, in which case $f(\mathbf{z}^T\beta_0)^{-1}$ is the combined response/coverage probability.

We denote $N_{ij},$ $\mathcal{S}_{ij},$ and r_{ij} to be the population size, respondent sample, and respondent sample size in classification $(i, j).$ Although N_{ij} is assumed known, our methodology does not require the knowledge of the row and column classifications of nonrespondents.

We define $N_{i\cdot} = \sum_j N_{ij}$ and analogously define $N_{\cdot j}.$

We will use estimators for a total T_y of the form

$$\hat{t}_y = \frac{N}{n}\sum_i \sum_j \sum_{k \in \mathcal{S}_{ij}} w_{ij}\, y_k \tag{4}$$

where the adjustment weights w_{ij} are defined as below. These are all special cases of equations (2) and (3) when we use $\mathbf{V} = \mathbf{I}.$

The *calibrated margins* estimator uses $f(\eta) = 1 + \eta$ and defines $\mathbf{x} = \mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. In this case $T_{\mathbf{x}}$ is a vector of $N_{i\cdot}$ and $N_{\cdot j}$. The adjustment weights $f(\mathbf{z}_k^T \beta)$ have the form $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ when $\mathbf{z}$ is the vector of indicator variables for membership in the i^{th} and j^{th} row and column classifications respectively. Since the number of equations (the dimension of $\mathbf{x}$) equals the number of unknowns (the dimension of $\hat{\beta}$), we expect to be able to solve the equations

$$T_{\mathbf{x}} = \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \beta)\mathbf{x}_k \tag{5}$$

exactly. Thus $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ solve the linear equations of rank $I + J - 1$

$$N_{i\cdot} = \frac{N}{n}\sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$

$$N_{\cdot j} = \frac{N}{n}\sum_i (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij},$$

which easily follows from (5).

The *calibrated cell counts* estimator uses $f(\eta) = 1+\eta$ and defines $\mathbf{x}$ to be the IJ indicator variables for the complete cross classification and $\mathbf{z}$ to be $I+J-1$ independent indicator variables for the marginal categories. In this case $T_{\mathbf{x}}$ is a vector of N_{ij} and, since $\mathbf{V} = \mathbf{I}$, the adjustment weights $w_{ij} = 1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}$ minimize the objective function

$$\sum_i \sum_j \left[N_{ij} - \frac{N}{n} \sum_i \sum_j (1 + \hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij} \right]^2 .$$

The *raking* estimator uses $f(\eta) = e^{\eta}$ and defines $\mathbf{x} = \mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. Its adjustment weights are $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ where $\hat{\beta}_{i\cdot}, \hat{\beta}_{\cdot j}$ solve the $I + J$ equations

$$N_{i\cdot} = \frac{N}{n} \sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$

$$N_{\cdot j} = \frac{N}{n} \sum_i \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}.$$

Since $\sum_i N_{i\cdot} = N = \sum_j N_{\cdot j}$, these $I+J$ equations yield only $I+J-1$ constraints. It should be noted, however, that if a constant c is added to each $\hat{\beta}_{i\cdot}$ and subtracted from each $\hat{\beta}_{\cdot j}$, the w_{ij} are not changed.

The *exponential calibrated cell counts* estimator uses $f(\eta) = e^{\eta}$ and defines $\mathbf{x}$ to be the IJ indicator variables for the complete cross classification and $\mathbf{z}$ to be $I + J - 1$ independent indicator variables for the marginal categories. Its adjustment weights $w_{ij} = \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j})$ minimize the objective function

$$\sum_i \sum_j \left[N_{ij} - \frac{N}{n} \sum_i \sum_j \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij} \right]^2 .$$

Chang and Kott (2008) give formulas for sample based estimation of the variance of $\hat{t}_y$. In the appendix, we apply these formulas to the four estimators above.

3. Empirical study

The population we use here is the data from the 2005 first quarter Quarterly Census of Employment and Wages (QCEW), restricted to five states and five industry groupings. The QCEW is compiled from mandatory reports to state employment offices and hence is virtually a census and the data we used is the complete QCEW for these five states and five industry groupings. This population has $N = 283{,}725$ firms, divided as in Table 1.

The response variables y are total employment and total (quarterly) wages. For these variables $T_y = 2{,}981{,}364$ for total employment and $T_y = 2{,}334{,}400$ (in tens of thousands of dollars) for total wages. In this study, we took 10,000 samples of sizes $n = 200, 1{,}000, 5{,}000$. For each of the 4 estimators, we report the estimated bias, standard error, and root mean square error. We also report square root of the mean of the estimated variances using the first term of equation (15). For purposes of comparison, we report the theoretical and empirical values for the unweighted estimator $N/n \sum_{k \in \mathcal{S}} y_k$. These results are reported in Table 2 for total employment and Table 3 for total wages.

For sample size $n = 5{,}000$, the expected sample size in the smallest cell (state E and industry group 5) is 9.07. While this might be a little small for poststratification, the probability that this cell has a sample size less than 2, the minimum size necessary for variance estimation, is 0.0011. In our simulations 9 runs had a cell with sample size less than 2. For this sample size, we also report the empirical behavior of poststratified estimator, excluding the 9 problem cases, using the variance estimate (7.6.5) of Särndal, Swensson and Wretman (1992) and its theoretical behavior using the variance approximation given in (7.6.6) of Särndal *et al.* (1992).

Table 1
Business entities by state and industry group

			industry group			
	1	**2**	**3**	**4**	**5**	**sum**
A	5,986	5,548	7,712	3,969	1,299	24,514
	(2.11%)	(1.96%)	(2.72%)	(1.40%)	(0.46%)	(8.64%)
B	18,782	31,572	22,012	4,982	4,504	81,852
	(6.62%)	(11.13%)	(7.76%)	(1.76%)	(1.59%)	(28.85%)
C	13,518	13,099	17,837	5,610	3,001	53,065
	(4.76%)	(4.62%)	(6.29%)	(1.98%)	(1.06%)	(18.70%)
D	30,428	36,017	32,541	10,963	5,399	115,348
	(10.72%)	(12.69%)	(11.47%)	(3.86%)	(1.90%)	(40.65%)
E	2,225	2,020	3,110	1,076	515	8,946
	(0.78%)	(0.71%)	(1.10%)	(0.38%)	(0.18%)	(3.15%)
sum	70,939	88,256	83,212	26,600	14,718	283,725
	(25.00%)	(31.11%)	(29.33%)	(9.38%)	(5.19%)	

Table 2
Empirical comparison of 4 estimators of total employment

estimator	bias	st. err.	rt. MSE	rt. est. var.
n = 200				
unweighted (theoretical)	0	1,113,220		
unweighted (empirical)	-1,280	1,068,944	1,068,945	1,059,463
cal. margins	-1,394	1,105,201	1,105,201	1,048,873
cal. cell cts.	-218,751	1,008,436	1,031,889	975,140
raking	-462	1,103,172	1,103,172	1,041,490
exp. cal. cell cts.	-227,578	1,000,154	1,025,719	962,153
n = 1,000				
unweighted (theoretical)	0	497,144		
unweighted (empirical)	-5,435	505,941	505,970	501,144
cal. margins	-6,212	506,239	506,277	498,946
cal. cell cts.	-56,118	493,611	496,790	488,222
raking	-4,854	507,938	507,961	499,237
exp. cal. cell cts.	-58,891	492,939	496,445	487,281
n = 5,000				
unweighted (theoretical)	0	220,751		
unweighted (empirical)	1,516	224,088	224,093	222,034
poststr. (theoretical)	0	220,315		
poststr. (empirical, 9 cases excluded)	1,234	223,225	223,228	221,094
cal. margins	1,649	223,091	223,098	220,833
cal. cell cts.	-8,606	222,170	222,337	220,347
raking	3,632	236,355	236,383	220,606
exp. cal. cell cts.	-10,643	223,472	223,725	220,207

Table 3
Empirical comparison of 4 estimators of total wages (tens of thousands of dollars)

estimator	bias	st. err.	rt. MSE	rt. est. var.
n = 200				
unweighted (theoretical)	0	1,682,571		
unweighted (empirical)	-11,119	1,551,186	1,551,226	1,543,483
cal. margins	-11,474	1,582,383	1,582,425	1,510,413
cal. cell cts.	-214,323	1,451,931	1,467,664	1,413,411
raking	-11,220	1,579,842	1,579,882	1,501,170
exp. cal. cell cts.	-221,435	1,438,810	1,455,750	1,393,246
n = 1,000				
unweighted (theoretical)	0	751,406		
unweighted (empirical)	-2,911	772,495	772,501	768,878
cal. margins	-4,372	776,955	776,968	768,869
cal. cell cts.	-51,649	756,201	757,963	751,384
raking	-4,684	778,302	778,316	769,428
exp. cal. cell cts.	-54,305	754,963	756,913	749,832
n = 5,000				
unweighted (theoretical)	0	333,654		
unweighted (empirical)	2,678	336,057	336,068	337,239
poststr. (theoretical)	0	333,765		
poststr. (empirical, 9 cases excluded)	1,802	335,271	335,276	336,192
cal. margins	2,510	334,910	334,920	336,064
cal. cell cts.	-7,149	333,560	333,637	335,006
raking	-4,679	339,074	339,106	335,230
exp. cal. cell cts.	-9,251	334,365	334,493	334,755

The response variables, total employment and total wages, are strongly skewed right. There is one firm (in state C and industry group 4) whose total employment is more than double the total employment of the next largest firm and many hundreds times the mean employment of the remaining firms. We repeat this study using a population with this firm removed. The results are presented in Tables 4 and 5. In practice with this population, the sampling would normally sample this firm with certainty (a *self representing unit*) and samples constructed from the remaining firms. Thus Tables 4 and 5 are perhaps more indicative of the relative performance of these estimators in actual practice.

The samples used for Tables 4 and 5 are identical to those used for Tables 2 and 3 except that if the outlier was included in the sample, it was replaced by a new observation from the population. This was done to improve the comparability of the results of Tables 4 and 5 with those of Tables 2 and 3.

Table 4
Empirical comparison of 4 estimators of total employment: population with outlier removed

estimator	bias	st. err.	rt. MSE	rt. est. var.
		n = 200		
unweighted (theoretical)	0	950,688		
unweighted (empirical)	5,395	975,617	975,632	965,448
cal. margins	5,777	1,019,583	1,019,599	963,314
cal. cell cts.	-211,568	909,070	933,365	877,343
raking	6,688	1,018,383	1,018,405	956,867
exp. cal. cell cts.	-217,810	902,756	928,660	868,797
		n = 1,000		
unweighted (theoretical)	0	424,552		
unweighted (empirical)	-8,393	422,116	422,199	414,019
cal. margins	-9,430	418,153	418,259	408,577
cal. cell cts.	-58,808	408,391	412,603	399,961
raking	-8,135	419,938	420,016	408,611
exp. cal. cell cts.	-61,014	407,780	412,320	399,311
		n = 5,000		
unweighted (theoretical)	0	188,517		
unweighted (empirical)	702	191,631	191,632	188,089
poststr. (theoretical)	0	187,691		
poststr. (empirical, 9 cases excluded)	563	190,854	190,855	187,180
cal. margins	820	190,662	190,664	186,664
cal. cell cts.	-9,376	189,884	190,115	186,202
raking	2,933	205,924	205,944	186,618
exp. cal. cell cts.	-9,922	189,813	190,072	186,140

Table 5
Empirical comparison of 4 estimators of total wages: population with outlier removed

estimator	bias	st. err.	rt. MSE	rt. est. var.
		n = 200		
unweighted (theoretical)	0	1,330,930		
unweighted (empirical)	711	1,341,900	1,341,901	1,334,556
cal. margins	1,256	1,387,484	1,387,485	1,318,285
cal. cell cts.	-201,575	1,225,852	1,242,314	1,194,071
raking	1,473	1,386,978	1,386,979	1,311,353
exp. cal. cell cts.	-206,956	1,217,881	1,235,340	1,184,166
		n = 1,000		
unweighted (theoretical)	0	594,370		
unweighted (empirical)	-8,169	587,775	587,832	582,524
cal. margins	-10,093	583,606	583,693	576,251
cal. cell cts.	-56,429	569,158	571,948	563,022
raking	-10,529	584,532	584,626	576,282
exp. cal. cell cts.	-58,435	568,277	571,273	562,061
		n = 5,000		
unweighted (theoretical)	0	263,923		
unweighted (empirical)	1,185	266,779	266,782	264,110
poststr. (theoretical)	0	263,339		
poststr. (empirical, 9 cases excluded)	566	265,973	265,973	263,210
cal. margins	991	265,449	265,451	262,556
cal. cell cts.	-8,565	264,126	264,265	261,483
raking	-6,008	271,535	271,602	262,021
exp. cal. cell cts.	-9,070	264,038	264,194	261,394

Examining Tables 2 and 3, we see that the $P > Q$ methods, that is those that calibrate the cross classified cell counts using calibration weights which depend upon the marginal classifications, are clearly more biased than the other techniques. However the biases of these estimators relative to their standard deviations decrease with increasing sample sizes. We will show in the next section, that under a highly simplified model, the bias has order n^{-1} and the standard deviation has order $n^{-1/2}$. Consider, for example, the results for the "calibrated cell counts" estimator in Table 2. In this case, the bias divided by the standard error is 0.217, 0.114, 0.039 for $n = 200, 1{,}000, 5{,}000$ respectively. For these values of n, the values of $n^{-1/2}$ are 0.0707, 0.0316, 0.0141 and it appears that the former series of three ratios is approximately 3 times the latter series.

It also appears that the exponential back link function f performs slightly better than the linear choice for f. Computationally the former is much more expensive than the latter. We also notice that as the sample sizes increase, the estimators' performances appear to converge. This is to be expected: because there is no nonresponse, as $n \to \infty$, $\hat{\beta} \to 0$, so that the adjustment weights $w = f(\mathbf{z}^T \hat{\beta}) \to 1$.

Comparing the linear calibrated cell counts estimator to the empirical values of the unweighted estimator, the former is approximately 7.3% more efficient in MSE when $n =$ 200 for total employment and 11.7% more efficient for total wages. (This means, for example, that the empirical MSE of the unweighted estimator is 1.117 times the empirical MSE for the linear calibrated cell counts estimator when estimating total wages.) For the exponential calibrated cell counts estimator, the improvement in efficiency relative to the empirical MSE of the unweighted estimator is 8.6% for total employment and 13.5% for total wages. Comparison to the theoretical values for the unweighted estimator would be more favorable to the calibrated cell counts estimators, but we will use the empirical results for the unweighted estimator as the various estimators have all used the same Monte Carlo samples. The calibrated cell counts estimator and exponential calibrated cell counts estimator still have an advantage in MSE over the unbiased estimator at sample size $n = 1{,}000$.

When the single extreme outlier is removed, leaving 283,724 remaining elements of the population, the calibrated cell count estimators have somewhat better performance relative to the unweighted estimator. For $n =$ 200, the linear calibrated cell count estimator offers a 9.3% improvement in efficiency for total employment and a 16.7% improvement for total wages. The comparable ratios for the exponential calibrated cell count estimator are 10.4% for total employment and 18.0% for total wages.

Finally, the variance estimator in equation (15) has a slight downward bias.

4. Model based bias and variance of calibrated estimators

Kott and Chang (2010) derived the asymptotic properties of $\hat{t}_{y,\mathbf{z} f \mathbf{V}}$ under a different, model-based, probability structure. In that paper $\mathcal{S}$ is a sample selected with selection probabilities d_k^{-1} so that nonresponse is not an issue in $\mathcal{S}$. Rather, if P the number of benchmark variables $\mathbf{x}$ equals Q the number of model variables $\mathbf{z}$, Kott and Chang (2010) assume a *prediction* model

$$y_k = \mathbf{x}_k^T \theta + \varepsilon_k, k \in \mathcal{U}. \tag{6}$$

Here θ is a unknown fixed vector, ε_k are model independent errors subject to

$$E(\varepsilon_k \mid \mathbf{z}_j, I_j, j \in \mathcal{U}) = 0, \tag{7}$$

and I_k is a random variable defined by $I_k = 1$ if $k \in \mathcal{S}$ and $I_k = 0$ otherwise.

When $P > Q$, the model equation (6) must be replaced by

$$y_k = (\mathbf{A}_\infty \mathbf{x}_k)^T \theta + \varepsilon_k, k \in \mathcal{U} \tag{8}$$

for some limiting $Q \times P$ matrix $\mathbf{A}_\infty$ (which is defined in a suitable asymptotic framework, see Kott and Chang (2010)).

Thus when $\mathbf{x}$ represents indicator variables for the complete $I \times J$ cross classification, we have that $\mathbf{x}_k^T \theta$, for k in the $(i, j)^{\text{th}}$ classification, is the mean value of the response variable over the $(i, j)^{\text{th}}$ classification. Hence, by definition, $E(\varepsilon_k \mid \mathbf{x}_j, j \in \mathcal{U}) = 0$ and, since $\mathbf{z}$ is a function of $\mathbf{x}$, the model (6) and (7) automatically holds when the sampling (including nonresponse) is noninformative.

However, in our application of calibration, $P = IJ > Q = I + J - 1$ and the model equation (8) has no a priori reason to hold.

Motivated by Kott and Chang (2010) we examine the behavior of calibrated estimates under the following scenario:

1. The benchmark variables $\mathbf{x}$ are indicator variables for some partition of the population into classes $\mathcal{C}_r$. The model (6) automatically holds where the r^{th} component of θ is the population mean of $\mathcal{C}_r$. Let f_r denote the proportion of the population in $\mathcal{C}_r$ and $V_r = \mathrm{Var}(\varepsilon_k \mid k \in \mathcal{C}_r)$. We shall also use the notation $\mathrm{Var}(\mathbf{x}_k)$ for V_r when $k \in \mathcal{C}_r$.
2. The sample is a simple random sample of size n chosen *with replacement*.
3. The back link function $f(\eta)$ in the estimator $\hat{t}_{y,\mathbf{z} f \mathbf{V}}$ of equation (2) is $f(\eta) = 1 + \eta$.

Although these assumptions are unrealistic in practice, the main purpose of this section is to heuristically justify a choice, given in the next section, for the matrix $\mathbf{V}$. At this point, we no longer place any requirements on $\mathbf{z}$.

We note that in this situation $E(\varepsilon_k \mid \mathbf{x}_j, I_j, j \in \mathcal{U}) = 0.$ Note that (7) will hold if the components of the model variables $\mathbf{z}$ are functions of $\mathbf{x},$ that is each component of $\mathbf{z}$ is constant on each class. However if $P > Q,$ (8) will generally not hold. In any event, in this section we require neither (7) nor (8).

We let

$$\mu_{\mathbf{x}} = \frac{1}{N}\sum_{j\in\mathcal{U}} \mathbf{x}_j$$

$$\mu_{\mathbf{xz}} = \frac{1}{N}\sum_{j\in\mathcal{U}} \mathbf{x}_j\, \mathbf{z}_j^T$$

and the matrix $\mathbf{A}_\infty$ of equation (8) becomes

$$\mathbf{A}_\infty = \mathbf{V}^{-1}\mu_{\mathbf{xz}}.$$

Let $\hat{\mu}_{y,\mathbf{zV}} = N^{-1}\hat{t}_{y,\mathbf{zV}}$ where $\hat{t}_{y,\mathbf{zV}}$ is defined as in (2). We have suppressed the f in the notation $\hat{t}_{y,\mathbf{zV}}$ because, in this section, $f(\eta) = 1 + \eta.$ Letting $\bar{y}_s$ and $\bar{\mathbf{x}}_s$ denote the indicated sample means and using Kott and Chang (2010)

$$\hat{\beta} = \left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j \mathbf{x}_j^T\, \mathbf{A}_\infty\right)^{-1} \left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j\, y_j\right) + O_p(n^{-1/2})$$

$$\begin{aligned}\hat{\mu}_{y,\mathbf{zV}} &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V}\mathbf{A}_\infty)^{-1}\left(\frac{1}{n}\sum_{j\in\mathcal{S}} \mathbf{z}_j y_j\right) + O_p(n^{-1})\\ &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{A}_\infty (\mathbf{A}_\infty^T \mathbf{V}\mathbf{A}_\infty)^{-1}\mu_{\mathbf{xz}}^T\theta + O_p(n^{-1})\\ &= \bar{y}_s + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T \mathbf{V}^{-1}\mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T\mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\theta + O_p(n^{-1}). \quad (9)\end{aligned}$$

If $\hat{\mu}_{y,\mathbf{zV}}$ is bounded, as would occur if $f(\eta) = 1 + \eta$ were modified for large η to prevent large calibration weight adjustments, we would have

$$\begin{aligned}\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}}) &= \mathrm{E}(\bar{y}_s) + O(n^{-1}) = \mu_y + O(n^{-1})\\ \mathrm{E}(\hat{\mu}_{y,\mathbf{zV}} | \bar{\mathbf{x}}_s) &= \bar{\mathbf{x}}_s^T\theta + (\mu_{\mathbf{x}} - \bar{\mathbf{x}}_s)^T\mathbf{V}^{-1}\mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T\mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\theta\\ &\quad + O(n^{-1})\\ \mathrm{Var}[\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}} | \bar{\mathbf{x}}_s)] &= \frac{1}{n}\theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta\\ &\quad + o(n^{-1}),\end{aligned}$$

where $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the covariance matrix of $\mathbf{x}$ and

$$\mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}} = \mu_{\mathbf{xz}}(\mu_{\mathbf{xz}}^T\mathbf{V}^{-1}\mu_{\mathbf{xz}})^{-1}\mu_{\mathbf{xz}}^T\mathbf{V}^{-1}. \quad (10)$$

Now

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s) &= \mathrm{Var}(\bar{y}_s \mid \bar{\mathbf{x}}_s) + o(n^{-1})\\ &= \frac{1}{n^2}\sum_{j\in\mathcal{S}} V(\mathbf{x}_j) + o(n^{-1})\\ \mathrm{E}[\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)] &= \mathrm{E}[\mathrm{Var}(\bar{y}_s \mid \bar{\mathbf{x}}_s)] + o(n^{-1})\\ &= \frac{1}{n}\sum_r f_r V_r + o(n^{-1}).\end{aligned}$$

It is easily seen that

$$\mathrm{Var}[\mathrm{E}(\bar{y}_s \mid \bar{\mathbf{x}}_s)] = \frac{1}{n}\theta^T\boldsymbol{\Sigma}_{\mathbf{x}}\theta.$$

Since $\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}}) = \mathrm{Var}[\mathrm{E}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)] + \mathrm{E}[\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}} \mid \bar{\mathbf{x}}_s)]$ and similarly for $\mathrm{Var}(\bar{y}_s),$ $\mathrm{Var}(\hat{\mu}_{y,\mathbf{zV}}) < \mathrm{Var}(\bar{y}_s)$ to terms $o(n^{-1})$ when

$$\theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\boldsymbol{\Sigma}_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta < \theta^T\boldsymbol{\Sigma}_{\mathbf{x}}\theta. \quad (11)$$

The derivation also establishes that the square bias has an asymptotically trivial contribution to the mean square error of $\hat{\mu}_{y,\mathbf{zV}}.$

5. A proposed new weight matrix $\mathbf{V}^{-1}$

In this section we return to our original benchmark $\mathbf{x}$ and model $\mathbf{z}$ variables. When $\mathbf{V} = \mathbf{I},$ the identity matrix, we see from (10) that $\mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{I}}^T\,\theta = \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{I}}\,\theta$ is the projection of θ onto the span of the columns of $\mu_{\mathbf{xz}}.$ The left hand side of (11) will be zero if θ is in this column span.

For simplicity, we will write $\mu_{\mathbf{xz}}$ as a singular matrix, of rank $I + J - 1,$ with one row for each possible double classification cell (i, j) and one column for each row classification i and each column classification $j.$ Thus, the $(i, j)^{\text{th}}$ row of $\mu_{\mathbf{xz}}$ has $f_{ij} = N_{ij}/N$ in the columns corresponding to i and j and zero elsewhere. Thus θ will be in the column span of $\mu_{\mathbf{xz}}$ if and only if for each i and j

$$\frac{\theta_{ij}}{f_{ij}} = \alpha_i + \beta_j \quad (12)$$

for some α_i and $\beta_j.$ In other words, the θ_{ij} / f_{ij} satisfy a two way ANOVA model, without interaction, in the column and row classifications.

Recalling that θ_{ij} represents the mean value of the variable of interest y in the $(i, j)^{\text{th}}$ cell, (12) does not appear to be a very promising approximation to the truth. A more likely approximation would be the usual two way ANOVA model

$$\theta_{ij} = \alpha_i + \beta_j. \quad (13)$$

Suppose we change variables $\tilde{\mathbf{x}} = \mathbf{C}\mathbf{x}$ for some diagonal matrix $\mathbf{C}.$ Note that the rows and columns of $\mathbf{C}$ are doubly indexed by (i, j) and we will let c_{ij} denote the diagonal entry in the $(i, j)^{\text{th}}$ row and column. Let $\tilde{\theta} = \tilde{\mathbf{C}}^{-1}\theta$ so that model (6) can be rewritten as

$$y_k = \tilde{\mathbf{x}}_k^T\tilde{\theta} + \varepsilon_k.$$

Now the matrix $\mu_{\tilde{\mathbf{x}}\mathbf{z}}$ has $c_{ij}f_{ij}$ in the $(i, j)^{\text{th}}$ row and the columns corresponding to i and $j.$ Now $\tilde{\theta}$ will be in the column span of $\mu_{\tilde{\mathbf{x}}\mathbf{z}}$ if and only if

$$c_{ij}^{-1}\theta_{ij} = \tilde{\theta}_{ij} = c_{ij} f_{ij}(\alpha_i + \beta_j).$$

Thus (13) is equivalent to $c_{ij} = f_{ij}^{-1/2}$. It is easily checked that

$$\tilde{\theta}^T(\mathbf{I} - \mathbf{P}_{\mu_{\tilde{\mathbf{x}}\mathbf{z}},\mathbf{I}})\Sigma_{\tilde{\mathbf{x}}}(\mathbf{I} - \mathbf{P}_{\mu_{\tilde{\mathbf{x}}\mathbf{z}},\mathbf{I}})\tilde{\theta} = \theta^T(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})\Sigma_{\mathbf{x}}(\mathbf{I} - \mathbf{P}_{\mu_{\mathbf{xz}},\mathbf{V}})^T\theta$$

when $\mathbf{V} = \mathbf{C}^{-2}$. We thus propose using the diagonal matrix $\mathbf{V}_o$ whose diagonal entries are f_{ij}.

With this choice of $\mathbf{V}_o$, equation (9) suggests the estimator for simple random sampling

$$\hat{t}_{y,\mathbf{z}\mathbf{V}_o} = N\bar{y}_s + (T_{\mathbf{x}} - N\bar{x}_s)^T\mathbf{V}_{\mathbf{o}}^{-1}\hat{\mu}_{\mathbf{xz}}(\hat{\mu}_{\mathbf{xz}}^T\mathbf{V}_o^{-1}\hat{\mu}_{\mathbf{xz}})^{-1}\left(\frac{1}{n}\sum_{k\in\mathcal{S}}\mathbf{z}_k y_k\right) \quad (14)$$

where $\hat{\mu}_{\mathbf{xz}} = n^{-1}\sum_{k\in\mathcal{S}}\mathbf{x}_k\mathbf{z}_k^T$. In our case both $\mu_{\mathbf{xz}}$ and $\mu_{\mathbf{x}}$ are known from the N_{ij}, but in the spirit of ratio estimation it is preferable to use $\hat{\mu}_{\mathbf{xz}}$ in place of $\mu_{\mathbf{xz}}$. This heuristic observation has been demonstrated using simulations (not shown) with the QCEW population.

We shall call the estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ of equation (14) the *weighted calibrated cell counts* estimator.

Simulations with artificial response variables y, also not shown, demonstrate that when the model (13) holds, then weighted calibrated cell counts estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ performs markedly better than the other estimators considered here. Table 6 gives statistics for the estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ for the populations and variables studied in Tables 2 - 5.

Comparing to Tables 2 - 5, we see that in all cases $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ has the highest bias but the lowest MSE of the estimators considered. For $n = 200$ and the full population, $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ has a 14.8% gain in efficiency (as measured by MSE) relative to the empirical results for the unbiased estimator when estimating total employment and a 21.1% efficiency gain when estimating total wages. For $n = 200$ and the population with a single extreme outlier deleted, the corresponding gains are 14.2% and 21.7% for total employment and total wages respectively.

The Associate Editor suggested that we compare our estimators to a poststratified estimator using collapsed cells to avoid the problem of empty cells in the sample. We explored this question for sample size $n = 200$ where it is most likely that empty cells will occur. We constructed 14 poststrata. Nine of these poststrata are the nine largest cells in the original data. The other 5 poststrata are A1 and A2; A3, A5, and B4; A4, B5, and C4; C5 and D4; and all cells from state E together with D5. After these combinations, the 5 combined poststrata had sizes that ranged between 4.07% and 5.06% of the population and the 9 retained original cells had sizes in the range of 4.62% to 11.47%.

Table 6
Empirical statistics for $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ of equation (14)

n	bias	st. err.	rt. MSE	rt. est. var.
		Full population - total employment		
200	-244,749	967,066	997,556	923,492
1,000	-64,839	490,758	495,023	483,550
5,000	-10,767	221,702	221,964	219,408
		Full population - total wages		
200	-242,528	1,388,489	1,409,511	1,333,793
1,000	-62,091	752,603	755,160	744,315
5,000	-9,821	332,682	332,827	333,782
		Population with outlier deleted - total employment		
200	-236,812	881,844	913,088	842,191
1,000	-67,468	405,215	410,793	396,105
5,000	-11,482	189,501	189,848	185,483
		Population with outlier deleted - total wages		
200	-228,441	1,194,922	1,216,562	1,151,417
1,000	-66,765	565,008	568,939	557,676
5,000	-11,138	263,699	263,934	260,768

Unfortunately, the author no longer has access to the QCEW data base. Besides the cell counts in Table 1, the author has only the means, standard deviations, and maximum values by cell. The author constructed a pseudo population using the squares of randomly generated gamma variables. The square gamma variables were constructed to have the same cell means and standard deviations as the cell means and standard deviations in the original data. After doing this, the square gamma variables were rounded upwards to integer values. For these pseudo populations, $T_y = 3{,}149{,}491$ for employment and 2,305,273, in tens of thousands of dollars, for wages.

A square gamma distribution was used because the gamma distribution is insufficiently right skewed. Even so, in almost all cells the largest value in the original population exceeded the largest value in the pseudo population. Of course without the original data, we cannot distinguish between right skew and a tendency to produce outliers.

10,000 Monte Carlo samples were constructed were taken for each sample size. The results are shown in Table 7. For the poststratified estimator, 5 of the samples of size 200 had an empty poststratum and these runs were excluded from the results in Table 7.

Table 7
Empirical comparison of 4 estimators

estimator	bias	st. err.	rt. MSE	bias	st. err.	rt. MSE
		total employment			total wages	
			n = 200			
unweighted	644	1,006,956	1,006,956	-9,970	1,481,450	1,481,483
poststratified	-5,387	1,026,266	1,026,280	-2,149	1,548,833	1,548,834
cal. cell cts.	-224,198	942,164	968,472	-203,531	1,377,823	1,392,775
wtd. cal. cell cts.	-248,937	919,419	952,523	-232,558	1,326,234	1,346,469
			n = 1,000			
unweighted	-3,317	445,676	445,687	1,544	679,148	679,150
poststratified	-2,967	448,218	448,228	1,672	685,370	685,372
cal. cell cts.	-54,311	436,821	440,185	-44,942	665,799	667,314
wtd. cal. cell cts.	-63,327	432,396	437,008	-54,913	660,726	663,004
			n = 5,000			
unweighted	2,466	206,249	206,264	-2,539	304,852	304,863
poststratified	2,108	205,661	205,672	-2,705	304,751	304,763
cal. cell cts.	-8,265	204,693	204,859	-12,096	303,231	304,472
wtd. cal. cell cts.	-10,551	204,080	204,352	-14,697	302,311	302,668

Evidently the poststratification did not help. Even though no poststratum had an expected count below eight, the actual poststrata had quite variable sizes. In addition, the cell populations are quite skewed so that the poststrata sample means are quite variable.

The other conclusions for the pseudo populations reflect the conclusions from the actual populations. In particular, when $n = 200$ and for the employment pseudo population, the weighted calibrated cell counts estimator $\hat{t}_{y,\mathbf{z}\mathbf{V}_o}$ has an 11.8% gain in efficiency relative to the unbiased estimator. For the wages pseudo population and $n = 200$, the efficiency gain is 21.1%.

6. Concluding remarks

The use in (3) of weight matrices $\mathbf{V}(\beta)^{-1}$ which depend upon β has not been explored in this paper. Experimentation with the use of such a matrix was not encouraging. Computation time increased dramatically, and there were significant numbers of cases which failed to numerically converge, with no improvement in efficiency over the fixed $\mathbf{V}$ estimators considered here. Perhaps the authors did not try the right $\mathbf{V}(\beta)$.

Besides the exponential back link function, the authors tried the logistic back link $f(\eta) = (1 + e^{-\eta})^{-1}$. These runs also did not converge. On reflection, the reason is obvious: because in the simulations there was no nonresponse or noncoverage problems, the calibration weight adjustments $f(\mathbf{z}^T\beta) \to 1$ as $n \to \infty$. But 1 is not in the range of f. It should be noted that in Chang and Kott (2008) a logistic back link was used to correct for nonresponse.

Several obvious issues arise. For example, how would the results of this study change if a more complicated sampling design than simple random sampling were used, or if non response and/or non coverage occur and the calibration was used to correct for it. Falk (2010) considers these questions both theoretically and with further simulations using the QCEW population. Falk (2010) also considers non linear link functions.

There are obvious extensions to 3-way (and beyond) cross classified data. If I, J, K denote the number of cells in each of the 3 classifications, there are IJK fully classified cells whose totals can be used for benchmark $\mathbf{x}$ variables. There are $IJ + IK + JK - I - J - K + 1$ one-way and two-way marginal variables that can be used for model $\mathbf{z}$ variables. Clearly, one might not want to use the plethora of variables available.

In the context of linear calibration using the same $\mathbf{x}$ and $\mathbf{z}$ variables, several studies have been made on the choice of variables. Examples of such studies are Banker, Rathwell and Majkowski (1992), Silva and Skinner (1997), and Clark and Chambers (2008). The last paper remarks that too many variables can deteriorate the MSE of $\hat{T}_y$.

The alternatives to poststratification discussed here can be used in the presence of small and even empty cells. For example, in our simulations, the expected count in the state E, industry group 5, cell is 0.36 when $n = 200$. One might be tempted to collapse cells and use poststratification. Generally, however, it is not possible to do so and maintain the convenient doubly classified structure of the data. Our approaches, like poststratification, introduce weights for the purpose of sample balancing but avoid collapsing cells. These approaches generally increase bias but can offer substantial reductions in MSE.

Furthermore, in the presence of nonresponse or noncoverage, the inverse of the weight adjustments can be considered, under a quasi-randomization model for the response or coverage, as estimated probabilities of response and/or coverage. In our calibration approaches, these probabilities are assumed to be a function of the row and column classifications. When cells are collapsed without maintaining the double classification, these probabilities are harder to interpret.

Acknowledgements

The author would like to thank Phil Kott and John Eltinge for several very interesting insights. The author also thanks Larry Huff and Ken Robertson of the Bureau of Labor Statistics for help in obtaining and understanding the data.

Appendix

Here we derive, using Chang and Kott (2008) equations (16) and (17), sample based variance estimators for the 4 estimators studied in Section 2.

Let

$$\hat{\mathbf{H}}_y = \frac{\partial \hat{t}_{y,\mathbf{z}\,f\,\mathbf{V}}}{\partial \beta}(\hat{\beta}).$$

Here $\hat{t}_{y,\mathbf{z}\,f\,\mathbf{V}}$ is defined in (2). $\hat{\mathbf{H}}_y$ is a row vector with one entry for each $\mathbf{z}$ variable. In out case, $\hat{\mathbf{H}}_y$ has $(I + J - 1)$ entries, one for each of the $I + J - 1$ linearly independent indicator variables for the row and column classifications.

For the calibrated margins and calibrated cell counts estimators, $f(\eta) = 1 + \eta$. Define the constants s_{ij} and t_{ij} by

$$s_{ij} = \frac{N}{n} r_{ij}$$

$$t_{ij} = \frac{N}{n} \sum_{k \in \mathcal{S}_{ij}} y_k.$$

Then a simple calculation shows that if a entry exists in $\hat{\mathbf{H}}_y$ for the i^{th} row classification, we place in that entry $\sum_j t_{ij}$. Similarly if a entry exists for the j^{th} column classification, we place in that entry $\sum_i t_{ij}$. Here we use the convention that if the i^{th} row or j^{th} column is not one of the chosen $I + J - 1$ linearly independent indicator variables then corresponding $\beta_{i\cdot}$ or $\beta_{\cdot j}$ is 0.

For the raking and exponential calibrated cell counts estimators, $f(\eta) = e^{\eta}$ and we can similarly calculate $\hat{\mathbf{H}}_y$ using instead

$$s_{ij} = \frac{N}{n} \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) r_{ij}$$

$$t_{ij} = \frac{N}{n} \sum_{k \in \mathcal{S}_{ij}} \exp(\hat{\beta}_{i\cdot} + \hat{\beta}_{\cdot j}) y_k.$$

Here we use the convention that if the i^{th} row or j^{th} column is not one of the chosen $I + J - 1$ linearly independent indicator variables then corresponding $\beta_{i\cdot}$ or $\beta_{\cdot j}$ is 1.

Analogously to (2), let

$$\hat{t}_{\mathbf{x},\mathbf{z}\,f\,\mathbf{V}} = \sum_{k \in \mathcal{S}} d_k\, f(\mathbf{z}_k^T \hat{\beta})\, x_k.$$

$\hat{t}_{\mathbf{x},\mathbf{z}\,f\,\mathbf{V}}$ is a column vector with one entry for each $\mathbf{x}$ variable. Define the $\hat{\mathbf{H}}$ matrix to be

$$\hat{\mathbf{H}} = \frac{\partial \hat{t}_{\mathbf{x},\mathbf{z}\,f\,\mathbf{V}}}{\partial \beta}(\hat{\beta}).$$

$\hat{\mathbf{H}}$ is a matrix with one row for each $\mathbf{x}$ variable and one column for each $\mathbf{z}$ variable.

For the calibrated cell counts and exponential calibrated cell counts estimators the matrix $\hat{\mathbf{H}}$ has dimensions $IJ \times (I + J - 1)$. Each of the rows of $\hat{\mathbf{H}}$ corresponds to a pair (i, j) of row and column classifications. We place s_{ij} in the row corresponding to (i, j) and the columns corresponding to the i^{th} row classification and the j^{th} column classification (whenever these columns exist). All other entries of $\hat{\mathbf{H}}$ are set to zero.

For the calibrated margins and raking estimators the matrix $\hat{\mathbf{H}}$ has dimensions $(I + J - 1) \times (I + J - 1)$. If a row (and hence a column) of $\hat{\mathbf{H}}$ exists for the i^{th} row classification we put $\sum_j s_{ij}$ in the corresponding diagonal entry of $\hat{\mathbf{H}}$. Similarly, if a row and column exist for the j^{th} column classification, we put $\sum_i s_{ij}$ on the diagonal of $\hat{\mathbf{H}}$. We place s_{ij} in the entry whose row corresponds to the i^{th} row classification and whose column corresponds to j^{th} column classification (whenever these exist). We also place s_{ij} in the entry whose column corresponds to the i^{th} row classification and whose row corresponds to j^{th} column classification (again whenever these exist). All other entries of $\hat{\mathbf{H}}$ are set to zero.

Let $\mathbf{B} = \hat{\mathbf{H}}_{\mathbf{y}}^{\mathbf{T}} (\hat{\mathbf{H}}^{\mathbf{T}} \mathbf{V}^{-1} \hat{\mathbf{H}})^{-1} \hat{\mathbf{H}}^{\mathbf{T}} \mathbf{V}^{-1}$ where currently we are using an identity matrix for $\mathbf{V}$. $\mathbf{B}$ has dimensions $1 \times (I + J - 1)$ for the calibrated margins and raking estimators and $1 \times IJ$ for the calibrated cell counts and the exponential calibrated cell counts estimators. In the former cases, we will denote the entries of $\mathbf{B}$ by $b_{i\cdot}$ or $b_{\cdot j}$, and, for the single case when a column or row index does not correspond to one of the $I + J - 1$ independent indicator variables, we will set the corresponding b to zero. In the latter cases, we will denote the entries of $\mathbf{B}$ by b_{ij}. For $k \in \mathcal{S}_{ij}$, let $u_k = w_{ij}(y_k - b_{i\cdot} - b_{\cdot j})$ for the calibrated margins and raking estimators and $u_k = w_{ij}(y_k - b_{ij})$ for the calibrated cell counts and exponential calibrated cell counts estimators.

Essentially Chang and Kott (2008) showed that, asymptotically, the calibrated estimator has the same form as a regression estimator of the form Särndal *et al.* (1992)

equation (6.6.1) where the above $\mathbf{B}$ plays the role of $\mathbf{B}$ in (6.6.1) and the sampling weights d_k are replaced by $d_k\, f(\mathbf{z}_k^T\, \hat{\beta})$. For non replacement designs, they propose to estimate the variance of $\hat{t}_{y,\mathbf{z}\, f\, \mathbf{V}}$ using the analogous changes to Särndal *et al.* (1992) equation (6.6.3).

