Catalogue no. 12-001-X

Survey Methodology

December 2010

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. 12-001-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, Catalogue no. 12-001-X, 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

December 2010

Published by authority of the Minister responsible for Statistics Canada

December 2010

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

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

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 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 36, Number 2, December 2010

Contents

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, December 2010
Vol. 36, No. 2, p. 121
Statistics Canada, Catalogue No. 12-001-X

Waksberg Invited Paper Series

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg, who has made many important contributions to survey methodology. Each year a prominent survey researcher is chosen to author an article as part of the Waksberg Invited Paper Series. The paper reviews the development and current state of a significant topic within the field of survey methodology, and reflects the mixture of theory and practice that characterized Waksberg's work.

Please see the announcements at the end of the Journal for information about the nomination and selection process of the 2012 Waksberg Award.

This issue of *Survey Methodology* opens with the tenth paper of the Waksberg Invited Paper Series. The editorial board would like to thank the members of the selection committee Leyla Mohadjer (Chair), Daniel Kasprzyk, Elisabeth A. Martin and Wayne Fuller for having selected Ivan P. Fellegi as the author of this year's Waksberg Award paper.

2010 Waksberg Invited Paper

Author: Ivan P. Fellegi

Ivan P. Fellegi is Chief Statistician of Canada Emeritus at Statistics Canada. He was the Chief Statistician of Canada from 1985 to 2008, and it was during that period that Statistics Canada was ranked by *The Economist* as the best statistical office in the world. Dr. Fellegi contributed significantly both to survey methodology and to the effective management of a large organization during his long career at Statistics Canada.

He has published extensively on statistical methods, on the social and economic applications of statistics and on the successful management of statistical agencies. Some of his methodology papers have become landmarks: topics covered include sample design, edit and imputation, record linkage, and the analysis of survey data. He has actively participated on several committees: he was chair, Conference of European Statisticians of the United Nations Economic Commission for Europe (1993-97); Chair of the Committee on Statistics of the Organisation for Economic Cooperation and Development (2004-2008); past President of the International Statistical Institute, the International Association of Survey Statisticians, and the Statistical Society of Canada; and past chair of the Board of Governors, Carleton University (1995-97). He has a long list of honours that include: Officer of the Order of Canada; recipient of the Outstanding Achievement Award of the Public Service of Canada; the Order of Merit of the Hungarian Republic; the Career Achievement Award of the Canadian Policy Research Initiative, La Médaille de la ville de Paris, Member of the Hungarian Academy of Sciences, Gold Medal of the Statistical Society of Canada and the Robert Schuman medal of the European Community. He is the recipient of Honorary Doctorates from Université de Montréal, Université du Québec (Institut national de la recherche scientifique), Simon Fraser University, McMaster University, Carleton University, and the University of Ottawa. He is an Honorary Member of the International Statistical Institute, Honorary Fellow of the Royal Statistical Society.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 123-130
Statistics Canada, Catalogue No. 12-001-X

The organisation of statistical methodology and methodological research in national statistical offices

Ivan P. Fellegi [1]

Abstract

The paper explores and assesses the approaches used by statistical offices to ensure effective methodological input into their statistical practice. The tension between independence and relevance is a common theme: generally, methodologists have to work closely with the rest of the statistical organisation for their work to be relevant; but they also need to have a degree of independence to question the use of existing methods and to lead the introduction of new ones where needed. And, of course, there is a need for an effective research program which, on the one hand, has a degree of independence needed by any research program, but which, on the other hand, is sufficiently connected so that its work is both motivated by and feeds back into the daily work of the statistical office. The paper explores alternative modalities of organisation; leadership; planning and funding; the role of project teams; career development; external advisory committees; interaction with the academic community; and research.

Key Words: Methodology; Official statistics; Statistical organisation; Research; Relevance; Independence.

1. Introduction

It is a great honour to accept an award named after Joe Waksberg. Joe has been a close personal friend, as well a good friend of Statistics Canada.

I came to know Joe during his latter years in the Bureau of the Census when Morris Hansen asked me to become a member of what was then a most imposing methodology advisory committee of the Bureau chaired by Bill Cochran. Subsequently, in the late 1970s, when Statistics Canada had serious problems of image and of internal management, Statistics Canada asked a group of prominent statisticians to review what was wrong. At my recommendation, Joe was one of the three wise men asked to take part (the others being Richard Ruggles and the chairman, Claus Moser). Joe immediately agreed and in his inimitable low-key manner made invaluable contributions to Statistics Canada; the very helpful basic message being that while we had serious management problems, there was nothing much wrong with our methodology.

A few years ago the Census Bureau honoured me by asking to give one of their annual "wise elders" lectures. While I objected strongly on the grounds that I neither considered myself "wise", nor "elder", in the end I accepted their kind invitation. With typical grace, Joe took the time to show up for my talk, even though he was well into the middle of his eighties but still very busy as chairman of the board of WESTAT. We had a really good chat; and that was the last time I saw him. What a career; what a life!

So it is not only a professional honour to accept the Waksberg Award, but also a personal pleasure to be associated with Joe one more time.

I was told that generally the recipients of the Waksberg Award give an overview of an area of methodology. But while, as you know, I did spend the first half of my career as a methodologist, I stopped being a practitioner some decades ago – although I am still an ardent advocate (see Fellegi 2004). So I thought I would join the first half of my career – methodology – to the second half – management of statistical offices. I shall therefore, talk about the lessons I learnt about the organisation of applied methodological work and methodology research in national statistical offices; what works well and what less so (I assume that the basic conditions for an effective methodology function exist: there is a supply of trained statisticians in the country, the statistical office has a functioning infrastructure, salaries, if they are not competitive, are at least within sight of what is offered in the private sector, and so on).

I have two overall themes. Managing the tension between independence and relevance is one of them: generally, methodologists must work closely with the rest of the statistical organisation for their work to be relevant. Indeed, they must strive to serve the objectives of external clients, represented inside the office by subject matter experts. However, for them to be effective they must enjoy the necessary independence to question the use of existing methods, and to champion new ones if they believe they could reduce costs or increase statistical quality.

But the effectiveness of methodology also depends on a strong methodology research capacity which, on the one hand, has the necessary independence needed by any research program, but which, on the other hand, is sufficiently connected to on-going work so that it is both motivated by and feeds back into the daily practice of the

1. Ivan P. Fellegi is Chief Statistician Emeritus at Statistics Canada. E-mail: ivan.p.fellegi@statcan.gc.ca.

statistical office. The organisation of methodology research will be my second them.

But first I want to define what I mean in the present context by the terms *methodology*, *relevance* and *independence*.

2. Some definitions

Methodology

The unique service performed by *methodology* is to maximise statistical quality given an imposed budget (or conversely). They do so through the application of statistical practice that is either based on statistical theory or on organized empirical observation. In other words methodologists are wizards of the relevant statistical theories; but also of "organised empirical observation" where formal theory abandons us. By organised empirical evidence I mean designed experiments or analytically assessed experience. So I am including all organized knowledge about the use of methods and approaches that result in the objective of maximising quality within a budget – or conversely, minimising the budget needed to achieve a stated quality level.

This would include such things as sample design, estimation, data editing, imputation, exploitation of administrative data, record linkage, seasonal adjustment, questionnaire design, measurement of accuracy and quality assurance of censuses and surveys, the use of experimental designs, and so on.

Methodologists are predominantly mathematical statisticians and they work on the applied end of their subject. Because of the interdisciplinary nature of official statistics they interact with survey managers, experts in data collection, IT personnel, geographers, sociologists, economists, *etc*.

Relevance

Methodology is *relevant* if the day to day practice of the statistical office is actually based on sound methodology. A major issue in the organization of methodology is how to balance the intrinsically service nature of methodology against the need for the function to provide strong and effective guidance. Much of the paper will deal with all those arrangements needed to ensure the objective of relevance.

In the case of methodological *research*, relevance means that the research is both motivated by and informs applied work.

Independence

The notion of independence of methodology means the ability to provide sound methodological guidance to projects, irrespective of the hierarchical arrangement of line organisations that *can be debated but not ignored*; and that this debate is based on evidence, not authority. So my definition of independence is not that methodologists should be able to "do their own thing" but rather that they should have an authoritative voice.

Independence is frequently contrasted with relevance. Since relevance is about embedding methodology into practice, this is often attempted by building methodological services right into the fabric of subject matter organisations. By contrast, independence is thought to be enhanced by giving methodologists their own organisation(s). In this sense, therefore, there is a tension between the two. However, I would argue that relevance cannot be achieved if methodological guidance is ignored, so appropriate arrangements to ensure independence are necessary for relevance.

Independence of *methodological research* is different: it is generally meant to refer to an environment in which researchers have predominant say in the choice of their topics. Clearly, providing researchers with such an environment does create a permanent tension with the need to be relevant at all times, particularly when it is not at all obvious in the short term where the relevance lies.

In my discussion of how to balance relevance and independence of both the applied methodology function and of methodology research I will describe not only organisational arrangements, but a wide variety of tools and arrangements that should be considered in the pursuit of this objective. I shall use Statistics Canada as a concrete illustration. What I wish to emphasize is that the issue is much more complicated than what the terms "centralisation" and "decentralisation" denote for whichever of these basic organisational arrangements is adopted, many additional tools are needed to offset their disadvantages while maintaining their intrinsic advantages. Indeed, I have organised the rest of the paper around a discussion of the main tools (in choosing these tools for discussion, I borrowed from the paper by Brackstone 1997) involved under the following headings:

- Organisation;
- Leadership;
- Planning and funding;
- Project teams;
- Career development;
- Advisory Committees;
- Interaction with the academic community; and
- Research.

3. Organisation

General thoughts

National statistical offices differ in the way they organise their methodology functions. In some it is distributed to

individual parts of the agency, each responsible for a given subject (*e.g.*, labour). In other agencies decentralisation is only partial, *e.g.*, to broader subject matter areas (such as demography or business statistics). The US Bureau of the Census, for example has largely decentralised its methodology function. By contrast, Statistics Canada and the Australian Bureau of Statistics have largely centralised it. Many factors influence the organizational choice. For example, in France and in India where all professionals share similar background in statistics and are largely recruited from a single teaching institution the accent is obviously on centralizing training and to a lesser extent research.

The traditional arguments are that decentralisation favours relevance and centralisation favours independence. However, the aim should be to have both. That being the case the question is how we can enhance independence in the case of decentralised methodology organisations, and relevance in the case of centralised ones.

Decentralisation, while potentially serving to underscore relevance, has some built-in disadvantages. Since each unit to which methodology is decentralized is necessarily smaller than it would be in more centralized options, it is less likely to facilitate specialisation and research. It is also less likely to encourage cross-fertilisation by methodologists working on other issues. Also, since the line organisations to which methodology is decentralised are typically not headed by methodologists, this model tends to result in lower hierarchical positions for the heads of the decentralised methodology units. In case of "conflicts" – and these will be inevitable because of different perceptions of priority, cost, quality and so on - other things being equal it will be more difficult for methodologists to defend their professional advice. If left without a counterweight, this kind of organization could get out of balance.

A critical counterweight could be a "chief methodologist" who reports directly to the head of the statistical office and inevitably is called upon to play an important role in long term planning and resource allocation. The "Chief Methodologist" could have his hand strengthened if given direct line responsibility for a strong research and development function which could serve as the "intellectual home base" for the decentralised methodology staff.

Project teams, brought together for large developments, are another important tool to enhance independence in the case of centralised organisations. Such projects – which if at all significant are necessarily multi-disciplinary – are carried out by ad hoc project teams which operate off-line from the agency's line organization. The organization of project teams is a matter to which Statistics Canada devoted considerable attention and it has been refined over time. Among its elements there is the feature that whenever professional disputes within the teams arise and the team believes that their solution requires outside intervention, the dispute is referred to a senior group of which someone from the staff of the "chief methodologist" is a member (this is automatically the case if the methodologist comes from a centralised group). It is this senior steering group that can contribute to protecting independence.

Consideration might also be given to providing some additional tools for the "chief methodologist": he could be authorised and funded to develop a strong methodology training program; he could be given a strong role in the allocation and career development of the methodology staff; he could be supported by a strong external advisory committee; and so on. These features recognize that the role of "chief methodologist" is particularly delicate and could become more so if his place in the hierarchy were dependent on the size of the staff he controls directly without provision – as there is in some countries – to have his level of access and place in the ladder depend on his personal prestige rather than on the size or level of supporting staff.

Centralisation: the Statistics Canada model

Many years ago Statistics Canada opted for the centralised model (see Fellegi 1996) and that option was never seriously challenged (it was challenged for a brief period of time in the late seventies but in concrete terms the challenge did not get anywhere), and put in place a number of practices designed to reduce the threat that centralisation might result in diminished relevance.

1. Project teams: These are inter-disciplinary and include as a matter of course a methodologist but they are headed by a project manager whose association with the project is subject matter and who is likely to assume operational responsibility for the completed project.
2. Funding: much of the funding for the methodology function is controlled by the rest of Statistics Canada. Program areas (within limits that I will describe further) are free to spend their money on buying methodology services or not so long as they do not fall foul of the agency's quality norms and accepted standards. With their budget largely on the line year after year, this accountability means that it is very much in the interest of methodologists to be responsive to the needs of the Agency' s Programs.
3. Organisation of the methodology function: it largely parallels the organisation of Statistics Canada. There are four methodology divisions: three of them provide methodology input to three different areas of the agency, while the fourth is devoted to research. In

fact, the three applied methodology divisions are themselves organised by subject matter in parallel with the manner in which the bureau is organised (regular rotation of methodology staff ensures broad development opportunities for methodologists).

4. Co-location of methodology staff: methodologists are occasionally physically moved to the offices of the subject matter areas whose surveys they help to design. This is an additional measure taken to ensure that they focus on the right issues.
5. Finally, as a matter of sound practice, methodologists conduct – and follow up on the results of – client satisfaction surveys which provide feedback on all aspects of their performance and first and foremost on the relevance thereof.

4. Leadership

General thoughts

Leadership is crucial. The leader of the methodology function, in addition to a proper academic background and a great deal of experience in methodology, must possess a strategic vision and a personality that inspires confidence. This is an intrinsically difficult function. In the overwhelming majority of offices operational and subject matter considerations are the ones that receive the most attention. In such an environment an authoritative voice for methodology is needed to ensure adequate resources for the methodology function itself, but even more importantly to lead the *entire agency* in directions that are technically sound, and conversely to hold back initiatives that cannot be supported by sound methodology. "Soundly based" involves more than good survey design that uses the best available current knowledge. It also includes the notion of strategic planning of research, experiments and pilot surveys so as to improve the likelihood that whatever knowledge will be needed in the future will be available. For the opinions of methodologists to make a proper impact they must be supported by a leader whose unchallenged personal competence is combined with a seat at the statistical agency's most senior table

If methodologists do not belong to a central organization within the statistical agency it is all the more important for their senior representative to be highly placed in the hierarchy since under a decentralized scheme he would not have direct line authority for (the bulk of) methodology resources.

Centralisation: *the Statistics Canada model*

Centralisation provides another lever to enable the leader of the methodology function to carry out his proper role as it enables him to make rational and authoritative assignments of the resources under his direction to the most strategic projects. The top advocate of sound methodology in Statistics Canada has the status of Assistant Chief Statistician (ACS) – the rank immediately below that of the Chief Statistician of Canada. In order to secure such a high position in a government bureaucracy, the line responsibility of the ACS (Methodology) includes statistical standards (classifications and central registers), as well as informatics (IT). While the position is therefore responsible for more than methodology, it is by long tradition (over 35 years) filled by someone who is a noted expert on methodology and can therefore speak at the top table authoritatively about its importance in general as well as in the context of particular projects.

5. Planning and funding

General thoughts

The effective functioning of methodology (as indeed the entire statistical office) greatly depends on the existence of a proper planning system (see Fellegi 1992 and Brackstone 1991):

- Planning is a necessary condition to ensure that resources are allocated rationally at all times.
- It also serves to mark explicitly the beginning and the end of development projects and therefore constitutes the ideal opportunity for methodology to "sign off" on the proposed design of new projects.
- Lastly, the planning system creates an opportunity for methodology to make an explicit judgement on whether a planned new venture can respect simultaneously its budgetary constraints, the agency's quality standards, and the expected maintenance bill. In fact, the planning system also provides an opportunity for all representatives of the disciplines involved in the creation of a new project (its planning or its implementation) to "sign off" as a mark of assuming professional responsibility for the adequacy of its funding or for the integrity of its functioning.

Such a planning system is essential where the main disciplines (methodology, systems development, data collection, *etc*.) are centralised for otherwise the organisations responsible cannot make provisions for the needed resources. But, for more subtle reasons, decentralised offices need it just as much: to provide an explicit forum for the leaders of methodology (and, indeed, other key disciplines), to make their input during the critical formative stages of new projects.

Centralisation: the Statistics Canada model

Every new project or major redesign is approved within Statistics Canada's planning system. In preparation for its consideration, a comprehensive budget is developed and all major disciplines which are required to contribute sign off on the appropriateness of the proposed design and operational modalities. If the project is approved, its budget is divided up and distributed to participating disciplines, including methodology. In turn, these organisations "contract" to deliver the agreed contributions within the approved budgets. A project manager oversees both progress and expenditures, with authority to reassign resources, if necessary.

The budget of the Methodology organization is composed of five distinct sources. These are designed, on the one hand, to facilitate the sound planning of the use of methodology and its thorough integration into the work of the Agency, and on the other to secure for it the needed funding.

1. The contribution of methodology to *developmental* projects is guaranteed by the planning process of Statistics Canada, as indicated above. The financial contribution to the methodology budget from these sources may vary from year to year, but there is a reasonable overall stability (facilitating the hiring and development of permanent staff). They account for almost 30 per cent of the total methodology budget. These projects typically involve major redesigns, often with significant experimentation and innovation.
2. But methodological contributions are also needed for maintenance (quality control, monitoring of various errors including variance estimation where relevant, minor design adjustments, *etc*.). For these activities there are core resources set aside and more or less permanently allocated by broad subject matter. This constitutes the second component of the methodology budget and it accounts for somewhat less than 25%. While for methodology this "on-going" work accounts for less than 25% of their workload, for Statistics Canada as a whole "on-going" work accounts for over 90% of our budget. This is because of the innovative nature of methodology work.
3. A third component comes from supplementary resources funded directly by the beneficiary subject matter divisions who, in effect, make savings from their other expenditures to avail themselves of additional methodology contributions. These supplementary funds account for a by no means negligible 20% or so of the methodology budget. The very fact that subject matter divisions consider methodology sufficiently valuable to fund methodological advice directly says a lot about the health of the relationship and of the extent to which it is valued. The funds in question are for a mixture of projects including enhancements short of a major redesign of on-going projects. They also strengthen the awareness of methodology staff of the need to remain relevant for their users. The kind of service they provide has a direct bearing on the amount of resources that are made available to them.
4. The fourth part of the methodology budget (about 20 per cent) comes from externally funded projects, typically from the budget of surveys funded by other departments. No more needs to be said about them.
5. The final part (7 per cent) is for research. This is a "block fund", meaning that a certain fixed amount of funds is allocated for the research function. The annual allocation is governed by a mechanism described below.

The intricacies of the funding mechanism and the multiplicity of funding sources are a reflection of the care exercised in the agency to balance the virtues of independence with those of relevance.

6. Project teams

General thoughts

The use of project teams in developmental projects helps to strengthen relevance without it being necessarily at the expense of independence. But project teams are not a universal panacea as everything depends on establishing appropriate checks and balances. In centralised organisations project teams, most often headed by a project manager from the sponsoring subject matter area, help to nudge the participating methodology staff to pay proper attention to the objectives and constraints of projects. Nonetheless there remains an inherent danger that the project manager will not give sufficient weight to the considered advice of methodologists.

Project teams in decentralised organisations are just as important to ensure that the views of methodologists are given appropriate weight. Here, however, the dice are clearly weighted in favour of relevance and against independence. Moreover, an exaggerated emphasis on "relevance" has its danger as well since it can lead to local optimisation. Local optimisation is a situation where surveys are optimised without regard to agency wide objectives. An example might be a situation where surveys are customised to an extent such that the introduction of important efficiencies through the use of agency-wide

standards and general systems becomes difficult (the widespread use of generalized approaches, systems and tools can be a source of considerable agency-wide efficiencies: they shorten implementation times, reduce the expenditure on both systems development and maintenance, facilitate staff rotation, *etc*. However, generalized systems might lack some features which could enhance the efficiency of any given operation. Decentralized organizations are more likely to favour such locally developed solutions in preference to agency-wide standard tools, even though the latter might lead to substantial *long-run* efficiencies).

Centralisation: the case of Statistics Canada

In Statistics Canada project teams working on major development projects are accountable and report to steering committees typically composed of the heads of the participating disciplines. A steering committee approves the broad project strategy, and serves, if needed, as a forum to which issues can be referred that could not be resolved within the team itself. In practice such appeals are rare and are restricted to cases where professional principles or truly strategic issues are involved. Steering committees ensure that issues do not get resolved within the project team on the basis of rank but rather on the basis of professional merit.

Methodologists serving on project teams carry out a dual function:

- At a strategic level, they help ensure that the overall survey design achieves the project's substantive objectives, while striking a balance between reliability, cost, timeliness and respondent burden. While striving for this balance concerns the entire project team it is the methodologists who provide the framework and techniques that must be considered in seeking the optimum balance.
- At a tactical level the methodologists provide the statistical methods and tools that are incorporated into the overall survey design: the sample design, the estimation and weighting approach, quality control, editing and imputation strategies, coverage checks, analytic methods and the like.

Project teams function best in an organisation dedicated to making decisions on the basis of merit; where everyone can pose questions and expect reasoned answers; one that is devoted to making maximum use of the expertise of everyone involved.

7. Career development

General considerations

Career development is essential for all professional groups, and it involves both formal training as well as formal and informal approaches to facilitate on-the-job learning. Methodology staff, in my view, requires special attention in this respect. The reason is that universities in general offer few, if any, courses in survey methodology (there is an increasing number of exceptions, although their numbers are still far from overwhelming. A most notable one is the Joint Program in Survey Methodology, University of Maryland. But there are also degree programs on official statistics in the UK, Ireland and New Zealand which include survey methodology). Since a thorough professional knowledge is essential for both relevance and independence, most statistical offices wanting to maintain a strong methodology staff have no alternative to having a carefully designed career development program – whether methodology is organised in a centralised or decentralised manner.

For the courses to be relevant, it is desirable that a substantial portion of courses should be taught by staff members who are themselves active practitioners. This is easier arranged in centralised organisations where the senior methodologists can not only deploy staff to do teaching (typically on a part time basis), but can also arrange suitable replacements for them in their current project work.

The broader aspects of career development are also easier arranged in centralised organisations: they can more readily manage the periodic assignment of staff to different types of survey work, attendance at scientific conferences, the provision of research opportunities to those interested in and capable of doing part-time research work, and most importantly the service of apprenticeships under more experienced methodologists.

The case of Statistics Canada

Training, not only in methodology, is emphasized by Statistics Canada (see Statistics Canada 1995). Overall, expenses on training amount to about 3% of its budget (or $15 million) on formal training – plus a great deal more on various means of career development. But, in line with the centrality of training in methodology, the percentage of methodology budget spent on it is almost twice as much (bordering on 6 per cent in the 2008-09 fiscal year).

Training is provided in formal courses within Statistics Canada's Training Institute which currently (in 2009) offers some 20 courses in methodology, ranging in level from introductory courses to graduate level material. Most courses are taught by in-house staff, occasionally university personnel, mostly from local universities, are engaged if they are interested to teach and/or help develop our staff in other ways (*e.g.*, consultation) (in the latter modality we have been particularly fortunate in having had the contributions of Professor J.N.K. Rao of Carleton University over a period of some decades).

All recruits have to take a basic six weeks course which teaches (and provides practice in) survey design, survey operations, processing and analysis. This introductory training serves a multiplicity of purposes. Since the same basic six-week course in survey work is taken by *all* new professionals, it helps early on to inculcate in everyone a basic knowledge of all that is involved in survey work ; and, even more importantly, to drive home the critical importance of inter-disciplinary team work. It is also at this stage that new recruits from other disciplines are exposed for the first time to the requirements of methodology in survey design

Career development involves much more than training. The staff, particularly at the earlier stage of their career, is regularly given opportunities to work on different types of work: demographic, socio-economic, business surveys, use of administrative records, record linkage, *etc*. Significant numbers also attend scientific conferences. For example, during the last several years some 17 per cent of the methodology staff attended various Canadian and international professional conferences per annum. Staff is also encouraged to work on research projects and publish findings in peer reviewed journals, including Statistics Canada's *Survey Methodology*. Finally, for many years now Statistics Canada has organised an international methodology symposium to which leading research personnel from around the world are invited. These symposia are, of course, open to all Statistics Canada personnel and most methodologists choose to attend them.

8. Advisory Committee

General considerations

A Methodology Advisory Committee can serve a most useful function (a) ensuring sound methodology practices, (b) integrating these practices into the daily work of statistical organisations, and (c) training staff. But the Committee can only be effective if (a) its advice is sought on significant issues of methodology and (b) there are mechanisms to ensure that the Committee's views are given due weight. I have observed Methodology Advisory Committees playing an equally useful role in a centralised office (Statistics Canada) and in a decentralised one (the Bureau of the Census in the 1960s).

The case of Statistics Canada

Statistics Canada's Methodology Advisory Committee plays a key role. There are several factors that contribute to its usefulness and standing:

- The personal standing of the Committee's members is part of the reason.
- Every significant project of Statistics Canada is referred to the Committee for advice.
- The Committee's review is facilitated by the preparation of a paper for each item of the agenda which is introduced by a brief oral presentation by staff.
- Designated members of the Committee serve as formal discussants of each item on the agenda. The discussants present their views formally. Given that most of the papers are prepared by mid-career staff, these discussions make not only a substantive contribution to the projects that are discussed, but also to the training of the staff concerned – and that of the audience.
- Meetings of the Committee are attended not only by a large number of the relevant methodologists, but also by senior personnel of the subject matter division concerned, including often the Chief Statistician as well as one or two of his assistants.
- The Committee meets regularly: twice a year, for a day and a half on each occasion.
- The Committee regularly reviews the follow-up arising from its conclusions and formal recommendations; this helps ensure that their advice is taken seriously.

9. Research

General considerations

I am taking it for granted that for this audience I do not need to spend time underscoring the intrinsic importance of research in a statistical agency. But let me stress the following points:

- Careful thought should be given to organising the research function in a manner that maximises both its relevance and the likelihood that its benefits will be successfully transmitted into daily practice. It is crucial to avoid the twin dangers of research being self-serving, or alternatively so completely task-oriented that it becomes pedestrian.
- Research needs to be adequately funded.
- In-house research needs to develop and to maintain close links with relevant extramural research.

The case of Statistics Canada

One of the four methodology divisions is formally devoted to full time research. But the research is organised in a particular manner. Even though the research budget provides for the equivalent of 22 full time research staff, the research division itself has only six full time members. The remaining budget is assigned to finance the part-time research work of some other 70 methodologists. This arrangement serves a variety of purposes. First, it contributes to the relevance of research. Secondly, it

contributes to the adoption of the results of research. And thirdly, it helps morale for while not everyone wants to do research (or is able to do so), many want to try their hand at it. And the very act of conducting some research, by those capable of it, leads to more open mindsets and a better informed practice.

We are trying to ensure that the particular projects approved for research are in line with the broad research priorities of Statistics Canada, but at the same time leave some scope for self-initiated research. We do this by establishing broad priorities each year and inviting proposals in those areas from staff. The proposals are subject to formal adjudication: the best ones are selected and staff are assigned to work on them. Senior advice and guidance is provided by the Director of the Statistical Research and Innovation Division and its small permanent staff.