For simple random sampling, and in the absence of nonresponse or noncoverage, the variance estimator works out to

$$\widehat{\mathbf{V}} = \frac{N^2}{n}(1 - n/N)\, s_u^2 \qquad (15)$$

where s_u^2 is the sample variance of the u_k.

In the presence of nonresponse, if one assumes that the respondents $\mathcal{S}$ are a Poisson sample from the original simple random sample with Poisson probabilities $f(\mathbf{z}^T \beta_0)^{-1}$, the variance estimator becomes

$$\hat{\mathbf{V}} = \frac{N^2}{n}(1 - n/N)\, s_u^2 + \frac{N}{n}\sum_i \sum_j (1 - w_{ij}) \sum_{k \in \mathcal{S}_i j} u_k^2 \qquad (16)$$

where s_u^2 is the sample variance of the u_k. The same formula works for noncoverage where $f(\mathbf{z}^T \beta_0)^{-1}$ represents the combined coverage and response probability in a three stage model in which the covered universe is assumed to be a Poisson sample from the desired universe, the sample is a simple random sample from the covered universe, and the respondents are a Poisson sample from the original sample.

References

Banker, M.D., Rathwell, S. and Majkowski, M. (1992). Two step generalized least squares estimation in the 1991 canadian census. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 764-769.

Chang, T., and Kott, S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 557-571.

Clark, R., and Chambers, R.L. (2008). Adaptive Calibration for Prediction of Finite Population Totals. University of Wollongong (on line working paper).

Falk, G. (2010). *Calibration Adjustment for Nonresponse in Cross-Classified Data*, University of Virginia (dissertation).

Fuller, W.A., Loughin, M.M. and Baker, H.D. (1994). Regression weighting in the Presence of Nonresponse with application to the 1987-1988 Nationalwide Food Consumption Survey. *Survey Methodology*, 20, 75-85.

Kott, S., and Chang, T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, 105, 1265-1275.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Silva, P.L.D.N., and Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 23-32.

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 43-52
Statistics Canada, Catalogue No. 12-001-X

On variances of changes estimated from rotating panels and dynamic strata

Paul Knottnerus and Arnout van Delden [1]

Abstract

Many business surveys provide estimates for the monthly turnover for the major Standard Industrial Classification codes. This includes estimates for the change in the level of the monthly turnover compared to 12 months ago. Because business surveys often use overlapping samples, the turnover estimates in consecutive months are correlated. This makes the variance calculations for a change less straightforward. This article describes a general variance estimation procedure. The procedure allows for yearly stratum corrections when establishments move into other strata according to their actual sizes. The procedure also takes into account sample refreshments, births and deaths. The paper concludes with an example of the variance for the estimated yearly growth rate of the monthly turnover of Dutch Supermarkets.

Key Words: Births; Business surveys; Conditional covariances; Deaths; Overlapping samples; Stratum corrections.

1. Introduction

In many surveys a changing population is repeatedly sampled so that the level and the change in the level of a characteristic between two occasions can be estimated. For example, in many countries a monthly business survey is held to estimate the level of the monthly turnover and the change in that level compared to a month or a year ago; see Konschnik, Monsour and Detlefsen (1985). Another example is the labour force survey in which the population is sampled on a monthly basis to estimate the number of unemployed persons and the unemployment rate. Variance estimation is needed to judge whether the observed changes are statistically significant. Variance estimation is also needed in the design stage of the survey, to determine the optimal sample size and allocation or to determine the optimal estimator.

In repeated surveys, changes are often estimated by using a stratification of the population. Businesses are extremely heterogeneous in terms of size and type of economic activity. Therefore, business surveys are usually designed as a stratified simple random sample selected without replacement (STSRS); see Smith, Pont and Jones (2003). In surveys for households or individuals the sample is usually not stratified because households are less heterogeneous. Some social surveys, such as labour force surveys, however, use poststratification to reduce the variance and bias of the estimator.

In deriving formulas for the variance of an estimated change in a population with dynamic strata, one has to pay attention to three complicating factors. Firstly, the change in a level is the result of two components. One component is due to the change in the population mean of units that remain in the same stratum on both occasions. The other component is caused by the change in the stratum composition between two occasions resulting from births and deaths in the population and from population units that migrate between strata; see Holt and Skinner (1989). Secondly, due to the migration of population units between strata, the estimated mean of stratum h at occasion t may be correlated with the mean of stratum ℓ at occasion $t+1$. Thirdly, another complicating factor is that the population is repeatedly sampled, resulting in partially overlapping samples between two occasions. Different rotating panel designs may be used in business surveys.

Various authors have derived formulas for design-based variance estimators for the estimation of changes. Assuming a large population without births and deaths, Kish (1965) derived an expression for the variance of an estimated change based on overlapping samples. Tam (1984) removed the assumption of a large population. Elaborating on Tam's results, Qualité and Tillé (2008) compare several variance estimators of an estimated change. Wood (2008) generalizes Tam's results for surveys with unequal probabilities. Lowerre (1979) and Laniel (1987) deal with the variance estimation of a change in dynamic populations, but they do not take stratification into account. Hidiroglou, Särndal and Binder (1995) deal with dynamic populations and stratification, but not with changing strata. Nordberg (2000) and Berger (2004) derived formulas for the more complicated situation of a dynamic population with units that move between strata. For the Swedish sampling design Nordberg (2000) derives formulas using inclusion indicators which requires some algebra. Assuming that the size of the overlap of two samples at two different occasions is fixed, Berger (2004) derives his formulas based on Poisson sampling conditional on the sample size per stratum which requires some matrix algebra.

[1] Paul Knottnerus and Arnout van Delden, Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands. E-mail: pkts@cbs.nl and adln@cbs.nl.

In this paper, we derive the expressions for STSRS sampling in a more straightforward manner without assuming that sizes of overlaps are fixed. Furthermore, unlike the Swedish design, the Dutch one doesn't require time-consuming calculations for estimating one of the variance components for a change. In addition, we propose an alternative estimation method for sampling designs with such a non-zero component. In order to clarify the variance estimation procedure, we describe its application to the yearly growth rates of the turnover of Dutch Supermarkets of 4-week periods.

The outline of the paper is as follows. Section 2 briefly describes the Dutch business survey for monthly turnover, including the sampling design. The variance formulas for the estimator of a change are derived in section 3. Section 4 illustrates the variance estimation procedure by comparing the variances of two different estimators for the yearly growth rate of the monthly turnover of Dutch Supermarkets in the period 2003-2004. Section 5 summarizes the main results and conclusions.

2. The sampling design of the Dutch business surveys

Every month Statistics Netherlands estimates the monthly turnover for some of the major SIC codes. The publication includes the 12-month growth rates of the monthly turnover, *i.e.*, the relative change in the monthly level of turnover compared to 12 months ago. Throughout this paper we will refer to this growth rate as the yearly growth rate.

All statistical units or *establishments* are listed in the General Business Register (GBR) that is maintained by Statistics Netherlands. The register is updated each month for births and deaths from administrative sources, while once a year, on December 31, the size category and the type of economic activity (SIC code) are updated. Note that the registration in the GBR may lag behind the changes in the population (births, deaths, size class changes *etc.*). Moreover, the (unknown) deaths in the frame may lead to a biased estimate of the level of the turnover. In order to avoid this kind of bias, it is important to quickly detect and remove deaths from the frame. Deaths detected in the sample may play a role here. However, a further analysis and correction of these errors are beyond the scope of this paper on variance estimation for growth rates. For estimating these variances, we assume that the population units and their characteristics in the register are correct. Likewise, we assume that there is zero non-response among the surveys.

Every first day of the month an STSRS-like sample from the GBR is conducted to estimate the turnover of the current month. In fact, a rotating sample is used. The sample is stratified by size and by type of economic activity. The actual probability of selection depends on size and economic activity. The probability of selection increases with the size of establishment, with the largest establishments being included in the sample with probability 1. For some SICs there are not only survey data available but also data from administrative sources. The units already present in the administrative files are considered as a separate stratum. The estimates from this stratum have a zero variance.

The sample is updated in two ways. Every month the sample is updated to correct for births and deaths in the population. Once a year, in January, 10% of the sample units are replaced and stratum corrections are carried out. We will discuss the monthly and yearly updates in more detail.

2.1 Monthly update

Each month t $(t = 1, 2, ...)$ a fixed proportion f_h of the N_h^t units in stratum U_h^t is sampled $(h = 1, ..., H)$. This results in a sample s_h^t of size $n_h^t = f_h N_h^t$. Hence, the actual number of units in the sample may change from month to month due to births and deaths in the population. Note that apart from minor round-off errors the actual sampling fraction f_h does not depend on month t. In fact, the update procedure for s_h^t in month t is as follows. Define $U_{0h}^{t-1,t}$ as the set of births in stratum h in month $t-1$ and denote its size by $N_{0h}^{t-1,t}$. The number of sampled units from $U_{0h}^{t-1,t}$ in month t is $n_{0h}^{t-1,t} = f_h N_{0h}^{t-1,t}$. In addition, denote the further required difference $n_h^t - n_{0h}^{t-1,t}$ by $n_{h,\mathrm{REQ}}^{t-1,t}$ and define $s_{h,\mathrm{PRE}}^t$ by $s_{h,\mathrm{PRE}}^t = s_h^{t-1} \cap U_h^t$, that is the set of units in s_h^{t-1} that still exist in month t. Let $n_{h,\mathrm{PRE}}^t$ denote the size of $s_{h,\mathrm{PRE}}^t$. When $n_{h,\mathrm{PRE}}^t \geq n_{h,\mathrm{REQ}}^{t-1,t}$, randomly drop the difference, otherwise select the difference from $U_h^t \backslash U_{0h}^{t-1,t} \backslash s_{h,\mathrm{PRE}}^t$. Note that units dropped from the sample in month $t-1$ or earlier may be re-selected in month t.

2.2 Yearly update

Each January, the sample is updated to account for both a re-stratification of the units and a sample replacement of 10%. All sample units of December that still exist in January are stratified according to their actual size, *i.e.*, the number of employees and the SIC-code of January. The size class boundaries themselves remain unchanged. Consequently, the resulting sample from a stratum according to the new January stratification may consist of units with different inclusion probabilities because units move between strata with different sampling fractions.

In order to correct for possibly different inclusion probabilities in stratum ℓ, denote the substratum consisting of units that belonged to stratum h in December and in January to stratum ℓ by $U_{h\ell}^{\mathrm{dec,\,jan}}$ and denote its size by $N_{h\ell}^{\mathrm{dec,\,jan}}$ $(h, \ell = 1, ..., H)$. In analogy with the monthly update procedure define $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$ by $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}} = s_h^{\mathrm{dec}} \cap U_{h\ell}^{\mathrm{dec,\,jan}}$ and let $n_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$ denote the size of $s_{h\ell,\mathrm{PRE}}^{\mathrm{jan}}$. Since the required size of sample $s_{h\ell,\text{REQ}}^{\text{dec, jan}}$ from $U_{h\ell}^{\text{dec, jan}}$ in January is $n_{h\ell,\text{REQ}}^{\text{dec, jan}} = f_\ell N_{h\ell}^{\text{dec, jan}}$, the yearly update of sample $s_{h\ell,\text{PRE}}^{\text{jan}}$ is carried out as follows.

Firstly, when $n_{h\ell,\text{PRE}}^{\text{jan}} \geq n_{h\ell,\text{REQ}}^{\text{dec, jan}}$, randomly drop the difference from $s_{h\ell,\text{PRE}}^{\text{jan}}$. In addition, 10% of the $n_{h\ell,\text{REQ}}^{\text{dec, jan}}$ remaining units in $s_{h\ell,\text{PRE}}^{\text{jan}}$ is replaced by units from $U_{h\ell}^{\text{dec, jan}} \setminus s_{h\ell,\text{PRE}}^{\text{jan}}$ provided that the latter set contains enough units. When there are not enough units available, the number of replaced units is only $N_{h\ell}^{\text{dec, jan}} - n_{h\ell,\text{PRE}}^{\text{jan}}$. Secondly, when $n_{h\ell,\text{PRE}}^{\text{jan}} < n_{h\ell,\text{REQ}}^{\text{dec, jan}}$, select the difference from $U_{h\ell}^{\text{dec, jan}} \setminus s_{h\ell,\text{PRE}}^{\text{jan}}$. Subsequently, an additional replacement of $n_{h\ell,\text{PRE}}^{\text{jan}} - 0.9 n_{h\ell,\text{REQ}}^{\text{dec, jan}}$ units in $s_{h\ell,\text{REQ}}^{\text{dec, jan}}$ takes place when this difference is positive and enough new units are available. This procedure is done for all substrata $h\ell$, including $h = \ell$. Thirdly, similar to the monthly update procedure the number of sampled units in January from substratum $U_{0\ell}^{\text{dec, jan}}$ of new births in stratum ℓ is $n_{0\ell}^{\text{dec, jan}} = f_\ell N_{0\ell}^{\text{dec, jan}}$. In addition, note that this approach can also be followed when class size boundaries or sampling fractions are changed in January.

Apart from the stratum corrections in January, the resulting sample in month t can be considered more or less as a set of SRS samples from the strata U_h^t. When the population and the strata h are stable over the years, the procedure described so far amounts to a standard STSRS sampling design for month t. Therefore, Statistics Netherlands uses the familiar variance formulas for the STSRS sampling design for estimating the variance of the level of the monthly turnover. In the next section we show how the variance for a change of the level can be estimated under such an STSRS assumption.

3. Variance of the yearly growth rate of monthly turnover

3.1 Variance of the yearly growth rate

Let O^t denote the total turnover of all establishments in the population in month t and $g^{t,s}$ the relative change in the level of turnover between months t and s, *i.e.*,

$$g^{t,s} = \frac{O^t}{O^s} - 1 \quad (t > s).$$

For the corresponding estimates it holds by definition that

$$\hat{g}^{t,s} = \frac{\hat{O}^t}{\hat{O}^s} - 1, \tag{1}$$

where a "hat" indicates an estimate; for an estimator we use the same notation. Furthermore, define

$$G^{t,t-12} \equiv \frac{O^t}{O^{t-12}} = 1 + g^{t,t-12}.$$

In order to estimate the variance of the yearly growth rate of the monthly turnover, we use the first-order Taylor series expansion of a ratio of two estimators. That is,

$$\begin{aligned}
&\operatorname{var}(\hat{g}^{t,t-12}) \\
&= \operatorname{var}\left\{\frac{\hat{O}^t}{\hat{O}^{t-12}}\right\} \\
&\approx \frac{\operatorname{var}(\hat{O}^t - G^{t,t-12}\hat{O}^{t-12})}{(O^{t-12})^2} \\
&= \frac{\operatorname{var}(\hat{O}^t) + (G^{t,t-12})^2 \operatorname{var}(\hat{O}^{t-12}) - 2G^{t,t-12}\operatorname{cov}(\hat{O}^{t-12}, \hat{O}^t)}{(O^{t-12})^2}.
\end{aligned} \tag{2}$$

The major problem is the estimation of $\operatorname{cov}(\hat{O}^{t-12}, \hat{O}^t)$. In the next sections we examine this term and its estimation.

3.2 The covariance term of the yearly growth rate

Using the stratified sampling design, we can write $\operatorname{cov}(\hat{O}^{t-12}, \hat{O}^t)$ from (2) as

$$\begin{aligned}
\operatorname{cov}(\hat{O}^{t-12}, \hat{O}^t) &= \operatorname{cov}\left(\sum_{h=1}^{H} N_h^{t-12}\,\bar{o}_h^{t-12}, \sum_{\ell=1}^{H} N_\ell^t\,\bar{o}_\ell^t\right) \\
&= \sum_{h=1}^{H}\sum_{\ell=1}^{H} N_h^{t-12} N_\ell^t \operatorname{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^t),
\end{aligned} \tag{3}$$

where $\bar{o}_h^{t-m}$ stands for the sample mean of the turnover in stratum h in month $t-m$ $(m = 0, 12)$. Note that the stratification of the units in month $t-12$ may differ from that in month t. As we have seen in section 2.2, the standard refreshment of the panel takes place in January. Furthermore, each establishment is allocated to the correct stratum h according to its actual number of employees in January $(h = 1, ..., H)$. To take these design features into account, define

$N_{h\ell}^{t-12,t}$: size of substratum $U_{h\ell}^{t-12,t}$, *i.e.*, the set of units that in month $t-12$ belonged to stratum h and in month t to stratum ℓ $(h, \ell = 1, ..., H)$;

$O_{h\ell}^{t-m}$: the substratum population total of the turnover in $U_{h\ell}^{t-12,t}$ in month $t-m$ $(m = 0, 12)$;

$\bar{O}_{h\ell}^{t-m}$: the substratum population mean of the turnover in $U_{h\ell}^{t-12,t}$ in month $t-m$ [*i.e.*, $\bar{O}_{h\ell}^{t-m} = O_{h\ell}^{t-m}/N_{h\ell}^{t-12,t}$ $(m = 0, 12)$];

$n_{h\ell}^{t-m}$: size of sample $s_{h\ell}^{t-m}$, *i.e.*, the actual sample from $U_{h\ell}^{t-12,t}$ in month $t-m$ $(0 \leq m \leq 12)$;

$o_{h\ell}^{t-m}$: the sample total of the turnover in $s_{h\ell}^{t-m}$ $(m = 0, 12)$;

$\bar{o}_{h\ell}^{t-m}$: the sample mean of the turnover in $s_{h\ell}^{t-m}$ [*i.e.*, $\bar{o}_{h\ell}^{t-m} = o_{h\ell}^{t-m}/n_{h\ell}^{t-m}$ $(m = 0, 12)$];

$n_{h\ell}^{t-12,t}$: number of units in the overlap $s_{h\ell}^{t-12,t} \equiv s_{h\ell}^{t-12} \cap s_{h\ell}^t$;

$\bar{o}_{h\ell,\mathrm{OLP}}^{t-m}$: the sample mean of the turnover in the overlap $s_{h\ell}^{t-12,t}$ in month $t-m$. [*i.e.*, $\bar{o}_{h\ell,\mathrm{OLP}}^{t-m} = o_{h\ell,\mathrm{OLP}}^{t-m} / n_{h\ell}^{t-12,t}$ $(m = 0,12)$].

In addition to the notation in section 2, define the auxiliary stratum 0 for the *births* in months $t-12, ..., t-1$ and likewise, stratum $H+1$ for the *deaths* in that period. Then $\bar{o}_h^{t-12}$ and $\bar{o}_\ell^{t}$ can be written as

$$\bar{o}_h^{t-12} = \sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}$$

$$\bar{o}_\ell^{t} = \sum_{k=0}^{H} \frac{n_{k\ell}^{t}}{n_\ell^{t}} \bar{o}_{k\ell}^{t},$$

respectively $(1 \le h, \ell \le H)$. Consequently, the covariances in (3) can be rewritten as

$$\mathrm{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^{t}) = \mathrm{cov}\left(\sum_{g=1}^{H+1} \frac{n_{hg}^{t-12}}{n_h^{t-12}} \bar{o}_{hg}^{t-12}, \sum_{k=0}^{H} \frac{n_{k\ell}^{t}}{n_\ell^{t}} \bar{o}_{k\ell}^{t} \right) \quad (4a)$$

$$= \frac{1}{n_h^{t-12} n_\ell^{t}} \mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{t}), \quad (4b)$$

where we used $\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{k\ell}^{t} \bar{o}_{k\ell}^{t}) = 0$ $(k \ne h)$ and $\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{t}) = 0$ $(g \ne \ell)$. The latter covariance is zero because

$$\begin{aligned} &\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{t}) \\ &\quad = E\,\mathrm{cov}(n_{hg}^{t-12} \bar{o}_{hg}^{t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{t} \mid n_{hg}^{t-12}, n_{h\ell}^{t}) \\ &\qquad + \mathrm{cov}\{E(n_{hg}^{t-12} \bar{o}_{hg}^{t-12} \mid n_{hg}^{t-12}, n_{h\ell}^{t}), E(n_{h\ell}^{t} \bar{o}_{h\ell}^{t} \mid n_{hg}^{t-12}, n_{h\ell}^{t})\} \\ &\quad = 0 + \bar{O}_{hg}^{t-12}\, \bar{O}_{h\ell}^{t}\, \mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0. \end{aligned}$$

In the last line we also used that for $1 \le g \le H+1$

$$\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0. \quad (5)$$

For a justification and the underlying assumptions of (5), see Appendix A. Moreover, in Appendix A we propose an alternative estimation method when this covariance is non-negligible. The covariance in (4b) can be expressed as

$$\begin{aligned} &\mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{t}) \\ &\quad = E\{\mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \\ &\qquad + \mathrm{cov}\{E(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12} \mid \nu_{h\ell}), E(n_{h\ell}^{t} \bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \quad (6) \end{aligned}$$

where $\nu_{h\ell} = (n_{h\ell}^{t-12}, n_{h\ell}^{t-12,t}, n_{h\ell}^{t})$. The first component on the right-hand side is

$$\begin{aligned} &E\{\mathrm{cov}(n_{h\ell}^{t-12} \bar{o}_{h\ell}^{t-12}, n_{h\ell}^{t} \bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \\ &\quad = E\{n_{h\ell}^{t-12} n_{h\ell}^{t} \mathrm{cov}(\bar{o}_{h\ell}^{t-12}, \bar{o}_{h\ell}^{t} \mid \nu_{h\ell})\} \\ &\quad = E\left\{ n_{h\ell}^{t-12} n_{h\ell}^{t} \left(\frac{n_{h\ell}^{t-12,t} / n_{h\ell}^{t-12}}{n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}. \quad (7) \end{aligned}$$

In the last line we used (26) in Appendix B. Furthermore,

$$S_{h\ell}^{t-12,t} = \frac{1}{N_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{N_{h\ell}^{t-12,t}} (O_{h\ell i}^{t-12} - \bar{O}_{h\ell}^{t-12})(O_{h\ell i}^{t} - \bar{O}_{h\ell}^{t}). \quad (8)$$

The second component on the right-hand side of (6) is equal to $\bar{O}_{h\ell}^{t-12} \bar{O}_{h\ell}^{t} \mathrm{cov}(n_{h\ell}^{t-12}, n_{h\ell}^{t}) = 0$ on account of (5). It therefore follows from (4) and (6) that

$$\begin{aligned} &\mathrm{cov}(\bar{o}_h^{t-12}, \bar{o}_\ell^{t}) = \\ &\quad E\left\{ \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_h^{t-12} n_\ell^{t}} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) S_{h\ell}^{t-12,t} \right\}. \quad (9) \end{aligned}$$

3.3 Estimation of the covariance term of the yearly growth rate

Expression (9) can be estimated from the overlapping sample $s_{h\ell}^{t-12,t}$ by

$$\hat{\mathrm{cov}}(\bar{o}_h^{t-12}, \bar{o}_\ell^{t}) = \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_h^{t-12} n_\ell^{t}} \left(\frac{n_{h\ell}^{t-12,t}}{n_{h\ell}^{t-12} n_{h\ell}^{t}} - \frac{1}{N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell,\mathrm{OLP}}^{t-12,t}, \quad (10)$$

where

$$\hat{S}_{h\ell,\mathrm{OLP}}^{t-12,t} = \frac{1}{n_{h\ell}^{t-12,t} - 1} \sum_{i=1}^{n_{h\ell}^{t-12,t}} (o_{h\ell i}^{t-12} - \bar{o}_{h\ell,\mathrm{OLP}}^{t-12})(o_{h\ell i}^{t} - \bar{o}_{h\ell,\mathrm{OLP}}^{t}).$$

Note that (10) is unbiased for estimating (9) because

$$E(\hat{S}_{h\ell,\mathrm{OLP}}^{t-12,t} \mid \nu_{h\ell}) = S_{h\ell}^{t-12,t}.$$

Although (10) results in reasonable estimates for sufficiently large $n_{h\ell}^{t-12,t}$, a disadvantage of the covariance estimator $\hat{S}_{h\ell,\mathrm{OLP}}^{t-12,t}$ in (10) is that for small $n_{h\ell}^{t-12,t}$ it may lead to a negative estimate of $\mathrm{var}(\hat{O}^{t} - G^{t,t-12} \hat{O}^{t-12})$ in the numerator of (2). Recall that this variance is estimated by

$$\begin{aligned} \hat{\mathrm{var}}(\hat{O}^{t} - G^{t,t-12} \hat{O}^{t-12}) &= \hat{\mathrm{var}}(\hat{O}^{t}) + (\hat{G}^{t,t-12})^2 \hat{\mathrm{var}}(\hat{O}^{t-12}) \\ &\quad - 2\hat{G}^{t,t-12} \hat{\mathrm{cov}}(\hat{O}^{t}, \hat{O}^{t-12}). \quad (11) \end{aligned}$$

Therefore, we propose an alternative estimator to $\hat{S}_{h\ell,\mathrm{OLP}}^{t-12,t}$ in (10). Define the standard deviations

$$\hat{S}_{h\ell}^{t-m} = \sqrt{\frac{1}{n_{h\ell}^{t-m} - 1} \sum_{i=1}^{n_{h\ell}^{t-m}} (o_{h\ell i}^{t-m} - \bar{o}_{h\ell}^{t-m})^2} \quad (m = 0, 12).$$

We propose the following modified estimator for $S_{h\ell}^{t-12,t}$

$$\hat{S}_{h\ell}^{t-12,t} = \hat{\rho}_{h\ell,\mathrm{OLP}}^{t-12,t} \hat{S}_{h\ell}^{t-12} \hat{S}_{h\ell}^{t}, \quad (12)$$

where $\rho_{h\ell}^{t-12,t}$ is the correlation between the variables o^{t} and o^{t-12} in $U_{h\ell}^{t-12,t}$ and $\hat{\rho}_{h\ell,\mathrm{OLP}}^{t-12,t}$ is its estimate from $s_{h\ell}^{t-12,t}$. According to (10) and (12) covariance (3) can be estimated by

$$\begin{aligned} &\hat{\mathrm{cov}}(\hat{O}^{t-12}, \hat{O}^{t}) = \\ &\quad \sum_{h=1}^{H} \sum_{\ell=1}^{H} \frac{N_h^{t-12} N_\ell^{t}}{n_h^{t-12} n_\ell^{t}} n_{h\ell}^{t-12,t} \left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}} \right) \hat{S}_{h\ell}^{t-12,t}. \quad (13) \end{aligned}$$

For the estimate $\hat{\rho}_{h\ell,\text{OLP}}^{t-12,t}$, $\hat{\rho}_{h\ell,\text{OLP}}^{t-12,t} \le 1$ always holds whereas using (10) may lead implicitly to an estimated correlation larger than 1 and a possibly negative outcome of (11). See the next section for an example. In all applications met so far, negative outcomes of (11) could be explained by the fact that unlike (12) use of (10) leads implicitly to an estimated correlation larger than 1. This is in line with the findings of Berger (2004, page 462) that an overestimation of the correlation between $\hat{O}^{t-12}$ and $\hat{O}^{t}$ may lead to a serious underestimation of the variance of a change. Nevertheless, in some extraordinary circumstances, the use of (12) might lead to a negative outcome of (11) as well. Sufficient conditions that the use of (12) leads to a nonnegative variance estimator with probability 1 are available from the authors upon request. For a general review of variance estimation methods in business surveys, see Brodie (2003).

Applying (12), a special problem may arise when $n_{h\ell}^{t} = 1$ or $n_{h\ell}^{t-12} = 1.$ In order to evaluate the required sample variances, one may borrow the sample variance from a related substratum or from the same substratum in an earlier month. Alternatively, one may impute a variance when it emerges from the data that there is a relationship of the form $S_{h\ell}^{2} \approx \sigma^{2} \bar{O}_{h\ell}^{\beta};$ see Särndal, Swensson and Wretman (1992, page 461). In addition, the corresponding covariance term might be ignored when its (expected) contribution to the total variance is small. This is often the case when the sampling fractions in strata h and ℓ are small, that is in strata with relatively small units and, consequently, with small variances compared to the strata with larger units. Similar remarks apply to the imputed $\rho_{h\ell}^{t-12,t}$ when $n_{h\ell}^{t-12,t} \le 2$ and $n_{h\ell}^{t-m} \ge 2$ $(m = 0, 12).$ Since the $\rho_{h\ell}^{t-12,t}$ are often fairly high, this seems to be a viable way. In the example given in section 4 the $\rho_{h\ell}^{t-12,t}$ have an overall mean of 0.90 and a variance of 0.0074 so that the impact of the imputed $\rho_{h\ell}^{t-12,t}$ on the final results is likely to be moderate.

Furthermore, note that when $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12),$ the corresponding covariance term in (13) can be neglected without affecting its unbiasedness, provided that the remaining $S_{h\ell}^{t-12,t}$ are estimated in an unbiased way. Under this assumption such a term with $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12)$ can be neglected because the expectation of

$$n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t} \qquad (14)$$

from (13) is equal to

$$E\left[E\left\{n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right)\hat{S}_{h\ell}^{t-12,t} \mid \nu_{h\ell}\right\}\right] =$$

$$E\left\{n_{h\ell}^{t-12,t}\left(1 - \frac{n_{h\ell}^{t-12} n_{h\ell}^{t}}{n_{h\ell}^{t-12,t} N_{h\ell}^{t-12,t}}\right) S_{h\ell}^{t-12,t}\right\},$$

and the expectation on the right-hand side is the parameter to be estimated. Moreover, when $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12)$ and consequently $n_{h\ell}^{t-12,t} = 0,$ the outcome of (14) is zero and the estimator $\hat{S}_{h\ell}^{t-12,t}$ for $S_{h\ell}^{t-12,t}$ becomes irrelevant. Therefore, ignoring such a term when $n_{h\ell}^{t-m} = 0$ $(m = 0$ or $m = 12)$ does not affect the expectations of (13) and (14).

3.4 A comparison with Nordberg's results

Using the standard formalism of inclusion indicators δ_{hi}^{t} for each stratum, Nordberg (2000) derives a different expression for the first component in (6). However, it can be shown after some algebra that our expression (9) is equivalent to Nordberg's (3.4); a proof is available from the authors upon request. In addition, Nordberg derives a nonzero expression for the second component in (6), *i.e.*, the covariance between the two corresponding conditional expectations. Note that the Swedish sampling design is somewhat different from ours.

According to Nordberg (2000, page 370) the estimation of the second component for the Swedish sampling design requires a computer-intensive procedure which includes simulation of the sampling mechanism. However, since all $n_{h\ell}^{t}$, $n_{h\ell}^{t-12}$ and $n_{h\ell}^{t-12,t}$ are ancillary statistics, an alternative might be to condition on these statistics so that the second component can be ignored. Recall that a statistic is called ancillary when its marginal distribution doesn't depend on the target parameters to be estimated; see Cox and Hinkley (1974, pages 31-35). Such an alternative approach without the second component is to be recommended especially when $\hat{g}_{\text{STSRS}}^{t-12,t} \approx \hat{g}_{\text{PS,sub}}^{t-12,t}$ where $\hat{g}_{\text{PS,sub}}^{t-12,t}$ is the poststratified estimator based on the substrata $h\ell$. However, when the difference between $\hat{g}_{\text{STSRS}}^{t-12,t}$ and $\hat{g}_{\text{PS,sub}}^{t-12,t}$ is non-negligible, the calculation of the unconditional variance seems to be indispensable, including the estimation of the second component according to Nordberg. For a different approach to the estimation problem of the second component, see Appendix A.

For a justification of the use of a conditional (co)variance, see Holt and Smith (1979). An important advantage of the conditional (co)variance is that the corresponding confidence interval has better coverage properties than the one based on the unconditional variance. Denote the standard conditional 95% confidence interval for an arbitrary parameter θ by $(\hat{\theta}_{l}, \hat{\theta}_{u} \mid \nu)$ where ν denotes the vector consisting of all (ancillary) statistics involved in the conditional (co)variances. Then under the normality assumption and some mild conditions it holds that the actual 95% confidence level (CL) equals the nominal confidence level because

$$\begin{aligned} CL &= \textstyle\sum_{\nu \in \Omega_\nu} P(\nu) P(\hat{\theta}_l < \theta < \hat{\theta}_u | \nu) \\ &= 0.95 \textstyle\sum_{\nu \in \Omega_\nu} P(\nu) = 0.95, \end{aligned}$$

where Ω_ν stands for the set of all possible outcomes of the random vector ν. When unconditional (co)variances are used, the confidence intervals thus obtained may be quite inaccurate for a given sample allocation. Moreover, when averaged over all allocations *CL* may differ from 0.95; for an example, see Knottnerus (2003, pages 133-135). Note that in the planning stage before the sample is drawn, unconditional variances are always useful for examining Kish's design effect for a comparison of different sampling designs. In addition, note that for evaluating a conditional confidence interval for $g^{t-12,t}$ the underlying variances of $\hat{O}^{t-m}_{\text{PS,sub}}$ should also be taken conditional on the $\nu_{h\ell}$ $(m = 0, 12)$.

Finally, the unbiased estimator proposed by Nordberg [2000, Equation (3.9)] for the first component in (6) is quite different than those described in the previous subsection. In fact, his estimator is based on the following procedure for estimating the covariance term $S^{t-12,t}_{h\ell}$. Firstly, estimate the underlying quantity $\sum_{i=1}^{N^{t-12,t}_{h\ell}} O^{t-12}_{h\ell i} O^{t}_{h\ell i}$ from the overlap $s^{t-12,t}_{h\ell}$. Secondly, estimate the corresponding turnover means from $s^{t-12}_{h\ell}$ and $s^{t}_{h\ell}$, respectively. Since the components thus estimated stem from different samples, a negative outcome of (11) cannot always be avoided. For a small example with real data, see the following section. In the remainder Nordberg's underlying estimator for $S^{t-12,t}_{h\ell}$ is denoted by $\hat{S}^{t-12,t}_{h\ell\text{NBG}}$. A derivation of the explicit expression for $\hat{S}^{t-12,t}_{h\ell\text{NBG}}$ is available from the authors upon request.

4. An application to the change of turnover in Dutch Supermarkets

4.1 Two estimators for the yearly change of turnover

For the impact on the variance estimators it is important to know that in January the turnover is estimated twice. The first estimate, denoted by $\hat{O}^{\text{janO}}$ (with *O* for *old*), is made before the yearly sample update and is used to estimate the monthly change of the turnover in January compared to that in December. The second estimate, denoted by $\hat{O}^{\text{janN}}$ (with *N* for *new*), is made after the yearly sample update and is used to estimate the monthly change of the turnover in February compared to January. This procedure implies that units of the old sample as well as those of the new sample receive a questionnaire in January.

Unlike estimator (1) the actual estimator used by Statistics Netherlands for the yearly change in the monthly turnover is based on a chain of 12 monthly changes in turnover

$$\begin{aligned} \hat{G}^{t,t-12}_{\text{act}} = 1 + \hat{g}^{t,t-12}_{\text{act}} &= \prod_{j=0}^{11} (1 + \hat{g}^{\,t-j,t-j-1}) \\ &= \frac{\hat{O}^{\,t}}{\hat{O}^{\,t-1}} \times \frac{\hat{O}^{\,t-1}}{\hat{O}^{\,t-2}} \times \,\ldots\, \times \frac{\hat{O}^{\,\text{feb}}}{\hat{O}^{\,\text{janN}}} \\ &\quad \times \frac{\hat{O}^{\,\text{janO}}}{\hat{O}^{\,\text{dec}}} \times \,\ldots\, \times \frac{\hat{O}^{\,t-11}}{\hat{O}^{\,t-12}} \\ &= \frac{\hat{O}^{\,t}}{\hat{O}^{\,t-12}} \times \frac{\hat{O}^{\,\text{janO}}}{\hat{O}^{\,\text{janN}}} \quad (t \neq \text{jan}). \end{aligned} \tag{15}$$

In this section we will compare the variances of estimators (1) and (15). Similar to (2) the variance formulas for $\hat{g}^{t,t-12}_{\text{act}}$ can be derived by a first-order Taylor series expansion.

4.2 Description of the data

The calculations for the variances and confidence intervals in this example are based on turnover data of Dutch Supermarkets of 4-week periods in 2003 and 2004 (*i.e.*, $t = 1, \ldots, 26$). Hence, there are 13 observations in one year and, consequently, we use slightly adjusted symbols such as $g^{t,t-13}$ in the remainder of this section.

The population consists of about 3,500 establishments. The turnover data stem from a stratified sample and administrative files. A gross STSRS sample of about 900 units stratified by size is drawn from the full list of population units of the GBR that includes the units of the administrative files as well. Establishments with 50 or more *employees* are included with probability 1. The other establishments are sampled with decreasing inclusion probability from 1:2 (20-49 employees per establishment) to 1:40 in the smallest size (1 employee per establishment). The administrative files contain about 950 units, present in all size classes. About 500 of the 900 units in the gross sample were already present in the administrative files, but they do not receive a questionnaire. Thus, the net sample contains about 400 units. In fact, the sample size for each stratum in this specific example is random. However, as explained in subsection 3.4, we estimate all (co)variances conditional on the n_h in such a case. Data from units within the administrative files are put into a separate stratum with the sampling fraction being unity.

4.3 Results

Table 1 gives the yearly growth rates and their 95% margins for $t = 16, \ldots, 24$. It emerges that the 95% margins for the estimated growth rates $\hat{g}^{t,t-13}_{\text{act}}$, currently used by Statistics Netherlands, vary between 0.8 and 1.0 (per cent point). For example, in the first period $(t = 16)$ the 95% confidence interval for the yearly growth rate is -1.3 to 0.7 per cent. As expected, the 95% margins for the more complicated estimator $\hat{g}^{\,t,\,t-13}_{\text{act}}$ are close to those for the simpler $\hat{g}^{\,t,\,t-13}$ from (1). The 95% margins of $\hat{g}^{\,t,\,t-13}$ vary between 0.7 and 1.0 (per cent point). The estimator for the growth rate to be preferred is $\hat{g}_{\text{act}}^{t,\,t-13}$ as it corrects for the yearly sample update in January. The estimation of its variance, however, can be simplified by using the variance estimator described in section 3 rather than the more laborious expression for $\hat{\text{var}}(\hat{g}_{\text{act}}^{t,\,t-13})$.

Table 1
Estimated growth rates with 95% margins

t	$\hat{g}_{\text{act}}^{t,\,t-13} \times 100\%$	$\hat{g}^{\,t,\,t-13} \times 100\%$
16	-0.3 (± 1.0) [1]	-0.4 (± 1.0)
17	-3.7 (± 1.0)	-3.8 (± 0.9)
18	1.6 (± 1.0)	1.5 (± 0.9)
19	-2.2 (± 0.9)	-2.3 (± 0.9)
20	0.5 (± 0.8)	0.4 (± 0.7)
21	-1.7 (± 0.8)	-1.8 (± 0.7)
22	-2.2 (± 0.8)	-2.3 (± 0.7)
23	0.0 (± 0.8)	-0.1 (± 0.7)
24	-2.3 (± 0.9)	-2.4 (± 0.9)

[1] The 95% margins are given between parentheses.

As described in section 3, we have used the estimated correlation $\hat{\rho}_{h\ell,\text{OLP}}^{t-13,t}$ from the overlap $s_{h\ell}^{t-13,t}$ to estimate covariance $S_{h\ell}^{t-13,t}$ in order to avoid negative outcomes of (11). Knottnerus and Van Delden (2006) evaluated the bias of $\hat{\rho}_{h\ell,\text{OLP}}^{t-13,t}$ for the Dutch Supermarket data and found a small underestimation of $\hat{\rho}_{h\ell,\text{OLP}}^{t-13,t}$ resulting in a minor, less than 5%, overestimation of $\text{var}(\hat{g}^{\,t,t-13})$.

The use of estimator $\hat{S}_{h\ell,\text{OLP}}^{t-13,t}$ in (10) may give a negative outcome of (11) and an estimated correlation $\hat{\rho}_{h\ell}^{t-13,t}$ larger than 1. For example, consider the specific population with $N = 50$ and $H = 1$ consisting of the units of substratum $h\ell = 65.$ From the panel data for this population, given in Table 2 for $t = 3$ and $t = 16,$ we obtain after some calculations $\hat{S}^{\,t-13} = 410.7,$ $\hat{S}^{\,t} = 394.3$ and $\hat{G}^{\,t,t-13} = 1.028.$ Note that in the remainder of this section the subscript $h\ell = 11$ is omitted in the symbols because there is only one stratum. Table 3 gives, for three different approaches, some additional estimates for the panel data in Table 2. For example, using $\hat{S}_{\text{OLP}}^{t-13,t}$ in (10) results in an estimated correlation $\hat{\rho}^{\,t-13,t} = 1.39.$ This then yields a negative variance estimate from (11) of minus 2.2 million. Likewise, for the same data the alternative estimator $\hat{S}_{\text{NBG}}^{t-13,t}$ of $S^{\,t-13,t}$ based on Nordberg (2000) results in minus 36.9 million as outcome of (11) because the corresponding estimate $\hat{\rho}_{\text{NBG}}^{t-13,t}$ becomes 1.64. In contrast, using the correlation estimated from the overlapping sample $s^{t-13,t}$ according to (12) yields $\hat{\rho}_{\text{OLP}}^{t-13,t} = 0.9997$ and the positive variance estimate from (11) becomes 52.1 million. In addition, for the panel data in Table 2 the outcome of Nordberg's estimator (3.9) for the covariance between $\hat{O}^{\,t-13}$ and $\hat{O}^{\,t}$ is 111.1 million whereas covariance estimator (13) proposed here yields 67.8 million.