The following are additional measures that help the quality of research carried out:

- The possibility of publishing papers in *Survey Methodology*, Statistics Canada's own publication, serves as an incentive. While the peer review of the articles is rigorously managed by an international editorial board, the existence of a local yet prestigious outlet for methodology research represents a visible commitment by senior management.
- We regularly co-author papers with well known external research personnel (both Canadian and non-Canadian).
- We hold regular methodology interchanges with methodology staff in the US Bureaus of the Census and of Labour Statistics.
- We participate actively in Canadian, American and international statistical organisations.

10. Concluding comments

As indicated in the introduction, the bulk of the paper was devoted to the tools that should be considered by statistical offices in establishing and supporting the methodology function and the associated research, tools that in appropriate combination can enhance both the professional independence as well as the relevance of the function. I want to emphasise, however, that this is not a cook book. More important than all the tools is the environment: whether the statistical office welcomes questioning and ensures that substantive questions are answered in substance; whether change is intrinsically frowned upon; whether it fosters collegiality; whether intelligent risk taking is encouraged or frowned upon; whether experiments are welcomed, assessed on their merits, and acted upon. These are the attributes that come from the top leadership of the statistical office and tools cannot substitute for them. Under the wrong leadership the best methodology staff (or, indeed, the best statistical office itself) will wither. But the contrary is not true: it is essential to have a careful understanding of the subtle balances advocated in this paper, as well as a careful deployment of the tools that give them effect. And even then, only a long term strategy can succeed.

I am completely certain that Joe would agree with my conclusion (see Waksberg 1998).

References

Brackstone, G.J. (1991). Shaping statistical services to satisfy user needs. *Statistical Journal of the United Nations Economic Commission for Europe*, 8, 3/4, 243-258.

Brackstone, G. (1997). Organization of a survey methodology service. *Enquêtes et sondages* : *Méthodes*, *modèles*, *applications*, *nouvelles approches*, (Eds., G. Brossier and A.-M. Dussaix), Rennes, France, June 19-20, 3, 118-134.

Fellegi, I.P. (1991). Maintaining public confidence in official statistics. *Journal of the Royal Statistical Society*, Series A (Statistics in Society), 154, Part 1, 1-6.

Fellegi, I.P. (1992). Planning and Priority Setting - the Canadian Experience. In *Statistics in the Democratic Process at the End of the 20th Century*, (Eds., Hölder, Malaguerra and Vukovich); Anniversary publication for the 40th Plenary Session of the Conference of European Statisticians. Published by the Federal Statistical Office, Wiesbaden, Federal Republic of Germany.

Fellegi, I.P. (1996). Characteristics of an effective statistical system. *International Statistical Review*, 64, 2, 165-199.

Fellegi, I.P. (2004). Maintaining the credibility of official statistics. *Statistical Journal of the United Nations*, ECE 21, 191-198.

Statistics Canada (1995). Training and Development at Statistics Canada. Statistics Canada Training Institute, March 1995.

Waksberg, J. (1998). The Hansen Era: Statistical research and its implementation at the U.S. Census Bureau, 1940-1970. *Journal of Official Statistics*, 14, 2, 119-135.

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 131-144
Statistics Canada, Catalogue No. 12-001-X

Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias

Carl-Erik Särndal and Sixten Lundström [1]

Abstract

This article develops computational tools, called indicators, for judging the effectiveness of the auxiliary information used to control nonresponse bias in survey estimates, obtained in this article by calibration. This work is motivated by the survey environment in a number of countries, notably in northern Europe, where many potential auxiliary variables are derived from reliable administrative registers for household and individuals. Many auxiliary vectors can be composed. There is a need to compare these vectors to assess their potential for reducing bias. The indicators in this article are designed to meet that need. They are used in surveys at Statistics Sweden. General survey conditions are considered: There is probability sampling from the finite population, by an arbitrary sampling design; nonresponse occurs. The probability of inclusion in the sample is known for each population unit; the probability of response is unknown, causing bias. The study variable (the *y*-variable) is observed for the set of respondents only. No matter what auxiliary vector is used in a calibration estimator (or in any other estimation method), a residual bias will always remain. The choice of a "best possible" auxiliary vector is guided by the indicators proposed in the article. Their background and computational features are described in the early sections of the article. Their theoretical background is explained. The concluding sections are devoted to empirical studies. One of these illustrates the selection of auxiliary variables in a survey at Statistics Sweden. A second empirical illustration is a simulation with a constructed finite population; a number of potential auxiliary vectors are ranked in order of preference with the aid of the indicators.

Key Words: Calibration weighting; Nonresponse adjustment; Nonresponse bias; Auxiliary variables; Bias indicator.

1. Introduction

Large nonresponse is typical of many surveys today. This creates a need for techniques for reducing as much as possible the nonresponse bias in the estimates. Powerful auxiliary information is needed. Administrative data files are a source of such information. The Scandinavian countries and some other European countries, notably the Netherlands, are in an advantageous position. Many potential auxiliary variables (called *x*-variables) can be taken from high quality administrative registers where auxiliary variable values are specified for the entire population. Variables measuring aspects of the data collection is another useful type of auxiliary data. Effective action can be taken to control nonresponse bias. Beyond sampling design, *design for estimation* becomes, in these countries, an important component of the total design. Statistics Sweden has devoted considerable recourses to the development of techniques for selecting the best auxiliary variables.

Many articles discuss weighting in surveys with nonresponse and the selection of "best auxiliary variables". Examples include Eltinge and Yansaneh (1997), Kalton and Flores-Cervantes (2003), and Thomsen, Kleven, Wang and Zhang (2006). Weighting in panel surveys with attrition receives special attention in, for example, Rizzo, Kalton and Brick (1996), who suggest that "the choice of auxiliary variables is an important one, and probably more important than the choice of the weighting methodology". The review by Kalton and Flores-Cervantes (2003) provides many references to earlier work. As in this paper, a calibration approach to nonresponse weighting is favoured in Deville (2002) and Kott (2006).

Some earlier methods are special cases of the outlook in this article, which is based on a systematic use of auxiliary information by calibration at two levels. Recently the search for efficient weighting has emphasized two directions: (i) to provide a more general setting than the popular but limited cell weighting techniques, and (ii) to quantify the search for auxiliary variables with the aid of computable indicators. Särndal and Lundström (2005, 2008) propose such indicators, while Schouten (2007) uses a different perspective to motivate an indicator. An article of related interest is Schouten, Cobben and Bethlehem (2009).

This content of this article has four parts: The general background for estimation with nonresponse is stated in Sections 2 to 4. Indicators for preference ranking of **x**-vectors are presented in Sections 5 and 6, and the computational aspects are discussed. The linear algebra derivations behind the indicators is presented in Sections 7 and 8. The two concluding Sections 9 and 10 present two empirical illustrations. The first (Section 9) uses real data from a large survey at Statistics Sweden. The second (Section 10) reports a simulation carried out on a constructed finite population.

1. Carl-Erik Särndal, Professor and Sixten Lundström, Senior Methodological Advisor, Statistics Sweden. E-mail: carl.sarndal@scb.se.

2. Calibration estimators for a survey with nonresponse

A probability sample s is drawn from the population $U = \{1, 2, ..., k, ..., N\}$. The sampling design gives unit k the known inclusion probability $\pi_k = \Pr(k \in s) > 0$ and the known design weight $d_k = 1/\pi_k$. Nonresponse occurs. The response set r is a subset of s; how it was generated is unknown. We assume $r \subset s \subset U$, and r non-empty. The (design weighted) response rate is

$$P = \frac{\sum_r d_k}{\sum_s d_k} \tag{2.1}$$

(if A is a set of units, $A \subseteq U$, a sum $\sum_{k \in A}$ will be written as $\sum_A$). Ordinarily a survey has many study variables. A typical one, whether continuous or categorical, is denoted y. Its value for unit k is y_k, recorded for $k \in r$, not available for $k \in U - r$. We seek to estimate the population y-total, $Y = \sum_U y_k$. Many parameters of interest in the finite population are functions of several totals, but we can focus on one such total.

The auxiliary information is of two kinds. To these correspond two vector types, $\mathbf{x}_k^*$ and $\mathbf{x}_k^\circ$. *Population auxiliary information* is transmitted by $\mathbf{x}_k^*$, a vector value known for every $k \in U$. Thus $\sum_U \mathbf{x}_k^*$ is a known population total. Alternatively, we allow that $\sum_U \mathbf{x}_k^*$ is imported from an exterior source and that $\mathbf{x}_k^*$ is a known (observed) vector value for every $k \in s$. *Sample auxiliary information* is transmitted by $\mathbf{x}_k^\circ$, a vector value known (observed) for every $k \in s$; the total $\sum_U \mathbf{x}_k^\circ$ is unknown but is estimated without bias by $\sum_s d_k \mathbf{x}_k^\circ$. The auxiliary vector value combining the two types is denoted $\mathbf{x}_k$. This vector and the associated information is

$$\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}; \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}. \tag{2.2}$$

Tied to the k^{th} unit is the vector $(y_k, \mathbf{x}_k, \pi_k)$. Here, π_k is known for all $k \in U$, y_k for all $k \in r$, the component $\mathbf{x}_k^*$ of $\mathbf{x}_k$ carries population information, the component $\mathbf{x}_k^\circ$ of $\mathbf{x}_k$ carries sample information.

Many **x**-vectors can be formed with the aid of variables from administrative registers, survey process data or other sources. Among all the vectors at our disposal, we wish to identify the one most likely to reduce the nonresponse bias, if not to zero, so at least to a near-zero value.

We consider vectors having the property that there exists a constant non-null vector $\boldsymbol{\mu}$ such that

$$\boldsymbol{\mu}' \mathbf{x}_k = 1 \text{ for all } k \in U \tag{2.3}$$

"Constant" means that $\boldsymbol{\mu} \neq \mathbf{0}$ does not depend on k, nor on s or r. Condition (2.3) simplifies the mathematical derivations and does not severely restrict $\mathbf{x}_k$. Most **x**-vectors useful in practice are in fact covered. Examples include: (1) $\mathbf{x}_k = (1, x_k)'$, where x_k is the value for unit k of a continuous auxiliary variable x; (2) the vector representing a categorical x-variable with J mutually exclusive and exhaustive classes, $\mathbf{x}_k = \boldsymbol{\gamma}_k = (\gamma_{1k}, ..., \gamma_{jk}, ..., \gamma_{Jk})'$, where $\gamma_{jk} = 1$ if k belongs to group j, and $\gamma_{jk} = 0$ if not, $j = 1, 2, \ldots, J$; (3) the vector $\mathbf{x}_k$ used to codify two categorical variables, the dimension of $\mathbf{x}_k$ being $J_1 + J_2 - 1$, where J_1 and J_2 are the respective number of classes, and the 'minus-one' is to avoid a singularity in the computation of weights calibrated to the two arrays of marginal counts; (4) the extension of (3) to more than two categorical variables. Vectors of the type (3) and (4) are especially important in statistics production in statistical agencies (the choice $\mathbf{x}_k = x_k$, not covered by (2.3), leads to the nonresponse ratio estimator, known to be a usually poor choice for controlling nonresponse bias, compared with $\mathbf{x}_k = (1, x_k)'$, so excluding the ratio estimator is no great loss).

The calibration estimator of $Y = \sum_U y_k$, computed on the data y_k for $k \in r$, is

$$\hat{Y}_{\text{CAL}} = \sum_r w_k y_k \tag{2.4}$$

with $w_k = d_k \{1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k\}$. The weights w_k are calibrated on both kinds of information: $\sum_r w_k \mathbf{x}_k = \mathbf{X}$, which implies $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$ and $\sum_r w_k \mathbf{x}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$. We assume throughout that the symmetric matrix $\sum_r d_k \mathbf{x}_k \mathbf{x}_k'$ is nonsingular (for computational reasons, it is prudent to impose a stronger requirement: The matrix should not be ill-conditioned, or near-singular). In view of (2.3), we have $\hat{Y}_{\text{CAL}} = \sum_r w_k y_k$ with weights $w_k = d_k v_k$ where $v_k = \mathbf{X}' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k$. The weights satisfy $\sum_r d_k v_k \mathbf{x}_k = \mathbf{X}$, where $\mathbf{X}$ has one or both of the components in (2.2).

A closely related calibration estimator is based on the same two-tiered vector $\mathbf{x}_k$ but with calibration only to the sample level:

$$\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k \tag{2.5}$$

where

$$m_k = \left(\sum_s d_k \mathbf{x}_k \right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \mathbf{x}_k. \tag{2.6}$$

The calibration equation then reads $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$, where $\mathbf{x}_k$ has the two components as in (2.2). The auxiliary vector $\mathbf{x}_k$ serves two purposes: To achieve a low variance and a low nonresponse bias. From the variance perspective alone, $\hat{Y}_{\text{CAL}}$ is usually preferred to $\tilde{Y}_{\text{CAL}}$ because the former profits from the input of a known population total $\sum_U \mathbf{x}_k^*$. But this paper studies the bias. From that perspective, we are virtually indifferent between $\hat{Y}_{\text{CAL}}$ and

$\tilde{Y}_{\text{CAL}}$, and we focus on the latter. Under liberal conditions, the difference between the bias of $N^{-1}\hat{Y}_{\text{CAL}}$ and that of $N^{-1}\tilde{Y}_{\text{CAL}}$ is of order n^{-1}, thereby of little practical consequence even for modest sample sizes n, as discussed for example in Särndal and Lundström (2005).

An alternative expression for (2.5) is

$$\tilde{Y}_{\text{CAL}} = \left(\sum_s d_k \mathbf{x}_k\right)' \mathbf{B}_{\mathbf{x}} \tag{2.7}$$

where

$$\mathbf{B}_{\mathbf{x}} = \mathbf{B}_{\mathbf{x}|r;d} = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k \mathbf{x}_k y_k \tag{2.8}$$

is the regression coefficient vector arising from the (d_k-weighted) least squares fit based on the data $(y_k, \mathbf{x}_k)$ for $k \in r$.