Table 2
Panel data[1] from a population with N = 50 and H = 1

period	turnover per unit (in thousand euros)				
	1	2	3	4	5
$t = 3$		493.9	264.3	1,179.1	380.0
$t = 16$	475.3	472.0	267.0	1,169.0	

[1] Actually, the panel data belonged to substratum $h\ell$ = 65.

5. Conclusions

The variance formulas obtained in this paper are useful for calculating the variance of an estimated yearly growth rate of monthly turnover. The use of (13) as an estimator for $\text{cov}(\hat{O}^{\,t-12}, \hat{O}^{\,t})$ results in reasonable estimates of the covariance of change in particular. The variance estimation procedure allows for rotating panels, births, deaths, and units that migrate between strata.

Furthermore, we recommend estimating a population covariance according to (12) based on the corresponding correlation estimated from the overlap and on the corresponding variances estimated from the larger separate samples. This may help to avoid a serious underestimation or a negative outcome of the variance estimator for the yearly growth rate. The resulting estimated covariances are only slightly biased.

Table 3
Estimates from three different approaches

approach		parameters to be estimated		
		$S^{\,t-13,t}$	$\rho^{t-13,t}$	$\text{var}(\hat{O}^{\,t} - G^{\,t,t-13}\hat{O}^{\,t-13})$
Nordberg (2000)	estimator	$\hat{S}_{\text{NBG}}^{t-13,t}$	$\dfrac{\hat{S}_{\text{NBG}}^{t-13,t}}{\hat{S}^{\,t-13}\hat{S}^{\,t}}$	Eq. (11)
	result	265.2×10^3	1.64	-36.9×10^6
Eq. (10)	estimator	$\hat{S}_{\text{OLP}}^{t-13,t}$	$\dfrac{\hat{S}_{\text{OLP}}^{t-13,t}}{\hat{S}^{\,t-13}\hat{S}^{\,t}}$	Eq. (11)
	result	225.0×10^3	1.39	-2.2×10^6
Eq. (12)	estimator	$\hat{\rho}_{\text{OLP}}^{t-13,t}\,\hat{S}^{\,t-13}\hat{S}^{\,t}$	$\hat{\rho}_{\text{OLP}}^{t-13,t}$	Eq. (11)
	result	161.9×10^3	1.00[1]	52.1×10^6

[1] In fact, 0.9997.

For the sampling design of the Dutch Supermarkets the second covariance term in (6) is negligible due to the fact that $n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ is fixed. In contrast, for the SAMU design in Sweden this term is non-negligible and its estimation is time-consuming; the word SAMU (SAMordnade Urval) is a Swedish acronyme for coordinated samples. In Appendix A we propose an alternative method for estimating this covariance. However, under the condition that $\hat{g}^{\,t,t-12} \approx \hat{g}_{\mathrm{PS,sub}}^{\,t,t-12}$ it suffices in our opinion to only use the first covariance. This simplifies the estimation procedure considerably. Moreover, under the normality assumption the conditional confidence interval has better coverage properties compared to the unconditional interval.

The example of the Dutch Supermarkets shows one of the practical applications of the variance formulas: determining which estimator has the smallest variance. The results confirm that the variance of the simple estimator $\hat{g}^{\,t,t-13}$ is close to that of $\hat{g}_{\mathrm{act}}^{\,t,t-13}$ from section 4 which corrects for the sample refreshment in January. Hence, for the Dutch Supermarkets $\hat{\mathrm{var}}(\hat{g}^{\,t,t-13})$ might be used for estimating $\mathrm{var}(\hat{g}_{\mathrm{act}}^{\,t,t-13})$. For branches with another SIC code it needs to be checked whether $\mathrm{var}(\hat{g}_{\mathrm{act}}^{\,t,t-13}) \approx \mathrm{var}(\hat{g}^{\,t,t-13})$ since the impact of the refreshment in January need not be negligible.

Acknowledgements

The views expressed in the paper are those of the authors and do not necessarily reflect the policy of Statistics Netherlands. The authors would like to thank the Associate Editor and two anonymous referees for their useful comments and suggestions, which have led to a significant improvement of this paper.

Appendix A

Justification of (5)

Firstly, consider the case of strata without births and deaths. Apart from the yearly update in January, there are now no monthly updates. Hence, $n_{h\ell}^{t} = n_{h\ell,\mathrm{REQ}}^{\mathrm{dec,\,jan}}$ is fixed from which (5) follows. This case applies to the Dutch Supermarkets because that population has been quite stable over the years. Secondly, in case of births and deaths among the strata we can write $n_{h\ell}^{t}$ as

$$n_{h\ell}^{t} = n_{\ell}^{t} - n_{0\ell}^{t-12,t} - \sum_{k \neq h}^{H} n_{k\ell}^{t}, \qquad (16)$$

where $n_{0\ell}^{t-12,t}$ or, for short, $n_{0\ell}^{t}$ stands for the number of births in months $t-12, \ldots, t-1$ among s_{ℓ}^{t}. Because the sampling procedure among the new births after month $t-12$ is independent of the n_{hg}^{t-12}, the random variables $n_{0\ell}^{t-12,t}$ and n_{hg}^{t-12} have a zero covariance. Furthermore, using $\mathrm{cov}(n_{hg}^{t-12}, n_{k\ell}^{t}) = 0$ for $k \neq h$, it is seen from (16) that $\mathrm{cov}(n_{hg}^{t-12}, n_{h\ell}^{t}) = 0$ $(h = 1, \ldots, H)$.

In fact, it is assumed so far that the distribution of $n_{k\ell}^{t}$ $(k \neq h)$ can be described by a hypergeometric distribution with parameters $(N_{\ell}^{t}, N_{k\ell}^{t-12,t}, n_{\ell}^{t})$ irrespective of the values of the n_{hg}^{t-12}. A similar remark applies to $n_{0\ell}^{t-12,t}$. However, it can be argued that in practice these assumptions lead to a minor, second-order error in the variance formulas. In order to trace this error, we assume for simplicity's sake and without loss of generality that (*i*) births and deaths do not migrate between strata, (*ii*) there are no deaths among the births, (*iii*) $n_{0h}^{t} = f_h N_{0h}^{t}$ is fixed, (*iv*) after their first month in the population births are irrelevant for the monthly updates during the rest of the study period and (*v*) deaths are not selected in or removed from the sample by the monthly updates; so a third-order error is still ignored. Under these assumptions we now look more closely at the second covariance component for $\ell = h$, say $C_{hh,\,\mathrm{sec}}$, from (4a). In analogy with (6) $C_{hh,\,\mathrm{sec}}$ can be written as

$$\begin{aligned} C_{hh,\,\mathrm{sec}} &\equiv \frac{1}{n_h^{t-12} n_h^{t}} \\ &\qquad \mathrm{cov}\left\{ E\left(\sum_{g=1}^{H+1} n_{hg}^{t-12}\, \bar{o}_{hg}^{\,t-12} \middle| \nu_h \right), E\left(\sum_{k=0}^{H} n_{kh}^{t}\, \bar{o}_{kh}^{\,t} \middle| \nu_h \right) \right\} \\ &= \frac{1}{n_h^{t-12} n_h^{t}} \sum_{g=1}^{H+1} \sum_{k=1}^{H} \bar{O}_{hg}^{\,t-12}\, \bar{O}_{kh}^{\,t}\, \mathrm{cov}(n_{hg}^{t-12}, n_{kh}^{t}), \qquad (17) \end{aligned}$$

where $\nu_h = (n_{h1}^{t-12}, \ldots, n_{h,H+1}^{t-12}, n_{1h}^{t}, \ldots, n_{Hh}^{t})$. Note that under the above assumptions $C_{h\ell,\mathrm{sec}} = 0$ for $\ell \neq h$.

To estimate the covariances in (17), consider the formula for the conditional expectation of y given $x = x_0$ when y and x follow a bivariate normal distribution. That is, in standard notation,

$$E(y \mid x_0) = \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}(x_0 - \mu_x).$$

In addition, for a given change Δx_0 of x the conditional expectation of the change of y is equal to $E(\Delta y \mid \Delta x_0) = \sigma_{yx} \Delta x_0 / \sigma_x^2$ or, equivalently,

$$\sigma_{yx} = \frac{E(\Delta y \mid \Delta x_0)}{\Delta x_0} \sigma_x^2. \qquad (18)$$

So for estimating, for instance, $\mathrm{cov}(n_{h,H+1}^{t-12}, n_{kh}^{t})$ in (17) under normality it suffices to evaluate the expected effect on $y = n_{kh}^{t}$ caused by a change of the future deaths $x = n_{h,H+1}^{t-12}$ in s_h^{t-12}.

Let $\Delta n_{h,H+1}^{t-12}$ denote an additional (positive) change of these deaths in s_h^{t-12}. Define $p_{h,H+1}^{\mathrm{jan},t}$ by $p_{h,H+1}^{\mathrm{jan},t} = N_{h,H+1}^{\mathrm{jan},t} / N_{h,H+1}^{t-12}$ where $N_{h,H+1}^{\mathrm{jan},t}$ is the number of deaths in stratum h between January and month t. Also, $p_{hg}^{t-12} = N_{hg}^{t-12,t} / N_h^{t-12}$ $(g = 1, \ldots, H+1)$. Using assumption (*v*), the expected number of additional deaths in the sample of January before the refreshment can be estimated by $p^{\text{jan},t}_{h,\,H+1}\Delta n^{t-12}_{h,\,H+1}$. Subsequently, the expected number of additional deaths in the sample after the refreshment can be estimated by

$$\begin{aligned}&\gamma^{\text{jan}}_{\text{red}}\; p^{\text{jan},t}_{h,H+1}\,\Delta n^{t-12}_{h,H+1};\\ &\gamma^{\text{jan}}_{\text{red}} = (0.9 - f_h)/(1 - f_h),\end{aligned} \qquad (19)$$

where $\gamma^{\text{jan}}_{\text{red}}$ is the reduction factor due to the refreshment in January. For the derivation of (19), see the end of this appendix. The corresponding monthly updates between January and month t due to these additional deaths in the sample from stratum h lead to the following estimate of the expected increase of *incoming* units n^t_{kh} from stratum k $(k \neq h)$ in the sample of month t

$$E(\Delta n^t_{kh} \mid \Delta n^{t-12}_{h,H+1}) = \gamma^{\text{jan}}_{\text{red}}\; p^{\text{jan},t}_{h,H+1}\,\Delta n^{t-12}_{h,H+1}\; p^t_{kh}, \qquad (20)$$

where $p^t_{kh} = N^{t-12,t}_{kh}/(N^t_h - N^t_{0h})$. Recall from subsection 2.1 that an update in month s occurs only when $d^{s-1}_h \neq f_h D^{s-1}_h$, where D^s_h (d^s_h) stands for the number of deaths in U^s_h (s^s_h), and that $n^t_{kh} = f_h N^{t-12,t}_{kh}$ is fixed when $N^{\text{jan},t}_{h,H+1} = 0$ $(k \neq h)$. Furthermore, note that births are excluded in the definition of p^t_{kh} in (20) because of assumption (*iv*).

Next, define for $m = 0, 12$

$$\begin{aligned}
\bar{O}^{t-m}_h &= \frac{1}{N^{t-m}_h}\sum_{i=1}^{N^{t-m}_h} O^{t-m}_{hi};\\
(S^{t-m}_h)^2 &= \frac{1}{N^{t-m}_h - 1}\sum_{i=1}^{N^{t-m}_h}(O^{t-m}_{hi} - \bar{O}^{t-m}_h)^2;\\
p^{t-12}_{h,\le H} &= 1 - p^{t-12}_{h,H+1};\\
p^t_{in,h} &= 1 - p^t_{hh};\\
\bar{O}^{t-12}_{h,\le H} &= \sum_{g=1}^{H}\frac{p^{t-12}_{hg}}{p^{t-12}_{h,\le H}}\,\bar{O}^{t-12}_{hg};\\
\bar{O}^t_{in,h} &= \sum_{\substack{k=1\\k\neq h}}^{H}\frac{p^t_{kh}}{p^t_{in,h}}\,\bar{O}^t_{kh}.
\end{aligned}$$

Now using (18) and (20), we obtain for $k \neq h$ the following covariance approximation

$$\begin{aligned}
&\text{acov}(n^{t-12}_{h,H+1}, n^t_{kh})\\
&\quad = \frac{E(\Delta n^t_{kh} \mid \Delta n^{t-12}_{h,H+1})}{\Delta n^{t-12}_{h,H+1}}\,\text{var}(n^{t-12}_{h,H+1})\\
&\quad \approx \gamma^{\text{jan}}_{\text{red}}\; p^{\text{jan},t}_{h,H+1}\, p^t_{kh}\, n^{t-12}_h\, p^{t-12}_{h,H+1}(1 - p^{t-12}_{h,H+1})(1 - f_h)\\
&\quad = n^{t-12}_h\, p^t_{kh} A_h / p^t_{in,h};
\end{aligned} \qquad (21)$$

$$A_h = \gamma^{\text{jan}}_{\text{red}}\; p^t_{in,h}\; p^{\text{jan},t}_{h,H+1}\; p^{t-12}_{h,H+1}(1 - p^{t-12}_{h,H+1})(1 - f_h),$$

where, for simplicity, we omitted the term $N^{t-12}_h/(N^{t-12}_h - 1)$ in the second line. Because n^{t-12}_h is fixed, it holds that $\text{cov}(n^{t-12}_{h,H+1}, n^t_{kh}) = -\text{cov}(n^{t-12}_{h1} + \ldots + n^{t-12}_{hH}, n^t_{kh})$. Hence, in analogy with the multihypergeometric distribution we can use for $1 \le g \le H$ and $k \neq h$ the following relationship for an approximation of $\text{cov}(n^{t-12}_{hg}, n^t_{kh})$

$$\text{acov}(n^{t-12}_{hg}, n^t_{kh}) = -\frac{p^{t-12}_{hg}}{p^{t-12}_{h,\,\le H}}\,\text{acov}(n^{t-12}_{h,\,H+1}, n^t_{kh}) \qquad (22a)$$

$$= -\,n^{t-12}_h \frac{p^{t-12}_{hg}}{p^{t-12}_{h,\,\le H}}\frac{p^t_{kh}}{p^t_{in,\,h}} A_h, \qquad (22b)$$

where (21) is used as well. Alternatively, note that

$$\begin{aligned}\text{cov}(n^{t-12}_{h,\,H+1}, n^t_{kh}) &= -\text{cov}(n^{t-12}_{h,\,\le H}, n^t_{kh})\\ &= -\sum_{g\le H}\sum_{i\in U^{t-12,\,t}_{hg}}\text{cov}(\delta^{t-12}_{hgi}, n^t_{kh}),\end{aligned}$$

where

$$\delta^{t-12}_{hgi} = \begin{cases}1 & \text{if } i^{\text{th}} \text{ unit in } U^{t-12,\,t}_{hg} \text{ is included in sample } s^{t-12}_h\\ 0 & \text{otherwise.}\end{cases}$$

Hence, by symmetry, $\text{cov}(\delta^{t-12}_{hgi}, n^t_{kh}) = -\text{cov}(n^{t-12}_{h,\,H+1}, n^t_{kh})/N^{t-12,\,t}_{h,\,\le H}$ from which (22a) follows $(1 \le g \le H)$. Likewise, for $k = h$ we obtain from (21) and (22b)

$$\begin{aligned}\text{acov}(n^{t-12}_{h,\,H+1}, n^t_{hh}) &= -n^{t-12}_h A_h;\\ \text{acov}(n^{t-12}_{hg}, n^t_{hh}) &= n^{t-12}_h\, p^{t-12}_{hg} A_h / p^t_{h,\,\le H},\end{aligned} \qquad (23)$$

respectively $(1 \le g \le H)$. Now substituting (21)-(23) into (17), we get the approximation

$$C_{hh,\,\text{sec}} = A_h(\bar{O}^{t-12}_{h,\,H+1} - \bar{O}^{t-12}_{h,\,\le H})(\bar{O}^t_{in,\,h} - \bar{O}^t_{hh})/n^t_h. \qquad (24)$$

Assuming that the two terms between parentheses in (24) are absolutely smaller than S^t_h, it follows from (24) that

$$\left|C_{hh,\,\text{sec}}\right| \le \frac{\gamma^{\text{jan}}_{\text{red}}\; p^{\text{jan},\,t}_{h,H+1}\; p^t_{in,h}\; p^{t-12}_{h,H+1}(1 - p^{t-12}_{h,H+1})(1 - f_h)}{n^t_h}(S^t_h)^2.$$

Hence, when $p^t_{in,h}, p^{t-12}_{h,H+1} \le 0.1$, we may conclude that under the above assumptions the contribution of the second covariance component is less than 1% of $\text{var}(\bar{o}^t_h)$ so that (5) can be used without severely affecting the results. When $C_{hh,\text{sec}}$ is non-negligible, it can be estimated from the sample according to (24) by

$$\begin{aligned}&\hat{C}_{hh,\text{sec}} =\\ &A_h\{(\bar{o}^{t-12}_{h,H+1} - \bar{o}^{t-12}_{h,\le H})(\bar{o}^t_{in,h} - \bar{o}^t_{hh}) - \widehat{\text{cov}}(\bar{o}^{t-12}_{h,\le H}, \bar{o}^t_{hh})\}/n^t_h,\end{aligned} \qquad (25)$$

where in analogy with (10) and (12) $\widehat{\text{cov}}(\bar{o}^{t-12}_{h,\le H}, \bar{o}^t_{hh})$ is defined by

$$\widehat{\text{cov}}(\bar{o}^{t-12}_{h,\le H}, \bar{o}^t_{hh}) = \frac{n^{t-12}_{hh}}{n^{t-12}_{h,\le H}}\left(\frac{n^{t-12,t}_{hh}}{n^{t-12}_{hh} n^t_{hh}} - \frac{1}{N^{t-12,t}_{hh}}\right)\hat{\rho}^{t-12,t}_{hh,\text{OLP}}\hat{S}^{t-12}_{hh}\hat{S}^t_{hh}.$$

We used in (25) that (*i*) for two arbitrary (unbiased) estimators $\hat{a}$ and $\hat{b}$, $E(\hat{a}\hat{b}) = ab + \text{cov}(\hat{a}, \hat{b})$ and (*ii*) $\text{cov}(\bar{o}^{t-12}_{hg}, \bar{o}^t_{kh}) = 0$ $(g \neq h$ or $k \neq h)$.

We conclude this appendix with the derivation of (19). The expected number of additional deaths remaining in the sample of January during the refreshment is $0.9\,p^{\text{jan},\,t}_{h,\,H+1}\,\Delta n^{t-12}_{h,\,H+1}$. The number of deaths outside the sample just before the refreshment can be estimated by $N_{h,H+1}^{\text{jan},t} - n_{h,H+1}^{\text{jan}} - p_{h,H+1}^{\text{jan},t}\Delta n_{h,H+1}^{t-12}$. Hence, the number of *new* deaths in the sample due to the refreshments in all substrata $U_{hg}^{t-12,t}$ $(1 \le g \le H)$ in January can be estimated by

$$0.1(n_h^{\text{jan}} - n_{0h}^{\text{jan}})\frac{N_{h,H+1}^{\text{jan},t} - n_{h,H+1}^{\text{jan}} - p_{h,H+1}^{\text{jan},t}\Delta n_{h,H+1}^{t-12}}{N_h^{\text{jan}} - N_{0h}^{t-12,\text{jan}} - (n_h^{\text{jan}} - n_{0h}^{\text{jan}})}.$$

Now using $n_{0h}^{\text{jan}} = f_h N_{0h}^{t-12,\text{jan}}$ according to the above assumptions, it is seen that after the refreshments the final number of additional deaths in the sample due to $\Delta n_{h,H+1}^{t-12}$ can be estimated by

$$\left\{0.9 - \frac{0.1(n_h^{\text{jan}} - n_{0h}^{\text{jan}})}{N_h^{\text{jan}} - N_{0h}^{t-12,\text{jan}} - (n_h^{\text{jan}} - n_{0h}^{\text{jan}})}\right\} p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12}$$
$$= \frac{0.9 - f_h}{1 - f_h}\, p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12} = \gamma_{\text{red}}^{\text{jan}}\, p_{h,H+1}^{\text{jan},t}\,\Delta n_{h,H+1}^{t-12}.$$

Appendix B

Some useful covariance formulas for overlapping samples

Let s_{123} denote a mother sample consisting of three mutually disjoint SRS subsamples s_1, s_2 and s_3. Let the variable *x* be observed in s_{12} and the variable *y* in s_{23}. The corresponding sample means are denoted by $\overline{x}_{12}$ and $\overline{y}_{23}$, respectively. Denote the size of s_k by n_k $(k = 1, 2, 3, 12, 23)$. Define $\lambda = n_2 / n_{12}, \mu = n_2 / n_{23}$ and $f_k = n_k / N$. Furthermore, define S_{xy} by

$$S_{xy} = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \overline{X}_p)(Y_j - \overline{Y}_p) .$$

Then the covariance between $\overline{x}_{12}$ and $\overline{y}_{23}$ is equal to

$$\text{cov}(\overline{x}_{12}, \overline{y}_{23}) =$$
$$\left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}. \quad (26)$$

This can be shown as follows

$$\begin{aligned}
&\text{cov}(\overline{x}_{12}, \overline{y}_{23})\\
&\quad = \text{cov}\{(1-\lambda)\overline{x}_1 + \lambda\overline{x}_2, \mu\overline{y}_2 + (1-\mu)\overline{y}_3\}\\
&\quad = (1-\lambda)\text{cov}(\overline{x}_1, \overline{y}_{23}) + \lambda\mu\,\text{cov}(\overline{x}_2, \overline{y}_2)\\
&\qquad + \lambda(1-\mu)\text{cov}(\overline{x}_2, \overline{y}_3)\\
&\quad = -(1-\lambda)\frac{S_{xy}}{N} + \lambda\mu\left(\frac{1}{n_2} - \frac{1}{N}\right)S_{xy} - \lambda(1-\mu)\frac{S_{xy}}{N}\\
&\quad = \left(\frac{\lambda\mu}{n_2} - \frac{1}{N}\right)S_{xy} = \left(\frac{\mu}{n_{12}} - \frac{1}{N}\right)S_{xy} = \left(\frac{\lambda}{n_{23}} - \frac{1}{N}\right)S_{xy}.
\end{aligned}$$

In the third line we used that $\text{cov}(\overline{x}_1, \overline{y}_{23}) = \text{cov}(\overline{x}_2, \overline{y}_3) = -S_{xy} / N$. This follows from the conditional covariance formula

$$\begin{aligned}
\text{cov}(\overline{x}_2, \overline{y}_3) &= E\{\text{cov}(\overline{x}_2, \overline{y}_3 | s_2)\} + \text{cov}\{E(\overline{x}_2 | s_2), E(\overline{y}_3 | s_2)\}\\
&= 0 + \text{cov}\left\{\overline{x}_2, \frac{\overline{Y}_p - f_2\overline{y}_2}{1 - f_2}\right\}\\
&= -\frac{f_2}{1 - f_2}\text{cov}(\overline{x}_2, \overline{y}_2) = -\frac{S_{xy}}{N}.
\end{aligned}$$

For an alternative proof based on the sampling autocorrelation coefficient, see Knottnerus (2003, page 375).

References

Berger, Y.G. (2004). Variance estimation for measures of change in probability sampling. *The Canadian Journal of Statistics*, 32, 451-467.

Brodie, P. (2003). Review of recent work on variance estimation methods in business surveys. Unpublished report, Office for National Statistics, London.

Cox, D.R., and Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman and Hall.

Hidiroglou, M.A., Särndal, C.-E. and Binder, D.A. (1995). Weighting and estimation in business surveys. In *Business Survey Methods*, (Eds., B.G. Cox *et al.*). New York: John Wiley & Sons, Inc.

Holt, D., and Skinner, C.J. (1989). Components of change in repeated surveys. *International Statistical Review*, 57, 1-18.

Holt, D., and Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, A, 142, 33-46.

Kish, L. (1965). *Survey sampling.* New York: John Wiley & Sons, Inc.

Knottnerus, P. (2003). *Sample Survey Theory*: *Some Pythagorean Perspectives.* New York: Springer-Verlag.

Knottnerus, P., and Van Delden, A. (2006). Estimation of changes in repeated surveys and their significance, http://www.iser.essex.ac.uk/ulsc/mols2006/programme/data/paper/Knottnerus.doc.

Konschnik, C.A., Monsour, N.J. and Detlefsen, R.E. (1985). Constructing and maintaining frames and samples for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 113-122.

Laniel, N. (1987). Variances for a rotating sample from a changing population. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 496-500.

Lowerre, J.M. (1979). Sampling for change. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 343-347.

Nordberg, L. (2000). On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers. *Journal of Official Statistics*, 16, 363-378.

Qualité, L., and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 173-181.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Smith, P., Pont, M. and Jones, T. (2003). Developments in business survey methodology in the Office for National Statistics, 1994-2000. *The Statistician*, 52, 257-295.

Tam, S.M. (1984). On covariances from overlapping samples. *The American Statistician*, 38, 288-289.

Wood, J. (2008). On the covariance between related Horvitz-Thompson estimators. *Journal of Official Statistics*, 24, 53-78.

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 53-62
Statistics Canada, Catalogue No. 12-001-X

Variance inflation factors in the analysis of complex survey data

Dan Liao and Richard Valliant [1]

Abstract

Survey data are often used to fit linear regression models. The values of covariates used in modeling are not controlled as they might be in an experiment. Thus, collinearity among the covariates is an inevitable problem in the analysis of survey data. Although many books and articles have described the collinearity problem and proposed strategies to understand, assess and handle its presence, the survey literature has not provided appropriate diagnostic tools to evaluate its impact on regression estimation when the survey complexities are considered. We have developed variance inflation factors (VIFs) that measure the amount that variances of parameter estimators are increased due to having non-orthogonal predictors. The VIFs are appropriate for survey-weighted regression estimators and account for complex design features, *e.g.*, weights, clusters, and strata. Illustrations of these methods are given using a probability sample from a household survey of health and nutrition.

Key Words: Cluster sample; Collinearity diagnostics; Linearization variance estimator; Survey-weighted least squares; Stratified sample.

1. Introduction

Collinearity of predictor variables in a linear regression refers to a situation where explanatory variables are correlated with each other. The terms, multicollinearity and ill conditioning, are also used to denote the same situation. Collinearity is worrisome for both numerical and statistical reasons. The estimates of slope coefficients can be numerically unstable in some data sets in the sense that small changes in the $\mathbf{X}$'s or the $\mathbf{Y}$'s can produce large changes in the values of these estimates. Statistically, correlation among the predictors can lead to slope estimates with large variances. In addition, when $\mathbf{X}$'s are strongly correlated, the R^2 in a regression can be large while the individual slope estimates are not statistically significant. Even if slope estimates are significant, they may have signs that are the opposite of what are expected (Neter, Kutner, Wasserman and Nachtsheim 1996). Collinearity may also affect forecasts (Smith 1974; Belsley 1984).

In experimental designs, it may be possible to create situations where the explanatory variables are orthogonal to each other. But, in many surveys, variables that are substantially correlated are collected for analysis. For example, total income and its components (*e.g.*, wages and salaries, capital gains, interest and dividends) are collected in the Panel Survey of Income Dynamics (http://psidonline.isr.umich.edu/) to track economic well-being over time. When one explanatory variable is a linear combination of the others, this is known as perfect collinearity (or multicollinearity) and is easy to identify. Cases that are of interest in practice are ones where collinearity is less than perfect but still affects the precision of estimates (Kmenta 1986, section 10.3).

Although there is a substantial literature on regression diagnostics for non-survey data, there is considerably less for survey data. A few articles in the last decade introduced techniques for the evaluation of the quality of regression on complex survey data, mainly on identifying influential points and influential groups with abnormal data values or survey weights. Elliot (2007), for instance, developed Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. Li (2007a, b); Li and Valliant (2009, 2011) adapted and extended a series of traditional diagnostic techniques to regression on complex survey data, mainly on identifying influential observations and influential groups of observations. Li's research covers residuals and leverages, DFBETA, DFBETAS, DFFIT, DFFITs, Cook's Distance and the forward search approach. Although an extensive literature in applied statistics provides valuable suggestions and guidelines for data analysts to diagnose the presence of collinearity (*e.g.*, Farrar and Glauber 1967; Theil 1971; Belsley, Kuh and Welsch 1980; Fox 1984; Belsley 1991), none of this research touches upon diagnostics for collinearity when fitting models with survey data.

The variance inflation factor (VIF) described in section 2, is one of the most popular conventional collinearity diagnostic techniques, and is mainly aimed at ordinary or weighted least squares regressions. A VIF measures the inflation of the variance of a slope estimate caused by the nonorthogonality of the predictors over and above what the variance would be with orthogonality. In section 3, we consider the case of an analyst who estimates model parameters using survey-weighted least squares (SWLS) and derive VIFs appropriate to SWLS. The components of the VIF can be estimated using the ingredients of a variance

[1] Dan Liao, RTI International, 701 13th Street, N.W., Suite 750, Washington DC, 20005. E-mail: dliao@rti.org; Richard Valliant, University of Michigan and University of Maryland, Joint Program in Survey Methodology, 1218 Lefrak Hall, College Park, MD, 20742.

estimator that is in common usage in software packages for analyzing survey data. In the case of linear regression, a type of sandwich variance estimator will estimate both the model variance and design variance of the SWLS slope estimator. As we will show in section 3, the model or design variance of $\hat{\beta}_k$, an estimator of slope associated with the predictor $\mathbf{x}_k$, is inflated somewhat when different predictors are correlated with each other compared to what the variance would be if $\mathbf{x}_k$ were orthogonal to the other predictors. The measure of inflation, the VIF, is composed of terms that must be estimated from the sample. Our approach has been to substitute estimators that have both a model and design interpretation as described in section 3.5.

The fourth section presents an empirical study using data from the United States National Health and Nutrition Examination Survey. The application of our new approach is demonstrated and the newly-derived VIF values for SWLS are compared to the ones for OLS or WLS, which can be obtained from the standard statistical packages. The comparisons show that VIF values are different for different regression methods and a VIF specific to complex sample should be used to evaluate the harmfulness of collinearity in the analysis of survey data.

2. Collinearity diagnostics in ordinary least squares estimation

Suppose the sample s has n units, on each of which p $\mathbf{x}$'s or predictors and one analysis variable Y are observed. The standard linear model in a nonsurvey setting is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is an $n \times 1$ vector of observations on a response or dependent variable; $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ is an $n \times p$ design matrix of fixed constants with $\mathbf{x}_k$, the $n \times 1$ vector of values of explanatory variable k for the n sample units; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated; and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of statistically independent error terms with zero mean and constant variance σ^2. We assume, for simplicity, that $\mathbf{X}$ has full column rank. The ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T\mathbf{Y}$, for which the model variance is $\text{Var}_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Here, we use the subscript M to denote expectation under the model.

Collinearities of explanatory variables inflate the model variance of the regression coefficients compared to having orthogonal $\mathbf{X}$'s. This effect can be seen in the formula for the variance of a specific estimated non-intercept coefficient $\hat{\beta}_k$ (Theil 1971),

$$\text{Var}_M(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i \in s} x_{ik}^2} \frac{1}{1 - R_k^2} \tag{1}$$

where R_k^2 is the square of the multiple correlation from the regression of the k^{th} column of $\mathbf{X}$ on the other columns. This R-square defined as $R_k^2 = \hat{\beta}_{(k)} \mathbf{X}_{(k)}^T \mathbf{X}_{(k)} \hat{\beta}_{(k)} / \mathbf{x}_k^T \mathbf{x}_k$, where $\hat{\beta}_{(k)}$ is OLS estimate of the slope when $\mathbf{x}_k$ is regressed on the other $\mathbf{x}$'s and $\mathbf{X}_{(k)}$ is the $\mathbf{X}$ matrix with the k^{th} column removed. The term $\sigma^2 / \sum x_{ik}^2$ is the model variance of $\hat{\beta}_k$ if the k^{th} predictor were orthogonal to all the other predictors. The value of R_k^2 may be nonzero because the k^{th} predictor is correlated with one other explanatory variable or because of a more complex pattern of dependence between $\mathbf{x}_k$ and several other predictors. Consequently, the collinearity between $\mathbf{x}_k$ and some other explanatory variables can result in the inflation of the variance of $\hat{\beta}_k$ beyond what would be obtained with orthogonal $\mathbf{X}$'s. The second term in (1), $(1 - R_k^2)^{-1}$, is called the variance-inflation factor (VIF) (Theil 1971).

A basic reference on collinearity and other OLS diagnostics is Belsley *et al.* (1980). Collinearity diagnostics are covered in many other textbooks including Fox (1984) and Neter *et al.* (1996). In some cases, it is desirable to weight cases differentially in a regression analysis to incorporate a nonconstant residual variance. This form of weighting is model-based and is called weighted least squares (WLS). Most of current statistical software packages, (*e.g.*, SAS, Stata, S-Plus and R), use $(1 - R_{k(\text{WLS})}^2)^{-1}$ as VIF for WLS, where $R_{k(\text{WLS})}^2$ is the square of the multiple correlation from the WLS regression of the k^{th} column of $\mathbf{X}$ on the other columns. Fox and Monette (1992) also generalized this concept of variance inflation as a measure of collinearity to a subset of parameters in $\mathbf{b}$ and derived a *generalized variance-inflation factor* (GVIF). Furthermore, some interesting work has developed VIF-like measures, such as *collinearity indices* in Steward (1987) that are simply the square roots of the VIFs and *tolerance* defined as the inverse of VIF in Simon and Lesage (1988).

3. VIF in survey weighted least squares regression

3.1 Survey-weighted least squares estimators

Suppose the underlying structural model in the superpopulation is $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$, where the error terms in the model have a general variance structure $\mathbf{e} \sim (0, \sigma^2\mathbf{V})$ with known $\mathbf{V}$ and σ^2. Define $\mathbf{W}$ to be the diagonal matrix of survey weights. We assume throughout that the survey weights are constructed in such a way that they can be used for estimating finite population totals. The survey weighted least squares (SWLS) estimator is $\hat{\boldsymbol{\beta}}_{\text{SW}} = (\mathbf{X}^T \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$, assuming $\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X}$ is invertible. Fuller (2002) describes the properties of this estimator.

The estimate $\hat{\boldsymbol{\beta}}_{\text{SW}}$ is a model unbiased estimator of $\boldsymbol{\beta}$ under the model $\mathbf{Y} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{e}$ regardless of whether

$\mathrm{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$ is specified correctly or not, and is approximately design-unbiased for the census parameter $\mathbf{B}_U = (\mathbf{X}_U^T\mathbf{X}_U)^{-1}\mathbf{X}_U^T\mathbf{Y}_U$, in the finite population of N units. The subscript U stands for the finite population, $\mathbf{Y}_U = (Y_1, ..., Y_N)^T$, and $\mathbf{X}_U = (\mathbf{x}_1, ..., \mathbf{x}_p)$ with $\mathbf{x}_k$ as the $N \times 1$ vector of values for covariate k.

3.2 Model variance of coefficient estimates

The model variance of the parameter estimator $\hat{\boldsymbol{\beta}}_{\mathrm{SW}}$, assuming $\mathrm{Var}_M(\mathbf{e}) = \sigma^2\mathbf{V}$, can be expressed as

$$\begin{aligned}\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{SW}}) &= \sigma^2(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\\ &= \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\sigma^2 = \mathbf{G}\sigma^2, \qquad (2)\end{aligned}$$

where $\tilde{\mathbf{X}} = \mathbf{W}^{1/2}\mathbf{X}$, $\tilde{\mathbf{V}} = \mathbf{W}^{1/2}\mathbf{V}\mathbf{W}^{1/2}$, $\mathbf{A} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$, $\mathbf{B} = \tilde{\mathbf{X}}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}$, and $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$.

If the columns of $\mathbf{X}$ are orthogonal, then $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \mathrm{diag}(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$ and $\mathbf{A}^{-1} = \mathrm{diag}(1/\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)$, where $\tilde{\mathbf{x}}_k = \mathbf{w}_k^{1/2}\mathbf{x}_k$. The ij^{th} element of $\mathbf{G}$ then becomes $\tilde{\mathbf{x}}_i^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_j/(\tilde{\mathbf{x}}_i^T\tilde{\mathbf{x}}_i)^2$. Thus, when the $\mathbf{X}$'s are orthogonal, the model variance of $\hat{\beta}_{\mathrm{SW}_k}$ is

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{SW}_k}) = \sigma^2\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k/(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)^2, \qquad (3)$$

a fact we will use later. More generally, the model variance of $\hat{\beta}_{\mathrm{SW}_k}$, the coefficient estimate for the k^{th} explanatory variable, is

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{SW}_k}) = \mathbf{i}_k'\,\mathrm{Var}_M(\hat{\boldsymbol{\beta}}_{\mathrm{SW}})\mathbf{i}_k = \sigma^2\mathbf{i}_k'\mathbf{G}\mathbf{i}_k = \sigma^2 g^{kk} \qquad (4)$$

where $\mathbf{i}_k$ is a $p \times 1$ vector with 1 in position k and 0's elsewhere, and g^{kk} is the k^{th} diagonal element of matrix $\mathbf{G}$.

3.3 Model-based VIF

As shown in Appendix A, the model variance of $\hat{\beta}_{\mathrm{SW}_k}$ in (4) can be written as:

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{SW}_k}) = g^{kk}\sigma^2 = \frac{\zeta_k\rho_k}{1 - R^2_{\mathrm{SW}(k)}}\frac{\sigma^2\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k}{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k)^2}, \qquad (5)$$

where

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{V}}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}} = \frac{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}},$$

with $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{\mathrm{SW}(k)}$ being the residual from SWLS regressing $\mathbf{x}_k$ on $\mathbf{X}_{(k)}$ and $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\mathrm{SW}(k)} = \mathbf{W}^{1/2}\mathbf{e}_{xk}$,

$$\rho_k = \frac{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k} = \frac{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}\mathbf{V}\mathbf{W}\mathbf{x}_k}$$

and $R^2_{\mathrm{SW}(k)}$, defined in Appendix A, is the square of the multiple correlation from the weighted regression of the k^{th} column of $\mathbf{X}$ on the other columns. Hence, ζ_k and ρ_k depend on $\mathbf{W}$ and $\mathbf{V}$. The variance under orthogonality in (3) is inflated

$$\mathrm{VIF}_k = \frac{\zeta_k\rho_k}{1 - R^2_{\mathrm{SW}(k)}} \qquad (6)$$

times when incorporating the other $p - 1$ explanatory variables in SWLS. The model-based VIF in SWLS includes not only the multiple correlation coefficient $R^2_{\mathrm{SW}(k)}$ but also two adjustment coefficients, ζ_k and ρ_k, that are not present in the OLS and WLS cases.

Using the singular value decomposition of $\tilde{\mathbf{V}}$, we can bound the factor $\zeta_k\rho_k$, which is the adjustment to the VIF in WLS. Based on the extrema of the ratio of quadratic forms (Lin 1984), the term ζ_k is bounded in the range of $\mu_{\min}(\tilde{\mathbf{V}}) \le \zeta_k \le \mu_{\max}(\tilde{\mathbf{V}})$, and ρ_k is bounded in the range of

$$\frac{1}{\mu_{\max}(\tilde{\mathbf{V}})} \le \rho_k \le \frac{1}{\mu_{\min}(\tilde{\mathbf{V}})},$$

where $\mu_{\min}(\tilde{\mathbf{V}})$ and $\mu_{\max}(\tilde{\mathbf{V}})$ are the minimum and maximum singular values of the matrix $\tilde{\mathbf{V}}$. Combining these results, the joint coefficient $\zeta_k\rho_k$ is bounded in the range of:

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k\rho_k \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

Notice that when $\tilde{\mathbf{V}} = \mathbf{I}$, $\zeta_k = \rho_k = 1$ and (6) reduces to

$$\frac{1}{1 - R^2_{\mathrm{SW}(k)}}\frac{\sigma^2}{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k},$$

which is the model variance of the WLS estimates when $\mathbf{V}$ is diagonal and $\mathbf{W}$ is correctly specified as $\mathbf{W} = \mathbf{V}^{-1}$. In that unusual case, the VIF currently computed by software packages will be appropriate for SWLS. However, rarely will it be reasonable to think that $\mathbf{W} = \mathbf{V}^{-1}$ in survey estimation. If $\tilde{\mathbf{V}} \ne \mathbf{I}$, then ζ_k and ρ_k are not equal to 1 and a specialized calculation of the VIF is still needed. When $\mathbf{V} = \mathbf{I}$, which is the usual application considered by analysts,

$$\tilde{\mathbf{V}} = \mathbf{W},\ \zeta_k = \frac{\tilde{\mathbf{e}}_{xk}^T\mathbf{W}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}},\ \rho_k = \frac{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k}{\tilde{\mathbf{x}}_k^T\mathbf{W}\tilde{\mathbf{x}}_k}$$

and

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}},$$

where $w_{\min}$ is the minimum value of the survey weights and $w_{\max}$ is their maximum value. In this case, the range of $\zeta_k\rho_k$ is bounded by

$$\left[\frac{w_{\min}}{w_{\max}}, \frac{w_{\max}}{w_{\min}}\right].$$

When all the survey weights are constant, $\zeta_k\rho_k = 1$ and the VIF produced by standard software, $(1 - R^2_{\mathrm{SW}})^{-1}$, does not need to be adjusted in SWLS; however, when the range of the survey weights is large, $\zeta_k\rho_k$ can be very small or large and can be either above or below 1. In this case the VIF produced by standard software is not appropriate and a special calculation is needed. These facts will be illustrated in our experimental studies.