A remark on the notation: When needed for emphasis, a symbol has two indices separated by a semicolon. The first shows the set of units over which the quantity is computed and the second indicates the weighting, as in $\mathbf{B}_{\mathbf{x}|r;d}$ given by (2.8), and in weighted means such as $\bar{y}_{r;d} = \sum_r d_k y_k / \sum_r d_k$. If the weighting is uniform, the second of the two indices is dropped as in $\bar{y}_U = (1/N)\sum_U y_k$.

3. Points of reference

The most primitive choice of vector is the constant one, $\mathbf{x}_k = 1$ for all k. Although inefficient for reducing nonresponse bias, it serves as a benchmark. Then $m_k = 1/P$ for all k, where P is the survey response rate (2.1), and $\tilde{Y}_{\text{CAL}}$ is the expansion estimator:

$$\tilde{Y}_{\text{EXP}} = (1/P)\sum_r d_k y_k = \hat{N}\,\bar{y}_{r;d} \tag{3.1}$$

where $\hat{N} = \sum_s d_k$ is design unbiased for the population size N. The bias of $\tilde{Y}_{\text{EXP}}$ can be large.

At the opposite end of the bias spectrum are the unbiased, or nearly unbiased, estimators obtainable under full response, when $r = s$. They are hypothetical, not computable in the presence of nonresponse. Among these are the GREG estimator with weights calibrated to the known population total $\sum_U \mathbf{x}_k^*$,

$$\hat{Y}_{\text{FUL}} = \sum_s d_k g_k y_k$$

where $g_k = 1 + (\sum_U \mathbf{x}_k^* - \sum_s d_k \mathbf{x}_k^*)'(\sum_s d_k \mathbf{x}_k^* \mathbf{x}_k^{*'})^{-1}\mathbf{x}_k^*$, and FUL refers to full response. The unbiased HT estimator (obtained when $g_k = 1$ for all k) is

$$\tilde{Y}_{\text{FUL}} = \sum_s d_k y_k = \hat{N}\,\bar{y}_{s;d}. \tag{3.2}$$

It disregards the information $\sum_U \mathbf{x}_k^*$, which may be important for variance reduction. But for the study of bias in this paper, we are indifferent between $\hat{Y}_{\text{FUL}}$ and $\tilde{Y}_{\text{FUL}}$. The difference in bias between the two is of little consequence, even for modest sample sizes. We can focus on $\tilde{Y}_{\text{FUL}}$.

4. The bias ratio

For a given outcome (s, r), consider the estimates $\tilde{Y}_{\text{CAL}}$, $\tilde{Y}_{\text{EXP}}$ and $\tilde{Y}_{\text{FUL}}$ given by (2.5), (3.1) and (3.2) as three points on a horizontal axis. Both $\tilde{Y}_{\text{EXP}}$ (generated by the primitive $\mathbf{x}_k = 1$) and $\tilde{Y}_{\text{CAL}}$ (generated by a better **x**-vector) are computable, but biased. As the **x**-vector improves, $\tilde{Y}_{\text{CAL}}$ will distance itself from $\tilde{Y}_{\text{EXP}}$ and may come near the unbiased but unrealized ideal $\tilde{Y}_{\text{FUL}}$. We consider therefore three deviations: $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$, $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}$ and $\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$, of which only the middle one is computable. The unknown "deviation total", $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$, is decomposable as "deviation accounted for" (by the chosen **x**-vector) plus "deviation remaining":

$$\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) + (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}). \tag{4.1}$$

If computable, $\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}$ would be of particular interest, as an estimate of the bias remaining in $\tilde{Y}_{\text{CAL}}$ (and in $\hat{Y}_{\text{CAL}}$), whereas $\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}$ would estimate the usually much larger bias of the benchmark, $\tilde{Y}_{\text{EXP}}$. The bias ratio for a given outcome (s, r) sets the estimated bias of $\tilde{Y}_{\text{CAL}}$ in relation to that of $\hat{Y}_{\text{EXP}}$:

$$\text{bias ratio} = \frac{\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}}{\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}}. \tag{4.2}$$

We scale the three deviations by the estimated population size $\hat{N} = \sum_s d_k$ and use the notation $\Delta_T = \Delta_A + \Delta_R$, where T suggests "total", A "accounted for" and R "remaining". Noting that $\sum_r d_k (y_k - \mathbf{x}_k'\mathbf{B}_{\mathbf{x}}) = 0$, we have

$$\Delta_T = \hat{N}^{-1}(\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}}) = \bar{y}_{r;d} - \bar{y}_{s;d};$$

$$\Delta_R = \hat{N}^{-1}(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) = \bar{\mathbf{x}}_{s;d}'\mathbf{B}_{\mathbf{x}} - \bar{y}_{s;d}$$

$$\Delta_A = \hat{N}^{-1}(\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}}$$

where $\bar{\mathbf{x}}_{s;d} = \sum_s d_k \mathbf{x}_k / \sum_s d_k$, $\bar{\mathbf{x}}_{r;d} = \sum_r d_k \mathbf{x}_k / \sum_r d_k$, and $\bar{y}_{s;d}$ and $\bar{y}_{r;d}$ are the analogously defined means for the y-variable. Then (4.2) takes the form

$$\text{bias ratio} = \frac{\Delta_R}{\Delta_T} = 1 - \frac{\Delta_A}{\Delta_T} = 1 - \frac{(\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}}}{\bar{y}_{r;d} - \bar{y}_{s;d}}. \tag{4.3}$$

We have bias ratio = 1 for the primitive vector $\mathbf{x}_k = 1$. Ideally, we want the auxiliary vector $\mathbf{x}_k$ for $\tilde{Y}_{\text{CAL}}$ to give bias ratio ≈ 0. For a given outcome (s, r) and a given y-variable, we take steps in that direction by finding an **x**-vector that makes the computable numerator $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}}$ large (in absolute value). This is within our

reach. But whatever our final choice of **x**-vector, the remaining bias of $\tilde{Y}_{\text{CAL}}$ is unknown. Even with the best available **x**-vector, considerable bias may remain. We have then attempted to do the best possible, under perhaps unfavourable circumstances.

To summarize, for a given outcome (*s*, *r*) and a given *y*-variable, the three deviations have the following features: (i) $\Delta_T = \bar{y}_{r;d} - \bar{y}_{s;d}$ is an unknown constant value, depending on both unobserved and observed *y*-values; (ii) Δ_A is computable; it depends on y_k for $k \in r$ and on the values $\mathbf{x}_k$ for $k \in s$ of the chosen **x**-vector; (iii) Δ_R cannot be computed; it depends on unobserved values y_k, and on $\mathbf{x}_k$ for $k \in s$.

To follow the progression of the estimates when the **x**-vector improves, consider a given outcome (*s*, *r*). The deviation Δ_T can have either sign. Suppose $\Delta_T > 0$, indicating a positive bias in $\tilde{Y}_{\text{EXP}}$, as when large units respond with greater propensity than small ones. When the **x**-vector in $\tilde{Y}_{\text{CAL}}$ becomes progressively more powerful by the inclusion of more and more *x*-variables, Δ_A tends to increase away from zero and will, ideally, come near Δ_T, indicating a desired closeness of $\tilde{Y}_{\text{CAL}}$ to the unbiased $\tilde{Y}_{\text{FUL}}$. As long as the **x**-vector remains relatively weak, $\Delta_A < \Delta_T$ is likely to hold. When the **x**-vector becomes increasingly powerful, Δ_A moves closer to the fixed Δ_T, a sign of bias nearing zero. It could even "move beyond", so that an "over-adjustment", $\Delta_A > \Delta_T$, has occurred. This not a detrimental feature; although $\Delta_R = \Delta_T - \Delta_A$ is then negative, it is ordinarily small (the analyst can only work with Δ_A; it is unknown to him/her whether Δ_A and Δ_T are close, or whether the over-adjustment $\Delta_A > \Delta_T$ has occurred). These points are illustrated by the simulation in Section 10. If $\Delta_T < 0$, these tendencies are reversed.

The form of (4.3) may suggest an argument which can however be misleading: Suppose that a vector $\mathbf{x}_k$ has been suggested, containing variables thought to be effective, along with an assumption that $y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$, where ε_k is a small residual. Then $\bar{y}_{r;d} - \bar{y}_{s;d} \approx (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}} \approx (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\boldsymbol{\beta}$, and consequently bias ratio ≈ 0, sending a message, often false, that the postulated vector $\mathbf{x}_k$ is efficient. One weakness of the argument stems from the well-known fact that nonresponse (unless completely random) will cause $\mathbf{B}_{\mathbf{x}}$ to be biased for a regression vector that describes the *y*-to-**x** relationship in the population. Further comments on this issue are given in Section 8.

Finally, there is the practical consideration that a typical survey has many *y*-variables. To every *y*-variable corresponds a calibration estimator, and a bias ratio given by (4.3). The ideal **x**-vector is one that would be capable of controlling bias in all those estimators. This is usually not possible without compromise, as we discuss later.

5. Expressing the deviation accounted for

The responding unit *k* receives the weight $d_k m_k$ in the estimator $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$. The nonresponse adjustment factor $m_k = (\sum_s d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1}\mathbf{x}_k$ expands the design weight d_k. We can view m_k as the value of a derived variable, defined for a particular outcome (*r*, *s*) and choice of $\mathbf{x}_k$, independent of all *y*-variables of interest, and computable for $k \in s$ (but used in $\tilde{Y}_{\text{CAL}}$ only for $k \in r$). Using (2.3), we have

$$\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k; \sum_r d_k m_k = \sum_s d_k;$$

$$\sum_r d_k m_k^2 = \sum_s d_k m_k. \qquad (5.1)$$

Two weighted means are needed:

$$\bar{m}_{r;d} = \frac{\sum_r d_k m_k}{\sum_r d_k} = \frac{\sum_s d_k}{\sum_r d_k} = \frac{1}{P}; \bar{m}_{s;d} = \frac{\sum_s d_k m_k}{\sum_s d_k} \qquad (5.2)$$

where *P* is the response rate (2.1). Thus the average adjustment factor in $\tilde{Y}_{\text{CAL}} = \sum_r d_k m_k y_k$ is $1/P$, regardless of the choice of **x**-vector. Whether a chosen **x**-vector is efficient or not for reducing bias will depend on higher moments of the m_k. The weighted variance of the m_k is

$$S_m^2 = S_{m|r;d}^2 = \sum_r d_k (m_k - \bar{m}_{r;d})^2 \Big/ \sum_r d_k. \qquad (5.3)$$

The simpler notation S_m^2 will be used. A development of (5.3) and a use of (5.1) and (5.2) gives

$$S_m^2 = \bar{m}_{r;d}(\bar{m}_{s;d} - \bar{m}_{r;d}). \qquad (5.4)$$

The coefficient of variation of the m_k is

$$\text{cv}_m = \frac{S_m}{\bar{m}_{r;d}} = \sqrt{\frac{\bar{m}_{s;d}}{\bar{m}_{r;d}} - 1}. \qquad (5.5)$$

The weighted variance of the study variable *y* is given by

$$S_y^2 = S_{y|r;d}^2 = \sum_r d_k (y_k - \bar{y}_{r;d})^2 \Big/ \sum_r d_k \qquad (5.6)$$

(when the response probabilities are not all equal, $S_y^2 = S_{y|r;d}^2$ is not unbiased for the population variance $S_{y|U}^2$, but this is not an issue for the derivations that follow). We need the covariance

$$\text{Cov}(y,m) = \text{Cov}(y,m)_{r;d} =$$

$$\frac{1}{\sum_r d_k} \sum_r d_k (m_k - \bar{m}_{r;d})(y_k - \bar{y}_{r;d}) \qquad (5.7)$$

and the correlation coefficient, $R_{y,m} = \text{Cov}(y,m)/(S_y S_m)$, satisfying $-1 \le R_{y,m} \le 1$.

The deviation $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_{\mathbf{x}}$ is a crucial component in the bias ratio (4.3). We seek an **x**-vector that

makes Δ_A large. The factors that determine Δ_A are seen in (5.8) to (5.10). Computational tools (indicators) to assist the search for effective *x*-variables are given in (5.11) and (5.12). Their derivation by linear algebra is deferred to Section 7, which may be bypassed by readers more interested in the practical use of these tools in the search for *x*-variables, as illustrated in the empirical Sections 9 and 10. We can factorize Δ_A/S_y as

$$\Delta_A/S_y = -R_{y,m} \times \mathrm{cv}_m. \qquad (5.8)$$

Two simple multiplicative factors determine Δ_A/S_y: The coefficient of variation cv_m, which is free of y_k and computed on the known $\mathbf{x}_k$ alone, and the (positive or negative) correlation coefficient $R_{y,m}$. Another factorization in terms of simple concepts is

$$\Delta_A/S_y = F \times R_{y,\mathbf{x}} \times \mathrm{cv}_m \qquad (5.9)$$

where $R_{y,\mathbf{x}} = \sqrt{R^2_{y,\mathbf{x}}}$ is the coefficient of multiple correlation between y and $\mathbf{x}$, $R^2_{y,\mathbf{x}}$ is the proportion of the *y*-variance S^2_y explained by the predictor $\mathbf{x}$, and $F = -R_{y,m}/R_{y,\mathbf{x}}$ (formula (7.8) states the precise expression for $R^2_{y,\mathbf{x}}$). As Section 7 also shows, $|R_{y,m}| \le R_{y,\mathbf{x}}$ for any $\mathbf{x}$-vector and *y*-variable; consequently $-1 \le F \le 1$.

In (5.8) and (5.9), cv_m and $R_{y,\mathbf{x}}$ are non-negative terms, while $R_{y,m}$ and F can have either sign (or possibly be zero). Hence

$$|\Delta_A|/S_y = |R_{y,m}| \times \mathrm{cv}_m = |F| \times R_{y,\mathbf{x}} \times \mathrm{cv}_m. \qquad (5.10)$$

All of S_y, cv_m, $R_{y,\mathbf{x}}$, $R_{y,m}$ and F are easily computed in the survey. Both cv_m and $R_{y,\mathbf{x}}$ increase (or possibly stay unchanged) when further *x*-variables are added to the $\mathbf{x}$-vector; $R_{y,m}$ does not have this property.

To illustrate with the aid of fairly typical numbers, if $F = 0.5; R_{y,\mathbf{x}} = 0.6$ and $\mathrm{cv}_m = 0.4$, then $\Delta_A/S_y = 0.12$, implying that $\tilde{Y}_{\mathrm{CAL}}/N = \tilde{Y}_{\mathrm{EXP}}/N - 0.12 \times S_y$. That is, the estimated *y*-mean $\tilde{Y}_{\mathrm{CAL}}/\hat{N}$ has become adjusted by 0.12 standard deviations down from the primitive estimate $\tilde{Y}_{\mathrm{EXP}}/\hat{N}$. The adjustment can be large compared to the standard deviation of the estimated *y*-mean, especially when the survey sample size is in the thousands. It remains unknown whether or not that adjustment has cured most of the biasing effect of nonresponse.

It follows from (5.8) that $0 \le |\Delta_A|/S_y \le \mathrm{cv}_m$ whatever the *y*-variable. A shaper inequality is $|\Delta_A|/S_y \le R_{y,\mathbf{x}} \times \mathrm{cv}_m$, but it depends on the *y*-variable. Further, if the correlation ratio *F* stays roughly constant when the $\mathbf{x}$-vector changes, so that $F \approx F_0$, then $|\Delta_A|/S_y \approx |F_0| \times R_{y,\mathbf{x}} \times \mathrm{cv}_m$.

Although computable for any $\mathbf{x}$-vector and any outcome (s, r), Δ_A does not reveal the value of the bias ratio. But Δ_A suggests computational tools, called indicators, for comparing alternative $\mathbf{x}$-vectors. By (5.8), let

$$H_0 = \Delta_A/S_y = -R_{y,m} \times \mathrm{cv}_m. \qquad (5.11)$$

As borne out by theory in Section 8 and by the empirical work in Section 10, over a long run of outcomes (s, r), the average of H_0 tracks the average deviation $\tilde{Y}_{\mathrm{CAL}} - Y$ (which measures the bias of $\tilde{Y}_{\mathrm{CAL}}$) in a nearly perfect linear manner when the $\mathbf{x}$-vector changes. This holds independently of the response distribution that generates *r* from *s*. Since H_0 can have either sign, it is practical to work with its absolute value denoted H_1; in addition we consider two other indicators, H_2 and H_3, inspired by (5.9) to (5.10):

$$H_1 = |\Delta_A|/S_y = |R_{y,m}| \times \mathrm{cv}_m;$$

$$H_2 = R_{y,\mathbf{x}} \times \mathrm{cv}_m; H_3 = \mathrm{cv}_m. \qquad (5.12)$$

Our main alternatives are H_1 and H_3. Of these, H_1 is motivated by its direct link to Δ_A, which we want to make large, for a given *y*-variable. A strong reason to consider H_3 is its independence of all *y*-variables in the survey. The indicator H_2 is an *ad hoc* alternative; although H_2 contains a familiar concept, the multiple correlation coefficient $R_{y,\mathbf{x}}$, it is less appropriate than H_1 because the correlation coefficient ratio $F = -R_{y,m}/R_{y,\mathbf{x}}$ may vary considerably from one $\mathbf{x}$-vector to another. Both H_2 and H_3 increase when further *x*-variables are added to the $\mathbf{x}$-vector, something which does not hold in general for H_1. The use of these indicators is illustrated in the empirical Sections 9 and 10.

6. Preference ranking of auxiliary vectors

The methods in this paper are intended for use primarily with the large samples that characterize government surveys. The sample size is ordinarily much larger than the dimension of the $\mathbf{x}$-vector. The variance of estimates is ordinarily small, compared to the squared bias. However, for categorical auxiliary variables, no group size should be allowed to be "too small". It is recommended that all group sizes be at least 30, if not at least 50, in order to avoid instability. The crossing of categorical variables (to allow interactions) implies a certain risk of small groups. It is preferable to calibrate on marginal counts, rather than on frequencies for small crossed cells.

In a number of countries, the many available administrative registers provide a rich source of auxiliary information, particularly for surveys on individuals and households. These registers contain many potential *x*-variables from which to choose. Many different $\mathbf{x}$-vectors can be composed. The indicators in (5.12) provide computational tools for obtaining a preference ordering, or a ranking, of potential $\mathbf{x}$-vectors, with the objective to reduce

as much as possible the bias remaining in the calibration estimator.

Scenario 1: Focus on a specific *y*-variable. The bias remaining in the calibration estimator depends on the *y*-variable; some are more bias prone than others. We identify one specific *y*-variable deemed to be highly important in the survey, and we seek to identify an **x**-vector that reduces the bias for this variable as much as possible (if more than one *y*-variable needs to be taken into account, a compromise must be struck, which suggests Scenario 2 below). For this purpose, we use the *y*-variable dependent indicator $H_1 = |\Delta_A|/S_y = |R_{y,m}| \times \mathrm{cv}_m$ and choose the **x**-vector so as to make H_1 large. An *ad hoc* alternative is to use the indicator $H_2 = R_{y,\mathbf{x}} \times \mathrm{cv}_m$, and strive to make it as large as possible.

Scenario 2: The objective is to identify a general purpose **x**-vector, efficient for all or most *y*-variables in the survey. This suggests $H_3 = \mathrm{cv}_m$ as a compromise indicator, and to choose the **x**-vector that maximizes H_3. To that same effect, Särndal and Lundström (2005, 2008) used the indicator $S_m^2 = H_3^2 / P^2$. They showed that the derived variable m_k in (2.6) can be seen as a predictor of the inverse of the unknown response probability and that choosing the **x**-vector to make S_m^2 large signals a bias reduction in the calibration estimator, irrespective of the *y*-variable.

For each scenario we can distinguish two procedures:

All vectors procedure: A list of candidate **x**-vectors is prepared, based on appropriate judgment. We compute the chosen indicator for *every* candidate **x**-vector, and settle for the vector that gives the highest indicator value. The resulting **x**-vector may not be the same for H_1 (which targets a specific *y*-variable) as for H_3 (which seeks a compromise for all *y*-variables in the survey).

Stepwise procedure: There is a pool of available *x*-variables. We build the **x**-vector by a stepwise forward (or stepwise backward) selection from among the available *x*-variables, one variable at a time, using the successive changes (if considered large enough) in the value of the chosen indicator to signal the inclusion (or exclusion) of a given *x*-variable at a given step. The indicators H_1, H_2 and H_3 do not in general give the same selection of variables. Consider two **x**-vectors, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, such that $\mathbf{x}_{2k}$ is made up of $\mathbf{x}_{1k}$ and an additional vector $\mathbf{x}_{+k}$: $\mathbf{x}_{2k} = (\mathbf{x}'_{1k}, \mathbf{x}'_{+k})'$. The transition from $\mathbf{x}_{1k}$ to $\mathbf{x}_{2k}$ will increase the value of H_2 and H_3. In each step of a forward selection procedure we select the variable bringing the largest increase in H_2 or H_3. But the transition does not guarantee an increased value for the most appropriate indicator, H_1. However, H_1 may be used in stepwise selection in the manner described in Section 9.

7. Derivations

For given *y*-variable and outcome (*s*, *r*), we seek an **x**-vector to make the computable numerator $\Delta_A = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\mathbf{B}_\mathbf{x}$ in the bias ratio (4.3) large, in absolute value. In this section we prove the factorizations $\Delta_A/S_y = -R_{y,m} \times \mathrm{cv}_m = F \times R_{y,\mathbf{x}} \times \mathrm{cv}_m$ in (5.8) and (5.9). We note first that cv_m^2 is a quadratic form in the vector that contrasts the **x**-mean in the response set *r* with the **x**-mean in the sample *s*. Let

$$\mathbf{D} = \bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d};\ \boldsymbol{\Sigma} = \textstyle\sum_r d_k \mathbf{x}_k \mathbf{x}'_k \Big/ \sum_r d_k. \quad (7.1)$$

Then, with *P* given by (2.1),

$$\mathrm{cv}_m^2 = P^2 \times S_m^2 = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}. \quad (7.2)$$

This expression follows from (5.3) and a consequence of (2.3), namely,

$$\bar{\mathbf{x}}'_{r;d}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = \bar{\mathbf{x}}'_{r;d}\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{s;d} = 1. \quad (7.3)$$

The vector of covariances with the study variable *y* is

$$\mathbf{C} = \Big(\textstyle\sum_r d_k (\mathbf{x}_k - \bar{\mathbf{x}}_{r;d})(y_k - \bar{y}_{r;d})\Big) \Big/ \Big(\sum_r d_k\Big). \quad (7.4)$$

We can then write Δ_A as a bilinear form:

$$\Delta_A = \mathbf{D}'\mathbf{B}_\mathbf{x} = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C} \quad (7.5)$$

using that $\mathbf{D}'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = (\bar{\mathbf{x}}_{r;d} - \bar{\mathbf{x}}_{s;d})'\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}_{r;d} = 0$ by (7.3).

A useful perspective on Δ_A is gained from the geometric interpretation of **C** and **D** in (7.5) as vectors in the space whose dimension is that of $\mathbf{x}_k$. We have

$$\Delta_A = \Lambda\, (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2} (\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2} \quad (7.6)$$

where

$$\Lambda = \frac{\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}}{(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2} (\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}}. \quad (7.7)$$

For a specific *y*-variable and a specific **x**-vector, the scalar quantities $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2}$ and $(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}$ represent the respective vector lengths of **D** and **C** (following an orthogonal transformation based on the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}^{-1}$). The scalar quantity Λ represents the cosine of the angle between **D** (which is independent of *y*) and **C** (which depends on *y*); hence $-1 \le \Lambda \le 1$.

When the auxiliary vector $\mathbf{x}_k$ is allowed to expand by adding further available *x*-variables, both vector lengths $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2}$ and $(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^{1/2}$ increase. The change in the angle Λ may be in either direction; if $|\Lambda|$ stays roughly constant, (7.6) shows that $|\Delta_A|$ will increase.

A second useful perspective on Δ_A follows by decomposing the total variability of the study variable *y*, $\sum_r d_k (y_k - \bar{y}_{r;d})^2 = (\sum_r d_k)\, S_y^2$. Two regression fits need

to be examined, the one of y on the auxiliary vector $\mathbf{x}$, and the one of y on the derived variable m defined by (2.6). To each fit corresponds a decomposition of S_y^2 into explained y-variation and residual y-variation. The two explained portions have important links to the bias ratio (4.3). Result 7.1 summarizes the two decompositions.

Result 7.1. For a given survey outcome (s, r), let $\mathbf{D}$, $\boldsymbol{\Sigma}$ and $\mathbf{C}$ be given by (7.1) and (7.4). Then the proportion of the y-variance S_y^2 explained by the regression of y on $\mathbf{x}$ is

$$R_{y,\mathbf{x}}^2 = (\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/S_y^2. \qquad (7.8)$$

The coefficient of correlation between y and the univariate predictor m is

$$R_{y,m} = -(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/[(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})^{1/2} \times S_y]. \qquad (7.9)$$

Consequently, the proportion of S_y^2 explained by m is

$$R_{y,m}^2 = (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2/[(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D}) \times S_y^2]. \qquad (7.10)$$

The proportions $R_{y,\mathbf{x}}^2$ and $R_{y,m}^2$ satisfy $R_{y,m}^2 \le R_{y,\mathbf{x}}^2 \le 1$.

Proof. The proof of (7.8) uses the weighted least squares regression of y on $\mathbf{x}$ fitted over r. The residuals are $y_k - \hat{y}(\mathbf{x})_k$, where $\hat{y}(\mathbf{x})_k = \mathbf{x}_k'\mathbf{B}_\mathbf{x}$ with $\mathbf{B}_\mathbf{x}$ given by (2.8). The decomposition is

$$\sum_r d_k (y_k - \bar{y}_{r;d})^2 = \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 + \sum_r d_k (y_k - \hat{y}(\mathbf{x})_k)^2.$$

The mixed term is zero. A development of the term "variation explained" gives $\sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 = (\sum_r d_k)\, \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}$. Thus the proportion of variance explained is $R_{y,\mathbf{x}}^2 = \sum_r d_k (\hat{y}(\mathbf{x})_k - \bar{y}_{r;d})^2 / [(\sum_r d_k)\, S_y^2] = \mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C}/S_y^2$, as claimed in (7.8). To show (7.9) we note that the covariance (5.7) can be written with the aid of (7.5) as

$$\mathrm{Cov}(y, m) = -\Delta_A / P = -\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}/P.$$

It then follows from (7.2) that $R_{y,m} = \mathrm{Cov}(y, m)/(S_y S_m)$ has the expression (7.9). The residuals from the regression (with intercept) of y on the univariate explanatory variable m are $\hat{y}(m)_k = \bar{y}_{r;d} + B_m(m_k - \bar{m}_{r;d})$ with $B_m = \mathrm{Cov}(y, m)/S_m^2 = -P\,(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})/(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})$. The proportion of variance explained is $\sum_r d_k (\hat{y}(m)_k - \bar{y}_{r;d})^2 / [(\sum_r d_k)\, S_y^2]$, which upon development gives the expression for $R_{y,m}^2$ in (7.10). Finally, $R_{y,m}^2 \le R_{y,\mathbf{x}}^2$ follows from the Cauchy-Schwarz inequality for a bilinear form: $(\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C})^2 \le (\mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{D})\,(\mathbf{C}'\boldsymbol{\Sigma}^{-1}\mathbf{C})$.