The VIF in (6) is appropriate regardless of whether the model contains an intercept or not. An alternative version can also be written that assumes that an intercept is in the model when $\tilde{\mathbf{x}}_k$ is regressed on the other $\mathbf{x}$'s. The derivation of this form is in Liao (2010). We summarize the result below.

The variance of $\hat{\beta}_{\mathrm{SW}_k}$ in a model M2 that includes an intercept and in which $\mathbf{x}_k$ is orthogonal to the other $\mathbf{x}$'s is:

$$\mathrm{Var}_{M2}(\hat{\beta}_{\mathrm{SW}_k}) = \frac{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{\mathrm{SST}^2_{\mathrm{SW}_{m(k)}}} \tag{7}$$

where $\tilde{\mathbf{1}} = (w_1^{1/2}, ..., w_n^{1/2})$, $\bar{\tilde{x}}_k = \sum_{i\in s} w_i\, x_{ki} / \hat{N}$, $\hat{N} = \sum_{i\in s} w_i$, and $\mathrm{SST}_{\mathrm{SW}_{m(k)}} = \tilde{\mathbf{x}}_k^T\,\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2$. The variance of $\hat{\beta}_{\mathrm{SW}_k}$ can then be rewritten as

$$\mathrm{Var}_M(\hat{\beta}_{\mathrm{SW}_k}) = \frac{\zeta_k\rho_{mk}}{1 - R^2_{\mathrm{SW}m(k)}}\mathrm{Var}_{M2}(\hat{\beta}_{\mathrm{SW}_k}) \tag{8}$$

where $R^2_{\mathrm{SW}m(k)}$ is the SWLS R-square from regressing $\tilde{\mathbf{x}}_k$ on the $\mathbf{x}$'s in the remainder of $\tilde{\mathbf{X}}$ (excluding a column for the intercept). The term ζ_k was defined following (5) and

$$\rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}.$$

Most software packages will consistently provide $(1 - R^2_{\mathrm{SW}m(k)})^{-1}$ as the VIF as part of WLS regression output. Note that this is different from the VIF, $(1 - R^2_{\mathrm{SW}(k)})^{-1}$, introduced in section 3.3 which does not assume that an intercept is retained in the model. Software packages generally do not supply $(1 - R^2_{\mathrm{SW}(k)})^{-1}$.

Using arguments similar to those in the previous section, we can bound $\zeta_k\rho_{mk}$ by

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} \le \zeta_k\rho_{mk} \le \frac{\mu_{\max}(\tilde{\mathbf{V}})}{\mu_{\min}(\tilde{\mathbf{V}})}.$$

The model variance of $\hat{\beta}_{\mathrm{SW}_k}$ is inflated by

$$\mathrm{VIF}_{mk} = \frac{\zeta_k\,\rho_{mk}}{1 - R^2_{\mathrm{SW}m(k)}}$$

compared to its variance in the model (M2) with only the explanatory variable $\tilde{\mathbf{x}}_k$ and intercept. The new intercept-adjusted VIF_{mk} retains some properties of VIF_k in (6). When $\tilde{\mathbf{V}} = \mathbf{I}$, we have $\zeta_k = 1$, $\rho_{mk} = 1$ and the intercept-adjusted VIF in (8) for SWLS is equal to the conventional intercept-adjusted VIF: $(1 - R^2_{m(k)})^{-1}$. When $\mathbf{V} = \mathbf{I}$, we have $\tilde{\mathbf{V}} = \mathbf{W}$,

$$\zeta_k = \frac{\tilde{\mathbf{e}}_{xk}\mathbf{W}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\,\tilde{\mathbf{e}}_{xk}},\ \rho_{mk} = \frac{(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\mathbf{W}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}$$

and

$$\frac{\mu_{\min}(\tilde{\mathbf{V}})}{\mu_{\max}(\tilde{\mathbf{V}})} = \frac{w_{\min}}{w_{\max}}.$$

The range of $\zeta_k\rho_{mk}$ also depends on the range of survey weights as did $\zeta_k\rho_k$.

3.4 Estimating the VIF for a model with stratified clustering when V is unknown

In the previous sections, we used model-based arguments to derive VIFs. The VIFs contain terms, $\tilde{\mathbf{V}}$ in particular, that are unknown and must be estimated. In this section, we construct estimators of the components of the VIFs, again using model-based arguments. However, a standard, design-based linearization variance estimator also estimates the model variance, as shown below, and supplies the components needed to estimate the VIF. In the remainder of this section, we will present estimators that are appropriate for a model that has a stratified clustered covariance structure.

Suppose that in a stratified multistage sampling design, there are $h = 1, ..., H$ strata in the population, $i = 1, ..., N_h$ clusters in the corresponding stratum h and $t = 1, ..., M_{hi}$ units in cluster hi. We select $i = 1, ..., n_h$ clusters in stratum h and $t = 1, ..., m_{hi}$ units in cluster hi. Denote the set of sample clusters in stratum h by s_h and the sample of units in cluster hi as s_{hi}. The total number of sample units in stratum h is $m_h = \sum_{i\in s_h} m_{hi}$, and the total in the sample is $m = \sum_{h=1}^{H} m_h$. Clusters are assumed to be selected with replacement within strata and independently between strata. Consider this model:

$$\begin{aligned}
&E_M(Y_{hit}) = \mathbf{x}_{hit}^T\boldsymbol{\beta}\\
&\qquad h = 1, \ldots, H,\quad i = 1, \ldots, N_h,\quad t = 1, \ldots, M_{hi}\\
&\mathrm{Cov}_M(Y_{hit}, Y_{h'i't'}) = 0\\
&\qquad h \ne h', \text{ or, } h = h' \text{ and } i \ne i'.
\end{aligned} \tag{9}$$

Units within each cluster are assumed to be correlated but the particular correlation of the covariances does not have to be specified for this analysis. The estimator of the regression parameter is:

$$\hat{\boldsymbol{\beta}}_{\mathrm{SW}} = \sum_{h=1}^{H}\sum_{i\in s_h}\mathbf{A}^{-1}\mathbf{X}_{hi}^T\,\mathbf{W}_{hi}\,\mathbf{Y}_{hi} \tag{10}$$

where $\mathbf{X}_{hi}$ is the $m_{hi} \times p$ matrix of covariates for sample units in cluster hi, $\mathbf{W}_{hi} = \mathrm{diag}(w_t)$, $t \in s_{hi}$ is the diagonal matrix of survey weights for cluster hi and $\mathbf{Y}_{hi}$ is the $m_{hi} \times 1$ vector of response variables in cluster hi. The model variance of $\hat{\boldsymbol{\beta}}_{\text{SW}}$ is:

$$\begin{aligned}\text{Var}_M(\hat{\boldsymbol{\beta}}_{\text{SW}}) &= \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\sum_{i\in s_h}\mathbf{X}_{hi}^T\,\mathbf{W}_{hi}\,\mathbf{V}_{hi}\,\mathbf{W}_{hi}\,\mathbf{X}_{hi}\right]\mathbf{A}^{-1}\\ &= \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\mathbf{X}_{h}^T\,\mathbf{W}_{h}\,\mathbf{V}_{h}\,\mathbf{W}_{h}\,\mathbf{X}_{h}\right]\mathbf{A}^{-1}, \qquad (11)\end{aligned}$$

where $\mathbf{V}_{hi} = \text{Var}_M(\mathbf{Y}_{hi})$ and $\mathbf{V}_h = \text{Blkdiag}(\mathbf{V}_{hi}),\ i \in s_h$. Expression (11) is a special case of (2) with $\mathbf{X}^T = (\mathbf{X}_1^T, \mathbf{X}_2^T, \ldots, \mathbf{X}_H^T)$, $\mathbf{X}_h$ is the $m_h \times p$ matrix of covariates for sample units in stratum h, $\mathbf{W} = \text{diag}(\mathbf{W}_{hi})$, for $h = 1, \ldots, H$ and $i \in s_h$ and $\mathbf{V} = \text{Blkdiag}(\mathbf{V}_h)$.

Denote the cluster-level residuals as a vector, $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi}\,\hat{\boldsymbol{\beta}}_{\text{SW}}$. A design-based linearization estimator is:

$$\begin{aligned}\text{var}_L(\hat{\boldsymbol{\beta}}_{\text{SW}}) &= \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i\in s_h}(\mathbf{z}_{hi}-\overline{\mathbf{z}}_h)(\mathbf{z}_{hi}-\overline{\mathbf{z}}_h)^T\right]\mathbf{A}^{-1}\\ &= \mathbf{A}^{-1}\left[\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left(\sum_{i\in s_h}\mathbf{z}_{hi}\mathbf{z}_{hi}^T - n_h\overline{\mathbf{z}}_h\overline{\mathbf{z}}_h^T\right)\right]\mathbf{A}^{-1}, \qquad (12)\end{aligned}$$

where

$$\overline{\mathbf{z}}_h = \frac{1}{n_h}\sum\nolimits_{i\in s_h}\mathbf{z}_{hi}$$

and $\mathbf{z}_{hi} = \mathbf{X}_{hi}^T\mathbf{W}_{hi}\mathbf{e}_{hi}$ with $\mathbf{e}_{hi} = \mathbf{Y}_{hi} - \mathbf{X}_{hi}\,\hat{\boldsymbol{\beta}}_{\text{SW}}$. This expression can be reduced to the formula for a single-stage stratified design when the cluster sample sizes are all equal to 1, $m_{hi} = 1$. Expression (12) is used by the Stata and SUDAAN packages, among others. The estimator $\text{var}_L(\hat{\boldsymbol{\beta}}_{\text{SW}})$ is consistent and approximately design-unbiased under a design where clusters are selected with replacement (Fuller 2002). Li (2007a, b) showed that (12) is also an approximately model-unbiased estimator under model (11).

The term in brackets in (12) serves as an estimator of the matrix $\mathbf{B}$ in expression (2). The components of $\text{var}_L(\hat{\boldsymbol{\beta}}_{\text{SW}})$ can be used to construct estimators of ζ_k and ρ_k in (5) and ρ_{mk} in (8). In particular,

$$\hat{\zeta}_k = \frac{\mathbf{e}_{xk}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}_{xk}^T\mathbf{W}\mathbf{e}_{xk}}, \qquad (13)$$

where

$$\hat{\mathbf{V}} = \text{Blkdiag}\left[\frac{n_h}{n_h-1}\left(\hat{\mathbf{V}}_h - \frac{1}{n_h}\mathbf{e}_h\,\mathbf{e}_h^T\right)\right],\ h = 1, 2, \ldots, H,$$

with $\hat{\mathbf{V}}_h = \text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}_{hi}^T)$ and

$$\hat{\rho}_k = \frac{\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k}{\mathbf{x}_k^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{x}_k},$$

with $\mathbf{e}_{xk} = \mathbf{x}_k - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}$. The estimate of $\hat{\rho}_{mk}$, defined following (8), is

$$\hat{\rho}_{mk} = \frac{(\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k - \hat{N}\,\bar{\bar{x}}_k^2)}{(\mathbf{x}_k - \mathbf{1}\bar{\bar{x}}_k)^T\,\mathbf{W}\hat{\mathbf{V}}\mathbf{W}(\mathbf{x}_k - \mathbf{1}\bar{\bar{x}}_k)}. \qquad (14)$$

Given these component estimators VIF_k is estimated by

$$\widehat{\text{VIF}}_k = \frac{\hat{\zeta}_k\hat{\rho}_k}{1 - R_{\text{SW}(k)}^2}$$

and VIF_{mk} is estimated by

$$\widehat{\text{VIF}}_{mk} = \frac{\hat{\zeta}_k\,\hat{\rho}_{mk}}{1 - R_{\text{SW}m(k)}^2}.$$

4. Experimental study

We will now illustrate the proposed, modified collinearity diagnostics and investigate their behavior using dietary intake data from the National Health and Nutrition Examination Survey (NHANES) 2007-2008 (http://www.cdc.gov/nchs/nhanes/nhanes2007-2008/datadoc_changes_0708.htm). The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages. NHANES uses a complex, multistage, probability sampling design. Oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. Among the respondents who received the in-person interview in the mobile examination center (MEC), around 94% provided complete dietary intakes. The survey weights in this data were constructed by taking MEC sample weights and further adjusting for the additional nonresponse and the differential allocation by day of the week for the dietary intake data collection. These weights are more variable than the MEC weights. The data set used in our study is a subset of 2007-2008 data composed of female respondents aged 26 to 40. Observations with missing values in the selected variables are excluded from the sample which finally contains 672 complete respondents. The final weights in our sample range from 6,028 to 330,067, with a ratio of 55:1. The U.S. National Center for Health Statistics recommends that the design of the sample is approximated by the stratified selection with replacement of 32 PSUs from 16 strata, with 2 PSUs within each stratum.

For this empirical study, a linear regression of body weight(kg) is fitted using survey weighted least squares. The predictor variables considered include age, Black(race) and nine daily total nutrition intake variables, which are calorie(100kcal), protein(100gm), carbohydrate(100gm), sugar(100gm), total fat(100gm), total saturated fatty acids(100gm), total monounsaturated fatty acids(100gm), total polyunsaturated fatty acids(100gm) and alcohol(100gm). All the daily total nutrition intake variables are correlated with each other to different degrees as shown in Figure 1.

Three regression methods compared in this study. The first one uses *ordinary least squares* (OLS) method and ignores sampling complexities including the weighting. The second one uses *weighted least squares* (WLS), which incorporates the survey weights by assuming $\mathbf{V} = \mathbf{W}^{-1}$ but ignores all sampling complexities. The third one is *survey weighted least squares* (SWLS), which uses the actual complex sampling design as described in section 3.4. The weight matrices, coefficient variance estimators and collinearity diagnostics of these three methods are listed in Table 1.

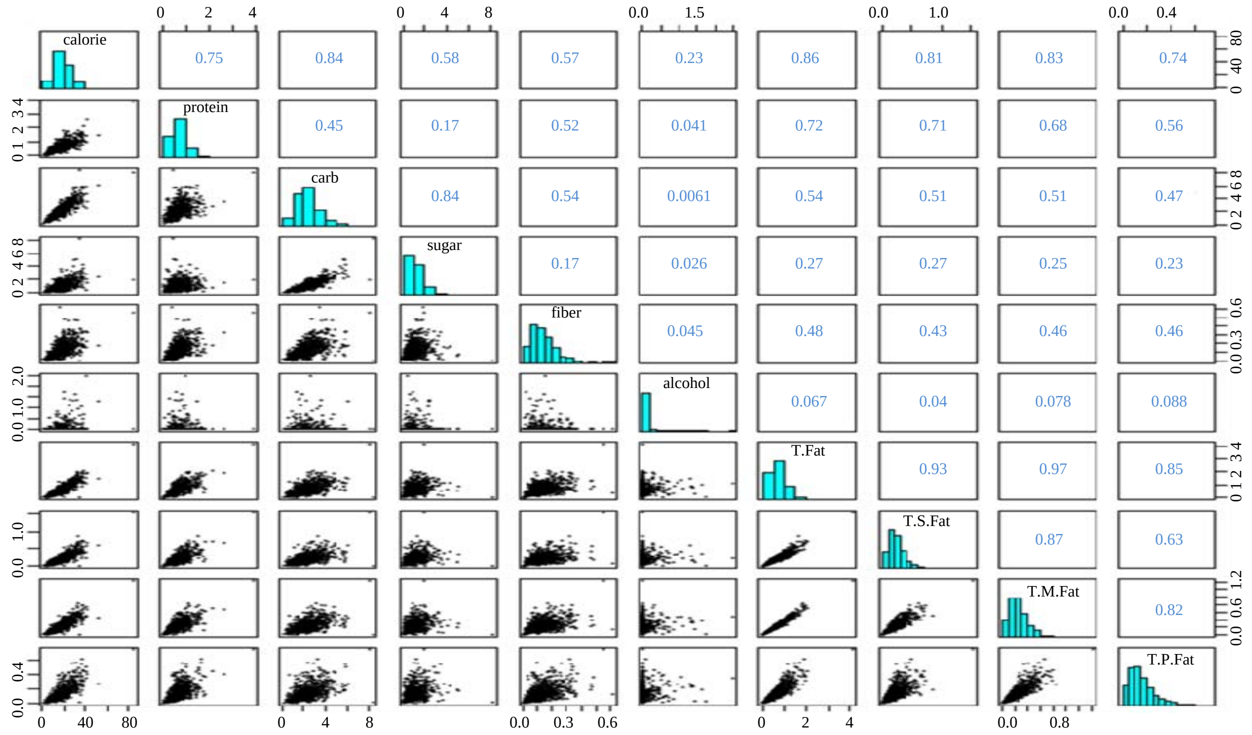
The results from fitting the model using three different regression methods are displayed in Table 2. The model with all the predictors is shown in the upper part of the table. In the lower tier of the table, a reduced model with less of the near-dependency problem is fitted with only three predictors: age, Black and calorie. In the reduced model, the value of the coefficient for calorie is positive and significant when WLS or SWLS is used, which seems logical and reflects the anticipated positive relationship between a respondent's body weight and her daily total calorie intake. However, when the other total nutrition intake variables are included in the model, the value of the calorie coefficient is negative and not significant due to its inflated variance. This is a typical example in which the variance of a coefficient is inflated, and its sign is illogical due to collinearity.

Table 3 reports the VIF values when the three different regression methods are used. The VIF formulas for these regression methods are listed in Table 1. When all the predictors are included in the model, calorie has the largest VIF values in all the regressions due to its high near-dependency with all the other total nutrition intake variables. As shown in Table 1, the VIF in SWLS can be obtained by multiplying the VIF from WLS with the adjustment coefficient $\zeta_k \rho_k$. In Table 3, the adjustment coefficients $\zeta_k \rho_k$ for all the non-fat total nutrition intake variables are all less than 1, especially the one for carbohydrate which is 0.46. This indicates that the VIF values for these variables in SWLS are much smaller than the ones in WLS and the collinearity among predictors in the model has less impact on the coefficient estimation when using SWLS, compared to using WLS. But for the fat-related nutrition intake variables, their $\zeta_k \rho_k$ are all larger than 1. Thus, the collinearity among the fat-related nutrition intake variables is more harmful to the coefficient estimation in SWLS than in WLS. To take a closer look at this problem, we also fitted a model that only contains two nutrition intake variables: total fat and total monounsatruated fatty acids. The SWLS VIF values are three times as large as the ones from OLS or WLS for these two nutrition variables. If an analyst is analyzing this survey data using SWLS but uses the unadjusted VIF values provided by standard statistical packages for either OLS (as shown in the first column) or WLS (as shown in the second column), the unadjusted VIFs will give somewhat misleading judgements on the severity of collinearity in this model. In summary, although the estimated slopes and predictions in regression using WLS and SWLS are the same, the VIFs can be underestimated or overestimated if survey complexities are ignored.

Table 1
Regression methods and their collinearity diagnostic statistics used in this experimental study

Regression Type	Weight Matrix $\mathbf{W}^a$	Variance Estimation of $\hat{\boldsymbol{\beta}}$	VIF fomula
OLS	$\mathbf{I}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$	$\text{VIF} = \frac{1}{1 - R^2_{m(k)}}{}^{b}$
WLS	$\mathbf{W}^c$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \frac{1}{1 - R^2_{\text{SW}m(k)}}$
SWLS	$\mathbf{W}$	$\hat{\sigma}^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}$	$\text{VIF} = \frac{\hat{\zeta}_k \hat{\rho}_{mk}}{1 - R^2_{\text{SW}m(k)}}$
		with	with $\hat{\zeta}_k = \frac{\mathbf{e}^T_{xk}\mathbf{W}\hat{\mathbf{V}}\mathbf{W}\mathbf{e}_{xk}}{\mathbf{e}^T_{xk}\mathbf{W}\mathbf{e}_{xk}}$,
		$\hat{\mathbf{V}} = \sum_{h=1}^{H} \frac{n_h}{n_h - 1}\left[\text{Blkdiag}(\mathbf{e}_{hi}\mathbf{e}^T_{hi}) - \frac{1}{n_h}\mathbf{e}_h\mathbf{e}^T_h\right]$	$\hat{\rho}_{mk} = \frac{(\tilde{\mathbf{x}}^T_k\tilde{\mathbf{x}}_k - \hat{N}\bar{\tilde{x}}^2_k)}{(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\hat{\mathbf{V}}(\tilde{\mathbf{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}$

[a] In all the regression models, the parameters are estimated by: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$.
[b] $R^2_{m(k)}$ is the OLS R-square from regressing $\mathbf{x}_k$ on the $\mathbf{x}$'s in the remainder of $\mathbf{X}$ (excluding a column for the intercept).
[c] $\mathbf{W}$ is the diagonal matrix with survey weights w_i on the main diagonal.



[a] T.Fat: total fat;
T.S.Fat: total saturated fatty acid;
T.M.Fat: total monounsaturated fatty acid;
T.P.Fat: total polyunsaturated fatty acid.

Figure 1 Pairwise scatterplots and correlation coefficients of nutrition variables[a]

Table 2
Parameter estimates with their associated standard errors using three different regression methods

	Full Model					
	OLS		WLS		SWLS	
Variable	**Beta**	**SE.**	**Beta**	**SE.**	**Beta**	**SE.**
Intercept	63.90***[a]	6.95	67.47***	6.36	67.47***	8.76
Age	0.26	0.19	0.08	0.18	0.08	0.25
Black	10.39***	2.07	10.59***	2.38	10.59***	2.20
Calorie	-6.41	5.76	-8.19	5.56	-8.19	5.75
Protein	25.72	24.76	40.98	23.60	40.98	25.38
Carbohydrate	26.67	23.93	32.31	22.96	32.31	22.65
Sugar	-1.90	3.06	-0.30	2.82	-0.30	4.06
Fiber	-41.17	20.23	-34.20	17.98	-34.20	19.05
Alcohol	38.84	39.45	49.37	38.28	49.37	40.10
Total Fat	150.25*	69.53	161.78*	72.12	161.78	94.76
Total Saturated Fatty Acids	-113.20*	49.81	-101.40	56.26	-101.40	82.71
Total Monounsaturated Fatty Acids	-72.05	48.03	-92.44	51.52	-92.44	83.51
Total Polyunsaturated Fatty Acids	-92.60*	46.13	-75.55	51.16	-75.55	78.76
	Reduced Model					
	OLS		WLS		SWLS	
Variable	**Beta**	**SE.**	**Beta**	**SE.**	**Beta**	**SE.**
Intercept	62.26***	6.88	67.52***	6.29	67.52***	8.48
Age	0.27	0.19	0.07	0.18	0.07	0.25
Black	12.54***	1.98	11.74***	2.32	11.74***	2.05
Calorie	0.15	0.10	0.23*	0.09	0.23*	0.10

[a] p values of significance: * $p = 0.05$; ** $p = 0.01$; *** $p = 0.005$.

Table 3
VIF values using three different regression methods

		Full Model		
	OLS	**WLS**	**SWLS**	
Variable	**VIF**	**VIF**	**VIF**	$\zeta_k\rho_k$
Age	1.02	1.03	0.96	0.94
Black	1.10	1.07	1.12	1.05
Calorie	3,411.61	3,562.70	2,740.83	0.77
Protein	123.12	127.35	103.50	0.81
Carbohydrate	1,074.87	1,007.40	462.08	0.46
Sugar	8.37	7.03	4.87	0.69
Fiber	4.59	3.94	2.37	0.60
Alcohol	120.56	115.67	89.92	0.78
Total Fat	1,190.24	1,475.27	2,513.69	1.70
Total Saturated Fatty Acids	76.80	112.61	202.91	1.80
Total Monounsaturated Fatty Acids	82.37	107.34	286.24	2.67
Total Polyunsaturated Fatty Acids	34.73	49.45	118.21	2.39
		Reduced Model		
	OLS	**WLS**	**SWLS**	
Variable	**VIF**	**VIF**	**VIF**	$\zeta_k\rho_k$
Age	1.00	1.00	0.98	0.98
Black	1.02	1.01	0.97	0.96
Total Fat	20.10	20.22	63.15	3.12
Total Monounsaturated Fatty Acids	20.16	20.26	61.57	3.04
		Reduced Model		
	OLS	**WLS**	**SWLS**	
Variable	**VIF**	**VIF**	**VIF**	$\zeta_k\rho_k$
Age	1.00	1.00	0.98	0.97
Black	1.00	1.03	1.00	1.00
Calorie	1.00	1.01	0.96	0.95

5. Conclusion

Regression diagnostics need to be adapted to be appropriate for models estimated from survey data to account for the use of weights and design features like stratification and clustering. In this paper we developed a new formulation for a variance inflation factor (VIF) appropriate for linear models. A VIF measures the amount by which the variance of a parameter estimator is inflated due to predictor variables being correlated with each other, rather than being orthogonal. Although survey-weighted regression slope estimates can be obtained from weighted least squares procedures in standard software packages, the VIFs produced by the non-survey routines are incorrect. The complex sample VIF is equal to the VIF from weighted least squares times an adjustment factor. The adjustment factor is positive but can be either larger or smaller than 1, depending on the nature of the data being analyzed.

In an empirical study, we illustrated the application of our new approach using data from the 2007-2008 National Health and Nutrition Examination Survey. We provided a simple example of how the collinearity among predictors affects the estimation of coefficients in linear regression and demonstrated that although the estimated coefficients (and fitted values) are the same when weighted least squares or survey-weighted least squares are used, their estimated variances and VIF values (reflecting the impact of collinearity on coefficient estimation) can be different.

The goals of an analysis must be considered in deciding how to use VIFs. If prediction is the main objective, then including collinear variables or selecting incorrect variables is less of a concern. If more substantive conclusions are desired, then an analyst should consider which variables should logically be included as predictors rather than relying on some automatic algorithm for variable selection. VIFs are a useful tool for identifying predictors whose estimated coefficients have variances that are unnecessarily large. Although VIFs might be considered as a tool for automatic variable selection, simulations in Liao (2010), not reported here, show that using VIFs is not a reliable way of identifying a true underlying model.

Acknowledgements

The authors thank the associate editor and referees whose comments led to important improvements.

Appendix A

Derivation of g^{kk}

Similar to the derivation of conventional OLS VIF in Theil (1971), the sum of squares and cross products matrix $\mathbf{A} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$, which can be partitioned as

$$\mathbf{A}_{p\times p} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T\tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{X}}_{(k)} \end{pmatrix} \qquad (15)$$

where the columns of $\tilde{\mathbf{X}}$ are reordered so that $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_k\ \tilde{\mathbf{X}}_{(k)})$ with $\tilde{\mathbf{X}}_{(k)}$ being the $n \times (p-1)$ matrix containing all columns except the k^{th} column of $\tilde{\mathbf{X}}$.

Using the formula for the inverse of a partitioned matrix, the upper-left element of $\mathbf{A}^{-1}$ can be expressed as:

$$\begin{aligned} a^{kk} &= \mathbf{i}_k^T\mathbf{A}^{-1}\mathbf{i}_k = \mathbf{i}_k^T(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\mathbf{i}_k \\ &= \frac{1}{\left(1-R^2_{\text{SW}_{(k)}}\right)\text{SST}_{\text{SW}_{(k)}}} \\ &= \frac{1}{\left(1-R^2_{\text{SW}_{(k)}}\right)\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k} \qquad (16) \end{aligned}$$

where

$$R^2_{\text{SW}_{(k)}} = \frac{\hat{\beta}^T_{\text{SW}_{(k)}}\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{X}}_{(k)}\hat{\beta}_{\text{SW}_{(k)}}}{\text{SST}_{\text{SW}_{(k)}}},$$

with $\hat{\beta}_{\text{SW}_{(k)}} = (\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{X}}_{(k)})^{-1}\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{x}}_k$, is the coefficient of determination corresponding to the regression of $\mathbf{x}_k$ on the $p-1$ other explanatory variables. The term $\text{SST}_{\text{SW}_{(k)}} = \tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k$, is the total sum of squares in this regression.

The term $(1-R^2_{\text{SW}_{(k)}})^{-1}$ in (16) is the VIF that will be produced by standard statistical packages when a weighted least squares regression is run. Under the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim (0, \sigma^2\mathbf{W}^{-1})$, expression (16) is equal to $\text{Var}_M(\hat{\beta}_{\text{SW}_k})/\sigma^2$. However, this is not appropriate for survey-weighted least squares regressions because the variance of $\hat{\boldsymbol{\beta}}_{\text{SW}}$ has the more complex form in (2).

The matrix $\mathbf{G} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$ can be expressed as:

$$\mathbf{G} = \begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix}\begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix}\begin{pmatrix} a^{kk} & \mathbf{a}^{k(k)} \\ \mathbf{a}^{(k)k} & \mathbf{A}^{(k)(k)} \end{pmatrix} \qquad (17)$$

where the inverse matrix is $\mathbf{A}^{-1} = [a^{hk}]$, $h, k = 1, ..., p$, $\mathbf{a}^{k(k)}$ is defined as the k^{th} row of $\mathbf{A}^{-1}$ excluding a^{kk}, $(a^{k1}, ..., a^{k(k-1)}, a^{k(k+1)}, ..., a^{kp})$, $\mathbf{a}^{(k)k} = [\mathbf{a}^{k(k)}]^T$ and $\mathbf{A}^{(k)(k)}$ is defined as the $(k-1)\times(k-1)$ part of matrix $\mathbf{A}^{-1}$ excluding the k^{th} row and column. The partitioned version of $\mathbf{B}$ is

$$\mathbf{B} = \begin{pmatrix} b_{kk} & \mathbf{b}_{k(k)} \\ \mathbf{b}_{(k)k} & \mathbf{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k & \tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)} \\ \tilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k & \tilde{\mathbf{X}}_{(k)}^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)} \end{pmatrix}. \qquad (18)$$

By virtue of the symmetry of $\mathbf{A}$ and $\mathbf{B}$, the k^{th} diagonal element of $\mathbf{G}$ is

$$g^{kk} = a^{kk}(a^{kk}b_{kk} + 2\mathbf{b}_{k(k)}\mathbf{a}^{(k)k}) + \mathbf{a}^{(k)kT}\mathbf{B}_{(k)(k)}\mathbf{a}^{(k)k}. \qquad (19)$$

Using the partitioned inverse of matrix $\mathbf{A}$, which represents $(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}$, it can be shown that

$$\mathbf{a}^{(k)k} = -a^{kk}(\tilde{\mathbf{X}}^T_{(k)}\tilde{\mathbf{X}}_{(k)})^{-1}\tilde{\mathbf{X}}_{(k)}\tilde{\mathbf{x}}_k = -a^{kk}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}. \qquad (20)$$

Substituting $a^{(k)k}$ in (19), g^{kk} can be compactly expressed in terms of a^{kk}, $\hat{\boldsymbol{\beta}}_{\text{SW}(k)}$ and the lower right component of matrix $\mathbf{B}$:

$$\begin{aligned} g^{kk} &= (a^{kk})^2\left(b_{kk} - 2\mathbf{b}_{k(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{SW}(k)}\mathbf{B}_{(k)(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right) \\ &= a^{kk}\times\frac{1}{1-R^2_{\text{SW}(k)}}\frac{1}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k} \\ &\quad \times\left(\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{x}}_k - 2\tilde{\mathbf{x}}_k^T\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)} + \hat{\boldsymbol{\beta}}^T_{\text{SW}(k)}\tilde{\mathbf{X}}_{(k)}\tilde{\mathbf{V}}\tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right) \\ &= a^{kk}\times\frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)^T\tilde{\mathbf{V}}\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)}{\left(1-R^2_{\text{SW}(k)}\right)\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k} \\ &= a^{kk}\times\frac{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)^T\tilde{\mathbf{V}}\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)}{\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)^T\left(\tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}\right)} \\ &= \frac{1}{1-R^2_{\text{SW}(k)}}\frac{1}{\tilde{\mathbf{x}}_k^T\tilde{\mathbf{x}}_k}\frac{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{V}}\tilde{\mathbf{e}}_{xk}}{\tilde{\mathbf{e}}_{xk}^T\tilde{\mathbf{e}}_{xk}}, \qquad (21) \end{aligned}$$

where $\tilde{\mathbf{e}}_{xk} = \tilde{\mathbf{x}}_k - \tilde{\mathbf{X}}_{(k)}\hat{\boldsymbol{\beta}}_{\text{SW}(k)}$ is the residual from regressing $\tilde{\mathbf{x}}_k$ on $\tilde{\mathbf{X}}_{(k)}$.

References

Belsley, D.A. (1984). Collinearity and forecasting. *Journal of Forecasting*, 38, 73-93.

Belsley, D.A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley & Sons, Inc.

Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics*: *Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York: Wiley Interscience.

Elliot, M. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, 33, 23-34.

Farrar, D.E., and Glauber, R.R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, 49, 92-107.

Fox, J. (1984). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York: John Wiley & Sons, Inc.

Fox, J., and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417), 178-183.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28(1), 5-23.

Kmenta, J. (1986). *Elements of Econometrics*. New York: Macmillan, 2nd Ed.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3341-3348.

Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland. Available at http://drum.lib.umd.edu/bitstream/1903/7598/1/umi-umd-4863.pdf.

Li, J., and Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, 35(1), 15-24.

Li, J., and Valliant, R. (2011). Linear regression influence diagnostics for unclustered survey data. *Journal of Official Statistics*, 20, 99-119.

Liao, D. (2010). *Collinearity Diagnostics for Complex Survey Data*. Unpublished doctoral dissertation, University of Maryland. Available at http://drum.lib.umd.edu/bitstream/1903/10881/1/Liao_umd_0117E_11537.pdf.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, 13, 1517-1520.

Neter, J., Kutner, M., Wasserman, W. and Nachtsheim, C. (1996). *Applied Linear Statistical Models*. New York: McGraw-Hill/Irwin, 4th Ed.

Simon, S.D., and Lesage, J.P. (1988). The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management*, 1, 137-152.

Smith, G. (1974). Multicollinearity and forecasting. Yale Univeristy Cowles Foundation Disscussion Paper No.383. Available at http://cowles.econ.yale.edu/P/cd/d03b/d0383.pdf.

Steward, G.W. (1987). Collinearity and least squares regression. *Statistical Science*, 2(1), 68-84.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons, Inc.

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 63-72
Statistics Canada, Catalogue No. 12-001-X

Estimating agreement coefficients from sample survey data

Hung-Mo Lin, Hae-Young Kim, John M. Williamson and Virginia M. Lesser [1]

Abstract

We present a generalized estimating equations approach for estimating the concordance correlation coefficient and the kappa coefficient from sample survey data. The estimates and their accompanying standard error need to correctly account for the sampling design. Weighted measures of the concordance correlation coefficient and the kappa coefficient, along with the variance of these measures accounting for the sampling design, are presented. We use the Taylor series linearization method and the jackknife procedure for estimating the standard errors of the resulting parameter estimates. Body measurement and oral health data from the Third National Health and Nutrition Examination Survey are used to illustrate this methodology.

Key Words: Clustering; Concordance correlation coefficient; Generalized estimating equations; Jackknife estimator; Kappa coefficient; Sample weighting; Stratification; Taylor series linearization.

1. Introduction

Surveys often collect multiple measures of latent conditions such as quality of life and aspiration for a college education, as well as multiple measures of difficult- to-classify conditions such as having chronic fatigue syndrome. When multiple measures are collected, interest naturally focuses on the agreement between the multiple measures and in obtaining confidence intervals on those agreement measures. Also, there may be interest in contrasting agreement across population subgroups and across alternate pairings of measurements. In this context, one might be interested in testing equality of agreement measures. This paper focuses on two measures of agreement between such multiple measures, the concordance correlation coefficient (CCC, ρ_c) and the kappa (κ) coefficient. The former is useful for continuous measurements with natural scales. If a measure of a latent concept has no natural scale, then it can be arbitrarily rescaled to have mean zero and unit variance. When this is possible, it is meaningless to talk about differences in marginal moments. However, if there is a natural scale, then rescaling is not desirable and a good measure of agreement will take into account both correlation and agreement of marginal moments. The kappa coefficient is most useful for binary classifications.

The CCC has been shown to be more appropriate for measuring agreement or reproducibility (Lin 1989; Lin 1992) than the Pearson correlation coefficient (ρ). It evaluates the accuracy between two readings by measuring the variation of the fitted linear relationship from the 45^0 line through the origin (the concordance line) and precision by measuring how far each observation deviates from the fitted line. Let Y_{i1} and Y_{i2} denote a pair of continuous random variables measured on the same subject i using two methods. The CCC for measuring the agreement of Y_{i1} and Y_{i2} is defined as follows:

$$\rho_c = 1 - \frac{E[(Y_{i1} - Y_{i2})^2]}{E_{\text{indep}}[(Y_{i1} - Y_{i2})^2]} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} \quad (1)$$

where $\sigma_1^2 = \text{var}(Y_{i1}), \sigma_2^2 = \text{var}(Y_{i2}),$ and $\sigma_{12} = \text{cov}(Y_{i1}, Y_{i2})$ (Lin 1989). As noted by Lin (1989), $\rho_c = 0$ if and only if $\rho = 0.$ It can also be shown algebraically that ρ_c is proportional to ρ and that $-1 \le -|\rho| \le \rho_c \le |\rho| \le 1$ (Lin 1989). Hence imprecision can be reflected by a smaller ρ and systematic bias can be reflected by a smaller ratio of ρ_c relative to $\rho.$ Together, information on ρ and ρ_c provide a set of tools to identify which corrective actions, either to improve accuracy and/or to improve precision, is most beneficial (Lin and Chinchilli 1997).

The intraclass correlation coefficient (ICC) is also a popular measure of agreement for variables measured on a continuous scale (Fleiss 1986). Suppose Y_{i1} and Y_{i2} can be described in a linear model as follows: $y_{ij} = \mu_j + \theta_i + e_{ij}$ where μ_j is the mean of the measurement from the j^{th} method, $\theta_i \sim (0, \sigma_\theta^2)$ is the latent variable for the i^{th} subject, and the $e_{ij} \sim (0, \sigma_e^2)$ are independent errors terms. Carrasco and Jover (2003, page 850) used a model with variance components to demonstrate that the CCC is the intraclass correlation coefficient (ICC) when one takes into account the difference in averages of the methods:

$$\rho_{\text{ICC}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_e^2 + \sigma_\mu^2} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2}.$$

[1] Hung-Mo Lin, Department of Anesthesiology, Mount Sinai School of Medicine, One Gustave L. Levy Place, Box 1010, New York, NY 10029, U.S.A. E-mail: hung-mo.lin@mountsinai.org; Hae-Young Kim, Center for Statistical Analysis and Research, New England Research Institutes, 9 Galen Street, Watertown, MA 02472, U.S.A.; John M. Williamson, Center for Global Health Research, Centers for Disease Control and Prevention/Kenya Medical Research Institute, 1578 Kisumu-Busia Road, Kisian Kisumu, Kenya; Virginia M. Lesser, Department of Statistics and Survey Research Center, Oregon State University, 44 Kidder Hall, Corvallis, OR 97731-4606, U.S.A.

Therefore, one can estimate the CCC using the variance components of a mixed effects model or the common method of moments. Because of its superiority to the Pearson correlation coefficient and its link to the ICC, application of the CCC has gained popularity in recent years (Chinchilli, Martel, Kumanyika and Lloyd 1996; Zar 1996). In 2009 and the 2010, the CCC was used as a measure of agreement in more than 60 medical publications in areas such as respiratory illness (Dixon, Sugar, Zinreich, Slavin, Corren, Naclerio, Ishii, Cohen, Brown, Wise and Irvin 2009; Kocks, Kerstjens, Snijders, de Vos, Biermann, van Hengel, Strijbos, Bosveld and van der Molen 2010), sleep (Khawaja, Olson, van der Walt, Bukartyk, Somers, Dierkhising and Morgenthaler 2010), pediatrics (Liottol, Radaelli, Orsi1, Taricco, Roggerol, Giann, Consonni, Mosca1 and Cetin 2010), neurology (MacDougall, Weber, McGarvie, Halmagyi and Curthoys 2009), and radiology (Mazaheri, Hricak, Fine, Akin, Shukla-Dave, Ishill, Moskowitz, Grater, Reuter, Zakian, Touijer and Koutcher 2009).

The kappa coefficient (κ) (Cohen 1960) and the weighted kappa coefficient (Cohen 1968) are the most popular indices for measuring agreement for discrete and ordinal outcomes, respectively (Fleiss 1981). Let Y_{i1} and Y_{i2} denote two binary random variables taking values 0 and 1 with probabilities denoted by $\pi_1 = \Pr(Y_{i1} = 1)$ and $\pi_2 = \Pr(Y_{i2} = 1)$. Kappa corrects the percentage of agreement between raters by taking into account the proportion of agreement expected by chance (calculated under independence), and is defined as follows:

$$\kappa = \frac{P_o - P_e}{1.0 - P_e}, \qquad (2)$$

where P_e is the probability that the pair of binary responses are equal assuming independence $(\pi_1\pi_2 + (1-\pi_1)(1-\pi_2))$ and P_o is the probability that the pair are equal (Cohen 1960). The difference $P_o - P_e$ is the excess of agreement over chance agreement. A value of 0 for κ indicates no agreement beyond chance and a value of 1 indicates perfect agreement (Fleiss 1981). Disadvantages of kappa are that is a function of the marginal distribution of the raters (Fleiss, Nee and Landis 1979; Tanner and Young 1985) and its range depends on the number of ratings per subject (Fleiss *et al.* 1979). Robieson (1999) noted that the CCC computed from ordinal scaled data is equivalent to the weighted kappa when integer scores are used. Kappa has been used to measure the validity and reproducibility of the similarity between twins (Klar, Lipsitz and Ibrahim 2000), different epidemiologic tools (Maclure and Willett 1987), and control-informant agreement from case-control studies (Korten, Jorm, Henderson, McCusker and Creasey 1992).