The inequality $R_{y,m}^2 \le R_{y,\mathbf{x}}^2 \le 1$ can also be deduced by the fact that, among all predictions $\hat{y}_k = \mathbf{x}_k'\boldsymbol{\beta}$ that are linear in the $\mathbf{x}$-vector, those that maximize the variance explained are $\hat{y}(\mathbf{x})_k = \mathbf{x}_k'\mathbf{B}_\mathbf{x}$, so the predictions $\hat{y}(m)_k$, which are linear in $\mathbf{x}_k$ via m_k, cannot yield a greater variance explained than that maximum.

Now from (7.9), (7.2) and (7.5), $-R_{y,m}\mathrm{cv}_m = \mathbf{D}'\boldsymbol{\Sigma}^{-1}\mathbf{C}/S_y = \Delta_A/S_y$, as claimed by formula (5.8). Moreover, (7.7), (7.8) and (7.9) imply $-R_{y,m}/R_{y,\mathbf{x}} = \Lambda$, so the correlation coefficient ratio F in (5.9) equals the angle Λ defined by (7.7).

8. Comments: Goodness of fit, properties of the bias and a related selection procedure

Three issues are examined in this section: (i) The relationship between bias and goodness of fit, (ii) the linear relation between the expected value of $\Delta_A = \hat{N}^{-1}(\tilde{Y}_{\mathrm{EXP}} - \tilde{Y}_{\mathrm{CAL}})$ and the bias of $\tilde{Y}_{\mathrm{CAL}}$ or $\hat{Y}_{\mathrm{CAL}}$, and (iii) the alternative method for selection of auxiliary variables proposed by Schouten (2007).

For the issue (i), recall that the total deviation in Section 4 is $\Delta_T = \Delta_A + \Delta_R$, where Δ_A is computable but Δ_T and Δ_R are not. If computable, $\hat{N}\Delta_R = \tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}}$ would be an estimate of the bias of $\tilde{Y}_{\mathrm{CAL}}$ (and of that of $\hat{Y}_{\mathrm{CAL}}$). A small Δ_R is desirable. The question arises: Is this achieved when $y_k = \boldsymbol{\beta}'\mathbf{x}_k + \varepsilon_k$ (with a given vector $\mathbf{x}_k$) fits the data well? We need to distinguish two aspects: (a) The computable fit to the data $(y_k, \mathbf{x}_k)$ observed for $k \in r$; and (b) The hypothetical fit to the data $(y_k, \mathbf{x}_k)$ for $k \in s$, some observed, some not.

A good fit for the respondents, $k \in r$, does not guarantee a small Δ_R: The weighted LSQ fit using the observed data $(y_k, \mathbf{x}_k)$ for $k \in r$ gives the residuals $e_{k|r;d} = y_k - \mathbf{x}_k'\mathbf{B}_{\mathbf{x}|r;d}$, computable for $k \in r$, with the property $\sum_r d_k e_{k|r;d} = 0$ (here, the detailed notation $\mathbf{B}_{\mathbf{x}|r;d}$ specified in (2.8) is preferable to the simplified notation $\mathbf{B}_\mathbf{x}$). For $k \in s - r$, $e_{k|r;d}$ is not computable; it has an unknown non-zero mean $\bar{e}_{s-r;d} = \sum_{s-r} d_k e_{k|r;d} / \sum_{s-r} d_k$. We have

$$\Delta_R = (\tilde{Y}_{\mathrm{CAL}} - \tilde{Y}_{\mathrm{FUL}})/\hat{N} = -(1-P)\,\bar{e}_{s-r;d} \ne 0. \qquad (8.1)$$

Regardless of whether the fit is good (small residuals $e_{k|r;d}$; $R_{y,\mathbf{x}}^2$ near one) or poor (large residuals $e_{k|r;d}$; $R_{y,\mathbf{x}}^2$ near zero), the deviation Δ_R given by (8.1) may be large, and $\tilde{Y}_{\mathrm{CAL}}$ far from unbiased. Even with a perfect fit for the respondents ($e_{k|r,d} = 0$ for all $k \in r$, and $R_{y,\mathbf{x}}^2 = 1$), there is no guarantee that the bias is small.

A similar inadequacy affects imputation based on the respondent data. If the regression imputations $\hat{y}_k = \mathbf{x}_k'\mathbf{B}_{\mathbf{x}|r;d}$ are used to fill in for the values y_k missing for $k \in s - r$, the imputed estimator is

$$\hat{Y}_{\mathrm{imp}} = \sum_r d_k y_k + \sum_{s-r} d_k \hat{y}_k.$$

Then $\hat{Y}_{\mathrm{imp}} = \tilde{Y}_{\mathrm{CAL}}$, so $\hat{Y}_{\mathrm{imp}}$ has the same exposure to bias as $\tilde{Y}_{\mathrm{CAL}}$, as is easily understood: When the nonresponse

causes a skewed selection of y-values, the imputed values computed on that skewed selection will misrepresent the unknown y-values that characterize the sample s or the population U.

Consider now the aspect (b) of the fit, that is, the hypothetical weighted LSQ regression fit to the data $(y_k, \mathbf{x}_k)$ for $k \in s$. The regression coefficient vector would be $\mathbf{B}_{\mathbf{x}|s;d} = (\sum_s d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_s d_k \mathbf{x}_k y_k$, and the residuals $e_{k|s;d} = y_k - \mathbf{x}_k' \mathbf{B}_{\mathbf{x}|s;d}$ for $k \in s$ satisfy $\sum_s d_k e_{k|s;d} = 0$. Using that $\sum_r d_k m_k \mathbf{x}_k / \hat{N} = \overline{\mathbf{x}}_{s;d}$ and $\sum_r d_k m_k y_k / \hat{N} = \overline{\mathbf{x}}_{s;d}' \mathbf{B}_{\mathbf{x}|r;d}$, we have

$$\Delta_R = \hat{N}^{-1}(\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}}) = (1/\hat{N}) \sum_r d_k m_k e_{k|s;d}. \quad (8.2)$$

Suppose the model is "true for the sample s", with a perfect fit, so that $e_{k|s;d} = 0$ for all $k \in s$. Then, by (8.2) we do have $\Delta_R = 0$, so the nonresponse adjusted estimator $\tilde{Y}_{\text{CAL}}$ agrees with the unbiased estimator $\tilde{Y}_{\text{FUL}}$. A belief that the bias is small hinges on an unverifiable assumption.

Turning to the issue (ii), we now explain the essentially linear relation between the bias of $\tilde{Y}_{\text{CAL}}$ and the expected value of the indicator $H_0 = \Delta_A / S_y = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{CAL}}) / \hat{N} S_y$. For a given outcome (s, r), a fixed y-variable and a fixed $\mathbf{x}$-vector we have

$$(\tilde{Y}_{\text{CAL}} - Y) / \hat{N} S_y = (\tilde{Y}_{\text{EXP}} - Y) / \hat{N} S_y - H_0.$$

Let E_{pq} denote the expectation operator with respect to all outcomes (s, r), that is, $E_{pq}(\cdot) = E_p(E_q(\cdot|s))$, where $p(s)$ and $q(r|s)$ are, respectively, the known sampling design and the unknown response distribution. We denote $\text{bias}(\tilde{Y}_{\text{CAL}}) = E_{pq}(\tilde{Y}_{\text{CAL}}) - Y$, $\text{bias}(\tilde{Y}_{\text{EXP}}) = E_{pq}(\tilde{Y}_{\text{EXP}}) - Y$ and $C = E_{pq}(\hat{N} S_y)$. Using the usual large sample replacement of the expected value of a ratio by the ratio of the expected values, we have $E_{pq}[(\tilde{Y}_{\text{CAL}} - Y) / \hat{N} S_y] \approx [E_{pq}(\tilde{Y}_{\text{CAL}}) - Y] / E_{pq}(\hat{N} S_y)$ and analogously for $\tilde{Y}_{\text{EXP}}$, so

$$\text{bias}(\tilde{Y}_{\text{CAL}}) \approx \text{bias}(\tilde{Y}_{\text{EXP}}) - C \times E(H_0). \quad (8.3)$$

Here $\text{bias}(\tilde{Y}_{\text{EXP}})$ and C do not depend on the choice of $\mathbf{x}$-vector, whereas $\text{bias}(\tilde{Y}_{\text{CAL}})$ and $E(H_0)$ do. Therefore, as the $\mathbf{x}$-vector changes, $\text{bias}(\tilde{Y}_{\text{CAL}})$ and $E(H_0)$ are essentially linearly related. No particular forms of $p(s)$ and $q(r|s)$ need to be specified for (8.3) to hold. As a consequence, when two auxiliary vectors, $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$, are compared, the difference in bias is, to close approximation, proportional to the change in the expected value of H_0:

$$\text{bias}(\tilde{Y}_{\text{CAL}}(\mathbf{x}_{1k})) - \text{bias}(\tilde{Y}_{\text{CAL}}(\mathbf{x}_{2k})) \approx -C(E_1 - E_2) \quad (8.4)$$

where $E_i = E_{pq}(H_0(\mathbf{x}_{ik}))$ for $i = 1, 2$. The properties (8.3) and (8.4) are validated by the Monte Carlo study in Section 10.

Note that formula (8.3) does not guarantee that $\tilde{Y}_{\text{CAL}}$ based on a certain vector $\mathbf{x}_k$ will have zero or near-zero bias. It does not state that a comparatively large value of $|\Delta_A|$ guarantees a small bias in $\tilde{Y}_{\text{CAL}}$. What (8.3) says is that $\text{bias}(\tilde{Y}_{\text{CAL}})$ is linearly related to the expectation of the indicator $H_0 = \Delta_A / S_y$. Therefore, to assess available $\mathbf{x}$-vectors in terms of the indicator H_0 (or the indicator $H_1 = |\Delta_A| / S_y$) is consistent with the objective of bias reduction.

Turning to the issue (iii), we comment on the alternative method for selection of auxiliary variables proposed by Schouten (2007). His indicator for the step-by-step selection of variables differs from our indicators; it will usually not select exactly the same set of variables. In a list of say 30 available categorical x-variables, the first ten to enter will not be the same set of ten as with our indicators H_0 to H_3. The order in which variables are selected will not necessarily be the same either. For comparison, we compared, in some of our empirical work, with the variable selection realized by Schouten's method. In some cases we noted a considerable congruence between the two sets of "first ten" picked in the two procedures.

The differences between the two approaches are best appreciated by a comparison of their background and derivation. Our indicators H_0 and H_1 originate in the notion of separation (or distance), for a given outcome (s, r), between the adjusted estimator $\tilde{Y}_{\text{CAL}}$ and the primitive one, $\tilde{Y}_{\text{EXP}}$, and in the idea that this separation will ordinarily increase when the $\mathbf{x}$-vector becomes more powerful. The probability sampling design is taken into consideration; no assumptions are made on the response distribution.

Schouten uses a superpopulation argument; sampling weights do not appear to enter into consideration. An expression for the model-expected bias of an estimator of the population mean is found to be proportional to the correlation (at the level of the population) between the y-variable and the 0-1 indicator for response. It is shown that this correlation (and consequently the bias) can be bounded inside an interval. In particular, the generalized regression estimator is considered and it is shown that its maximum absolute bias equals the width of the bias interval. This width depends on the true unknown regression vector $\boldsymbol{\beta}$ for the regression (at the population level) between y and $\mathbf{x}$. This unknown $\boldsymbol{\beta}$ is replaced by an estimate based on the respondents, thus subject to some bias because of the nonresponse. Schouten emphasizes that a missing-at-random assumption is not needed for his method, which is in that respect similar to our method.

9. Auxiliary variable choice for the Swedish pilot survey on gaming and problem gambling

We identified a real survey data set to illustrate the use of the indicators H_1, H_2 and H_3 in building the **x**-vector. In 2008, The Swedish National Institute of Public Health (*Svenska Folkhälsoinstitutet*) conducted a pilot survey to study the extent of gambling participation and the characteristics of persons with gambling problems. Sampling and weight calibration was carried out by Statistics Sweden. We illustrate the use of the indicators in this survey, for which a stratified simple random sample s of n = 2,000 persons was drawn from the Swedish Register of Total Population (RTP). The strata were defined by the cross classification of region of residence by age group. Each of the six regions was defined as a cluster of postal code areas deemed similar in regard to variables such as education level, purchasing power, type of housing, foreign background. The four age groups were defined by the brackets 16-24; 25-34; 35-64 and 65-84.

The overall unweighted response rate was 50.8%. The nonresponse, more or less pronounced in the different domains of interest, interferes with the accuracy objective. An extensive pool of potential auxiliary variables was available for this survey, including variables in the RTP, in the Education Register and a subset of those in another extensive Statistics Sweden data base, LISA. For this illustration, we prepared a data file consisting of 13 selected categorical variables. Twelve of these were designated as *x*-variables, and one, the dichotomous variable *Employed*, played the role of the study variable. The values of all variables are available for all units $k \in s$. Response $(k \in r)$ or not $(k \in s - r)$ to the survey is also indicated in the data file.

Variables that are continuous by nature were used as grouped; all 12 *x*-variables are thus categorical and of the $\mathbf{x}_k^\circ$ type, as defined in Section 2 (because most of the variables are available for the full population, they are potentially of the type $\mathbf{x}_k^*$, but since the effect on bias is of little consequence, we used them as $\mathbf{x}_k^\circ$-variables). The study variable value, $y_k = 1$ if k is *employed* and $y_k = 0$ otherwise, is known for $k \in s$, so the unbiased estimate $\tilde{Y}_{\text{FUL}}$ defined by (3.2) can be computed and used as a reference. We also computed $\tilde{Y}_{\text{EXP}}$ defined by (3.1), as well as $\tilde{Y}_{\text{CAL}}$ defined by (2.5) for different **x**-vectors built by stepwise selection from the pool of 12 *x*-variables with the aid of the indicators H_1, H_2 and H_3 defined by (5.12).

We carried out forward selection as follows: The auxiliary vector in Step 0 is the trivial $\mathbf{x}_k = 1$, and the estimator is $\tilde{Y}_{\text{EXP}}$. In Step 1, the indicator value is computed for every one of 12 presumptive auxiliary variables; the variable producing the largest value of the indicator is selected. In Step 2, the indicator value is computed for all 11 vectors of dimension two that contain the variable selected in Step 1 and one of the remaining variables. The variable that gives the largest value for the indicator is selected in Step 2, and so on, in the following steps. A new variable always joins already entered variables in the "side-by-side" (or "+") manner. Interactions are thereby relinquished. The order of selection is different for each indicator.

The values of H_2 and H_3 that identify the next variable for inclusion are by mathematical necessity increasing in every step. This does not hold for H_1. In a certain step j, we used the rule to include the *x*-variable with the largest of computed H_1-values. That value can be smaller than the H_1-value that identified the variable entering in the preceding step, $j-1$. The series of H_1-values for inclusion will increase up to a certain step, then begin to decline, as Table 9.1 illustrates.

The unbiased estimate is $\tilde{Y}_{\text{FUL}} = 4,265$; the primitive estimate is $\tilde{Y}_{\text{EXP}} = 4,719$ (both in thousands). This suggests a large positive bias in $\tilde{Y}_{\text{EXP}}$, whose relative deviation (in %) from $\tilde{Y}_{\text{FUL}}$ is $\text{RDF} = (\tilde{Y}_{\text{EXP}} - \tilde{Y}_{\text{FUL}})/\tilde{Y}_{\text{FUL}} \times 10^2 = 10.7$. Adding categorical *x*-variables one by one into the **x**-vector will successively change this deviation, although when a few variables have been admitted, the change is not always in the direction of a smaller value. In each step we computed the indicator, $\tilde{Y}_{\text{CAL}}$ and $\text{RDF} = (\tilde{Y}_{\text{CAL}} - \tilde{Y}_{\text{FUL}})/\tilde{Y}_{\text{FUL}} \times 10^2$.

Table 9.1 shows the stepwise selection with the indicator H_1 (the number of categories is given in parenthesis for each selected variable). First to enter is the variable *Income class*; this brings a large reduction in RDF from 10.7 to 4.5. The next five selections take place with increased H_1-values, and the value of RDF is reduced, but by successively smaller amounts. Step six, where *Marital status* is selected, brings about a turning point, indicated by the double line in Table 9.1: The value of H_1 then starts to decline, and $\tilde{Y}_{\text{CAL}}$ and RDF start to increase. At step 6, RDF is at its lowest value, 0.5, then starts to rise, illustrating that inclusion of all available *x*-variables may not be best. The turning point of H_1 and the point at which RDF is closest to zero happen to agree in this example. This is not generally the case. Moreover, in a real survey setting, RDF is unknown, as is the step at which RDF is closest to zero.

Table 9.2 shows the stepwise selection with indicator H_3. Its value increases at every step, but at a rate that levels off, and successive changes in $\tilde{Y}_{\text{CAL}}$ become negligible. This suggests to stop after six steps, at which point RDF = 2.8. In none of the 12 steps does RDF come as close to zero as the value RDF = 0.5 obtained with H_1 after six steps. In this respect H_1 is better than H_3, in this example. With all 12 *x*-variables selected, RDF attains in both tables the final value 2.6.

Table 9.1
Stepwise forward selection, indicator H_1, dichotomous study variable *Employed*. Successive values of $H_1 \times 10^3$, of $\tilde{Y}_{CAL}$ in thousands, and of RDF $= (\tilde{Y}_{CAL} - \tilde{Y}_{FUL})/\tilde{Y}_{FUL} \times 10^2$. For comparison, $\tilde{Y}_{EXP} \times 10^{-3} = 4,719$; $\tilde{Y}_{FUL} \times 10^{-3} = 4,265$

Auxiliary variable entered	$H_1 \times 10^3$	$\tilde{Y}_{CAL} \times 10^{-3}$	RDF
Income class (3)	76	4,458	4.5
Education level (3)	107	4,350	2.0
Presence of children (2)	114	4,326	1.4
Urban centre dwelling (2)	118	4,310	1.1
Sex (2)	123	4,296	0.7
Marital status (2)	125	4,286	0.5
Days unemployed (3)	121	4,301	0.9
Months with sickness benefits (3)	120	4,305	1.0
Level of debt (3)	115	4,322	1.3
Cluster of postal codes (6)	109	4,343	1.8
Country of birth (2)	103	4,363	2.3
Age class (4)	99	4,377	2.6

Table 9.2
Stepwise forward selection, indicator H_3, dichotomous study variable *Employed*. Successive values of $H_3 \times 10^3$, of $\tilde{Y}_{CAL}$ in thousands, of RDF $= (\tilde{Y}_{CAL} - \tilde{Y}_{FUL})/\tilde{Y}_{FUL} \times 10^2$. For comparison, $\tilde{Y}_{EXP} \times 10^{-3} = 4,719$; $\tilde{Y}_{FUL} \times 10^{-3} = 4,265$

Auxiliary variable entered	$H_3 \times 10^3$	$\tilde{Y}_{CAL} \times 10^3$	RDF
Education level (3)	186	4,520	6.0
Cluster of postcode areas (6)	250	4,505	5.6
Country of birth (2)	281	4,498	5.5
Income Class (3)	298	4,369	2.4
Age class (4)	354	4,399	3.1
Sex (2)	364	4,384	2.8
Urban centre dwelling (2)	374	4,378	2.6
Level of debt (3)	381	4,364	2.3
Months with sickness benefits (3)	384	4,380	2.7
Presence of children (2)	387	4,379	2.7
Marital status (2)	388	4,379	2.7
Days unemployed (3)	388	4,377	2.6

The set of the first six variables to enter with H_3 has three in common with the corresponding set of six with H_1. There is no contradiction in the quite different selection patterns, because H_1 is geared to the specific *y*-variable *Employed*, while H_3 is a compromise indicator, independent of any *y*-variable. To save space, the step-by-step results for indicator H_2 are not shown. Its selection pattern resembles more that of H_3 than that of H_1. Out of the first six variables to enter with H_2, four are among the first six with H_3. As a general comment, we believe that in many practical situations the use of more than six variables is unnecessary, and the selection of the first few becomes crucially important.

10. Empirical validation by simulation for a constructed population

The theory presented in earlier sections makes no assumptions on the response distribution. It is unknown. The sampling design is arbitrary; its known inclusion probabilities are taken into account. For the experiment in this section, we specify several different response distributions with a specified positive value for the response probability θ_k for every $k \in U$. That is, with specified probability θ_k, the value y_k gets recorded in the experiment; with probability $1 - \theta_k$, it goes missing. We find that the indicators H_0 (or $H_1 = |H_0|$) defined in (5.11) ranks the different **x**-vectors in the correct order of preference for all participating response distributions, consistent with the theoretical results (8.3) and (8.4). We confirm that, over a long run of outcomes (s, r), the average of $H_0 = \Delta_A / S_y = -R_{y,m} \times \mathrm{cv}_m$ tracks the bias of the calibration estimator, measured by the average of $\tilde{Y}_{CAL} - Y$, in an essentially perfectly linear manner, when the **x**-vector moves through 16 different formulations. We also examine the indicators H_2 and H_3 defined in (5.12), and find in this experiment that they also have strong relationship to the bias of $\tilde{Y}_{CAL}$.

We experimented with several created populations; the conclusions were similar. We report here results for one constructed population of size $N = 6,000$, with created values $(y_k, \mathbf{x}_k, \theta_k)$ for $k = 1, 2, \ldots, N = 6,000$, for 16 alternative categorical formulations of $\mathbf{x}_k$, and four different ways to assign the θ_k.

The 16 alternative categorical auxiliary **x**-vectors were obtained by grouping the generated values x_{1k} and x_{2k} of two continuous auxiliary variables, x_1 and x_2. The values (y_k, x_{1k}, x_{2k}) for $k = 1, 2, ..., 6,000$ were created in three steps as follows. Step 1 (the variable x_1): The 6,000 values x_{1k} were obtained as independent outcomes of the gamma distributed random variable $\Gamma(a, b)$ with parameter values $a = 2$, $b = 5$. The mean and variance of the 6,000 realized values x_{1k} was 10.0 and 49.9, respectively. Step 2 (the variable x_2): For unit *k*, with value x_{1k} fixed by Step 1, a value x_{2k} is realized as an outcome of the gamma random variable with parameters such that the conditional expectation and variance of x_{2k} are $\alpha + \beta x_{1k} + K h(x_{1k})$ and $\sigma^2 x_{1k}$, respectively, where $h(x_{1k}) = x_{1k}(x_{1k} - \mu_{x_1})(x_{1k} - 3\mu_{x_1})$ with $\mu_{x_1} = 10$. We used the values $\alpha = 1$, $\beta = 1$, $k = 0.001$ and $\sigma^2 = 25$. The polynomial term $K h(x_{1k})$ gives a mild non-linear shape to the plot of (x_{2k}, x_{1k}), to avoid an exactly linear relationship. The mean and variance of the 6,000 realized values x_{2k} were 11.0 and 210.0, respectively. The correlation coefficient between x_1 and x_2, computed on the 6,000 couples (x_{1k}, x_{2k}), was 0.48. Step 3 (the study variable *y*): For unit *k*, with values x_{1k} and x_{2k} fixed by Steps 1 and 2, a value y_k is realized as an outcome of the gamma random variable with parameters such that the conditional expectation and variance of y_k are $c_0 + c_1 x_{1k} + c_2 x_{2k}$ and $\sigma_0^2 (c_1 x_{1k} + c_2 x_{2k})$, respectively. We used $c_0 = 1$, $c_1 = 0.7$, $c_2 = 0.3$ and $\sigma_0^2 = 2$. The mean and the variance of the 6,000 realized

values y_k were 11.4 and 86.5, respectively. The correlation coefficient between y and x_1, computed on the 6,000 couples (y_k, x_{1k}), was 0.76; that between y and x_2, computed on the 6,000 couples (y_k, x_{2k}), was 0.73.

Each of the two x-variables was then transformed into four alternative group modes, denoted 8G, 4G, 2G and 1G, yielding $4 \times 4 = 16$ different auxiliary vectors $\mathbf{x}_k$. The 6,000 values x_{1k} of variable x_1 were size ordered; eight equal-sized groups were formed. Group 1 consists of the units with the 750 largest values x_{1k}, group 2 consists of the next 750 units in the size ordering, and so on, ending with group 8. In this mode 8G of x_1, unit k is assigned the vector value $\boldsymbol{\gamma}_{(x_1;8)k}$, of dimension eight with seven entries "0" and a single entry "1" to code the group membership of k. Next, successive group mergers are carried out, so that two adjoining groups always define a new group, every time doubling the group size. Thus for mode 4G, the merger of groups 1 and 2 puts the units with the 1,500 largest x_{1k}-values into a first new group; groups 3 and 4 merge to form the second new group of 1,500, and so on; the vector value associated with unit k is $\boldsymbol{\gamma}_{(x_1;4)k}$. In mode 2G, unit k has the vector value $\boldsymbol{\gamma}_{(x_1;2)k} = (1,0)'$ for the 3,000 largest x_1-value units and $\boldsymbol{\gamma}_{(x_1;2)k} = (0,1)'$ for the rest. In the ultimate mode, 1G, all 6,000 units are put together, all x_1-information is relinquished, and $\boldsymbol{\gamma}_{(x_1;1)k} = 1$ for all k. The 6,000 values x_{2k} were transformed by the same procedure into the group modes 8G, 4G, 2G and 1G. Corresponding group membership of unit k is coded by the vectors $\boldsymbol{\gamma}_{(x_2;8)k}, \boldsymbol{\gamma}_{(x_2;4)k}, \boldsymbol{\gamma}_{(x_2;2)k}$ and $\boldsymbol{\gamma}_{(x_2;1)k} = 1$. The $4 \times 4 = 16$ different auxiliary vectors $\mathbf{x}_k$ take into account both kinds of group information; the two $\boldsymbol{\gamma}$-vectors are placed side by side (as opposed to crossed), the result being a calibration on two margins, as indicated by the "+" sign. Thus for the case denoted 8G + 8G, unit k has the auxiliary vector value $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;8)k}, \boldsymbol{\gamma}'_{(x_2;8)k})'_{(-1)}$, where (-1) indicates that one category is excluded in either $\boldsymbol{\gamma}_{(x_1;8)k}$ or $\boldsymbol{\gamma}_{(x_2;8)k}$ to avoid a singular matrix in the computation, giving $\mathbf{x}_k$ the dimension $8 + 8 - 1 = 15$. The case 8G + 8G has the highest information content. At the other extreme, the case 1G + 1G disregards all the x-information and $\mathbf{x}_k = 1$ for all k. There are 14 intermediate cases of information content. For example, 4G + 2G has $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;4)k}, \boldsymbol{\gamma}'_{(x_2;2)k})'_{(-1)}$ of dimension $4 + 2 - 1 = 5$; 4G + 1G has $\mathbf{x}_k = (\boldsymbol{\gamma}'_{(x_1;4)k}, 1)'_{(-1)} = \boldsymbol{\gamma}_{(x_1;4)k}$ of dimension 4 (there is non-negligible interaction between x_1 and x_2 in this experiment, but we restrict the experiment to **x**-vectors without interactions, causing no risk of small group counts).

We discuss here the results for four response distributions. Their response probabilities θ_k, $k = 1, 2, ..., N = 6{,}000$, were specified as follows:

IncExp$(10 + x_1 + x_2)$, with $\theta_k = 1 - e^{-c(10+x_{1k}+x_{2k})}$ where $c = 0.04599$

IncExp$(10 + y)$, with $\theta_k = 1 - e^{-c(10+y_k)}$ where $c = 0.06217$

DecExp$(x_1 + x_2)$, with $\theta_k = e^{-c(x_{1k}+x_{2k})}$ where $c = 0.01937$

DecExp(y), with $\theta_k = e^{-cy_k}$ where $c = 0.03534$.