The value of sample surveys have been well recognized and estimation for data collected from sample surveys has been widely documented (Hansen, Hurwitz and Madow 1953; Cochran 1963; Kish 1965). For example, a number of federal studies conducted in the U.S. to obtain estimates of the health of the population are based on national surveys, such as the National Health Interview Survey (NHIS), the Behavioral Risk Factor Surveillance System (BRFSS), and the National Health and Nutrition Examination Surveys (NHANES). Each of these studies incorporates complex survey design structure, namely oversampling of subpopulations, stratification and clustering. These designs are often used to improve precision, provide estimates for subpopulations, or reduce costs associated with frame development. In order to draw design-based inference to the targeted population for complex survey designs, estimators and their variances include sampling weights and account for the design structure to obtain unbiased estimates. In addition, by including the sampling weights and incorporating the sample design in analyses, any potential correlation from the clusters in a multistage design is taken into account so that the standard errors of the estimators are not underestimated.

Often researchers are not interested in testing whether their estimation of agreement using either the CCC or kappa is significantly different from zero. Their interest is to report the confidence intervals along with their estimates (*e.g.*, Dixon *et al.* 2009; Mazaheri *et al.* 2009). Similar to the Pearson correlation coefficient, there is no target value that can be used to judge if agreement is strong. Therefore, it is essential that judgment of agreement between any test and reference methods should be made with an established degree of certainty. In some situations, studies are conducted that require hypothesis testing or comparisons of agreement indexes for more than one new methods against a reference method. For examples, Khawaja *et al.* (2010) tested the equality of two CCCs that compared the apnea hypopnea index (AHI) from the first 2 and 3 hours of sleep with the gold standard AHI from FN-PSG (FN-AHI). In radiology research, associations between volume measurements of prostate tumor from imaging and also from pathologic examination were assessed by comparing CCCs. The two imaging methods were tested for equality of agreement with the pathologic results (Mazaheri *et al.* 2009). Tests of equal kappa have been used to compare visual assessment and computerized planimetry in assessing cervical ectopy (Gilmour, Ellerbrock, Koulos, Chiasson, Williamson, Kuhn and Wright 1997; Williamson, Manatunga and Lipsitz 2000), and in comparing monozygotic and dizygotic twins in terms of cholesterol levels (Feinleib, Garrison, Fabsitz, Christian, Hrubec, Borhani, Kannel, Roseman, Schwartz and Wagner 1977).

As illustrated in the two NHANES III examples in Section 3, large differences can exist between the weighted and unweighted estimates of parameter estimate standard errors in survey studies. Failure to include sampling weights and take into account the sample design in analyses will result in underestimation of standard errors and incorrect inference. This is especially important for surveys repeated every few years, and researchers often have a special interest in comparing changes among domains or subpopulations. For instance, in the first NHANES III application, we compare the agreement between self reported and measured body weights at examination in adolescents. Computing accurate standard errors (confidence intervals) are necessary if interest is to compare the CCC across domains, such as normal weight and obese subgroups.

We provide weighted measures of the CCC and kappa coefficient, along with the variance estimators of these measures accounting for the sampling design. In Section 2, we present a generalized estimating equations approach for estimating these two agreement coefficients from sample survey data. In Section 3, we illustrate our method with data collected from the NHANES III study. We use body measurement data to estimate ρ_c for assessing the agreement between self-reported and actual weight. We also use oral health data to estimate κ for assessing the agreement between two definitions of periodontal disease. We account for stratification and clustering, and incorporate weights of the survey design in both examples. We conclude with a short discussion.

2. Methods

We propose a general approach for estimating the CCC and kappa from sample survey data using two GEE approaches. For the CCC, three sets of estimating equations are required. A first set of estimating equations models the distribution of the continuous responses. Following Barnhart and Williamson (2001), a second set of estimating equations is used to estimate the variances of the continuous responses. A third set of estimating equations estimates the CCC by modeling the covariance between the paired continuous responses and the estimates of the means and variances from the first two sets of estimating equations. For κ, only two sets of estimating equations are required. A first set of estimating equations models the marginal distribution of the binary responses. Following Lipsitz, Laird and Brennan (1994), a second set of estimating equations is introduced to estimate κ by modeling a binary random variable depicting agreement between two responses on a subject.

In order to account for variable selection probabilities, weight matrices are incorporated into each set of estimating equations. Standard error estimation of the proposed $\hat{\rho}_c$ and $\hat{\kappa}$ from sample survey data are conducted with the Taylor series linearization method. We also show how standard error estimation of the proposed estimators can be accomplished by using the jackknife approach.

Assume a sample survey is conducted with stratification, clustering, and unequal probabilities of selection. Let Y_{hij} denote the response variable for the j^{th} member ($j = 1, ..., m_{hi}$) of the i^{th} cluster ($i = 1, ..., n_h$) of the h^{th} stratum ($h = 1, ..., H$). Averaging over all possible samples, the corresponding expected value is $E[Y_{hij}] = \mu_{hij}$ if Y_{hij} is a continuous response, and the corresponding probability $E[Y_{hij}] = \Pr[Y_{hij} = 1] = \pi_{hij}$ if Y_{hij} is a binary response. The sampling weight w_{hij} is the inverse of the probability of selection for the j^{th} member of the i^{th} cluster of the h^{th} stratum.

2.1 The concordance correlation coefficient

Liang and Zeger (1986) developed moment-based methods for analyzing correlated observations from the same cluster (*e.g.*, repeated measurements over time on the same individual or observations on multiple members of the same family). The GEE approach results in consistent marginal parameter estimation, even with misspecification of the correlation structure by using a robust "sandwich" estimator of variance. We use the GEE approach to analyze sample survey data by additionally incorporating a sampling weight matrix as follows:

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} \mathbf{D}'_{hi}\, \mathbf{W}_{hi}\, \mathbf{V}_{hi}^{-1} (\mathbf{Y}_{hi} - \boldsymbol{\mu}_{hi}(\hat{\mu})) = \mathbf{0},$$

where $\mathbf{D}'_{hi}$ is the $(q \times m_{hi})$ derivative matrix $d[\boldsymbol{\mu}_{hi}]'/d\boldsymbol{\mu}$, $\mathbf{W}_{hi}$ is a $(m_{hi} \times m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij}, $\mathbf{V}_{hi}$ is a $(m_{hi} \times m_{hi})$ working variance-covariance matrix for the within-cluster responses, $\mathbf{Y}_{hi}$ is a $(m_{hi} \times 1)$ response vector consisting of the responses Y_{hij}, and $\boldsymbol{\mu}_{hi} = E[\mathbf{Y}_{hi}]$ is possibly a function of the $(q \times 1)$ parameter vector $\boldsymbol{\beta}$. The GEE can then be solved non-iteratively, resulting in the usual estimate

$$\hat{\mu} = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\, Y_{hij} \right) \Bigg/ \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \right)$$

if we are estimating a common mean $\mu = \boldsymbol{\beta}$ $(q = 1)$ and are using an independence working covariance matrix.

Assume a pair of continuous responses are observed for the j^{th} member of the i^{th} cluster of the h^{th} stratum, Y_{hij1} and Y_{hij2}, and their expected values are μ_{hij1} and μ_{hij2}. Again, assume we are estimating common means μ_1 and μ_2 without covariates for the pair of within-subject continuous responses, which can be estimated by using the above generalized estimating equation.

Barnhart and Williamson (2001) demonstrated how three sets of generalized estimating equations can be used to model the CCC defined in (1) using correlated data. We extend Barnhart and Williamson's (2001) second set of GEE equations to estimate the variances of the continuous responses by again incorporating a weight matrix as follows:

$$\nu_2(\hat{\sigma}_1^2, \hat{\sigma}_1^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \mathbf{F}'_{hi}\, \mathbf{W}_{hi}\, \mathbf{H}_{hi}^{-1}(\mathbf{Y}_{hi}^2 - \delta_{hi}^2(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = \mathbf{0},$$

where $\mathbf{F}'_{hi}$ is the $(2 \times 2m_{hi})$ derivative matrix $d[\delta_{hi}^2]'/d\boldsymbol{\sigma}^2$ with $\boldsymbol{\sigma}^2 = [\sigma_1^2, \sigma_2^2]'$, $\mathbf{W}_{hi}$ is a $(2m_{hi} \times 2m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij}, $\mathbf{H}_{hi}$ is a $(2m_{hi} \times 2m_{hi})$ working variance-covariance matrix for the within-cluster squared responses, $\mathbf{Y}_{hi}^2 = [Y_{hi11}^2, Y_{hi12}^2, Y_{hi21}^2, Y_{hi22}^2, ..., Y_{him_{hi}1}^2, Y_{him_{hi}2}^2]'$ is a $(2m_{hi} \times 1)$ response vector of the continuous variables, and $\delta_{hi}^2 = E[\mathbf{Y}_{hi}^2]$. Although δ_{hi}^2 is a function of both the variance terms σ_1^2 and σ_2^2 and the means μ_1 and μ_2, it is assumed that the means are fixed in δ_{hi}^2 and one only takes derivatives of δ_{hi}^2 with respect to the variances. Again we choose the $(2m_{hi} \times 2m_{hi})$ matrix $\mathbf{H}_{hi}$ to be the "independence" working variance-covariance matrix and the $(2m_{hi} \times 1)$ column vector $\delta_{hi}^2 = [\sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2, ..., \sigma_1^2 + \mu_1^2, \sigma_2^2 + \mu_2^2]'$ because we are assuming common variances and means across all strata and clusters. The above GEE can thus be solved non-iteratively:

$$\hat{\sigma}_p^2 = \left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} w_{hijp}\, Y_{hijp}^2\right) \Bigg/ \left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} w_{hijp}\right) - \hat{\mu}_p^2,$$

for the p^{th} measurement in the pair, $p = 1, 2$.

The CCC can be estimated in a third set of estimating equations by using the pairwise products of the responses to model σ_{12}, once the means and variances are estimated. Let $\mathbf{U}_{hi} = [Y_{hi11}Y_{hi12}, Y_{hi21}Y_{hi22}, ..., Y_{him_{hi}1}Y_{him_{hi}2}]'$ be a $(m_{hi} \times 1)$ vector of pairwise products of the responses and denote $\boldsymbol{\theta}_{hi} = E[\mathbf{U}_{hi}]$, which is a function of the means, variances, and CCC. We solve for $\hat{\rho}_c$ in a third set of estimating equations:

$$\nu_3(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_1^2, \hat{\mu}_1, \hat{\mu}_2) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} \mathbf{C}'_{hi}\, \mathbf{W}_{hi}\, \mathbf{K}_{hi}^{-1}(\mathbf{U}_{hi} - \boldsymbol{\theta}_{hi}(\hat{\rho}_c, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\mu}_1, \hat{\mu}_2)) = 0,$$

where $\mathbf{C}'_{hi}$ is a $(1 \times m_{hi})$ derivative vector $= \partial\boldsymbol{\theta}_{hi}/\partial\rho_c$, $\mathbf{W}_{hi}$ is a $(m_{hi} \times m_{hi})$ main diagonal matrix consisting of the person-specific sampling weights w_{hij}, and $\mathbf{K}_{hi}$ is a $(m_{hi} \times m_{hi})$ working covariance matrix that we choose to be the "independence" covariance matrix. The above GEE can be solved non-iteratively:

$$\hat{\rho}_c = \frac{2\hat{\sigma}_{12}}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2 + (\hat{\mu}_1 - \hat{\mu}_2)^2},$$

where

$$\hat{\sigma}_{12} = \frac{\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} w_{hij12}\, Y_{hij1}\, Y_{hij2}\right)}{\left(\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{j=1}^{M_{hi}} w_{hij12}\right)} - \hat{\mu}_1\hat{\mu}_2.$$

2.2 Linearization estimator of variance

The usual robust estimators of variance for the means and CCC from the GEE approach are invalid here because they do not take into account the sampling structure, only the correlation of observations made on the same individual. We propose standard error estimation using the Taylor series linearization method (Binder 1983; Binder 1996). The first derivatives of ρ_c (equation 1) with respect to μ_1, μ_2, σ_1^2, σ_2^2, and σ_{12} are:

$$\frac{\partial\rho_c}{\partial\mu_1} = \frac{-4\sigma_{12}(\mu_1 - \mu_2)}{D^2},$$
$$\frac{\partial\rho_c}{\partial\mu_2} = \frac{-4\sigma_{12}(\mu_2 - \mu_1)}{D^2},$$
$$\frac{\partial\rho_c}{\partial\sigma_1^2} = \frac{-2\sigma_{12}}{D^2},$$
$$\frac{\partial\rho_c}{\partial\sigma_2^2} = \frac{-2\sigma_{12}}{D^2},$$
$$\frac{\partial\rho_c}{\partial\sigma_{12}} = \frac{2}{D},$$

where $D = \sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2$. Thus

$$\begin{aligned}
\hat{\rho}_c - \rho_c \approx{}& \left(\frac{\partial\rho_c}{\partial\mu_1}\right)(\hat{\mu}_1 - \mu_1) + \left(\frac{\partial\rho_c}{\partial\mu_2}\right)(\hat{\mu}_2 - \mu_2) \\
&+ \left(\frac{\partial\rho_c}{\partial\sigma_1^2}\right)(\hat{\sigma}_1^2 - \sigma_1^2) \\
&+ \left(\frac{\partial\rho_c}{\partial\sigma_2^2}\right)(\hat{\sigma}_2^2 - \sigma_2^2) + \left(\frac{\partial\rho_c}{\partial\sigma_{12}}\right)(\hat{\sigma}_{12} - \sigma_{12}) \\
={}& \frac{-4\sigma_{12}(\mu_1 - \mu_2)}{D^2}(\hat{\mu}_1 - \mu_1) \\
&+ \frac{-4\sigma_{12}(\mu_2 - \mu_1)}{D^2}(\hat{\mu}_2 - \mu_2) \\
&+ \frac{-2\sigma_{12}}{D^2}(\hat{\sigma}_1^2 - \sigma_1^2) + \frac{-2\sigma_{12}}{D^2}(\hat{\sigma}_2^2 - \sigma_2^2) \\
&+ \frac{2}{D}(\hat{\sigma}_{12} - \sigma_{12}).
\end{aligned}$$

The above equation can be rearranged into two parts, one involving the parameter estimates $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$, and $\hat{\sigma}_{12}$ and the other involving only parameters which does not contribute to the variance estimation of $\hat{\rho}_c$. Thus the first part becomes

$$\begin{aligned}
&-\frac{4\sigma_{12}(\mu_1-\mu_2)}{D^2}\hat{\mu}_1-\frac{4\sigma_{12}(\mu_2-\mu_1)}{D^2}\hat{\mu}_2\\
&\quad-\frac{2\sigma_{12}}{D^2}\hat{\sigma}_1^2-\frac{2\sigma_{12}}{D^2}\hat{\sigma}_2^2+\frac{2}{D}\hat{\sigma}_{12}\\
&=-\frac{2\sigma_{12}}{D^2}(2(\mu_1-\mu_2)(\hat{\mu}_1-\hat{\mu}_2)+\hat{\sigma}_1^2+\hat{\sigma}_2^2)+\frac{2}{D}\hat{\sigma}_{12}\\
&=-\frac{2\sigma_{12}}{D^2}\left(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}2(\mu_1-\mu_2)(w_{hij}^*Y_{hij1}-w_{hij}^*Y_{hij2})\right.\\
&\qquad\qquad\left.+\,w_{hij}^*(Y_{hij1}-\mu_1)^2+w_{hij}^*(Y_{hij2}-\mu_2)^2\right)\\
&\quad+\frac{2}{D}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}^*(Y_{hij1}-\mu_1)(Y_{hij2}-\mu_2) \qquad (3)
\end{aligned}$$

where $w_{hij}^*=w_{hij}/(\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij})$. Equation (3) becomes a linear function of the data after the summation is moved to the front, which we can then express as $\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}^*\,z_{hij}$, where

$$\begin{aligned}
z_{hij}&=-\frac{2\sigma_{12}}{D^2}(2(\mu_1-\mu_2)(Y_{hij1}-Y_{hij2})\\
&\quad+(Y_{hij1}-\mu_1)^2+(Y_{hij2}-\mu_2)^2)\\
&\quad+\frac{2}{D}(Y_{hij1}-\mu_1)(Y_{hij2}-\mu_2). \qquad (4)
\end{aligned}$$

One then creates a random variable $\hat{z}_{hij}$ based on equation (4) that replaces the parameters with their respective estimates. The variance of this new estimator $\hat{z}_{hij}$ is an approximation for the variance of $\hat{\rho}_c$, which can be estimated using standard survey software (see Appendix).

2.3 Jackknife estimator of variance

We also use the jackknife technique for standard error estimation of the parameters following Rust and Rao (1996, Section 2.1) for comparison with the linearization estimates. The jackknife technique is implemented by calculating a set of replicate estimates and estimating the variance using them. A replicate data set is created for each cluster by deleting all observations from the given cluster from the sample. The weights of all other observations in the stratum containing the cluster are inflated by a factor $n_h/(n_h-1)$. Weights in the other strata remain unchanged. Thus, the new weights for the replicated data set created by removing cluster i from stratum h are:

$$\begin{array}{lll}
\omega_{klj}^{(hi)}= & w_{klj} & \text{if } k\neq h \text{ (different strata)}\\[1ex]
\omega_{hlj}^{(hi)}= & w_{hlj}n_h/(n_h-1) & \text{if } l\neq i\\
 & & \text{(same strata but different clusters)}\\[1ex]
\omega_{hij}^{(hi)}= & 0 & \text{(for the cluster being removed).}
\end{array}$$

The resulting jackknife variance estimator for $\hat{\rho}_c$ is

$$v_J(\hat{\rho}_c)=\sum_{h=1}^{H}\left(\frac{n_h-1}{n_h}\right)\sum_{i=1}^{n_h}(\hat{\rho}_{c(hi)}-\hat{\rho}_c)^2$$

where $\hat{\rho}_{c(hi)}$ is estimated in the same way as $\hat{\rho}_c$, but using the recalculated weights $\omega^{(hi)}$ instead of the weights ω. The jackknife estimators for the means are similarly calculated.

2.4 The kappa coefficient

Assume a pair of binary responses are observed for the j^{th} member of the i^{th} cluster of the h^{th} stratum, Y_{hij1} and Y_{hij2}, and their expected values are the probabilities π_{hij1} and π_{hij2}. Again assume we are estimating common probabilities π_1 and π_2 without covariates for the pair of within-subject binary responses. Lipsitz *et al.* (1994) demonstrated how two sets of generalized estimating equations can be used to develop simple non-iterative estimates of the κ-coefficient that can be used for unbalanced data as previous estimates of kappa and its variance were only proposed for balanced data. They defined the binary random variable $U_{hij}=Y_{hij1}Y_{hij2}+(1-Y_{hij1})(1-Y_{hij2})=1$ if both responses in the pair agree and 0 otherwise. Accordingly, $E[U_{hij}]=P_o$, which denotes the probability of observed agreement and is assumed here to be constant over all strata, clusters, and pairs of observations. Now let $E[Y_{hij1}Y_{hij2}]=\Pr[Y_{hij1}=Y_{hij2}=1]=\omega$. The probability of observed agreement can be expressed as $P_o=1-\pi_1-\pi_2+2\omega$. The probability of expected agreement by chance is defined as $P_e=\pi_1\pi_2+(1-\pi_1)(1-\pi_2)$ and is estimated by $\hat{P}_e=\hat{\pi}_1\hat{\pi}_2+(1-\hat{\pi}_1)(1-\hat{\pi}_2)$, where $\hat{\pi}_1$ and $\hat{\pi}_2$ are calculated in the first set of estimating equations.

We can derive estimates of κ from sample survey data following the approach for the CCC in Section 2.1. We can incorporate the survey weight matrices into Lipsitz *et al.*'s (1994) two sets of GEE equations for estimating kappa. Then, by choosing "independence" working covariance matrices for the two sets of equations as in Lipsitz *et al.*'s (1994) approach, we arrive at the following non-iterative estimate of kappa for sample survey data:

$$\hat{\kappa}=\frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}U_{hij}-\hat{P}_e\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}-\hat{P}_e\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}w_{hij}}. \qquad (5)$$

This estimator is identical to Lumley's (2010), which can be computed using the R software survey package and svykappa function.

Standard error estimation of $\hat{\kappa}$ can be conducted similarly to that of $\hat{\rho}_c$ using the Taylor series linearization method. The first derivatives of kappa with respect to $P_o,\pi_1,$ and π_2 are:

$$\frac{\partial\kappa}{\partial P_o} = \frac{1}{1-P_e},$$
$$\frac{\partial\kappa}{\partial\pi_1} = \frac{(1-P_o)(1-2\pi_2)}{(1-P_e)^2},$$
$$\frac{\partial\kappa}{\partial\pi_2} = \frac{(1-P_o)(1-2\pi_1)}{(1-P_e)^2}.$$

Thus

$$\begin{aligned}\hat{\kappa} - \kappa &\approx \left(\frac{\partial\kappa}{\partial P_o}\right)(\hat{P}_o - P_o) \\ &\quad + \left(\frac{\partial\kappa}{\partial\pi_1}\right)(\hat{\pi}_1 - \pi_1) + \left(\frac{\partial\kappa}{\partial\pi_2}\right)(\hat{\pi}_2 - \pi_2) \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}^{*}\, z_{hij},\end{aligned}$$

where $w_{hij}^{*} = w_{hij} / (\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij})$ and

$$\begin{aligned}z_{hij} &= \left(\frac{\partial\kappa}{\partial P_o}\right)U_{hij} + \left(\frac{\partial\kappa}{\partial\pi_1}\right)Y_{hij1} + \left(\frac{\partial\kappa}{\partial\pi_2}\right)Y_{hij2} \\ &= \frac{U_{hij}}{1-P_e} + \frac{(1-P_o)[Y_{hij1}(1-2\pi_2) + Y_{hij2}(1-2\pi_1)]}{(1-P_e)^2}. \quad (6)\end{aligned}$$

Replacing the parameters in (6) with their respective estimates, one then treats $\hat{z}_{hij}$ as a random variable and estimates its variance using standard survey software that accounts for the sampling design. The variance of this new estimator $\hat{z}_{hij}$ is an approximation for the variance of $\hat{\kappa}$. The jackknife method can also be used to estimate the variance of $\hat{\kappa}$.

3. NHANES III survey

We used data from the Third National Health and Nutrition Examination Survey to illustrate our method. NHANES III was conducted by the National Center for Health Statistics of the Centers for Disease Control and Prevention and was designed as a six-year survey divided into two phases (1988-1991 and 1991-1994). The data were collected using a complex, multistage, probability sampling design to select participants representative of the civilian, non-institutionalized US population. Details of the survey design and analytic and reporting guidelines were published in the NHANES III reference manuals and reports (National Center for Health Statistics 1996).

3.1 The adolescent weight study

Obesity is a rapidly increasing public health problem with surveillance most often based on self-reported values of height and weight. A series of recent studies and systemic reviews have attempted to assess the agreement between self-reported and measured weight, especially in the adolescent population. The general findings suggest that self-reported weight was slightly lower than measured weight, and that a significant number of normal weight adolescents misperceive themselves as overweight and are engaging in unhealthy weight control behaviors (Field, Aneja and Rosner 2007; Gorber, Tremblay, Moher and Gorber 2007; Sherry, Jefferds and Grummer-Strawn 2007). Therefore, researchers have suggested that obesity prevention programs should address weight misperceptions and the harmful effects of unhealthy weight control methods even among normal weight adolescents (Talamayan, Springer, Kelder, Gorospe and Joye 2006). A similar Canadian study from the 2005 Canadian Community Health Survey that focused on adult individuals also showed that associations between obesity and health conditions may be overestimated if self-reported weight is used (Shield, Gorber and Tremblay 2008). We use data obtained from the Body Measurements (Anthropometry) component of the NHANES III study to estimate the CCC that measures agreement between self-reported and measured weight (pounds) obtained from adolescents (aged 12 through 16 years).

The self-reported weight was obtained just prior to the actual measurement of weight. We use data from the entire six-year survey period (both 1988-1991 and 1991-1994). For simplicity, we excluded one stratum which only had one PSU. Hence, there were 48 strata and each stratum had two PSUs. The sample weight labeled wtpfex6 accounting for the differential selection probability was used in our analyses. There were 1,651 subjects with complete data for both weight measurements. The estimates of the self-reported and actual weights (in pounds) were 135.5 (s.e. = 1.8) and 136.3 (s.e. = 1.8), respectively, calculated using PROC SURVEYMEANS in SAS. The estimates of the standard errors based on the jackknife approach are the same as above.

The CCC is a natural choice for assessing the agreement between the two weight measurements because they are measured on the same scale and their ranges are similar (self-reported weight: 78 lbs ~ 350 lbs and actual weight: 73 lbs ~ 372 lbs) (Lin and Chinchilli 1997). The estimate of the CCC for measuring the agreement between the two definitions of weight using the proposed method is 0.93. The standard error of the estimate is 0.021 using the Taylor series linearization method. The jackknife standard error of 0.021 agrees closely with the linearization standard error. These statistics are summarized in Table 1 along with their values computed when the sampling structure is ignored. The standard errors for the estimates incorporating the sampling structure are much larger than the unweighted estimates.

Table 1
Unweighted and weighted average, CCC, and respective standard errors for adolescent self-reported and actual weight in pounds

	Self-reported	Actual	CCC
Unweighted Estimate	135.31	136.96	0.890
SE	0.76	0.80	0.0005
Weighted Estimate	135.47	136.30	0.926
SE	1.75	1.82	0.0205

Similar to the CCC, the usual Pearson correlation coefficient between the self-reported and the actual weight measures is also 0.93. In this case, the mean difference between the two weight measurements is just less than one pound. When subpopulations are examined, differences are noted in the CCC and the Pearson correlation coefficient. Consider a subpopulation of those individuals that had a measured weight $>$ 200 lbs at examination. Summarizing the data for this subpopulation, the self-reported weight is on average 8 pounds less than the measured weight (223.2 lbs *vs* 231.4 lbs). There is a slight departure of the CCC (0.72) from the Pearson correlation coefficient (0.76). The discrepancy between the two measures increases in the more obese subgroup. In the subpopulation where measured weight is $>$ 220 lbs, the means of self-reported and measured weights are 231.9 lbs and 248.8 lbs, respectively. The CCC is 0.67, whereas the Pearson correlation coefficient is 0.85. In this situation, the CCC reflects both the reproducibility and differences between the self-reported and measured means. Therefore, the CCC is informative and advantageous when considering these comparisons, particularly in domain analysis within a complex survey.

3.2 The oral health study

Slade and Beck (1999) used extent of pocket depth and loss of attachment as indices of periodontal conditions. Prevalence of periodontal disease using previously reported thresholds of pocket depth $\geq$ 4 mm and attachment loss $\geq$ 3 mm were estimated by Slade and Beck (1999, Table 1). Pocket depth may be reflective of inflammation rather than chronic periodontal disease and, thus, attachment level may be a more meaningful measure of periodontal destruction. However, pocket depth remains the recommended measurement in clinical practice (Winn, Johnson and Kingman 1999). Therefore, we compare the agreement of these two definitions of periodontal disease using the kappa coefficient.

We use the sample that was analyzed by Slade and Beck (1999). The data include 14,415 persons aged 13 or older who had complete pocket depth and attachment loss assessment by six designated dentists. We again use data from the entire six-year survey period (both 1988-1991 and 1991-1994). There were a total of 49 strata and each stratum had two PSUs. The variable labeled sample weight, wtpfex6, accounting for differential selection probability, was used in our analyses.

The first definition of periodontal disease is pocket depth $\geq$ 4 mm and the second is maximum attachment loss $\geq$ 3 mm. For both variables we are using the maximum values among all teeth in an individual's mouth. The probability estimates of the attachment loss and pocket depth variables are 0.358 (jackknife s.e. = 0.0088) and 0.212 (jackknife s.e. = 0.016), respectively, using the proposed method. The asymptotic standard errors based on the usual Taylor series expansion (Woodruff 1971, produced by PROC SURVEYFREQ in SAS, version 9.1) are 0.0088 and 0.015, respectively.

Kappa is a natural choice for assessing the agreement between two binary ratings as it corrects for chance agreement (Fleiss 1981). The estimate of kappa for measuring the agreement between the two definitions of periodontal disease (pocket depth of $\geq$ 4 mm and attachment loss of $\geq$ 3 mm) using the proposed method is 0.307. The standard error of 0.0158 was obtained by both the Taylor series linearization and jackknife methods. Table 2 compares these results to the measures when the complex sampling structure is ignored. The standard error of the kappa coefficient is larger when accounting for the survey structure.

Table 2
Unweighted and weighted average, kappa, and respective standard errors for attachment loss and pocket depth

	Attachment Loss	Pocket Depth	Kappa
Unweighted Estimate	0.393	0.283	0.334
SE	0.004	0.004	0.008
Weighted Estimate	0.358	0.212	0.307
SE	0.009	0.016	0.0158

4. Discussion

The CCC and kappa evaluate the agreement between two measurements for continuous and categorical responses, respectively. In this paper, we have proposed a generalized estimating equation approach for estimating the CCC for a pair of continuous variables, and kappa for a pair of binary variables, from sample survey data where the data have been collected using complex survey features such as stratification or clustering. The usual sandwich estimator of the variance only accounts for repeated measurements made on the same individual, and does not account for the sampling framework (*e.g.,* clustering, stratification, and weighting). In the GEE approach, standard error estimation of the estimators is conducted with the Taylor series linearization and jackknife approaches. If the data are not collected using complex survey features, the proposed estimators will be identical to the usual estimators. As is evident in the two examples from the NHANES III study, we have shown the need to incorporate sampling weights and the sampling design features so that the standard errors are not underestimated when data are collected from a complex sampling design. Tables 1 and 2 show that there were large differences in the standard errors between weighted and unweighted estimates of the standard errors for both CCC and kappa. Confidence intervals that incorporate weights and the design features will allow correct inference.

In the appendix, we show steps for calculating the weighted measures of the CCC and kappa, along with their standard errors using standard survey software that incorporates the sampling weights, clustering and stratification. The GEE approach is advantageous because it is a convenient framework for developing estimators of the agreement coefficients and is easily extended to multiple raters, multiple methods, covariate adjustment and unbalanced cluster sizes. This design-based approach results in correct standard error estimation without assuming an underlying model and accounting for the sampling structure. If one is interested in estimating the agreement between two ordinal variables with kappa then Williamson *et al.*'s (2000) generalized estimating equation approach can be extended similarly to the proposed method.

Acknowledgements

We thank the anonymous reviewers and editor for their helpful comments. In particular, the editor gave extremely valuable suggestions for the introduction section.

Appendix

Steps for calculating the CCC and its standard error using standard survey software

Step 1: Calculate the means of the continuous variables Y_{hij1} and Y_{hij2} using software for survey data that incorporates stratification, clustering, and sample weighting (*e.g.*, PROC SURVEYMEANS in SAS).

Step 2: Square the centered Y_{hij1} and Y_{hij2} values around their respective means.

Step 3: Calculate the means of the squared centered Y_{hij1} and Y_{hij2} values using standard software for survey data. These means are the variance estimates of Y_{hij1} and Y_{hij2}. Calculate the mean of the product of the centered Y_{hij1} and Y_{hij2} values using standard software for survey data. This mean is the estimated covariance of Y_{hij1} and Y_{hij2}.

Step 4: Calculate the CCC by substituting the estimated means and variances into equation (1). Create the new variable Z_{hij} based on equation (4).

Step 5: Calculate the standard error of Z_{hij} using standard software for survey data. The standard error of Z_{hij} estimates the standard error of $\hat{\rho}_c$.

SAS CODE:

Let $y1$ and $y2$ denote the variables for the pair of continuous responses, and s, c and w denote the variables for strata, cluster and weight:

```
PROC SURVEYMEANS DATA=dataset MEAN; /* Step 1 above */;
        STRATA s;
        CLUSTER c;
        WEIGHT w;
        VAR y1 y2;
        ODS OUTPUT STATISTICS=stat;
data _null_;
        set stat (where=(varname= 'y1' ));
        call symputx('muy1', mean);
data _null_;
        set stat (where=(varname= 'y2' ));
        call symputx('muy2', mean);
data dataset; set dataset;                       /* Step 2 above */;
        cy1 = y1 – &muy1;
        cy2 = y2 – &muy2;
        vary1 = cy1 * *2;
        vary2 = cy2 * *2;
        covy12 = cy1 * cy2;
PROC SURVEYMEANS MEAN;                           /* Step 3 above */;
        STRATA s;
        CLUSTER c;
        WEIGHT w;
        VAR vary1 vary2 covy12;
        ODS OUTPUT STATISTICS=stat;
run;
data _null_;
        set stat (where=(varname= 'vary1' ));
        call symputx('vary1', mean);
data _null_;
        set stat (where=(varname= 'vary2' ));
        call symputx('vary2', mean);
data _null_;
        set stat (where=(varname= 'covy12' ));
        call symputx('covy12', mean);
data dataset; set dataset;                       /* Step 4 above */;
        d = &vary1 + &vary2 + (&muy1 – &muy2) ** 2;
        CCC = 2 * &covy12 / d;
        z = (2 / d) * (cy1 * cy2) – (2 * &covy12 / d / d) * ((cy1 * *2) +
        (cy2 * *2) + 2 * (&muy1 – &muy2) * (y1 – y2));
PROC SURVEYMEANS MEAN;                           /* Step 5 above */ ;
        STRATA s;
        CLUSTER c;
        WEIGHT w;
        VAR CCC z;
run;
```

Steps for calculating kappa and its standard error using standard survey software

Step 1: Estimate the probabilities of the binary variables Y_{hij1} and Y_{hij2} using software for survey data that incorporates stratification, clustering, and sample weighting (*e.g.*, PROC SURVEYFREQ in SAS).

Step 2: Estimate $P_e (= \hat{\pi}_1 \hat{\pi}_2 + (1 - \hat{\pi}_1)(1 - \hat{\pi}_2))$.

Step 3: Create the new agreement variable $U_{hij} (= Y_{hij1} Y_{hij2} + (1 - Y_{hij1})(1 - Y_{hij2}))$.

Step 4: Calculate the sum of the sample survey weights and the sum of the weighted U_{hij} (*e.g.*, using PROC SURVEYMEANS in SAS). Estimate kappa using equation (2).

Step 5: Create a new variable z_{hij} using equation (6).

Step 6: Calculate the standard error of z_{hij} using standard software for survey data. The standard error of z_{hij} estimates the standard error of $\hat{\kappa}$.

References

Barnhart, H.X., and Williamson, J.M. (2001). Modeling concordance correlation via GEE to evaluate reproducibility. *Biometrics*, 57, 931-940.

Behavioral Risk Factor Surveillance System (BRFSS). http://www.cdc.gov/BRFSS.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 20, 37-46.

Binder, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.

Carrasco, J.L., and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, 59, 849-858.

Chinchilli, V.M., Martel, J.K., Kumanyika, S. and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics,* 52, 341-353.

Cochran, W.G. (1963). *Sampling Techniques*, 2nd Ed. New York: John Wiley & Sons, Inc.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Dixon, A.E., Sugar, E.A., Zinreich, S.J., Slavin, R.G., Corren, J., Naclerio, R.M., Ishii, M., Cohen, R.I., Brown, E.D., Wise, R.A. and Irvin, C.G. (2009). Criteria to screen for chronic sinonasal disease. *Chest*, 136 (5), 1324-1332.

Feinleib, M., Garrison, R.J., Fabsitz, R.R., Christian, J.C., Hrubec, Z., Borhani, N.O., Kannel, W.B., Roseman, R., Schwartz, J.T. and Wagner, J.O. (1977). The NHLBI Twin Study of cardiovascular disease risk factors: Methodology and summary of results. *American Journal of Epidemiology*, 106, 284-295.

Field, A.E., Aneja, P. and Rosner, B. (2007). The validity of self-reported weight change among adolescents and young adults. *Obesity*, 15, 2357-2364.

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions*, 2nd Edition. New York: John Wiley & Sons, Inc.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, Inc.

Fleiss, J.L., Nee, J.C.M. and Landis, J.R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.

Gilmour, E., Ellerbrock, T.V., Koulos, J.P., Chiasson, M.A., Williamson, J.M., Kuhn, L. and Wright, T.C. (1997). Measuring cervical ectopy: Direct visual assessment versus computerized planimetry. *American Journal of Obstetrics and Gynecology*, 176, 108-111.

Gorber, S.C., Tremblay, M., Moher, D. and Gorber, B. (2007). A comparison of direct *vs*. self-report measures for assessing height, weight and body mass index: A systematic review. *Obesity Review*, 8, 373-374.

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons, Inc. Vols I and II.

Khawaja, I.S., Olson, E.J., van der Walt, C., Bukartyk, J., Somers, V., Dierkhising, R. and Morgenthaler, T.I. (2010). Diagnostic accuracy of split-night polysomnograms. *Journal of Clinical Sleep Medicine*, 6 (4), 357-362.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Klar, N., Lipsitz, S.R. and Ibrahim, J.G. (2000). An estimating equations approach for modeling kappa. *Biometrical Journal*, 42, 45-58.

Kocks, J.W., Kerstjens, H.A., Snijders, S.L., de Vos, B., Biermann, J.J., van Hengel, P., Strijbos, J.H., Bosveld, H.E. and van der Molen, T. (2010). Health status in routine clinical practice: validity of the clinical COPD questionnaire at the individual patient level. Health and Quality of Life Outcomes. 8, 135-141.

Korten, A.E., Jorm, A.F., Henderson, A.S., McCusker, E. and Creasey, H. (1992). Control-informant agreement on exposure history in case-control studies of Alzheimer's disease. *International Journal of Epidemiology*, 21, 1121-1131.

Liang, K.Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

Lin, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45, 255-268.

Lin, L. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, 48, 599-604.

Lin, L., and Chinchilli, V. (1997). Rejoinder to the letter to the editor from Atkinson and Nevill. *Biometrics*, 53, 777-778.

Liottol, N., Radaelli, T., Orsi1, A., Taricco, E., Roggerol, P., Giann, M.L., Consonni, D., Mosca1, F. and Cetin, I. (2010). Relationship between in utero sonographic evaluation and subcutaneous plicometry after birth in infants with intrauterine growth restriction: An exploratory study. *Italian Journal of Pediatrics*, 36, 70-77.

Lipsitz, S.R., Laird, N.M. and Brennan, T.A. (1994). Simple moment estimates of the κ -coefficient and its variance. *Applied Statistics*, 43, 309-323.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R. Hoboken*. New Jersey: John Wiley & Sons, Inc.

MacDougall, H.G., Weber, K.P., McGarvie, L.A., Halmagyi, G.M. and Curthoys, I.S. (2009). The video head impulse test. Diagnostic accuracy in peripheral vestibulopathy. *Neurology*, 73, 1134-1141.

Maclure, M., and Willett, W.C. (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology*, 126, 161-169.

Mazaheri, Y., Hricak, H., Fine, S.W., Akin, O., Shukla-Dave, A., Ishill, N.M., Moskowitz, C.S., Grater, J.E., Reuter, V.E., Zakian, K.L., Touijer, K.A. and Koutcher, J.A. (2009). Prostate tumor volume measurement with combined T2-weighted imaging and diffusion-weighted MR: Correlation with pathologic tumor volume. *Radiology*, 252 (2), 449-457.

National Center for Health Statistics (2011). Third National Health and Nutrition Examination Survey, 1988-1994, NHANES III Examination data file (CD-ROM). http://www.cdc.gov/nchs/nhanes.htm.

National Health Interview Survey (NHIS) (2011). http://www.cdc.gov/nchs/ nhis.htm.

Robieson, W. (1999). On weighted kappa and concordance correlation coefficient. Ph.D. thesis, University of Illinois in Chicago/Graduate College/Mathematics.

Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Sherry, B., Jefferds, M.E. and Grummer-Strawn, L.M. (2007). Accuracy of adolescent self-report of height and weight in assessing overweight status: A literature review. *Archive of Pediatrics Adolescent Medicine*, 161, 1154-1161.

Shield, M., Gorber, S.C. and Tremblay, M.S. (2008). Effects of measurement on obesity and morbidity. *Health Reports*, 19, 77-84.

Slade, G.D., and Beck, J.D. (1999). Plausibility of periodontal disease estimates from NHANES III. *Journal of Public Health Dentistry*, 59, 67-72.

Talamayan, K.S., Springer, A.E., Kelder, S.H., Gorospe, E.C. and Joye, K.A. (2006). Prevalence of overweight misperception and weight control behaviors among normal weight adolescents in the United States. *The Scientific World Journal*, 6, 365-373.

Tanner, M.A., and Young, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association*, 80, 175-180.

Williamson, J.M., Manatunga, A.K. and Lipsitz, S.R. (2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics*, 1, 191-202.

Winn, D.M., Johnson, C.L. and Kingman, A. (1999). Periodontal disease estimates in NHANES III: Clinical measurement and complex sample design issues. *Journal of Public Health Dentistry*, 59, 73-78.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.

Zar, J.H. (1996). *Biostatistical Analysis. Upper Saddle River*. New Jersey: Prentice Hall International.

Combining synthetic data with subsampling to create public use microdata files for large scale surveys

Jörg Drechsler and Jerome P. Reiter [1]

Abstract

To create public use files from large scale surveys, statistical agencies sometimes release random subsamples of the original records. Random subsampling reduces file sizes for secondary data analysts and reduces risks of unintended disclosures of survey participants' confidential information. However, subsampling does not eliminate risks, so that alteration of the data is needed before dissemination. We propose to create disclosure-protected subsamples from large scale surveys based on multiple imputation. The idea is to replace identifying or sensitive values in the original sample with draws from statistical models, and release subsamples of the disclosure-protected data. We present methods for making inferences with the multiple synthetic subsamples.

Key Words: Confidentiality; Disclosure; Multiple imputation.

1. Introduction

National Statistical Institutes (NSIs) like the U.S. Census Bureau and Statistics Canada conduct large scale surveys that are highly valued by secondary data analysts, such as the American Community Survey (ACS) and the National Longitudinal Survey of Children and Youth (NLSCY). While these analysts desire access to as much data as possible, the NSI also must protect the confidentiality of survey participants' identities and sensitive attributes. A common strategy for reducing disclosure risks in large scale studies is to release subsamples of the original survey data; for example, the Census Bureau releases a subsample from the collected ACS data comprising 1% of all U.S. households (the collected ACS data comprise 2.5% of all households), and Statistics Canada releases a 20% sample of individuals from the NLSCY. See Willenborg and de Waal (2001) and Reiter (2005) for discussions of the confidentiality protection engendered by sampling. Typically, however, subsampling alone does not eliminate disclosure risks, particularly for units in the subsample with unusual combinations of characteristics. NSIs therefore alter data before dissemination. For example, in the ACS, the Census Bureau performs data swapping, topcoding of selected variables, aggregating of geography, and age perturbation; in the NLSCY, Statistics Canada uses data swapping and suppression.

When implemented with high intensity, as may be necessary to protect confidentiality in highly visible surveys, standard disclosure limitation strategies can seriously distort inferences (Winkler 2007; Elliott and Purdam 2007; Drechsler and Reiter 2010). Further, for many standard techniques it is difficult for data analysts - especially those without advanced statistical training - to properly account for the effects of the disclosure control in estimation. Motivated by these limitations, we propose a new approach for generating public use microdata samples from large scale surveys called subsampling with synthesis. The basic idea is to replace identifying or sensitive values in the original sample with multiple draws from statistical models estimated with the original data file, and release subsamples of the disclosure-protected data. The subsamples can comprise one common set of records, or they can be taken independently.

This approach is a variant of partially synthetic data (Little 1993; Reiter 2003), which has been used in the U.S. to create several public use data products, including the Survey of Income and Program Participation, the Longitudinal Business Database, the Survey of Consumer Finances, the American Community Survey group quarters data, and OnTheMap. The approach proposed here differs from partial synthesis because of the subsampling, which necessitates adjustments to the inferential methods of Reiter (2003); these are presented here. The approach also differs from the methods for creating synthetic public use microdata samples of census data developed recently by Drechsler and Reiter (2010). In subsampling with synthesis, the initial data come from a survey and not from a census; thus, inferences must account for the additional uncertainty that results from the initial sampling.

2. General approach

We now describe the data generation and inferential procedures for the two approaches to subsampling with

[1] Jörg Drechsler, Institute for Employment Research, Department for Statistical Methods, Regensburger Straße 104, 90478 Nürnberg, Germany. E-Mail: joerg.drechsler@iab.de; Jerome P. Reiter, Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708-0251. E-Mail: jerry@stat.duke.edu.

synthesis: releasing different (independent) subsamples, and releasing a common set of records in each subsample. The data generation methods, as well as methods for making valid inferences from the multiple datasets, depend on the subsampling approach. For both approaches, we let D denote the original survey data of n_1 units sampled from a population consisting of N units. We initially assume that the original sampling design is a simple random sample; we later extend to stratified sampling. We assume that all sampled units fully respond in D. Unlike for standard partial synthesis (Reiter 2004), methods have not been developed to handle missing data and synthesis with subsampling simultaneously. We focus here on general descriptions of the approaches and presentation of the inferential methods. We do not discuss synthesis model building strategies; see Drechsler and Reiter (2009) and the references therein for guidance.

2.1 Releasing different random subsamples

2.1.1 Summary of approach

To begin, the NSI creates m partially synthetic datasets, $D_{syn} = \{D_i\colon i = 1, \ldots, m\}$, for the original survey following the approach of Reiter (2003). Specifically, the NSI replaces identifying or sensitive values in D with multiple imputations. Synthesis models are estimated using only the records whose values will be synthesized. The synthesis is done independently m times, resulting in D_{syn}. The NSI then takes a simple random subsample of $n_2 < n_1$ records from each D_i. These m subsamples, $d_{syn} = \{d_i\colon i = 1, \ldots, m\}$, are released to the public.

The analyst of d_{syn} seeks inferences about some estimand Q, such as a population mean or regression coefficient. In each d_i, the analyst estimates Q with some point estimator q and estimates the variance of q with some estimator u, where the analyst specifies q and u acting as if d_i were the collected data. Here, u is specified ignoring any finite population correction factors; for example, when q is the sample mean, $u = s^2/n_2$, with s^2 being the sample variance. For $i = 1, \ldots, m$, let q_i and u_i be the values of q and u in d_i. The following quantities are needed for inferences.

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q_i - \bar{q}_m)^2 / (m-1) \tag{2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i / m. \tag{3}$$

The analyst then can use $\bar{q}_m$ to estimate Q and

$$T_d = (n_2/n_1 - n_2/N)\,\bar{u}_m + b_m/m \tag{4}$$

to estimate the variance of $\bar{q}_m$. Derivations of these estimates are presented in Section 2.1.2. We note that without subsampling, *i.e.*, $n_2 = n_1$, (4) equals the variance estimate for standard partial synthesis (Reiter 2003). For large n_2, inferences are based on a t-distribution, $(\bar{q}_m - Q) \sim t_{\nu_d}(0, T_d)$, with degrees of freedom $\nu_d = (m-1)(1 + (n_2/n_1 - n_2/N)\, m\bar{u}_m / b_m)^2$.

The inferential methods can be extended to stratified samples in which the NSI uses the same strata for the subsample and original sample. Let N_h be the population size in stratum h, where $h = 1, \ldots, H$. For each h, let $\bar{q}_{mh}$ and T_{dh} be the values of (1) and (4) computed using only the records in d_{syn} in stratum h. These estimates are used in inferences for population quantities in stratum h. For inferences about the entire population mean, the point estimate of Q is $\bar{q}_m = \sum_h (N_h/N)\,\bar{q}_{mh}$, and its estimated variance is $T_d = \sum_h (N_h/N)^2\, T_{dh}$. Point and variance estimates for nonlinear functions of means can be derived using Taylor series expansions. We note that NSIs should release the values of n_{2h}/n_{1h} for all strata to enable variance estimation.

2.1.2 Derivation of inferences for the different random subsamples approach

The analyst seeks $f(Q \mid d_{syn})$, which can be written as

$$f(Q \mid d_{syn}) = \int f(Q \mid D_{syn}, d_{syn})\, f(D_{syn} \mid d_{syn})\, dD_{syn}. \tag{5}$$

For all derivations in Section 2.1.2, we assume that the analyst's distributions are identical to those used by the NSI for creating D_{syn}. We also assume that the sample sizes are large enough to permit normal approximations for these distributions. Thus, we require only the first two moments for each distribution, which we derive using standard large sample Bayesian arguments. Diffuse priors are assumed for all parameters.

Let Q_i and U_i be the point estimate of Q and its variance that the analyst would compute with D_i (which is not available to the analyst). Let $\bar{Q}_m, \bar{U}_m$, and B_m be defined as in (1) - (3) but using Q_i and U_i. From standard partial synthesis results (Reiter 2003), we have $(Q \mid D_{syn}) \sim N(\bar{Q}_m, \bar{U}_m + B_m/m)$. We assume that $(q_i \mid D_i) \sim N(Q_i, (1 - n_2/n_1) u_i)$ and, as is typical in multiple imputation contexts, that $u_i \approx \bar{u}_m$. Thus, using standard Bayesian theory, we have $(\bar{Q}_m \mid d_{syn}) \sim N(\bar{q}_m, (1 - n_2/n_1)\bar{u}_m/m)$ and $((m-1)b_m/(B_m + (1 - n_2/n_1)\,\bar{u}_m) \mid d_{syn}) \sim \chi^2_{m-1}$. Hence, we have $f(Q \mid d_{syn}, B_m, \bar{U}_m) = N(\bar{q}_m, \bar{U}_m + B_m/m + (1 - n_2/n_1)\,\bar{u}_m/m)$.

To get $f(Q \mid d_{syn})$, we need to integrate out B_m and $\bar{U}_m$ from this distribution. We do so by substituting B_m and $\bar{U}_m$ with their approximate expected values. To approximate

$E(B_m \mid d_{syn})$, we use $b_m - (1 - n_2/n_1)\,\bar{u}_m$. To approximate $E(\bar{U}_m \mid d_{syn})$, we note that

$$\begin{aligned}\mathrm{Var}(Q \mid d_i) &= E[\mathrm{var}(Q \mid D_i) \mid d_i] + \mathrm{var}[E(Q \mid D_i) \mid d_i] \\ &= E(U_i \mid d_i) + \mathrm{var}(Q_i \mid d_i). \end{aligned} \tag{6}$$

Here, $\mathrm{var}(Q \mid d_i) = (1 - n_2/N)u_i$. Solving (6), we have $E(\bar{U}_m \mid d_{syn}) \approx (n_2/n_1 - n_2/N)\bar{u}_m$. After substitution of these expected values, we have $\mathrm{var}(Q \mid d_{syn}) = T_d$.

Since we use an estimated variance for Q, we approximate $f(Q \mid d_{syn})$ with a t-distribution with mean $\bar{q}_m$ and variance T_d. The degrees of freedom, ν_d, is derived by matching the first two moments of $(\nu_d T_d)/\{(n_2/n_1 - n_2/N)\bar{u}_m + B_m/m + (1 - n_2/n_1)\bar{u}_m/m\}$ to those of a $\chi^2_{\nu_d}$ distribution.

2.2 Releasing the same random subsample

At first glance, releasing a common set of records in each subsample looks like standard partial synthesis. However, Reiter's (2003) variance estimator can be positively biased in this context. To illustrate, suppose that D comprises one variable with sample mean $\bar{x}_1$. Also suppose that we create D_{syn} by replacing all values of x, and we randomly select a common set of n_2 records for the subsample. Let $m = \infty$, and let Q be the population mean of x. If replacements are simulated from the correct model, which is estimated with D, then $\bar{q}_\infty = \bar{x}_1$. Hence, $\mathrm{var}(\bar{q}_\infty)$ is identical to the variance of $\bar{x}_1$, which is $(1 - n_1/N)\,s_1^2/n_1$. However, Reiter's (2003) variance estimate includes $\bar{u}_m$ based on $(1 - n_2/N)s_2^2/n_2$, where $E(s_2^2) = s_1^2$. Hence, in general Reiter's (2003) variance will have positive bias for subsamples with synthesis.

In place of standard partial synthesis, we adopt the approach taken by Reiter (2008) for multiple imputation for missing data when records used for imputation are not used or disseminated for analysis. This setting is akin to subsampling the same records in each d_i because the models for the synthesis are estimated with D, but the analyst only has d_{syn} for analysis; that is, not all records used for imputation are disseminated for analysis.

For convenience, we summarize the methodology of Reiter (2008) here but do not include the derivations. First, as in standard partial synthesis, the NSI estimates the synthesis models using only the records whose values will be synthesized. Let θ be the parameters that govern the distribution of the synthetic data models. Second, the NSI samples m values of θ from its posterior distribution. Third, for each drawn $\theta^{(l)}$ where $l = 1, \ldots, m$, the NSI draws a replacement dataset $D^{(l,\,p)}$ from the synthesis models based on $\theta^{(l)}$. The NSI repeats this process r times for each $\theta^{(l)}$. Finally, the NSI releases the collection of $M = mr$ subsamples from these datasets, $d^* = \{d^{(l,\,p)}: l = 1, \ldots, m;\ p = 1, \ldots, r\}$. Each $d^{(l,\,p)}$ includes an index of its nest l.

For $l = 1, \ldots, m$ and $p = 1, \ldots, r$, let $q^{(l,\,p)}$ and $u^{(l,\,p)}$ be the estimate of Q and its estimated variance computed with $d^{(l,\,p)}$. Here, $u^{(l,\,p)}$ includes the finite population correction factor. The following quantities are used for inferences:

$$\bar{q}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} q^{(l,\,p)}/(mr) = \sum_{l=1}^{m} \bar{q}_r^{(l)}/m, \tag{7}$$

$$\bar{w}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} (q^{(l,\,p)} - \bar{q}_r^{(l)})^2 / \{m(r-1)\} = \sum_{l=1}^{m} w_r^{(l)}/m, \tag{8}$$

$$b_M = \sum_{l=1}^{m} (\bar{q}_r^{(l)} - \bar{q}_M)^2 / (m-1), \tag{9}$$

$$\bar{u}_M = \sum_{l=1}^{m}\sum_{p=1}^{r} u^{(l,\,p)}/(mr). \tag{10}$$

The analyst can use $\bar{q}_M$ to estimate Q and $T_s = \bar{u}_M - \bar{w}_M + (1 + 1/m)\,b_M - \bar{w}_M/r$ to estimate the variance of $\bar{q}_M$. When r is large, inferences are based on a t-distribution, $(\bar{q}_M - Q) \sim t_{\nu_s}(0, T_s)$, with degrees of freedom

$$\nu_s = \left(\frac{\{(1+1/m)b_M\}^2}{(m-1)T_s^2} + \frac{\{(1+1/r)\,\bar{w}_M\}^2}{\{m(r-1)\}T_s^2} \right)^{-1}. \tag{11}$$

It is possible that $T_s < 0$, particularly for small m and r. Instead, analysts can use the always positive but conservative variance estimator, $T_s^* = \lambda T_s + (1 - \lambda)(1 + 1/m)b_M$, where $\lambda = 1$ when $T_s > 0$ and $\lambda = 0$ otherwise. Motivation for this estimator is provided in Reiter (2008). Generally, negative values of T_s can be avoided by making m and r large. When $T_s < 0$, inferences are based on a t-distribution with $(m - 1)$ degrees of freedom, which comes from using only the first term and T_s^* in (11).

For stratified designs, the point estimate for whole population quantities is $\bar{q}_M = \sum_h (N_h/N)\bar{q}_{Mh}$, and its estimated variance is $T_s = \sum_h (N_h/N)^2 T_{sh}$, where $\bar{q}_{Mh}$ and T_{sh} are the point estimate and its variance in stratum h. The degrees of freedom in the t-distribution for stratified sampling is

$$\nu_{st} = \left\{ \frac{\left\{\sum_h ((N_h/N)^2(1+1/m)b_{Mh})\right\}^2}{(m-1)\sum_h (N_h/N)^2 T_{sh}^2} + \frac{\left\{\sum_h ((N_h/N)^2(1+1/k)\,\bar{w}_{Mh})\right\}^2}{\{m(r-1)\}\sum_h (N_h/N)^2 T_{sh}^2} \right\}^{-1}. \tag{12}$$

This is derived by moment matching to a χ^2 random variable.

3. Illustrative simulations using a stratified sampling design

In this section, we investigate the analytical properties of the inferential procedures for subsampling with synthesis for stratified simple random sampling. We generate a population of $N = 1{,}000{,}000$ records comprising five variables, $Y_1, \ldots, Y_5$, in $H = 4$ strata. Y_1 is a categorical variable with ten categories generated according to the distribution in Table 1. The distributions for $(Y_2, \ldots, Y_5)$ are displayed in Table 2, along with the stratum sizes.

Table 1
Empirical distribution of Y_1 in the generated population

	1	2	3	4	5	6	7	8	9	10
percentage	24.77	32.63	16.38	15.06	7.13	2.53	0.95	0.33	0.15	0.09

To create D, we randomly sample $n_{1h} = 7{,}500$ records from each stratum. Each subsample comprises $n_{2h} = 5{,}000$ records for each stratum. In practice, the NSI might use proportional allocation to set each n_{1h} and choose smaller sampling rates to set n_{2h}. We use a common sample size and large sampling fractions to illustrate that the variance formulas for subsampling with synthesis correctly handle non-trivial finite population correction factors, *e.g.*, 50% of the records are sampled in stratum 4.

We consider Y_4 and Y_5 to be the confidential variables and illustrate two synthesis scenarios. In the first, we synthesize all records' values of Y_4 and Y_5. To do so, in each stratum we simulate Y_{4h} using a regression of Y_{4h} on (Y_{1h}, Y_{2h}, Y_{3h}) estimated with D, and we simulate Y_{5h} using a regression of Y_{5h} on $(Y_{1h}, Y_{2h}, Y_{3h}, Y_{4h})$ estimated with D. Predictions of Y_{5h} are based on the synthesized values of Y_{4h}. In the second approach, in each stratum we replace Y_{4h} and Y_{5h} only for all records with $Y_{3h} > p_h$, where p_h is the 90th percentile of Y_3 in the population in stratum h. We generate replacement values by sampling from regression models; however, the models in each stratum are estimated only with those records satisfying $Y_{3h} > p_h$.

For the different subsamples approach, we generate $m = 5$ synthetic surveys as outlined in Section 2.1. For the same subsample approach, we first draw $m = 5$ values of θ, the regression coefficients and variances. For each $\theta^{(l)}$, we generate $r = 5$ synthetic datasets for every first stage nest.

For all scenarios, we repeat the process of (i) creating D by sampling from the population and (ii) generating subsamples with synthesis a total of 5,000 times. For each of these 5,000 runs, we obtain inferences for fifty quantities, including the population means and within-stratum means of Y_4 and Y_5, the coefficients from a regression of Y_3 on all other variables, and the coefficients from a regression of Y_5 on all other variables. The regressions are estimated separately in each stratum.

Table 2
Parameters for drawing $(Y_2, \ldots, Y_5)$ for the population

	Stratum size	Model	Distribution of the error term
Stratum 1	750,000	$Y_2 = Y_1 + e$ $Y_3 = Y_1 + Y_2 + e$ $Y_4 = Y_1 + Y_2 + Y_3 + e$ $Y_5 = Y_1 + Y_2 + Y_3 + Y_4 + e$	$e \sim N(0, 5)$
Stratum 2	200,000	$Y_2 = 2Y_1 + e$ $Y_3 = 2Y_1 + 0.5Y_2 + e$ $Y_4 = 2Y_1 + 0.5Y_2 + Y_3 + e$ $Y_5 = 2Y_1 + 0.5Y_2 + 0.5Y_3 - 0.25Y_4 + e$	$e \sim N(0, 10)$
Stratum 3	40,000	$Y_2 = -3Y_1 + e$ $Y_3 = -3Y_1 - 1.5Y_2 + e$ $Y_4 = -3Y_1 + Y_2 - 1/3Y_3 + e$ $Y_5 = -3Y_1 + Y_2 - 1/3Y_3 + 1/9Y_4 + e$	$e \sim N(0, 30)$
Stratum 4	10,000	$Y_2 = -2Y_1 + e$ $Y_3 = -Y_1 - 1.5Y_2 + e$ $Y_4 = -2Y_1 + Y_2 + 1/4Y_3 + e$ $Y_5 = 2Y_1 - Y_2 - 1/4Y_3 + 1/16Y_4 + e$	$e \sim N(0, 20)$

Figure 1 displays key results of the simulations. The left panel displays the ratios of the simulated average of T_d (and T_s) over the corresponding simulated $\text{var}(\bar{q}_m)$ for the fifty estimands. The median ratios are close to one in all scenarios, and the averages of T_d (and T_s) never differ by more than 10% from their actual variances. Thus, both T_s and T_d appear to be approximately valid variance estimators.

The middle panel of Figure 1 summarizes the percentages of the 5,000 synthetic 95% confidence intervals based on T_d (and on T_s) that cover their corresponding Q. The coverage rates are close to 0.95 except for the regression coefficients for the same subsampling approach with 100% synthesis. For these coefficients, $T_s < 0$ in up to 38% of the simulation runs, so that confidence intervals are based on the conservative T_s^*. The highest fraction of negative variances occurs in the smallest stratum which has a sampling rate of 50%. All variance estimates are positive when only 10% of the records are synthesized.

The right panel of Figure 1 displays the ratios of the simulated root mean squared error (RMSE) of $\bar{q}_m$ over the simulated RMSE from the subsamples without any synthesis. For the same subsampling approach, the RMSEs of the synthetic subsamples tend to be smaller than the RMSEs based on the subsamples without any synthesis, particularly for the 100% synthesis. The smaller RMSEs result because the synthesis models are determined with D, *i.e.*, the survey data before taking the subsample, so that they carry additional information that is not in the subsamples without synthesis. For the different synthetic subsamples, the RMSE ratios typically exceed one. Here, increased synthesis leads to greater loss in efficiency. We note that the RMSEs from the different sample and same sample approaches in Figure 1 are not directly comparable because they are based on different denominators.

To enable comparisons across the methods, as well as to illustrate the losses in efficiency from subsampling, we repeat the simulation design using $m = 25$ for the independent subsamples approach and $mr = 25$ for the same subsamples approach. The left panel of Figure 2 displays the simulated RMSE ratios for the fifty estimands in the different scenarios, where the denominators are the average RMSEs based on the original data before any confidentiality protection. The right panel of Figure 2 displays the ratios of simulated average lengths of the 95% confidence intervals, where the denominators are the average lengths based on the original data before any confidentiality protection. Based on the left panel, for a given total number of released datasets and given synthesis percentage, the independent sample approach results in more efficient estimates than the same sample approach. The right panel tells a similar story, although it is harder to see because of the scaling. Here, the same sample approach with 100% synthesis results in high fractions of negative variance estimates, so that the adjusted variance T_s^* is often used, thereby inflating the interval lengths. Figure 2 also includes results from synthesis without any subsampling, which generally provides more efficient estimates than either subsampling approach.

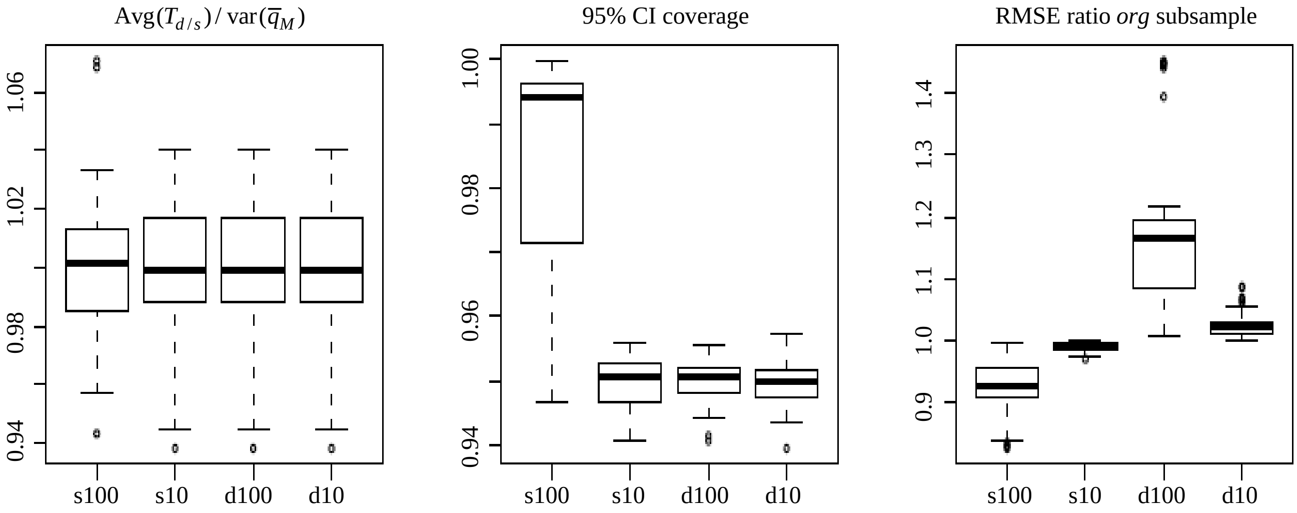


Figure 1 Simulation results for the stratified sampling design. In the labels, s and d indicate the same subsample and the different subsamples approach. The numbers indicate the percentage of records that are being synthesized. The denominators of the RMSE are based on the point estimates from the subsamples without synthesis. For the different subsamples approach, the RMSE is computed from the average of the *m* point estimates. Each box plot comprises fifty estimands

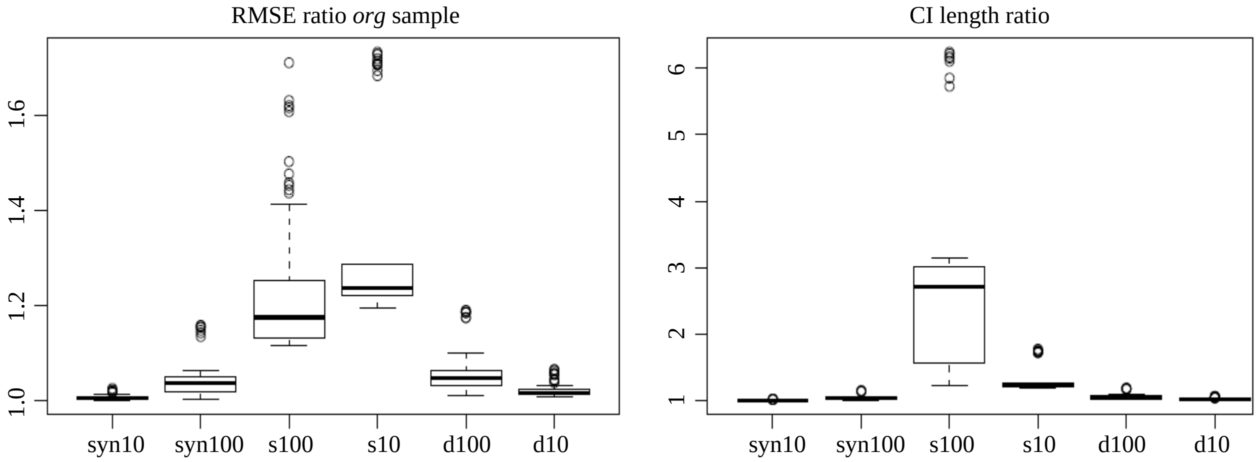


Figure 2 **Efficiency comparisons for the stratified sampling design. In the labels, *org* and *syn* indicate the original sample and the synthetic sample before subsampling; and, s, d, and the numbers are as in Figure 1. The denominators of the RMSE are based on the point estimates from the original sample without synthesis. Each box plot comprises fifty estimands**

4. Concluding remarks

The different subsamples and same subsamples approaches have competing advantages. For a fixed number of released datasets M, the different subsamples approach enables estimation with greater efficiency than the same subsamples approach - as evident in Figure 2 - since the released subsamples are independent rather than correlated. The different subsamples approach also guarantees positive variance estimates; the same subsample approach does not. However, with large M the different subsamples approach weakens the confidentiality protections of subsampling, since the combined datasets are likely to contain most of the records from the original survey. Hence, unless the subsampling rate is small (*e.g.*, 1% or 2%), the NSI may have to make m modest (*e.g.*, $m = 5$) to use the different subsamples approach. Because of this, the different samples approach is not viable when the original sample size is modest.

As an alternative to subsampling with synthesis, agencies could release partially synthetic data that include all records from the original sample, assuming that they are willing to release files of that size. Partial synthesis on the original data generally engenders estimates with lower variances than subsampling with synthesis - as evident in Figure 2 - since more records are released. However, partial synthesis on the original data generally engenders higher disclosure risks than subsampling with synthesis, since more at risk records are in the released data and since the additional protection from subsampling is absent. Agencies can compare the two options on disclosure risks using the methods of Drechsler and Reiter (2008), which account for the protection afforded by sampling, and on data utility by comparing inferences for representative analyses.

It is also possible that the process of subsampling may engender sufficient additional protection to enable lesser amounts of synthesis than would be necessary in a partial synthesis of the entire original dataset. Evaluating the data utility for subsampling with synthesis versus synthesis only for given disclosure risks is beyond the scope of this short note, but it is an interesting area for future research.

We have not developed subsampling with synthesis approaches for sampling designs other than (stratified) simple random samples. For the different subsamples approach, appropriate inferential methods require an approximately unbiased estimate of the variance from the first phase of sampling that can be computed from the subsample alone. This is elusive for complicated designs. For the same subsample approach, we conjecture that analysts can use the inferential methods presented in Section 2.2, provided that $\bar{u}_M$ appropriately accounts for the two phases of sampling. We note that the formulas for $\bar{w}_M$ and b_M remain the same for other designs. Evaluating this conjecture is a subject of future research.

Acknowledgements

This research was supported by U.S. National Science Foundation grant SES-0751671.

References

Drechsler, J., and Reiter, J.P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases*, (Eds., J. Domingo-Ferrer and Y. Saygin), New York: Springer, 227-238.

Drechsler, J., and Reiter, J.P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics*, 25, 589-603.

Drechsler, J., and Reiter, J.P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association*, 105, 1347-1357.

Elliott, M., and Purdam, K. (2007). A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymized Records. *Environment and Planning A*, 39, 1101-1118.

Little, R.J.A. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, 9, 407-426.

Reiter, J.P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-189.

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology*, 30, 235-242.

Reiter, J.P. (2005). Estimating identification risks in microdata. *Journal of the American Statistical Association*, 100, 1103-1113.

Reiter, J.P. (2008). Multiple imputation when records used for imputation are not used or disseminated for analysis. *Biometrika*, 95, 933-946.

Willenborg, L., and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.

Winkler, W.E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Tech. rep., Statistical Research Division, U.S. Bureau of the Census, Washington, DC.

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 81-93
Statistics Canada, Catalogue No. 12-001-X

A hierarchical Bayesian nonresponse model for two-way categorical data from small areas with uncertainty about ignorability

Balgobin Nandram and Myron Katzoff [1]

Abstract

We study the problem of nonignorable nonresponse in a two dimensional contingency table which can be constructed for each of several small areas when there is both item and unit nonresponse. In general, the provision for both types of nonresponse with small areas introduces significant additional complexity in the estimation of model parameters. For this paper, we conceptualize the full data array for each area to consist of a table for complete data and three supplemental tables for missing row data, missing column data, and missing row and column data. For nonignorable nonresponse, the total cell probabilities are allowed to vary by area, cell and these three types of "missingness". The underlying cell probabilities (*i.e.*, those which would apply if full classification were always possible) for each area are generated from a common distribution and their similarity across the areas is parametrically quantified. Our approach is an extension of the selection approach for nonignorable nonresponse investigated by Nandram and Choi (2002a, b) for binary data; this extension creates additional complexity because of the multivariate nature of the data coupled with the small area structure. As in that earlier work, the extension is an expansion model centered on an ignorable nonresponse model so that the total cell probability is dependent upon which of the categories is the response. Our investigation employs hierarchical Bayesian models and Markov chain Monte Carlo methods for posterior inference. The models and methods are illustrated with data from the third National Health and Nutrition Examination Survey.

Key Words: Metropolis-Hastings sampler; SIR algorithm; Nonignorable nonresponse model; Expansion model.

1. Introduction

In sample surveys, data are typically summarized in two-way categorical tables. We consider the problem of nonignorable nonresponse for many $r \times c$ categorical tables, each obtained from a single area. For many of these surveys, there are missing data and this gives rise to partial classification of the sampled individuals. Thus, for each two-way table there are both item nonresponse (one of the two categories is missing) and unit nonresponse (both categories are missing). One may not know how the data are missing and a model that includes some difference between the observed data and missing data (*i.e.*, nonignorable missing data) may be preferred. For a general $r \times c$ categorical table, we address the issue of estimation of the cell probabilities of the two-way tables when there is possibly nonignorable nonresponse but there is really no information about ignorability. In such a situation, we would like to express a degree of uncertainty about ignorability. Nandram and Choi (2002a, b) have described an expansion model appropriate for binary data when there are data from many small areas. We will extend this work to $r \times c$ categorical tables.

Letting x denote the covariates and y the response variable, Little and Rubin (2002) describe three types of missing-data mechanism. These types differ according to whether the probability of response (a) is independent of x and y; (b) depends on x but not on y; or (c) depends on y and possibly x. The missing data are missing completely at random (MCAR) in (a), missing at random (MAR) in (b) and the data are missing not at random (MNAR) in (c). Models for MCAR and MAR missing-data mechanisms are called ignorable if the parameters of the dependent variable and the response variable are distinct (Rubin 1976). Models for MNAR missing-data mechanisms are called nonignorable. The general difficulty with nonignorable nonresponse model is that the parameters are not identifiable [*e.g.*, see Nandram and Choi (2004, 2005, 2008, 2010) and Nandram, Han and Choi (2002)].

For a $r \times c$ categorical table, let $I_{ijkl} = 1$ if the l^{th} individual within the i^{th} area falls in the j^{th} row and k^{th} column and 0 otherwise. Also, let $J_{il} = 1$ if the l^{th} individual within the i^{th} area has complete information and 0 otherwise. Finally, let $P(J_{il} = 1 \mid I_{ijkl} = 1, I_{ij'k'l} = 0, j' \neq j, k' \neq k) = \pi_{ijk}$. For unit nonresponse, if $\pi_{ijk} = \pi_i$, the model is ignorable; for item nonresponse, if the columns are missing, row is observed and $\pi_{ijk} = \pi_{ij}$ (or $\pi_{ijk} = \pi_i$), the model is ignorable; and if the rows are missing but columns are observed and $\pi_{ijk} = \pi_{ik}$ (or $\pi_{ijk} = \pi_i$), the model is ignorable. All other models are nonignorable; see Rubin (1976) for further explanation.

Nandram and Choi (2002a, b) use an expansion model to study nonignorable nonresponse binary data. The expansion

[1] Balgobin Nandram, Department of Mathematical Sciences, Worcester Polytechnic Institue, 100 Institue Road, Worcester, MA 01609. E-mail: balnan@wpi.edu; Myron Katzoff, Office of research and Methodology, National Center for Health Statisitcs, CDC, 3311 Toledo Road, Hyattsville, MD 20782. E-mail: mjk5@cdc.gov.

model, a nonignorable nonresponse model, degenerates into an ignorable nonresponse model (in the spirit of Draper 1995) when a centering parameter is set to unity. This permits an expression of uncertainty about ignorability; see also Forster and Smith (1998).

We discuss the model of Nandram and Choi (2002a, b) for binary data from small areas. So that J_{il} denote the response indicators and I_{il} denote the binary response. Specifically, introducing the centering parameters γ_i for area i to incorporate uncertainty about ignorability, the model of Nandram and Choi (2002a, b) is

$$I_{il} \mid p_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p_i),$$

$$J_{il} \mid \{\pi_i, J_{il} = 0\} \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi_i), l = 1, \ldots, n_i, i = 1, \ldots, l,$$

$$J_{il} \mid \{\pi_i, \gamma_i, y_{il} = 1\} \overset{\text{iid}}{\sim} \text{Bernoulli}(\gamma_i \pi_i),\ 0 < \gamma_i \pi_i < 1.$$

When $\gamma_i = 1,$ the nonignorable nonresponse model degenerates to an ignorable nonresponse model. Here γ_i is the ratio of the odds of success among respondents to the odds of success among all individuals for the i^{th} area. The parameter γ_i describes the extent of nonignorability of the response mechanism for area *i*, and it is through the γ_i that uncertainty about ignorability is incorporated. Nandram and Choi (2002a, b) define $\delta_i = \pi_i \{\gamma_i p_i + (1 - p_i)\}$ to be the probability that an individual responds in area *i* in the entire population, and with a belief that all the areas are similar they take $(p_i, \delta_i, \gamma_i)$ to have a common distribution. A priori they take beta distributions for p_i and π_i respectively.

Here, the parameters are not identifiable. However, if $\gamma_i = 1,$ then all the parameters are identifiable. That is, identifiability of the parameters depend on the γ_i. Note, that when $\gamma_i = 1,$ we get an ignorable model for a MAR mechanism. As the parameters are identifiable in this model, it is quite sensible to use this model (or similar models) as a baseline model. However, note this model is still not justified because it assumes that missing data are like observed data. Thus, to add flexibility to this ignorable nonresponse model, we use the γ_i.

Let γ_{iuv} be the number of individuals with $I_{il} = u,$ $J_{il} = v\,(u, v = 0, 1)$ in the i^{th} area. Then, under the model,

$$(y_{i00}, y_{i01}, y_{i10}, y_{i11}) \mid \pi_i, p_i, \gamma_i \overset{\text{ind}}{\sim} \text{Multinomial}\{n_i, (1 - p_i)(1 - \pi_i), (1 - p_i)\ \pi_i, (1 - \gamma_i \pi_i)\ p_i, \gamma_i \pi_i p_i\}$$

with independence over areas. Here, only y_{i01} and y_{i11} are observed, and therefore all parameters are nonidentifiable if the γ_i are unknown. We obtain the likelihood function in a similar manner for the more complete $r \times c$ categorical table with missing data.

We start with a gamma distribution, and to permit centering on the ignorable nonresponse model, we must take each γ_i to have mean 1. However, we need to use a truncated gamma distribution because $0 < \pi_i < 1$ and $0 < \gamma_i \leq 1 / \pi_i$. An interesting idea of Nandram and Choi (2002a, b) is to model the centering as a truncated gamma

$$\gamma_i \mid \nu \overset{\text{iid}}{\sim} \text{Gamma}(\nu, \nu), 0 < \gamma_i < 1 / \pi_i, 0 < \pi_i < 1.$$

The model is complete with noninformative prior densities on all hyperparameters. One can use alternative distributions (*e.g.*, a truncated lognormal density) for the γ_i, but this is not a key issue and it would not matter much.

One can use an area level model with random effects in which, conditional on the observed data, the nonresponse is dependent upon area-level random effects. This can be formulated using a logit link function, but we have not developed our models in this direction partly because we are not using covariates here; see Nandram and Choi (2010) for the use of covariates and random effects.

The approach in Nandram and Choi (2002a, b) is attractive, but it does not apply immediately to the current $r \times c$ categorical table problem. Specifically, only one centering parameter per area is needed in Nandram and Choi (2002a, b). In our formulation, one now needs *rc* centering parameters per area; each of these parameters has to have a distribution centered at one to allow degeneration to the ignorable nonresponse model. There are also inequality constraints that must be included in the non-ignorable nonresponse model. In addition, one cannot rule out the possibility that these parameters are correlated. The methodology needed to apply the work of Nandram and Choi (2002a, b) to the $r \times c$ categorical table is not straightforward. Noting these difficulties Nandram, Liu, Choi and Cox (2005) (with a single supplemental table) and Nandram, Cox and Choi (2005) (with the three supplemental tables) use a simpler idea, but not quite as elegant as in Nandram and Choi (2002a, b), for centering; see also Nandram and Choi (2005).

Essentially, Nandram, Cox and Choi (2005) and Nandram, Liu, Cox and Choi (2005) assume an ignorable model, obtain samples of the response probabilities and use these sampled response probabilities to fit the response probabilities of a nonignorable nonresponse model while "controlling" its parameters. Of course, a possible alternative occurs when there is information about the degree of nonignorability. However, the problem of incorporating prior information about a systematic departure from ignorability is more complex for our problem, and it would need additional costly field work to obtain such information.

We discuss our philosophy about the nonignorable nonresponse problem, a fundamentally aliased problem. In fact, this problem is extremely difficult and we believe that there is really no solution to it, but we must try. Without any information one cannot tell how respondents and nonrespondents differ. An ignorable nonresponse model is short because it assumes that respondents and nonrespondents are similar, but the respondents and nonrespondents may differ. Statisticians must not confront imprecision (sampling error) only, but they must be bold enough to study subjectivity (ignorance arising from missing information). Unfortunately, as is well known, nonignorable nonresponse models have nonidentifiable parameters. We discuss how the key nonignorability parameters are identified. We know that if the respondents and nonrespondents are similar, then the γ_i are equal unity, and we get the ignorable nonresponse model with all parameters identified. We can now expand the ignorable nonresponse model into a nonignorable nonresponse model by putting a distribution on these γ_i centered at 1, still maintaining identifiability. One can formulate a nonignorable nonresponse model to add flexibility to the ignorable nonresponse model as we have done in our work; the flexibility is a form of sensitivity analysis, coherent in this case, and indeed it is a Bayesian uncertainty (risk) assessment (*e.g.*, Greenland 2009). This is what we have been doing or trying to do in our work.

In this paper we attempt to solve the difficult problem of Nandram and Choi (2002a, b) in its original form for $r \times c$ tables for many areas. The plan of this paper is as follows. In Section 2, we describe the hierarchical Bayesian model. Specifically we describe the nonignorable nonresponse mechanism and we construct an appropriate prior distribution. In Section 3, we show how to fit the model using the sampling importance resampling (SIR) algorithm to subsample from an approximate posterior density after an innovative collapsing of the complete joint posterior density. In Section 4, we illustrate our methodology with public-use data from thirteen states in the third National Health and Nutrition Examination Survey (NHANES III). Section 5 has concluding remarks.

2. The nonignorable nonresponse model

For the problem of nonresponse in a two-dimensional table, we can have both item and unit nonresponse. Thus, one may consider the full data array to consist of four tables: one for complete data and three supplemental tables - one for missing row information, one for missing column information and a table for which neither row nor column membership has been recorded. Throughout this paper, we index rows by $j = 1, \ldots, r;$ columns, by $k = 1, \ldots, c;$ and the four tables by $s = 1, 2, 3, 4.$ We index areas by $i = 1, 2, \ldots, A$ and individuals within areas by $l = 1, 2, \ldots, n_i$. We next describe the nonignorable nonresponse model (*i.e.*, the expansion model).

2.1 Sampling process

We adapt the terminology and definitions used in Nandram, Cox and Choi (2005) to our situation. For sample individual *l* in area *i*, let

$$I_{ijkl} = \begin{cases} 1, & \text{if the outcome category is } (j, k) \\ 0, & \text{otherwise,} \end{cases}$$

and let $\boldsymbol{J}_{il}$ denote one of the 4-tuples $(1, 0, 0, 0),$ $(0, 1, 0, 0), (0, 0, 1, 0), (0, 0, 0, 1).$ We assume that

$$\boldsymbol{I}_{il} \overset{\text{def}}{=} \text{vec}(\{I_{ijkl} \mid j = 1, \ldots, r; \\ k = 1, \ldots, c\}) \mid \boldsymbol{p}_i \overset{\text{iid}}{\sim} \text{Mult}\{1, \boldsymbol{p}_i\} \quad (1)$$

and

$$\boldsymbol{J}_{il} \mid \{I_{ijkl} = 1, I_{ij'k'l} = 0 \text{ for all } j' \neq j \\ \text{and } k' \neq k \mid \boldsymbol{\pi}_{ijk}\} \overset{\text{iid}}{\sim} \text{Mult}\{1, \boldsymbol{\pi}_{ijk}\}, \quad (2)$$

where $\boldsymbol{p}_i \overset{\text{def}}{=} \text{vec}(\{p_{ijk} | j = 1, 2, \ldots, r; k = 1, 2, \ldots, c\})$ is a vector of probabilities for the table of rc categories for the variable of observation which must sum to one and, for cell (j, k) in that two-dimensional table,

$$\boldsymbol{\pi}_{ijk} \overset{\text{def}}{=} \text{vec}(\{\pi_{isjk}\} \text{ for } s = 1,2,3,4)$$

is a vector of probabilities which must sum to one.

Next, we define cell counts y_{isjk}, for each table $s = 1, \ldots, 4$ for area *i* such that, for cell (j, k),

$$(y_{i1jk}, y_{i2jk}, y_{i3jk}, y_{i4jk}) = \sum_{l=1}^{n_i} I_{ijkl} \boldsymbol{J}_{il},$$

where y_{i1jk} are observed and y_{isjk}, for $s = 2, 3, 4,$ are latent variables which satisfy the observed constraints $\sum_k y_{i2jk} = u_{ij}, \sum_j y_{i3jk} = v_{ik}$ and $\sum_{j,k} y_{i4jk} = w_i.$ All inferences will be conditional on the observed quantities, u_{ij}, v_{ik} and w_i. But see Nandram (2009) for the analysis of a single $r \times c$ table under nonresponse when the margins are also random. We will denote the vector of the y_{i1jk} by $\boldsymbol{y}_1$, the vector of the $y_{isjk}, s = 2, 3, 4,$ by $\boldsymbol{y}_{(1)}$, and the complete vector by $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_{(1)})'$.

The parameters π_{isjk} are not identifiable. If the distributions of these parameters are known completely, then the nonidentifiability will disappear. Thus, the key issue is how to identify these parameters. We know that if the respondents and nonrespondents are similar (*i.e.*, the four patterns, complete and partially complete tables), then we can take $\pi_{isjk} = \pi_{is}$; this is the ignorable nonresponse model. The π_{is} can be estimated by the proportions of cases falling in the four tables for each area. This is a natural point to start. To expand the ignorable nonresponse model into a nonignorable model, and still maintain identifiability, first we need a simplification. We take $\pi_{ijks} = \psi_{ijk}\, \pi_{is}$, which gives a nonignorable nonresponse model in which the parameters ψ_{ijk} are not identifiable.

To center the nonignorable model on the ignorable model, we take

$$\pi_{isjk} = \begin{cases} \tilde{\psi}_{ijk}\pi_{is}, & \text{for } s = 1, \\ \psi_{ijk}\pi_{is}, & \text{for } s = 2,3,4, \end{cases} \tag{3}$$

and require that $\sum_{s=1}^{4} \pi_{is} = 1$. A little algebra then yields the relationship

$$\begin{aligned} \tilde{\psi}_{ijk}\pi_{i1} &= \left[1 + (1-\psi_{ijk})\left(\frac{1-\pi_{i1}}{\pi_{i1}}\right)\right]\pi_{i1} \\ &= a_{ijk}(\pi_{i1}, \psi_{ijk})\,\psi_{ijk}\pi_{i1}, \end{aligned} \tag{4}$$

where $a_{ijk}(\pi_{i1}, \psi_{ijk}) = \{\psi_{ijk}^{-1} + (\psi_{ijk}^{-1} - 1)(\pi_{i1}^{-1} - 1)\}$, from which it is clear that $\tilde{\psi}_{ijk} = 1$ if, and only if, $\psi_{ijk} = 1$. Note that since $0 \le \pi_{isjk} \le 1$ and $(1-\pi_{i1})^{-1} \le \min\{\pi_{is}^{-1}; s = 2,3,4\}$, it follows that $0 < \psi_{ijk} \le (1-\pi_{i1})^{-1}$.

By combining (1) and (2) and noting the definition of π_{isjk} in (3), similar to binary case, we get a multinomial distribution for $\boldsymbol{y}$ conditional on $\pi, \psi, \boldsymbol{p}$, and the likelihood function for the sample can now be seen to be

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\pi}, \psi, \boldsymbol{p}) &= \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \left[\prod_{j,k} (\tilde{\psi}_{ijk}\pi_{i1}p_{ijk})^{y_{i1jk}} \right] \right. \\ &\qquad \left. \prod_{s=2}^{4} \prod_{j,k} (\psi_{ijk}\pi_{is}p_{ijk})^{y_{isjk}} \right\} \\ &= \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \left\{ \prod_{s=1}^{4} \prod_{j,k} (\psi_{ijk}\pi_{is}p_{ijk})^{y_{isjk}} \right. \\ &\qquad \left. \prod_{j,k} \left[a_{ijk}(\pi_{i1}, \psi_{ijk})\right]^{y_{i1jk}} \right\}, \end{aligned} \tag{5}$$

where

$$\begin{aligned} \boldsymbol{y}_{is}^{rc\times 1} &= \text{vec}(\{y_{isjk} \mid j = 1,\ldots,r; k = 1,\ldots,c\}), \\ \boldsymbol{y} &= (\boldsymbol{y}'_{11}, \boldsymbol{y}'_{12}, \boldsymbol{y}'_{13}, \boldsymbol{y}'_{14}, \boldsymbol{y}'_{21}, \ldots, \boldsymbol{y}'_{24}, \ldots, \boldsymbol{y}'_{A1}, \boldsymbol{y}'_{A2}, \boldsymbol{y}'_{A3}, \boldsymbol{y}'_{A4})', \\ \boldsymbol{\pi}^{A4\times 1} &= (\pi_{11}, \ldots, \pi_{14}, \pi_{21}, \ldots, \pi_{24}, \ldots, \pi_{A1}, \ldots, \pi_{A4})', \\ \boldsymbol{\psi}^{Arc\times 1} &= (\psi_{111}, \ldots, \psi_{1rc}, \psi_{211}, \ldots \psi_{2rc}, \ldots, \psi_{A11}, \ldots, \psi_{Arc}), \end{aligned}$$

and

$$\boldsymbol{p}^{Arc\times 1} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_A)'.$$

Collecting factors which are powers of π_{is}, the likelihood function may also be expressed as

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{\pi}, \psi, \boldsymbol{p}) &= \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} \\ &\left\{ \prod_{s=1}^{4} \pi_{is}^{y_{is..}} \times \prod_{j,k} \{p_{ijk}\ \psi_{ijk}\}^{y_{i.jk}} \left[a_{ijk}(\pi_{i1}, \psi_{ijk})\right]^{y_{i1jk}} \right\}, \end{aligned} \tag{6}$$

where $0 \le \pi_{is} \le 1, \sum_s \pi_{is} = 1$ and $0 \le \psi_{ijk} \le (1-\pi_{i1})^{-1}$. Here we note that $y_{is..}$ and y_{i1jk} are observed variables but the $y_{i.jk}$ are latent variables.

2.2 Prior construction

The following assumptions describe the prior distributions for the nonignorable nonresponse model:

1. For the vector of cell probabilities $\boldsymbol{p}_i$, we assume that

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1\tau_1),$$

where $\mu_1 = (\mu_{111}, \mu_{112}, \ldots, \mu_{11k}, \mu_{121}, \ldots, \mu_{1rc})'; \mu_{1jk} \ge 0$; and $\sum_{j=1}^{r}\sum_{k=1}^{c} \mu_{1jk} = 1$. The parameter τ_1 informs us of similarity among the $\boldsymbol{p}_i$: the larger τ_1, the more alike the $\boldsymbol{p}_i$. This is true because large τ_1 means that the variances of the $\boldsymbol{p}_i$ are small, and because they have the same mean, this means that they are more similar with larger τ_1.

Thus, the density for $\boldsymbol{p}$ is

$$\begin{aligned} g_1(\boldsymbol{p} \mid \boldsymbol{\mu}_1, \tau_1) &= \prod_{i=1}^{A} g_{1i}(\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1) \\ &= \prod_{i=1}^{A} \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{1jk}\tau_1 - 1}}{\text{D}(\boldsymbol{\mu}_1\tau_1)} \right\}, \end{aligned} \tag{7}$$

where, for a k-tuple $\boldsymbol{c}$ and a scalar t

$$\text{D}(\boldsymbol{c}t) = \frac{\prod_{j=1}^{k} \Gamma(\boldsymbol{c}_j t)}{\Gamma(t)}$$

for $c_j > 0$ and $\sum_{j=1}^{k} c_j = 1$.

2. Independently of the $\boldsymbol{p}_i$, the $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3}, \pi_{i4})'$ follow the specification

$$\boldsymbol{\pi}_i \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2\tau_2),$$

with $\pi_{is} \ge 0$ and $\sum_s \pi_{is} = 1$, where $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22}, \mu_{23}, \mu_{24})', \mu_{2s} \ge 0, \sum_{s=1}^{4} \mu_{2s} = 1$ and τ_2 is a measure of similarity among the $\boldsymbol{\pi}_i$. Thus, the density for $\boldsymbol{\pi}_i$ is

$$g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2) = \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\text{D}(\boldsymbol{\mu}_2\tau_2)}. \tag{8}$$

3. For each i, let $\boldsymbol{\psi}_i = (\psi_{i11}, \ldots, \psi_{i1c}, \psi_{i21}, \ldots, \psi_{i2c}, \ldots, \psi_{irc})'$ so that $\boldsymbol{\psi} = (\boldsymbol{\psi}'_1, \ldots, \boldsymbol{\psi}'_A)'$. We assume for each i that the ψ_{ijk} are independently and identically

distributed in accordance with a distribution derived from the $\text{Gamma}(\beta, \beta)$, where the support is confined to the open interval $(0, (1-\pi_{i1})^{-1})$; in other words, the ordinary gamma distribution is truncated as

$$\psi_{ijk} \mid \beta, \boldsymbol{\pi}_i \overset{\text{ind}}{\sim} \text{Gamma}(\beta, \beta)$$
$$\text{such that } 0 < \psi_{ijk} < (1-\pi_{i1})^{-1}.$$

It is worth noting that these ψ_{ijk} are identically distributed over *j* and *k*. Again, one can use other distributions such as a truncated lognormal density, but this will make little difference. In this formulation, there is some information about β because the small areas are assumed to share a common effect.
Thus, for area *i*, the density for $\boldsymbol{\psi}_i$ is

$$g_{3i}(\boldsymbol{\psi}_i \mid \beta, \boldsymbol{\pi}_i) = \prod_{j=1}^{r}\prod_{k=1}^{c}\left\{ \frac{\beta^{\beta}\psi_{ijk}^{\beta-1}e^{-\beta\psi_{ijk}}}{\Gamma(\beta)} \Big/ \int_0^{(1-\pi_{i1})^{-1}} \frac{\beta^{\beta}\psi_{ijk}^{\beta-1}e^{-\beta\psi_{ijk}}}{\Gamma(\beta)} \mathrm{d}\psi_{ijk} \right\},$$

For $0 < \psi_{ijk} < (1-\pi_{i1})^{-1}$. Making the transformation $t_{ijk} = \beta\psi_{ijk}$, one can see that the normalizing constant in the denominator of each of the factors in $g_{3i}(\boldsymbol{\psi}_i \mid \beta, \boldsymbol{\pi}_i)$ is $G_{\beta}[\beta(1-\pi_{i1})^{-1}]$, where $G_{\beta}(\cdot)$ is the gamma function with scale parameter β. To eliminate the dependence of the range of integration on π_{i1}, let $\phi_{ijk} = (1-\pi_{i1})\psi_{ijk}$ and let $\boldsymbol{\phi}_i = (\phi_{i11}, \ldots, \phi_{i1c}, \phi_{i21}, \ldots, \phi_{i2c}, \ldots, \phi_{irc})'$. Then

$$g_{3i}(\boldsymbol{\phi}_i \mid \beta, \boldsymbol{\pi}_i) = \prod_{j=1}^{r}\prod_{k=1}^{c}\left\{ \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta(1-\pi_{i1})^{-1}]} \right\}, \quad (9)$$

for $0 < \phi_{ijk} < 1$. The joint prior for $\boldsymbol{\pi}_i$ and $\boldsymbol{\phi}_i$ is just the product of $g_{3i}(\boldsymbol{\phi}_i \mid \beta, \boldsymbol{\pi}_i)$ and $g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2)$. Thus, the joint prior for $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_A)'$ and $\boldsymbol{\pi}$ is

$$g^{*}(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta) \overset{\text{def}}{=} \prod_{i=1}^{A}\{g_{3i}(\boldsymbol{\phi}_i \mid \beta, \boldsymbol{\pi}_i) \cdot g_{2i}(\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2)\}.$$

That is

$$g^{*}(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta) = \prod_{i=1}^{A}\left\{ \frac{\prod_{s=1}^{4}\pi_{is}^{\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)} \prod_{j=1}^{r}\prod_{k=1}^{c} \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta(1-\pi_{i1})^{-1}]} \right\}. \quad (10)$$

The description of the model is completed by specifying the assumptions on the hyperparameters. As there are no conjugate priors, we use shrinkage priors for τ_1, τ_2 and β because these are proper and noninformative. Priors of the form $p(\tau_1) \propto 1/\tau_1$, and specifically proper diffused gamma priors, are discouraged; see, for example, Gelman (2006). Other alternatives are half Cauchy densities and gamma densities (one would need to specify the hyperparameters). Thus, we take

1. τ_1, τ_2 and β have independent shrinkage priors of the form
$$f(x) = \frac{a_0}{(a_0+x)^2}, \text{ for } x \geq 0,$$
where a_0 is specified; it is standard practice to take $a_0 = 1$.
2. We also assume that $\boldsymbol{\mu}_1 \sim \text{Dirichlet}(1, 1, \ldots, 1)$ and $\boldsymbol{\mu}_2 \sim \text{Dirichlet}(1, 1, 1, 1)$.

Let $\Omega = (\beta, \boldsymbol{\mu}_1, \tau_1, \boldsymbol{\mu}_2, \tau_2)$. Then the density for Ω is

$$p(\Omega) = \frac{a_0}{(a_0+\tau_1)^2} \cdot \frac{b_0}{(b_0+\tau_2)^2} \cdot \frac{c_0}{(c_0+\beta)^2}(rc-1)!3!$$

for τ_1, τ_2 and $\beta \geq 0, \sum_{j,k}\mu_{1jk} = 1$ and $\sum_{s=1}^{4}\mu_{2s} = 1$.

By Bayes' theorem, the joint posterior density is

$$\begin{aligned} &h(\Omega, \boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \propto \\ &\quad f(\boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{p})\, g_1(\boldsymbol{p} \mid \boldsymbol{\mu}_1, \tau_1)\, g^{*}(\boldsymbol{\pi}, \boldsymbol{\phi} \mid \boldsymbol{\mu}_2, \tau_2, \beta)\, p(\Omega) \\ &= \prod_{i=1}^{A}\Bigg[\binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}} (1-\pi_{i1})^{-y_{i\ldots}} \\ &\quad \left\{ \prod_{s=1}^{4}\pi_{is}^{y_{is..}} \times \prod_{j,k}\{p_{ijk}\phi_{ijk}\}^{y_{i.jk}} \left[a_{ijk}(\pi_{i1}, \phi_{ijk})\right]^{y_{i1jk}} \right\} \\ &\quad \times \left\{ \frac{\prod_{j,k} p_{ijk}^{\mu_{1jk}\tau_1-1}}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)} \right\} \left\{ \frac{\prod_{s=1}^{4}\pi_{is}^{\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)} \right. \\ &\quad \left. \prod_{j=1}^{r}\prod_{k=1}^{c} \frac{\beta^{\beta}}{(1-\pi_{i1})^{\beta}} \frac{\phi_{ijk}^{\beta-1} e^{-\frac{\beta\phi_{ijk}}{1-\pi_{i1}}}}{\Gamma(\beta)\, G_{\beta}[\beta(1-\pi_{i1})^{-1}]} \right\} \Bigg] \\ &\quad \times \frac{a_0}{(a_0+\tau_1)^2} \cdot \frac{b_0}{(b_0+\tau_2)^2} \cdot \frac{c_0}{(c_0+\beta)^2}, \end{aligned} \quad (11)$$

where, substituting $(1-\pi_{i1})^{-1}\phi_{ijk}$ for ψ_{ijk},

$$a_{ijk}(\pi_{i1}, \phi_{ijk}) = \left(\frac{1-\pi_{i1}}{\phi_{ijk}}\right)\left[1 + \frac{1}{\pi_{i1}}\{1-\pi_{i1}-\phi_{ijk}\}\right]. \quad (12)$$

To make inferences about the p_{ijk}, we will draw samples from $h(\Omega, \boldsymbol{p}, \boldsymbol{\pi}, \boldsymbol{\phi}, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ using Markov chain Monte Carlo methods. This procedure is described in Section 3.

3. Computations

We use the SIR algorithm to subsample a random sample from an approximate posterior density. There are three steps to accomplish this task. We collapse over the $\boldsymbol{p}_i$, $\boldsymbol{\pi}_i$ and ϕ_i, approximate the collapsed density by a simpler one and sample from it, and then subsample these samples to get samples from the original density. We show how to do these three steps in this section.

To obtain the approximation and to simplify the computations, in Appendix A we collapse over the $\boldsymbol{p}_i$, $\boldsymbol{\pi}_i$ and ϕ_i to get

$$h(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \pi_a(\Omega\, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \cdot \prod_{i=1}^{A} I_i,$$

where

$$I_i = \iiint_0^{\frac{\beta b_i}{1-\pi_{i1}}} \frac{G_{rc\beta}\left(\dfrac{\beta b_i}{1-\pi_{i1}}\right)}{\left[G_\beta\left(\dfrac{\beta}{1-\pi_{i1}}\right)\right]^{rc}} \prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)\sum_{s=2}^{4} y_{isjk}\left[\frac{1}{\phi^*_{ijk}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi^*_{ijk}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}$$
$$\frac{W_i^{rc\beta-1}e^{-W_i}}{\Gamma(rc\beta)G_{rc\beta}\left(\dfrac{\beta b_i}{1-\pi_{i1}}\right)}\,\mathrm{d}W_i\left\{\frac{\prod_{j,k}\phi^{*\ y_{i.jk}+\beta-1}_{ijk}}{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\boldsymbol{j})}\right\}$$
$$\left\{\frac{\prod_{s=1}^{4}\pi_{is}^{y_{is..}+\mu_2 s\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}\right\}\mathrm{d}\phi_i^*\,\mathrm{d}\boldsymbol{\pi}_i, \qquad (13)$$

with $b_i = \min\{\{1/\phi^*_{ijk}\},\ j=1,\ldots,r; k=1,\ldots,c\}$ and

$$\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$$
$$= \frac{a_0}{(a_0+\tau_1)^2}\cdot\frac{b_0}{(b_0+\tau_2)^2}\cdot\frac{c_0}{(c_0+\beta)^2}$$
$$\prod_{i=1}^{A}\frac{\Gamma(rc\beta)}{[\Gamma(\beta)]^{rc}}\mathrm{D}(\boldsymbol{y}_i^{(1)}+\beta\boldsymbol{j})$$
$$\times\prod_{i=1}^{A}\binom{n_i}{\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \boldsymbol{y}'_{i3}, \boldsymbol{y}'_{i4}}\frac{\mathrm{D}(\boldsymbol{y}_i^{(1)}+\boldsymbol{\mu}_1\tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1\tau_1)}\frac{\mathrm{D}(\boldsymbol{y}_i^{(2)}+\boldsymbol{\mu}_2\tau_2)}{\mathrm{D}(\boldsymbol{\mu}_2\tau_2)}. \qquad (14)$$

To evaluate I_i for each $i = 1, \ldots, A,$ we proceed as follows given $(\Omega, \boldsymbol{y}_{i(1)})$:

1. Draw independent samples of vectors $\boldsymbol{\pi}_i$ and ϕ_i^* from the Dirichlet$(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2\tau_2)$ and Dirichlet$(\boldsymbol{y}_i^{(1)} + \beta\,\boldsymbol{j})$, respectively. For each $\boldsymbol{\pi}_i$ and ϕ_i^*, draw a sample of values for W_i from the truncated gamma distribution on the interval $(0, \{\beta b_i/1-\pi_{i1}\})$ with parameter $rc\beta$.
2. For each $\boldsymbol{\pi}_i$, ϕ_i^* and W_i selected in step (1), compute R_1R_2, where

$$R_1 = G_{rc\beta}\left(\frac{\beta b_i}{1-\pi_{i1}}\right)\Bigg/\left[G_\beta\left(\frac{\beta}{1-\pi_{i1}}\right)\right]^{rc} \qquad (15)$$

and

$$R_2 = \prod_{j,k}\left\{\left(\frac{W_i}{\beta}\right)\sum_{s=2}^{4} y_{isjk}\right.$$
$$\left.\left[\frac{1}{\phi^*_{ijk}}\left(1+\frac{1-\pi_{i1}}{\pi_{i1}}\left\{1-\frac{W_i\phi^*_{ijk}}{\beta}\right\}\right)\right]^{y_{i1jk}}\right\}. \qquad (16)$$

3. Repeat steps (1) and (2) 1,000 times. Then compute the average of R_1R_2 over these 1,000 values.

The rest of our computation has two parts. First, we use the griddy Metropolis-Hastings sampler to draw from $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. We sample μ_1, μ_2, τ_1 and τ_2 from their conditional posterior densities using grids; this entails transforming τ_1 and τ_2 to the unit interval $(0, 1)$. For each distribution, 100 grids are used; see Nandram, Cox and Choi (2005) for a similar procedure. Here $\boldsymbol{y}_{(1)}$ is drawn by sampling from its conditional probability mass function component wise. Draws are made from the conditional posterior density of β using a Metropolis step in a manner similar to Nandram and Choi (2002a, b). We have performed this algorithm 11,000 times and we allowed a "burn-in" of 1,000 iterates. We found that the autocorrelations among the iterates was small, thereby indicating strong mixing of the sampler. We have also used the batch-means method to further assess the computation. We used batches of 25 to compute numerical standard errors.

Second, we use the SIR algorithm to subsample the sample of 10,000 iterates we obtained from π_a $(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$. For each of the 10,000 iterates we calculate the weights

$$w_m = \frac{h(\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})}{\pi_a(\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})},$$
$$m = 1, \ldots, M = 10{,}000, \qquad (17)$$

and we resample $\{\Omega^{(m)}, \boldsymbol{y}_{(1)}^{(m)}\}$ with probabilities proportional to the weights w_m for $m = 1, \ldots, M$ without replacement. We use a 10% sampling, and we subsample the 10,000 iterates to get 1,000 iterates; sampling without replacement is a good idea because it avoids repeated values which already exist because the Metropolis-Hastings sampler is not really an accept-reject sampler and it gives repeated values. As usual with sampling without replacement the weights are calculated every time a value is selected.

Finally, we can now make exact (within limitations of Markov chain Monte Carlo methods) inference about $\boldsymbol{p}_i$ a posteriori. Letting $y_{i\cdot jk} = \sum_{s=1}^{4} y_{isjk}$ and $\boldsymbol{y}_i^*$ denote the vector of $y_{i\cdot jk}$. Then,

$$\boldsymbol{p}_i \mid \boldsymbol{y}_i^*, \boldsymbol{\mu}_1, \tau_1 \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{y}_i^* + \boldsymbol{\mu}_1\tau_1),\ i = 1, \ldots, A.$$

Thus, for each value of $\boldsymbol{y}_i^*, \boldsymbol{\mu}_1,$ and τ_i we obtain from the SIR algorithm, we draw a value of $\boldsymbol{p}_i, i = 1, \ldots, A.$ Thus, we obtain a Rao-Blackwellized density for each of the $\boldsymbol{p}_i,$ and inference proceeds in the usual way.

4. An illustrative example

Our illustrative example is in health statistics. In Section 4.1 we briefly discuss the data we used from the third National Health and Nutrition Examination Survey (NHANES III). Specifically, we study the relationship between bone mineral density and family income; see Nandram, Cox and Choi (2005) for a discussion of this problem. In Section 4.2, after briefly discussing our computation, we present posterior inference on the cell probabilities. In section 4.3 using the Bayes factor we discuss the relation between BMD and FI.

4.1 NHANES III data

The sample design is a stratified multistage probability design which is representative of the total civilian non-institutionalized population, 2 months of age or older, in the United States. Further details of the NHANES III sample design are available (National Center for Health Statistics 1992, 1994). The NHANES III data collection consists of two parts: the first part is the sample selection and the interview of the members of a sampled household for their personal information, and the second part is the examination of those interviewed at the mobile examination center (MEC). The health examination has information on physical examination, tests and measurements performed by technicians, and specimen collection. The sample was selected from households in 81 primary units across the continental United States during the period from October 1988 through September 1994. The final data for this study is a part of the 35 largest primary sampling units with population at least 500,000, and we consider 13 subnational areas.

Nonresponse occurs in the interview and examination parts of the survey. The interview nonresponse arises from sampled persons who did not respond for the interview. Some of those who were already interviewed and included in the subsample for a health examination missed the examination at home or at the MEC, thereby missing all or part of the examinations.

Doctors believe that obese and overweight individuals do not generally turn up at the MEC. Cohen and Duffy (2002) point out that “Health surveys are a good example, where it seems plausible that propensity to respond may be related to health.” NHANES III is a good example.

Sampled persons in NHANES III can be categorized by many types of attributes, and researchers analyze such categorical tables for goodness of fit or independence. Here we study bone mineral density (BMD) and family income (FI). We note here that while FI is a discrete variable, we have classified BMD into three levels (normal, osteopenia and osteoporosis), and FI into three levels (low, medium and high). However, only partial classification of the individuals is available because some individuals are classified by only one attribute while others are not classified. About 62% of the households have both FI and BMD observed, 8% with only BMD observed, 29% with only income observed, 1% with neither income nor BMD among those participated in the examination stage. Our problem is to estimate the cell probabilities and to test for association between BMD and FI for each of 13 subnational areas using our expansion model that pools the data adaptively.

In Table 1 we present the 3×3 tables of BMD and FI for the aforementioned 13 areas. Note that areas 6 and 48 have enough data so that they can stand by themselves. However, the other areas are very small; the counts in the table with row totals are generally small except for area 17 and the counts in the table with just total are small. Even for the table with complete data the cell counts are generally small forcing us to use small area estimation techniques to borrow strength.

Table 1
Counts of the 3×3 tables of BMD and FI corresponding to 13 subnational areas in NHANES III

State	Complete Table									Column Total			Row Total			Total
4	21	14	9	8	7	3	2	2	0	11	5	6	4	0	1	1
6	257	127	106	92	51	32	32	5	7	178	54	82	65	28	4	20
12	33	18	21	22	4	4	15	5	0	18	11	16	5	6	2	1
17	25	15	13	8	5	3	0	0	1	18	10	16	17	2	2	4
25	9	7	12	6	5	9	2	1	0	9	6	12	1	4	5	1
26	18	11	18	6	5	9	2	1	1	10	5	11	4	3	0	1
29	9	4	10	3	2	4	3	1	2	9	2	9	0	2	4	1
36	42	17	27	32	13	18	9	6	1	43	21	42	9	7	6	1
39	8	6	14	2	5	4	3	0	1	9	7	5	2	3	0	0
42	14	8	11	12	8	4	8	1	2	35	15	24	3	1	0	0
44	12	9	6	8	5	0	5	1	0	19	4	12	7	1	0	1
48	159	44	22	51	11	13	9	6	2	88	12	23	16	8	2	14
53	14	10	15	10	10	14	3	1	1	9	4	8	2	4	1	0

Note: In the complete 3×3 table the first (second, third) set of three numbers is the first (second, third) row; the column (row) total refers to the 3×3 table with only column (row) totals; the total refers to the 3×3 table with only total.

4.2 Posterior inference of the cell probabilities

We discuss the performance of our computations for the expansion model, and then we discuss posterior inference about the cell probabilities. We use the posterior mean (PM), posterior standard deviation (PSD) and 95% credible interval for each of parameters of interest. We also present the numerical standard errors (NSE) to assess the repeatability of our computations.

In Table 2 we present summaries of the posterior distributions of $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$ and β, both before and after the application of the SIR algorithm. These summaries are very similar indicating that the SIR approximation $\pi_a(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ is not unreasonable. For example, a 95% credible interval for β before and after the application of the SIR algorithm are respectively (1.081, 1.940) and (1.086, 1.947), very good agreement. The estimates of τ_1 and τ_2 should show the largest discrepancies, but these are also reasonably close [*e.g.*, for τ_1 95% credible intervals with the approximation is (28.282, 64.204) and with the SIR algorithm it is (27.962, 64.425)]. In both cases the NSEs are small indicating that the computations are repeatable.

In Table 3 we have compared our expansion model (Model 3) with two other models. Model 1, an ignorable nonresponse model, and Model 2, a nonignorable nonresponse model (no centering), are described in Appendix B. For illustration we have selected three areas, a large area, a medium-sized area and a small area. There are differences among the three models. In general, the larger estimates tend to be smaller for Model 2, and even smaller than Model 1, than for Model 3 (*i.e.*, the estimates from Model 3 are naturally closest to Model 1, and not Model 2). Model 2 produces the largest variability; as expected, Model 3 gives slightly larger variability than Model 1. Because of space restrictions we have not presented the NSEs, but we note that they are all smaller than 0.005.

4.3 Bayes factor for evidence of association

We have also considered the association between BMD and FI. It appears doubtful whether such an association might exist, but it is interesting to look at this issue; see Nandram, Cox and Choi (2005) for a discussion on this problem. We use the Bayes factor (Kass and Raftery 1995) to measure the strength of the evidence of an association relative to no association in the $r \times c$ categorical table. We have done so for each of the thirteen areas and all areas combined.

We have used two procedures, one without extensive modeling and the other using our nonignorable nonresponse (expansion) model. The simple method is to fill in the cell counts using an ordinary raking procedure, and we assume there is no error in doing so. This is a common sense procedure that survey practitioners have used routinely. In the second procedure using our nonignorable nonresponse model, we have obtained 1,000 combined tables for each area as described in Section 3 on computations. For each area we have obtained the cell counts for all four tables, and we summed them to get a single table of all counts.

Table 2
NHANES data on 13 areas: Comparison of the approximate posterior density and the correct posterior density using the posterior means (PM), posterior standard deviations (PSD), numerical standard errors (NSE) and 95% credible intervals of the hyperparameters

	Approximation				Adjusted			
	PM	PSD	NSE	95% Int	PM	PSD	NSE	95% Int
μ_{21}	0.528	0.031	0.001	(0.463, 0.582)	0.525	0.031	0.008	(0.456, 0.578)
μ_{22}	0.131	0.021	0.001	(0.096, 0.181)	0.133	0.021	0.002	(0.094, 0.179)
μ_{23}	0.328	0.028	0.001	(0.274, 0.383)	0.328	0.028	0.005	(0.269, 0.383)
μ_{24}	0.013	0.006	0.000	(0.004, 0.027)	0.014	0.006	0.000	(0.004, 0.029)
τ_2	21.638	9.559	0.255	(8.347, 46.587)	20.078	8.632	0.303	(8.538, 38.625)
μ_{111}	0.280	0.023	0.001	(0.234, 0.324)	0.277	0.023	0.004	(0.228, 0.319)
μ_{112}	0.133	0.016	0.000	(0.102, 0.165)	0.134	0.017	0.002	(0.101, 0.165)
μ_{113}	0.200	0.019	0.000	(0.163, 0.238)	0.199	0.019	0.003	(0.162, 0.236)
μ_{121}	0.105	0.015	0.000	(0.078, 0.135)	0.107	0.015	0.002	(0.079, 0.135)
μ_{122}	0.065	0.011	0.000	(0.044, 0.088)	0.065	0.011	0.001	(0.044, 0.087)
μ_{123}	0.072	0.012	0.000	(0.050, 0.096)	0.073	0.012	0.001	(0.049, 0.097)
μ_{131}	0.061	0.011	0.000	(0.041, 0.083)	0.061	0.011	0.001	(0.040, 0.083)
μ_{132}	0.037	0.008	0.000	(0.023, 0.054)	0.036	0.008	0.001	(0.022, 0.054)
μ_{133}	0.048	0.009	0.000	(0.031, 0.068)	0.048	0.009	0.001	(0.031, 0.068)
τ_1	45.960	10.094	0.153	(28.282, 64.204)	45.177	10.562	0.679	(27.962, 64.423)
β	1.472	0.218	0.004	(1.081, 1.940)	1.449	0.208	0.022	(1.086, 1.947)

Note: The hyperparameters are $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \tau_1, \tau_2$ and β.

Table 3
Posterior means of the cell probabilities and 95% credible intervals (CI) for three areas (large, medium and small) by the three models

	Model 1			Model 2			Model 3		
Cell	**PM**	**PSD**	**95% CI**	**PM**	**PSD**	**95% CI**	**PM**	**PSD**	**95% CI**
a. Large									
(1,1)	0.239	0.044	(0.157, 0.326)	0.196	0.046	(0.117, 0.295)	0.259	0.038	(0.189, 0.335)
(1,2)	0.140	0.035	(0.078, 0.213)	0.127	0.035	(0.068, 0.200)	0.132	0.029	(0.082, 0.197)
(1,3)	0.240	0.044	(0.159, 0.332)	0.198	0.047	(0.118, 0.301)	0.248	0.037	(0.175, 0.322)
(2,1)	0.092	0.032	(0.039, 0.162)	0.098	0.040	(0.037, 0.188)	0.077	0.022	(0.039, 0.126)
(2,2)	0.074	0.028	(0.029, 0.136)	0.077	0.030	(0.030, 0.144)	0.056	0.020	(0.024, 0.099)
(2,3)	0.133	0.036	(0.070, 0.210)	0.121	0.042	(0.056, 0.219)	0.110	0.028	(0.058, 0.168)
(3,1)	0.036	0.020	(0.008, 0.083)	0.069	0.039	(0.013, 0.153)	0.047	0.018	(0.018, 0.086)
(3,2)	0.023	0.015	(0.003, 0.061)	0.043	0.025	(0.007, 0.100)	0.032	0.014	(0.009, 0.063)
(3,3)	0.025	0.017	(0.003, 0.066)	0.071	0.040	(0.010, 0.154)	0.042	0.016	(0.016, 0.079)
b. Medium									
(1,1)	0.233	0.034	(0.169, 0.302)	0.213	0.043	(0.141, 0.305)	0.254	0.032	(0.194, 0.318)
(1,2)	0.143	0.028	(0.093, 0.200)	0.127	0.032	(0.072, 0.196)	0.146	0.024	(0.102, 0.197)
(1,3)	0.190	0.031	(0.132, 0.254)	0.140	0.034	(0.084, 0.218)	0.208	0.027	(0.156, 0.259)
(2,1)	0.174	0.031	(0.118, 0.237)	0.160	0.042	(0.092, 0.249)	0.154	0.027	(0.106, 0.211)
(2,2)	0.043	0.018	(0.015, 0.083)	0.060	0.028	(0.017, 0.124)	0.032	0.012	(0.012, 0.059)
(2,3)	0.049	0.020	(0.017, 0.095)	0.065	0.031	(0.018, 0.136)	0.042	0.014	(0.020, 0.072)
(3,1)	0.112	0.025	(0.068, 0.167)	0.120	0.041	(0.059, 0.209)	0.092	0.020	(0.056, 0.134)
(3,2)	0.047	0.018	(0.018, 0.088)	0.059	0.026	(0.019, 0.118)	0.040	0.014	(0.018, 0.071)
(3,3)	0.010	0.009	(0.000, 0.033)	0.056	0.032	(0.006, 0.122)	0.032	0.012	(0.013, 0.059)
c. Small									
(1,1)	0.196	0.052	(0.103, 0.305)	0.164	0.055	(0.077, 0.288)	0.253	0.043	(0.175, 0.334)
(1,2)	0.081	0.034	(0.028, 0.158)	0.081	0.032	(0.030, 0.155)	0.091	0.028	(0.043, 0.152)
(1,3)	0.213	0.052	(0.118, 0.323)	0.175	0.055	(0.087, 0.300)	0.220	0.043	(0.137, 0.306)
(2,1)	0.093	0.041	(0.028, 0.186)	0.111	0.055	(0.029, 0.234)	0.073	0.028	(0.030, 0.139)
(2,2)	0.056	0.029	(0.012, 0.126)	0.066	0.031	(0.018, 0.136)	0.045	0.020	(0.014, 0.094)
(2,3)	0.115	0.045	(0.042, 0.215)	0.118	0.053	(0.038, 0.240)	0.092	0.030	(0.041, 0.158)
(3,1)	0.115	0.048	(0.036, 0.222)	0.113	0.056	(0.031, 0.239)	0.081	0.030	(0.033, 0.148)
(3,2)	0.044	0.028	(0.006, 0.113)	0.065	0.035	(0.013, 0.144)	0.043	0.020	(0.012, 0.086)
(3,3)	0.087	0.042	(0.022, 0.184)	0.107	0.055	(0.023, 0.227)	0.103	0.034	(0.047, 0.181)

Note: See Appendix B for a description of Models 1 and 2.

The raking procedure to obtain the cell counts is described as follows. Let n_{jk} denote the cell counts for the four tables combined. Let $n_{jk}^{(1)}$ denote the cell counts for the table with complete data, $n_{j,c+1}^{(2)}$ denote the table with row totals, $n_{r+1,k}^{(3)}$ denote the table with column totals and $n_{r+1,c+1}^{(4)}$ denote the table with total. The cell counts for the four tables are estimated as

$$n_{jk} = n_{jk}^{(1)} + \left(\frac{n_{jk}^{(1)}}{n_{j\cdot}^{(1)}}\right) n_{j,c+1}^{(2)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot k}^{(1)}}\right) n_{r+1,k}^{(3)} + \left(\frac{n_{jk}^{(1)}}{n_{\cdot\cdot}^{(1)}}\right) n_{r+1,c+1}^{(4)},$$

$j = 1, \ldots, r, k = 1, \ldots, c.$

In either case we denote the sum of the cell counts for each area by n_{jk}. For the raking procedure we have a single table for each area, and for the nonignorable nonresponse model we have a sample of 1,000 tables for each area. We also have a single combined table for all areas under the raking procedure and 1,000 tables for all areas combined. We obtain the Bayes factor for each table under a multinomial-Dirichlet model. It is worth noting that our method uses the expansion model so that the cell counts borrow strength from other areas unlike the raking procedure.

Then, for each table we take

$$\boldsymbol{n} \mid \boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi}) \text{ and } \boldsymbol{\pi} \sim \text{Dirichlet}(\mathbf{1}).$$

That is, we take a uniform prior for $\boldsymbol{\pi}$ with $\pi_{jk} > 0$ and $\sum_{j=1}^{r} \sum_{k=1}^{c} \pi_{jk} = 1$. Under the hypothesis of no association we have $\pi_{jk} = \pi_j \pi_k$, where $\pi_j > 0, \sum_{j=1}^{r} \pi_j = 1$ and $\pi_k > 0, \sum_{k=1}^{c} \pi_k = 1$. Thus, the hypothesis of association is that the π_{jk} are unrestricted (except that they are nonnegative and sum to unity) whereas for the hypothesis of no associate $\pi_{jk} = \pi_j \pi_k$.

The Bayes factor is the ratio of the marginal likelihood under association versus the marginal likelihood under no association. This measures the strength of evidence of association versus no association; see Kass and Raftery (1995). Let $p_a(\boldsymbol{n})$ denote the marginal likelihood under association and $p_0(\boldsymbol{n})$ denote the marginal likelihood under no association. Then, letting $n_{j\cdot} = \sum_{k=1}^{c} n_{jk}$ and $n_{\cdot k} = \sum_{j=1}^{r} n_{jk}$, it is easy to show that

$$p_a(\boldsymbol{n}) = p_0(\boldsymbol{n}) \left\{ \prod_{u=0}^{n-1} \frac{u+rc}{(u+r)(u+c)} \right\} \frac{\prod_{j=1}^{r} n_{j\cdot}! \prod_{k=1}^{c} n_{\cdot k}!}{\prod_{j=1}^{r} \prod_{k=1}^{c} n_{jk}!},$$

where $p_0(\boldsymbol{n}) = n! \prod_{u=0}^{n-1} (u + rc)^{-1}$. Observe that $p_0(\boldsymbol{n})$ is not a function of $\{n_{jk}\}$. Thus, as a measure of association it is the deviation of $\prod_{j=1}^{r} n_{j.}! \prod_{k=1}^{c} n_{.k}!$ from $\prod_{j=1}^{r} \prod_{k=1}^{c} n_{jk}!$ that matters. However, we note that for the classical Pearson statistic for independence it is the deviations of n_{jk} from $n_{j.}n_{.k}$ that matter. But note that this test cannot be applied because many of the expected cell counts are smaller than 5 under the hypothesis of no association and multinomial sampling.

We present our results in Table 4 and in Figure 1 corresponding to the data in Table 1 for the cross-classification of BMD and FI. We have presented the logarithms of the marginal likelihoods (base *e*) and the Bayes factors; these are to be interpreted using the rule of thumb of Kass and Raftery (1995).

In Figure 1 we can see that the box plots are all above zero except the one for the third area which provides no evidence for association; perhaps there is no evidence for association in area 42 (10 in figure) as well. A summary of these results are presented in Table 4. The Bayes factors show association in all areas, except area 12, and they are much larger under the nonignorable nonresponse model. Area 6 and all areas combined are elevated (336.3 *vs.* 5.8 and 3,798.2 *vs.* 0.183).

Table 4
NHANES data on 13 areas: Comparison of the negative marginal likelihoods and Bayes factors or association of BMD and F1 from the raking procedure and the expansion model by area

	Raking			Expansion	
area	$-\ln\{p_0(n)\}$	$-\ln\{p_a(n)\}$	***BF***	$-\ln\{p_a(n)\}$	***BF***
4	26.19	23.07	22.855	$23.5_{0.014}$	$14.78_{0.169}$
6	45.73	43.98	5.766	$40.5_{0.038}$	$336.27_{11.465}$
12	31.14	38.01	0.001	$33.4_{0.054}$	$0.37_{0.027}$
17	29.13	27.03	8.134	$27.0_{0.026}$	$10.27_{0.191}$
25	25.44	26.02	0.558	$23.8_{0.029}$	$9.55_{0.202}$
26	26.89	23.18	40.562	$23.9_{0.018}$	$24.71_{0.370}$
29	23.21	20.87	10.301	$21.3_{0.018}$	$8.40_{0.115}$
36	34.99	36.09	0.330	$33.1_{0.064}$	$21.13_{0.928}$
39	23.77	24.89	0.325	$23.6_{0.044}$	$2.24_{0.68}$
42	29.51	30.21	0.497	$30.3_{0.099}$	$4.33_{0.255}$
44	25.61	30.48	0.008	$24.4_{0.027}$	$5.19_{0.137}$
48	38.83	35.34	32.650	$39.1_{0.060}$	$2.15_{0.081}$
53	27.11	24.82	9.865	$24.2_{0.017}$	$19.40_{0.282}$
All	53.43	55.13	0.183	$46.1_{0.049}$	$3{,}798.24_{151.82}$

Note: Area 'all' refers to all areas combined; the notation a_b means that the average is *a* and the standard error is *b* over the 1,000 iterates; $\ln\{p_0(n)\}$ is the same for both procedures.

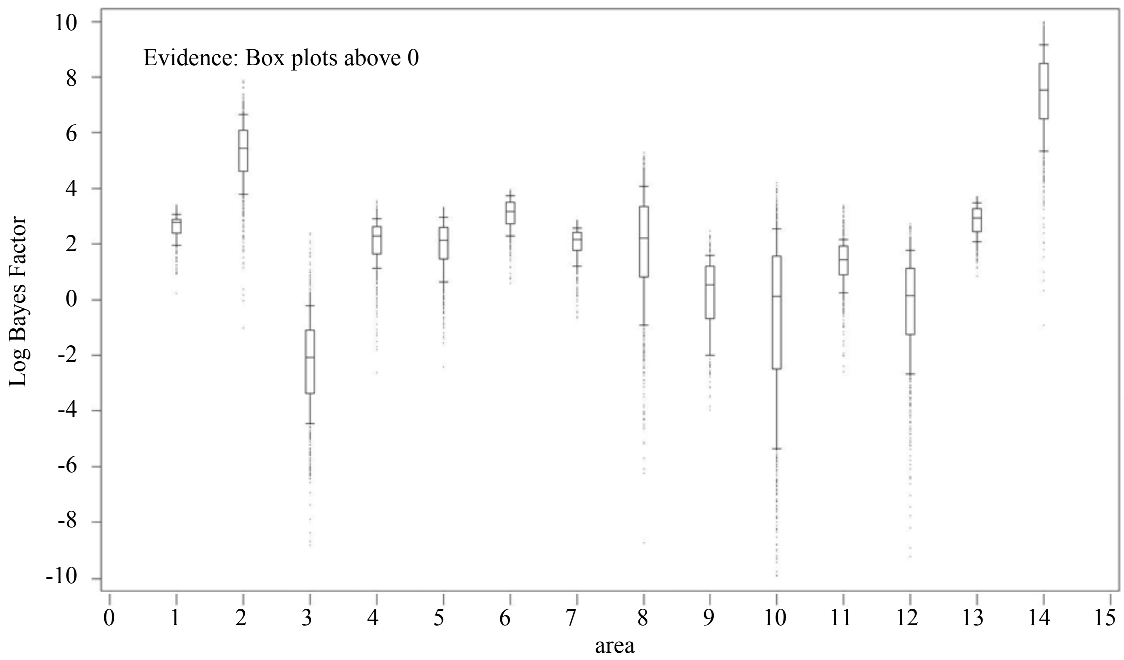


Figure 1 Box plots of log Bayes factors versus areas to measure evidence for association between BMD and FI

5. Concluding remarks

The purpose of this paper has been to develop a methodology to analyze data from incomplete two-way categorical tables, each table corresponding to an area. We have done so by extending the Bayesian methodology of Nandram and Choi (2002a, b) for binary data to $r \times c$ categorical tables for small areas. We have constructed a new Bayesian nonignorable nonresponse model (*i.e.*, expansion model) which is centered on the ignorable nonresponse model. We have used Markov chain Monte Carlo methods (specifically the griddy Metropolis-Hastings sampler) to fit the model. We have compared our model with an ignorable nonresponse model and a nonignorable nonresponse model. Finally, we have done an illustrative example on estimating the cell probabilities for the 3×3 table of BMD and income over thirteen subnational areas.

We have shown that there are differences among the three models. Using the data on BMD and FI, we have shown that our expansion model is a compromise between the ignorable nonresponse model and the nonignorable nonresponse model. Using the Bayes factor we have shown that there are differences between the tests of association for BMI and FI when the cell counts are estimated from our model and when using a raking procedure. In fact, owing to the borrowing of strength, we have seen that the evidence for association under our model is much stronger than the from the raking procedure.

There are three additional avenues that we can explore. First, we can construct a model to incorporate systematic departure from ignorability. This task will need more costly field work to get the much-needed information. Second, it is also interesting to relax the assumption that the margins of the categorical table are fixed; see, for example, Nandram (2009) who looked at a single large area. Third, there can be further improvement in calibration (*i.e.*, incorporating information about margins).

Acknowledgements

We are grateful to the two referees and the associate editor for their assistance in the presentation of the material.

Appendix A

Joint posterior density of the expansion model

First, integrating the joint posterior density over $\boldsymbol{p}$ we get that

$$h(\Omega, \boldsymbol{\pi}, \phi, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \propto \prod_{i=1}^{4} \frac{\mathrm{D}(\boldsymbol{y}_i^{(1)} + \boldsymbol{\mu}_1 \tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1 \tau_1)}$$

$$\left[\binom{n_i}{\boldsymbol{y}_{i1}', \boldsymbol{y}_{i2}', \boldsymbol{y}_{i3}', \boldsymbol{y}_{i4}'} (1 - \pi_{i1})^{-y_{i\ldots}} \right.$$

$$\times \left\{ \prod_{s=1}^{4} \pi_{is}^{y_{is..}} \prod_{j,k} \phi_{ijk}^{y_{i.jk}} [a_{ijk}(\pi_{i1}, \phi_{ijk})]^{y_{i1jk}} \right\}$$

$$\times \left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{\mu_{2s}\tau_2 - 1}}{\mathrm{D}(\boldsymbol{\mu}_2 \tau_2)} \prod_{j=1}^{r} \prod_{k=1}^{c} \frac{\beta^{\beta}}{(1 - \pi_{i1})^{\beta}} \right.$$

$$\left. \left. \frac{\phi_{ijk}^{\beta - 1} e^{-\frac{\beta \phi_{ijk}}{1 - \pi_{i1}}}}{\Gamma(\beta) G_{\beta}[\beta(1 - \pi_{i1})^{-1}]} \right\} \right]$$

$$\times \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2}, \qquad \text{(A.1)}$$

where the $rc \times 1$ vector

$$\boldsymbol{y}_i^{(1)} \overset{\text{def}}{=} (y_{i.11}, y_{i.12}, \ldots, y_{i.1c}, y_{i.21}, \ldots, y_{i.2c}, y_{i.r1}, \ldots, y_{i.rc})'.$$

Now let $\boldsymbol{j}$ denote an $rc \times 1$ vector of ones and let

$$\boldsymbol{y}_i^{(2)} \overset{\text{def}}{=} (y_{i1..}, y_{i2..}, y_{i3..}, y_{i4..})'.$$

Then, collapsing over $\boldsymbol{\pi}$ and ϕ, we have that

$$h(\Omega, \boldsymbol{y}_{(1)} \mid \boldsymbol{y}_1, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) \propto \frac{a_0}{(a_0 + \tau_1)^2} \cdot \frac{b_0}{(b_0 + \tau_2)^2} \cdot \frac{c_0}{(c_0 + \beta)^2}$$

$$\times \prod_{i=1}^{A} \binom{n_i}{\boldsymbol{y}_{i1}', \boldsymbol{y}_{i2}', \boldsymbol{y}_{i3}', \boldsymbol{y}_{i4}'} \frac{\mathrm{D}(\boldsymbol{y}_i^{(1)} + \boldsymbol{\mu}_1 \tau_1)}{\mathrm{D}(\boldsymbol{\mu}_1 \tau_1)} \frac{\mathrm{D}(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2 \tau_2)}{\mathrm{D}(\boldsymbol{\mu}_2 \tau_2)}$$

$$\times \prod_{i=1}^{A} \left[\frac{\beta^{\beta}}{\Gamma(\beta)} \right]^{rc} \mathrm{D}(\boldsymbol{y}_i^{(1)} + \beta \boldsymbol{j}) I_i, \qquad \text{(A.2)}$$

where

$$I_i = \iint \prod_{j,k} \left\{ \frac{\left[\left(\frac{1 - \pi_{i1}}{\phi_{ijk}} \right) \left(1 + \frac{1}{\pi_{i1}} \{1 - \pi_{i1} - \phi_{ijk}\} \right) \right]^{y_{i1jk}}}{(1 - \pi_{i1})^{y_{i.jk} + \beta} G_{\beta} \left(\frac{\beta}{1 - \pi_{i1}} \right)} \right\}$$

$$\times \left\{ \prod_{j,k} e^{-\frac{\beta \phi_{ijk}}{1 - \pi_{i1}}} \right\} \left\{ \frac{\prod_{j,k} \phi_{ijk}^{y_{i.jk} + \beta - 1}}{\mathrm{D}(\boldsymbol{y}_i^{(1)} + \beta \boldsymbol{j})} \right\}$$

$$\left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{y_{is..} + \mu_{2s}\tau_2 - 1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2 \tau_2)} \right\} \mathrm{d}\phi_i \, \mathrm{d}\boldsymbol{\pi}_i. \qquad \text{(A.3)}$$

Note that $0 \le \pi_{is} \le 1, \sum_{s=1}^{4} \pi_{is} = 1$ and $0 \le \phi_{ijk} \le 1.$ We simplify the computation for I_i in (A.3) in two steps.

First, in (A.3) we make the transformation

$$\phi_{ijk} = T_i\, \phi^*_{ijk} \qquad \sum_{j=1}^{r}\sum_{k=1}^{c} \phi_{ijk} = T_i.$$

The new variables ϕ^*_{ijk} satisfy the relationships $0 \le \phi^*_{ijk} \le 1, \sum_{j=1}^{r}\sum_{k=1}^{c} \phi^*_{ijk} = 1$ and the T_i are restricted so that $0 \le T_i \le 1/\phi^*_{ijk},$ for $j = 1, \ldots, r, k = 1, \ldots, c$ and $0 \le T_i \le rc.$ With this transformation we have

$$I_i = \iiint_0^{b_i} \prod_{j,k} \left\{ \left[\frac{T_i}{1-\pi_{i1}} \right]^{\sum_{s=2}^{4} y_{isjk}} \left[\frac{1}{\phi^*_{ijk}} \left(1 + \frac{1-\pi_{i1}}{\pi_{i1}} \left\{ 1 - \frac{T_i}{1-\pi_{i1}} \phi^*_{ijk} \right\} \right) \right]^{y_{i1jk}} \right\}$$

$$\times \left\{ \frac{\left(\dfrac{T_i}{1-\pi_{i1}} \right)^{rc\beta-1} e^{-\frac{\beta T_i}{1-\pi_{i1}}}}{(1-\pi_{i1}) \left[G_\beta \left(\dfrac{\beta}{1-\pi_{i1}} \right) \right]^{rc}} \right\} \left\{ \frac{\prod_{j,k} \phi^{*\ y_{i.jk}+\beta-1}_{ijk}}{\mathrm{D}(\boldsymbol{y}_i^{(1)} + \beta \boldsymbol{j})} \right\}$$

$$\times \left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{y_{is..}+\mu_{2s}\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2 \tau_2)} \right\} \mathrm{d}T_i\, \mathrm{d}\boldsymbol{\phi}_i^* \mathrm{d}\boldsymbol{\pi}_i,$$

where $b_i = \min\{\{1/\phi^*_{ijk}\}\, j = 1, \ldots, r; k = 1, \ldots, c\}.$

Second, letting $W_i = \{\beta T_i / 1 - \pi_{i1}\}$ and absorbing the factor $\beta^{rc\beta} / \Gamma(rc\beta)$ in $I_i,$ with some additional algebra, we have

$$I_i = \iiint_0^{\frac{\beta b_i}{1-\pi_{i1}}} \frac{G_{rc\beta}\left(\dfrac{\beta b_i}{1-\pi_{i1}} \right)}{\left[G_\beta \left(\dfrac{\beta}{1-\pi_{i1}} \right) \right]^{rc}}$$

$$\prod_{j,k} \left\{ \left(\frac{W_i}{\beta} \right)^{\sum_{s=2}^{4} y_{isjk}} \left[\frac{1}{\phi^*_{ijk}} \left(1 + \frac{1-\pi_{i1}}{\pi_{i1}} \left\{ 1 - \frac{W_i \phi^*_{ijk}}{\beta} \right\} \right) \right]^{y_{i1jk}} \right\}$$

$$\times \frac{W_i^{rc\beta-1} e^{-W_i}}{\Gamma(rc\beta) G_{rc\beta}\left(\dfrac{\beta b_i}{1-\pi_{i1}} \right)}$$

$$\mathrm{d}W_i \left\{ \frac{\prod_{j,k} \phi^{*\ y_{i.jk}+\beta-1}_{ijk}}{\mathrm{D}(\boldsymbol{y}_i^{(1)} + \beta \boldsymbol{j})} \right\} \left\{ \frac{\prod_{s=1}^{4} \pi_{is}^{y_{is..}+\mu_2 s\tau_2-1}}{\mathrm{D}(\boldsymbol{y}_i^{(2)} + \boldsymbol{\mu}_2 \tau_2)} \right\} \mathrm{d}\boldsymbol{\phi}_i^* \mathrm{d}\boldsymbol{\pi}_i. \quad \text{(A.4)}$$

Appendix B

Ignorable and nonignorable nonresponse models

Set $\psi_{ijk} \equiv 1$ in the expansion model to form the ignorable nonresponse model. For $i = 1, \ldots, A,$ we then take

$$\boldsymbol{\pi}_i \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2)$$

and independently

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1).$$

Also, $p(\tau_2) = \{1/(1+\tau_1)^2\},\ \tau_1 \ge 0, \boldsymbol{\mu}_1 \sim \text{Dirichlet}\,(\mathbf{1}),$ $p(\tau_2) = \{1/(1+\tau_2)^2\}, \tau_1 \ge 0$ and $\boldsymbol{\mu}_2 \sim \text{Dirichlet}\,(\mathbf{1}).$ Here we have independence at all levels and the vectors **1** are of the appropriate dimension with every coordinate equal to one. Note, that all the parameters of the ignorable model are identifiable and estimable.

Set $\pi_{isjk} = \pi_{is}\, \psi_{ijk}$ in the expansion model to form the nonignorable nonresponse model. In this case, for $i = 1, \ldots, A,$

$$\boldsymbol{\pi}_{ijk} \mid \boldsymbol{\mu}_2, \tau_2 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_2 \tau_2)$$

and independently

$$\boldsymbol{p}_i \mid \boldsymbol{\mu}_1, \tau_1 \overset{\text{iid}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_1 \tau_1).$$

In this model, the parameters $\boldsymbol{\pi}_{ijk}$ are not identifiable and we take $\tau_2 \sim \text{Gamma}\,(\alpha_0, \beta_0),$ where α_0 and β_0 are to be specified. The model specification is then completed by assigning $\tau_1, \boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ the same distributional properties as in the previous paragraph.

As in Nandram, Cox and Choi (2005), α_0 and β_0 are specified as follows. The ignorable nonresponse model is fit to obtain a sample from the posterior density of $\tau_2.$ Then α_0 and β_0 are obtained using the method of moments. Nandram, Cox and Choi (2005) found that inference about $\boldsymbol{p}_i$ is not very sensitive to the choice of these parameters.

References

Cohen, G., and Duffy, J.C. (2002). Are nonrespondents to health surveys less healthy than respondents? *Journal of Official Statistics*, 18, 13-23.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society*, Series B, 57, 45-97.

Forster, J.J., and Smith, P.W.F. (1998). Model-based inference for categorical survey data subject to non-ignorable nonresponse. *Journal of the Royal Statistical Society*, Series B, 60, 57-70.

Gelman, A. (2006). Prior distribution for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Greenland, S. (2009). Relaxation penalties and priors for plausible modeling of nonidentified bias sources. *Statistical Sciences*, 24, 195-210.

Kass, R.E., and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Second Edition. New York: John Wiley & Sons, Inc.

Nandram, B. (2009). Bayesian inference of the cell probabilities of a two-way categorical table under non-ignorability. *Communications in Statistics - Theory and Methods*, 38, 3015-3030.

Nandram, B., and Choi, J.W. (2002a). Hierarchical Bayesian nonresponse models for binary data from small areas with uncertainty about ignorability. *Journal of the American Statistical Association*, 97, 381-388.

Nandram, B., and Choi, J.W. (2002b). A Bayesian analysis of a proportion under nonignorable nonresponse. *Statistics in Medicine*, 21, 1189-1212.

Nandram, B., and Choi, J.W. (2004). A nonparametric Bayesian analysis of a proportion for a small area under nonignorable nonresponse. *Journal of Nonparametric Statistics*, 16, 821-839.

Nandram, B., and Choi, J.W. (2005). Hierarchical Bayesian nonignorable nonresponse regression models for small areas: An application to the NHANES data. *Survey Methodology*, 31, 73-84.

Nandram, B., and Choi, J.W. (2008). A Bayesian allocation of undecided voters. *Survey Methodology*, 34, 37-49.

Nandram, B., and Choi, J.W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, 105, 120-135.

Nandram, B., Cox, L.H. and Choi, J.W. (2005). Bayesian analysis of nonignorable missing categorical data: An application to bone mineral density and family income. *Survey Methodology*, 31, 213-225.

Nandram, B., Han, G. and Choi, J.W. (2002). A hierarchical Bayesian nonignorable nonresponse model for multinomial data from small areas. *Survey Methodology*, 28, 145-156.

Nandram, B., Liu, N., Choi, J.W. and Cox, L. (2005). Bayesian nonresponse models for categorical data from small areas: An application to BMD and age. *Statistics in Medicine*, 24, 1047-1074.

National Center for Health Statistics (1992). Third national health and nutrition examination survey. *Vital and Health Statistics*, Series 2, 113.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

Smith, A.F.M., and Gelfand, A.E. (1992). Bayesian statistics without tears: A sampling-resampling perspective. *The American Statistician*, 46, 84-88.

Survey Methodology, June 2012
Vol. 38, No. 1, pp. 95-99
Statistics Canada, Catalogue No. 12-001-X

Why one should incorporate the design weights when adjusting for unit nonresponse using response homogeneity groups

Phillip S. Kott [1]

Abstract

When there is unit (whole-element) nonresponse in a survey sample drawn using probability-sampling principles, a common practice is to divide the sample into mutually exclusive groups in such a way that it is reasonable to assume that each sampled element in a group were equally likely to be a survey nonrespondent. In this way, unit response can be treated as an additional phase of probability sampling with the inverse of the estimated probability of unit response within a group serving as an adjustment factor when computing the final weights for the group's respondents. If the goal is to estimate the population mean of a survey variable that roughly behaves as if it were a random variable with a constant mean within each group regardless of the original design weights, then incorporating the design weights into the adjustment factors will usually be more efficient than not incorporating them. In fact, if the survey variable behaved exactly like such a random variable, then the estimated population mean computed with the design-weighted adjustment factors would be nearly unbiased in some sense (*i.e.*, under the combination of the original probability-sampling mechanism and a prediction model) even when the sampled elements within a group are not equally likely to respond.

Key Words: Double protection; Prediction model; Probability sampling; Response model; Sampling phase; Stratified Bernoulli sampling.

1. Introduction

In the absence of nonresponse, it is possible to estimate the mean of a finite population from a survey sample without having to assume a statistical model which, no matter how reasonable, may not hold true. This is done by assigning each element of the population a positive probability of sample selection and creating estimators around this random-selection mechanism. Unfortunately, surveys taken in the real world often suffer from nonresponse.

Two different types of models can be used in the face of unit (whole-element) nonresponse. One is a prediction or outcome model in which the survey variable of interest is assumed to behave like a random variable with known characteristics but unknown parameters. The other is a response or selection model where the very act of an element's responding to a survey is treated as an additional phase of random sample selection.

Conventionally, survey statisticians prefer response models for two reasons. In addition to the convenience of response modeling allowing them to treat unit response as an additional phase of random sampling, a survey is usually designed to collect information on a number of variables from the sampled elements. Prediction modeling requires assuming a different model for each survey variable any one of which could fail. Response modeling, by contrast, requires only the assumption of a single model. This is no longer true when there is item (survey-variable-specific) nonresponse. Consequently, prediction modeling is more common when handling item nonresponse through imputation. That being said, item nonresponse is beyond the scope of this note.

Under an assumed response model, the element probabilities of response are treated as unknown, which means that they have to be estimated from the sample. Typically, the response mechanism is assumed to be independent across elements and not to depend on whether the element is in the sample (each element has an *a priori* probability of response which becomes operational if it is selected for the sample). The simplest and mostly commonly used response model separates the sample, and implicitly the entire population, into mutually exclusive groups, called "response homogeneity groups" by Särndal, Swensson and Wretman (1992) (the term "weighting classes" is more common; see, for example, Lohr (2009, pages 340-341)), and assumes that each element in a group is equally likely to be a unit respondent regardless of its probability of selection into the original sample, π_k. Thus, the response mechanism produces a stratified Bernoulli subsample with the groups serving as the strata.

Conditioned on the respondent sample sizes in the groups, a stratified Bernoulli subsample with unknown selection (response) probabilities is converted into a stratified simple random subsample with known selection probabilities: r_g / n_g for the elements in group g when that group has n_g sampled elements, r_g or which respond.

Although the conditional probabilities of response in group g under the stratified Bernoulli response model is r_g / n_g, we will see it is often better to multiply the design

[1] Phillip S. Kott, RTI International, Suite 902, 6100 Executive Blvd., Rockville, MD 20852, U.S.A. E-mail: pkott@rti.org.

weight, $d_k = 1/\pi_k$, for a responding element in the group not by n_g/r_g, but by

$$f_g = \frac{\sum\limits_{k\in S_g} d_k}{\sum\limits_{k\in R_g} d_k}, \tag{1}$$

where S_g is the original sample and R_g the respondent subsample in group g. This *adjustment factor* can be different from n_g/r_g when the d_k in group g vary.

Little and Vartivarian (2003) claim that using the f_g is what is usually done in practice. They argue, however, that incorporating design weights into the adjustment factor in this way can "add to the variance".

In section 2, we develop the notation for estimating the population mean of a survey variable. Using the n_g/r_g produces a double-expansion estimator, while using the f_g produces a reweighted-expansion estimator. We can express both using a formulation in Kim, Navarro and Fuller (2006). From that expression, it is possible to see that if the survey variable roughly behaves like a random variable with a constant mean within each group regardless of the design weights, then using the f_g will often be more efficient than using the n_g/r_g. In fact, if the survey variable behaved exactly like such a random variable, then the estimated population mean computed with the f_g would be nearly unbiased under the combination of the original sampling design and this prediction model even when the response model fails.

In Section 3, we show that empirical results in Little and Vartivarian (2003) are consistent with these arguments and offer some concluding remarks.

2. The two estimators

Suppose we want to estimate the population mean of a survey variable y_k:

$$\bar{y}_U = \frac{\sum\limits_{k\in U} y_k}{N} = \frac{\sum\limits_{g=1}^{G}\sum\limits_{k\in U_g} y_k}{\sum\limits_{g=1}^{G} N_g} = \frac{\sum\limits_{g=1}^{G} N_g \bar{y}_{U_g}}{\sum\limits_{g=1}^{G} N_g},$$

where the population U is divided into G groups, $U_1, \ldots, U_G$, each U_g contains N_g elements, and $N = N_1 + \ldots + N_G$. In the absence of nonresponse, each N_g is estimated in an unbiased fashion under probability-sampling theory by $\hat{N}_g = \sum_{k\in S_g} d_k$, and each $\bar{y}_{U_g}$ is estimated in a nearly (*i.e.*, asymptotically) unbiased fashion

$$\bar{y}_{S_g} = \frac{\sum\limits_{k\in S_g} d_k\, y_k}{\sum\limits_{k\in S_g} d_k}, \tag{2}$$

under mild conditions when n_g is sufficiently large. We assume both here.

For a formal statement of the conditions under which each $\bar{y}_{S_g}$ is consistent under probability sampling theory and therefore nearly unbiased, see Fuller (2009, page 115). The interested reader is directed to Fuller whenever a result in this note depends on assumptions about the design and population as the sample size grows arbitrarily large. A more rigorous treatment of much of what is to be discussed here under the response model can be found in Kim, Navarro and Fuller (2006).

Let us label the full-sample estimator for $\bar{y}_U$ we have been discussing $\bar{y}_S = \sum^G \hat{N}_g\, \bar{y}_{S_g}$. There are more direct ways to render $\bar{y}_S$, but this version will better serve our purposes.

If we adjust for nonresponse using the f_g in equation (1), we have the reweighted-expansion estimator:

$$\hat{\bar{y}}_{rw} = \frac{\sum\limits_{g=1}^{G}\left(f_g \sum\limits_{k\in R_g} d_k y_k\right)}{\sum\limits_{g=1}^{G}\left(f_g \sum\limits_{k\in R_g} d_k\right)}$$

$$= \frac{\sum\limits_{g=1}^{G}\left(\frac{\sum\limits_{k\in S_g} d_k}{\sum\limits_{k\in R_g} d_k}\sum\limits_{k\in R_g} d_k y_k\right)}{\sum\limits_{g=1}^{G}\sum\limits_{k\in S_g} d_k} = \frac{\sum\limits_{g=1}^{G}\left(\hat{N}_g \frac{\sum\limits_{k\in R_g} d_k y_k}{\sum\limits_{k\in R_g} d_k}\right)}{\sum\limits_{g=1}^{G}\hat{N}_g}.$$

Technically, $\hat{\bar{y}}_{rw}$ is the ratio of two reweighted-expansion estimators, but we use the simpler terminology here.

Employing the n_g/r_g results in the double-expansion estimator:

$$\hat{\bar{y}}_{de} = \frac{\sum\limits_{g=1}^{G}\left(\frac{n_g}{r_g}\sum\limits_{k\in R_g} d_k y_k\right)}{\sum\limits_{g=1}^{G}\left(\frac{n_g}{r_g}\sum\limits_{k\in R_g} d_k\right)}.$$

For our purposes, this estimator can also be expressed as

$$\hat{\bar{y}}_{de} = \frac{\sum\limits_{g=1}^{G}\left(\frac{\sum\limits_{k\in S_g} d_k p_k}{\sum\limits_{k\in R_g} d_k p_k}\sum\limits_{k\in R_g} d_k y_k\right)}{\sum\limits_{g=1}^{G}\left(\frac{\sum\limits_{k\in S_g} d_k p_k}{\sum\limits_{k\in R_g} d_k p_k}\sum\limits_{k\in R_g} d_k\right)} = \frac{\sum\limits_{g=1}^{G}\left(\hat{N}_g \frac{\sum\limits_{k\in R_g} d_k y_k}{\sum\limits_{k\in R_g} d_k p_k}\right)}{\sum\limits_{g=1}^{G}\left(\hat{N}_g \frac{\sum\limits_{k\in R_g} d_k}{\sum\limits_{k\in R_g} d_k p_k}\right)},$$

where

$$p_k = \frac{1}{d_k}\frac{\sum_{j\in S_g} d_j}{n_g} \quad \text{for } k \in S_g \qquad (3)$$

(so that $\sum_{S_g} d_k p_k = \sum_{S_g} d_k = \hat{N}_g$).

Both $\hat{\bar{y}}_{rw}$ and $\hat{\bar{y}}_{de}$ can now be written in the form:

$$\hat{\bar{y}}_{S,\mathbf{q}} = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k\in R_g} d_k y_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k q_k}\right)}. \qquad (4)$$

For the reweighted-expansion estimator, all $q_k = 1$, while for the double-expansion estimator, $q_k = p_k$ as defined by equation (3).

We will soon have use of the following for our two estimators:

$$\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_S = \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k\in R_g} d_k e_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k\in R_g} d_k}{\sum_{k\in R_g} d_k q_k}\right)} \approx \frac{\sum_{g=1}^{G} \left(\hat{N}_g \frac{\sum_{k\in R_g} d_k e_k}{\sum_{k\in R_g} d_k q_k}\right)}{\sum_{g=1}^{G} \hat{N}_g}, \qquad (5)$$

where $e_k = y_k - \bar{y}_S$. Equation (5) holds exactly when all $q_k = 1$. When $q_k = p_k$, the near equality depends on the r_g being sufficiently large and other mild conditions.

Now assume the following response model holds: Each element k in a group has an equal, positive probability of response that does not vary with π_k or with y_k. That is to say, the response indicator ρ_k, which is 1 when k responds if sampled and is 0 otherwise, is a Bernoulli random variable with a common mean in U_g regardless of the values of π_k and y_k.

By treating unit response as a second phase of probability sampling in this way, the added variance/mean-squared-error due to nonresponse given the original sample and the r_g for both estimators can be expressed as

$$A_{\mathbf{q}} = E_\rho[(\hat{\bar{y}}_{S,\mathbf{q}} - \bar{y}_s)^2 \mid S, \{r_g\}]$$
$$\approx \frac{\sum_{g=1}^{G} \hat{N}_g^{\,2}\, \text{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} | S_g, r_g)}{\left(\sum_{g=1}^{G} \hat{N}_g\right)^2}, \qquad (6)$$

where $\hat{\bar{e}}_{S_g,\mathbf{q}} = \hat{\bar{y}}_{S_g,\mathbf{q}} - \bar{y}_S$, $\bar{e}_{S_g} = \bar{y}_{S_g} - \bar{y}_S$, and

$$\text{Var}_\rho(\hat{\bar{e}}_{S_g,\mathbf{q}} \mid S_g, r_g)$$
$$\approx \left(\frac{n_g}{r_g} - 1\right) \frac{\sum_{k\in S_g} d_k^{\,2} (e_k - q_k \bar{e}_{S_g})^2}{\left(\sum_{k\in S_g} d_k q_k\right)^2}$$
$$= \left(\frac{n_g}{r_g} - 1\right) \frac{\sum_{k\in S_g} d_k^{\,2} ([y_k - \bar{y}_S] - q_k[\bar{y}_{S_g} - \bar{y}_S])^2}{\left(\sum_{k\in S_g} d_k q_k\right)^2}, \qquad (7)$$

under mild conditions on the population and original sampling design we assume to hold, including (again) that the r_g are sufficiently large. These conditions make both estimators nearly unbiased under quasi-probability sampling theory (probability theory augmented with a response model) and render the distinction between large-sample variance and mean squared error moot. Quasi-probability sampling theory is also known as “quasi-design-based” and “quasi-randomization-based” sampling theory.

Looking at equations (6) and (7), we see that at one extreme $\hat{\bar{y}}_{rw}$ has an added variance due to nonresponse of (approximately) zero when all the originally sampled y_k in a group are equal, while at the other $\hat{\bar{y}}_{de}$ has an added variance of zero when all the originally sampled $d_k e_k$ (or, put another way, the $d_k[y_k - \bar{y}_S]$) in a group are equal.

Heuristically, the reweighted-expansion estimator is more efficient than the double-expansion estimator when $\bar{e}_{S_g}$ is a better predictor of e_k for $k \in S_g$ than $p_k \bar{e}_{S_g}$. Thus, when the groups were constructed as advised in Little and Vartivarian (2003) and earlier in Little (1986) so that the y_k in a group are homogeneous (as opposed to the $d_k[y_k - \bar{y}_S]$ being homogeneous), then the reweighted-expansion estimated computed with the f_g will usually be more efficient than the double-expansion estimator computed with the n_g / r_g.

The heuristic observation can be formalized with an alternative justification for using the reweighted-expansion estimator. Suppose the following prediction model holds: Each y_k in U_g is a random variable with common mean, μ_g, regardless of π_k and ρ_k. Then $\hat{\bar{y}}_{rw}$ is nearly unbiased under mild conditions with respect to the combination of the original sampling mechanism (which treats the d_k as random, where $d_k = 0$ for $k \neq S$) the prediction model (which treats the y_k as random). That is to say, $\text{E}_d[\text{E}_y(\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] \approx 0$, since the double expectation of both $\hat{\bar{y}}_{rw}$ and $\bar{y}_U$ are nearly $\sum^G N_g \mu_g / \sum^G N_g$. This combined unbiasedness is exact when the design is such that $\sum_S d_k \equiv N$. Stratified, simple random sampling is an example of such a design. Unstratified sampling with unequal probabilities and many multistage designs are not.

It is not hard to see that $\hat{\bar{y}}_{rw}$ is also exactly unbiased with respect to this double expectation (*i.e.*, $\mathrm{E}_d\,[\mathrm{E}_y\,(\hat{\bar{y}}_{rw} - \bar{y}_U \mid S)] = 0$) when all the μ_g are equal. In fact, the prediction-model expectation of both $\hat{\bar{y}}_{rw}$ and $\hat{\bar{y}}_{de}$ equals this common mean, as does the prediction-model expectation of an estimator without any adjustment for unit nonresponse, that is, with the f_g in $\hat{\bar{y}}_{rw}$ replaced by 1. The advantage of $\hat{\bar{y}}_{rw}$ over $\hat{\bar{y}}_{de}$ under the prediction model obtains only when the μ_g vary, that is, when the survey variable has a different prediction mean across the groups.

Notice that if *either* the response model or the prediction model holds, then the reweighted-expansion estimator is nearly unbiased in some sense (*i.e.*, under the combination of the original design and the response model or under the original design and the prediction model). This property has been called "double protection" against nonresponse bias. See, for example, Bang and Robins (2005).

3. Concluding remarks

In this note, we discussed two distinct types of models. We stressed a response model, which treats the response indicators, ρ_k, as a Bernoulli random variable within each group but with unknown parameters. We also described a prediction model, which treats the survey values, y_k, as random variables with unknown means that could vary across groups but not within them.

As part of the response model, we assumed that within a group, the ρ_k do not depend on the y_k. Analogously as part of the prediction model, we assumed that within a group, the y_k do not depend on the ρ_k. When both ρ_k and y_k are treated as random variables the former assumption, that nonrespondents are *missing at random*, is equivalent to the latter assumption, that the response mechanism is *ignorable* (see, for example, Little and Rubin 1987). It should be understood, however, that the y_k need not be treated as random variables under the response model and the ρ_k need not be treated as random variables under the prediction model. The two concepts (missingness at random and ignorable nonresponse) may be equivalent in some sense but they are not identical.

The heart of Little and Vartivarian (2003) is a series of simulations featuring a binary survey variable, two potential response groups, and two original selection probabilities. Both the survey variable and response indicators are generated under five models. The expected value of each is a function of, 1, the response group only, 2, the selection probability only, 3, neither, or, 4 and 5, one of two equal combinations of response group and selection probability. This produces 25 scenarios, of which 10 are of primary interest to us. Those are the ones in which the survey variable is a function either of only the response group or of neither the response group nor the selection probability.

As our theory predicts when the survey variable is a function of neither the response group nor the selection probability, both the reweighted- and double-expansion estimators have empirical biases near zero (Table 5 in Little and Vartivarian) because both are nearly unbiased under the combination of the original sampling design and a valid prediction model: all population elements have the same mean. When the survey variable is a function of the response group *and* the response indicator is wholly or partly a function of the selection probability, only the reweighted-expansion estimator is nearly unbiased empirically since only it is unbiased under the combination of the original sampling design and a valid prediction model. As a result, $\hat{\bar{y}}_{rw}$ also has less empirical root mean squared error and significantly less average absolute error as an estimator for $\bar{y}_S$ (Tables 4 and 6 in Little and Vartivarian, respectively; the significance test treats the mean value across the simulations of $|\hat{\bar{y}}_{rw} - \bar{y}_S| - |\hat{\bar{y}}_{de} - \bar{y}_S|$ as asymptotically normal).

When both the survey variable and response indicators are functions of the response group only, the reweighted-expansion estimator has slightly less empirical root mean squared error and average absolute error than the double-expansion estimator but the latter is not significant.

It should not surprise us that the reduction in empirical root mean squared error is modest. The contribution to the variance from nonresponse under the response model mechanism expressed in equations (6) and (7) is conditioned on the original sample (technically, the contribution of non-response to the total quasi-probability variance of $\hat{\bar{y}}_{S,\mathbf{q}}$ is the expectation of $A_{\mathbf{q}}$ in equation (6) under the original sampling mechanism). In applications where the response rates are relatively large (in the simulations they averaged 0.5), this contribution can be dominated by the probability-sampling variance/mean squared error of the full-sample estimator, $\hat{\bar{y}}_U$.

Two warnings are in order. The respondent sample size within each group must be sufficiently large for the reweighted-expansion estimator to nearly unbiased under quasi-probability theory. For the double-expansion estimator, each r_g need only be positive. Moreover, that the reweighted-expansion estimator is doubly protected against nonresponse bias is only helpful when either the assumed response or prediction model is correct. If *both* the response probabilities and survey values vary with the design weights, then the reweighted-expansion estimator can be meaningfully biased. Despite the slant taken in this note, that is the take-away message Little and Vartivarian (2003) intended, and it cannot be disputed.

Acknowledgements

I thank the associate editor and two referees whose careful reading of earlier versions of the manuscript greatly improved the quality and accuracy of the resulting work. Any remaining errors in the text are wholly due to the author.

References

Bang, H., and Robins, J. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61, 962-972.

Fuller, W. (2009). *Sampling Statistics*, Hoboken, New Jersey: Wiley.

Kim, J.K., Navarro, A. and Fuller, W. (2006). Replication variance estimation for two-phase stratified sampling. *Journal of the American Statistical Association*, 101, 312-320.

Little, R. (1986). Survey nonresponse adjustments. *International Statistical Review*, 54, 139-157.

Little, R., and Rubin, D. (1987). *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, Inc.

Little, R., and Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, 22, 1589-1599.

Lohr, S (2009). *Sampling: Design and Analysis*, *Second Edition*, Boston: Brooks/Cole.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 27, No. 4, 2011

All inquires about submissions and subscriptions should be directed to jos@scb.se

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents
Volume 28, No. 1, 2012

All inquires about submissions and subscriptions should be directed to jos@scb.se

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 39, No. 4, December/décembre 2011

Errata

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 40, No. 1, March/mars 2012