The constant c was adjusted in all four cases to give a mean response probability of $\bar{\theta}_U = \sum_U \theta_k / N = 0.70$. In the first two, the value 10 (rather than 0) was used to avoid a high incidence of small response probabilities θ_k. These four options represent contrasting features for the response probabilities: increasing as opposed to decreasing, dependent on x-values only as opposed to dependent on y-values only. In the second and fourth option, the response is directly y-variable dependent, and could hence be called "purely non-ignorable".

We generated $J = 5{,}000$ outcomes (s, r), where s of size $n = 1{,}000$ is drawn from $N = 6{,}000$ by simple random sampling and, for every given s, the response set r is realized by each of the four response distributions. That is, for $k \in s$, a Bernoulli trial was carried out with the specified probability θ_k of inclusion in the response set r. The Bernoulli trials are independent.

For each response distribution, for each of the 16 **x**-vectors, and for every outcome (s, r), we computed the relative deviation $\mathrm{RD} = (\hat{Y}_{\mathrm{CAL}} - Y)/Y$, where $\hat{Y}_{\mathrm{CAL}}$ is given by (2.4) and $Y = \sum_U y_k$ is the targeted y-total, known in this experimental setting (alternatively, we used $\tilde{Y}_{\mathrm{CAL}}$ given by (2.5) but, as expected, the difference in bias compared with $\hat{Y}_{\mathrm{CAL}}$ is negligible). We also computed the indicators $H_i, i = 0, 1, 2, 3,$ given by (5.11) and (5.12). Summary measures were computed as

$$\mathrm{relbias} = \mathrm{Av}(\mathrm{RD}) = \frac{1}{J}\sum_{j=1}^{J} \mathrm{RD}_j;$$

$$\mathrm{Av}(H_i) = \frac{1}{J}\sum_{j=1}^{J} H_{ij} \quad \text{for} \quad i = 0,1,2,3$$

where j indicates the value computed for the j^{th} outcome, $j = 1, 2, \ldots, 5{,}000 = J$. For each response distribution, we thus obtain the value *relbias* (which is the Monte Carlo measure of the relative bias $(E_{pq}(\hat{Y}_{\mathrm{CAL}}) - Y)/Y)$ and 16 values of $\mathrm{Av}(H_i)$ (which is the Monte Carlo measure of $E_{pq}(H_i)$), $i = 0, 1, 2, 3,$ where p stands for simple random sampling, and q stands for one of the four response distributions.

Table 10.1 shows, for IncExp$(10 + x_1 + x_2)$, *relbias* in % and $\mathrm{Av}(H_1) \times 10^3$ for the 16 **x**-vectors. For the cell 1G + 1G, with vector $\mathbf{x}_k = 1$, all four Av-quantities are zero, and *relbias* is at its highest level, 13.2%. At the opposite extreme, the cell 8G + 8G represents the highest level of

information; it gives the highest value for $\text{Av}(H_1)$, and *relbias* is at its lowest value, 0.2%; virtually all bias is removed (except for a possible sign difference, $\text{Av}(H_0)$ and $\text{Av}(H_1)$ were equal for all cells).

The result (8.4), holding for any response distribution and any sampling design, states that the indicator H_0 will rank the $4 \times 4 = 16$ auxiliary vectors correctly for any response distribution (with response probabilities not all constant, as noted below). Table 10.1 illustrates (8.4) in terms of $H_1 = |H_0|$: The change, from any one cell to any other, in the value of $\text{Av}(H_1)$ (the Monte-Carlo estimate of the expected value of (H_1) is accompanied by a proportional change in the value of *relbias*. The same proportionality was noted for the other three response distributions. We could have chosen other response distributions to illustrate the same property.

Table 10.1
Relbias in % and, within parenthesis, the value of $\text{Av}(H_1) \times 10^3$ for 16 auxiliary vectors $\mathbf{x}_k$. Response distribution IncExp$(10 + x_1 + x_2)$

Groups based on x_{1k}	Groups based on x_{2k}							
	8G		4G		2G		1G	
8G	0.2	(101)	0.5	(99)	1.3	(93)	3.4	(76)
4G	0.5	(98)	0.9	(96)	1.8	(89)	4.1	(70)
2G	1.5	(91)	1.9	(88)	3.2	(78)	6.5	(52)
1G	4.1	(70)	5.0	(64)	7.3	(46)	13.2	(0)

The response distribution with a constant response probability θ_k for all k is a special case. The calibration estimator $\tilde{Y}_{\text{CAL}}$ based on any vector $\mathbf{x}_k$ then has zero bias (very nearly), and this includes the primitive estimator $\tilde{Y}_{\text{EXP}}$ with $\mathbf{x}_k = 1$. Result 8.3 continues to be valid, stating in that case that $E_{pq}(H_0) \approx \text{bias}(\tilde{Y}_{\text{CAL}}) \approx \text{bias}(\tilde{Y}_{\text{EXP}}) \approx 0$. In the context of the simulation in this section, if $\theta_k = 0.70$ for all k is taken to be an additional response distribution, Table 10.1 will in all 16 cells show nearly zero values of both *relbias in %* and $\text{Av}(H_1) \times 10^3$, from the weakest cell (1G + 1G) all the way to the cell of the most powerful **x**-vector (8G + 8G). There is no bias to be removed by an improvement of the **x**-vector. If in practice the indicator (H_1) does not react to an enlargement of the **x**-vector, there is no incentive to seek beyond the simplest vector formulation. It could signify one of three possibilities: The y-variable in question is not subject to nonresponse bias, or that the response probability is almost constant, or that none of the available **x**-vectors is capable of reducing an existing bias.

To save space we do not show the corresponding tables for $\text{Av}(H_2)$ and $\text{Av}(H_3)$. By mathematical necessity, both quantities increase in the nested transitions. Not shown either are the counterparts of Table 10.1 for the other three response distributions. The patterns are similar.

Table 10.2 for IncExp$(10 + x_1 + x_2)$ and Table 10.3 for IncExp$(10 + y)$ show how $\text{Av}(H_1)$, $\text{Av}(H_2)$ and $\text{Av}(H_3)$ rank the 16 **x**-vectors, represented by their value of *relbias*. To measure the success of ranking, we computed the Spearman rank correlation coefficient, denoted *rancor*, between *relbias* and the value of the indicator, based on the 16 values of each. For $\text{Av}(H_1)$, the bottom line of the two tables shows $|rancor| = 1$, for perfect ranking. For these data, $|rancor|$ is near one also for $\text{Av}(H_2)$ and $\text{Av}(H_3)$ (more generally, the ranking obtained with H_2 and H_3 may be good, but is data dependent).

Table 10.2
Value, in ascending order, of relbias in %, and corresponding value and rank of $\text{Av}(H_1) \times 10^3$, $\text{Av}(H_2) \times 10^3$ and $\text{Av}(H_3) \times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlations, *rancor*. Response distribution IncExp$(10 + x_1 + x_2)$

relbias	$\text{Av}(H_1) \times 10^3$		$\text{Av}(H_2) \times 10^3$		$\text{Av}(H_3) \times 10^3$	
0.2	101	(1)	127	(1)	232	(1)
0.5	99	(2)	119	(2)	225	(2)
0.5	98	(3)	118	(3)	224	(3)
0.8	96	(4)	109	(4)	217	(4)
1.3	93	(5)	109	(5)	216	(5)
1.5	91	(6)	105	(6)	213	(6)
1.8	89	(7)	98	(7)	207	(7)
1.9	88	(8)	94	(8)	205	(8)
3.2	78	(9)	80	(11)	192	(9)
3.4	76	(10)	90	(9)	188	(11)
4.1	70	(11)	84	(10)	190	(10)
4.1	70	(12)	77	(12)	175	(13)
5.0	64	(13)	70	(13)	179	(12)
6.4	52	(14)	52	(14)	146	(15)
7.3	46	(15)	46	(15)	156	(14)
13.2	0	(16)	0	(16)	0	(16)
Rancor		-1.00		-0.99		-0.99

There is one notable contrast between the results on *relbias* for the two response distributions in Tables 10.2 and 10.3. The best among the auxiliary vectors leave considerably more bias for the non-ignorable IncExp$(10 + y)$ than for IncExp$(10 + x_1 + x_2)$. This is not unexpected, and it is important to note that considerable bias reduction is obtained for the non-ignorable case as well.

In the simulation, the over-adjustment mentioned in Section 4, $\Delta_A > \Delta_T > 0$ (when $(\tilde{Y}_{\text{EXP}})$ has positive bias) or $\Delta_A < \Delta_T < 0$ (when $\tilde{Y}_{\text{EXP}}$ has negative bias), happens for some outcomes (s, r). The frequency varies with the strength of the auxiliary vector and is different for different response distributions. The cell for which this over-adjustment is most likely to occur is 8G + 8G, the most powerful of the 16 auxiliary vectors. For IncExp$(10 + x_1 + x_2)$, the bias is almost completely removed for cell 8G + 8G; *relbias* is only 0.2%. Hence $\tilde{Y}_{\text{CAL}}$ is close to the unbiased $\tilde{Y}_{\text{FUL}}$, Δ_A is near Δ_T, and $\Delta_A > \Delta_T$ happened for 45.6% of all outcome (s, r). By contrast, for the non-ignorable case IncExp$(10 + y)$, the incidence of $\Delta_A > \Delta_T$

was only 0.1% for the cell 8G + 8G. Although that cell brings considerable bias reduction (compared to the primitive 1G + 1G), there is bias remaining, and as a consequence, $\Delta_A > \Delta_T$ almost never happens.

We do not show the corresponding tables for DecExp($x_1 + x_2$) and DecExp(y). The lowest value of *rancor* was 0.94, recorded for $\text{Av}(H_3)$ in the case of DecExp($x_1 + x_2$).

A question not addressed in Tables 10.2 and 10.3 is: How often, over a long series of outcomes (s, r), does a given indicator $H(\mathbf{x}_k)$ succeed in pointing correctly to the preferred **x**-vector? To answer this, let $\mathbf{x}_{1k}$ and $\mathbf{x}_{2k}$ be two vectors selected for comparison. If the absolute value of the bias of $\hat{Y}_{\text{CAL}}(\mathbf{x}_{2k})$ is smaller than that of $\hat{Y}_{\text{CAL}}(\mathbf{x}_{1k})$, we would like to see that $H(\mathbf{x}_{2k}) \geq H(\mathbf{x}_{1k})$ holds for a vast majority of all outcomes (s, r), because then the indicator $H(\cdot)$ delivers with high probability the correct decision to prefer $\mathbf{x}_{2k}$. Because $H(\mathbf{x}_k)$ has sampling variability, its success rate (the rate of correct indication) depends on the sample size, and we expect it to increase with sample size.

Table 10.3
Value, in ascending order, of relbias in %, and corresponding value and rank of $\text{Av}(H_1) \times 10^3$, $\text{Av}(H_2) \times 10^3$ and $\text{Av}(H_3) \times 10^3$, for 16 auxiliary vectors. Bottom line: Value of Spearman rank correlations, *rancor*. Response distribution IncExp (10 + *y*)

relbias	$\text{Av}(H_1)\times 10^3$		$\text{Av}(H_2)\times 10^3$		$\text{Av}(H_3)\times 10^3$	
3.6	74	(1)	91	(1)	165	(1)
3.9	71	(2)	84	(2)	158	(2)
4.0	71	(3)	83	(3)	156	(3)
4.3	68	(4)	76	(5)	149	(5)
4.4	68	(5)	78	(4)	153	(4)
4.9	64	(6)	68	(8)	142	(3)
4.9	63	(7)	72	(6)	146	(6)
5.3	60	(8)	69	(7)	143	(7)
5.4	60	(9)	64	(9)	137	(9)
6.0	55	(10)	59	(10)	132	(10)
6.2	53	(11)	54	(11)	128	(11)
7.2	46	(12)	54	(12)	122	(12)
7.9	41	(13)	41	(14)	111	(13)
7.9	40	(14)	43	(13)	109	(14)
9.6	27	(15)	27	(15)	90	(15)
13.1	0	(16)	0	(16)	0	(16)
Rancor		-1.00		-0.99		-0.99

We threw some light on this question by extending the Monte Carlo experiment: 5,000 outcomes (s, r) were realized, first with sample size $n = 1{,}000$, then with sample size $n = 2{,}000$ (the response set r is realized according to one of the four response distributions, declaring unit k "responding" as a result of a Bernoulli trial with the specified probability θ_k). We computed the success rate as the proportion of all outcomes (s, r) in which the correct indication materializes in a confrontation of two different **x**-vectors. Several pairwise comparisons of this kind were carried out. Typical results are shown in Table 10.4, for IncExp($10 + x_1 + x_2$). The upper entry in a table cell shows the success rate in % for $n = 1{,}000$, the lower entry shows that rate for $n = 2{,}000$. Shown in parenthesis is the value of *relbias* for the vectors in question.

"Severe tests" are preferred, that is, confrontations of vectors with a small difference in absolute *relbias*, because the correct decision is then harder to obtain. There is a priori no reason why one of the indicators should always outperform the others in this study. In the five severe tests in Table 10.4, H_1 has, on the whole, better success rates than H_2 and H_3. The success rate of H_1 improves by doubling the sample size, and tends as expected to be greater when the *relbias* values are further apart. The case 4G + 8G *vs.* 8G + 8G compares nested **x**-vectors, so it is known beforehand that H_2 and H_3 give perfect success rates.

Table 10.4
Selected pairwise comparisons of auxiliary vectors; percentage of outcomes with correct indication, for the indicators H_1, H_2 and H_3. Within parenthesis, relbias in %. Upper entry: n = 1,000 lower entry: n = 2,000. Response distribution IncExp ($10 + x_1 + x_2$)

Cells compared	Percent outcomes with correct indication		
	H_1	H_2	H_3
4G + 8G(0.5) *vs.*	90.0	100.0	100.0
8G + 8G(0.2)	96.4	100.0	100.0
4G + 2G(1.8) *vs.*	66.8	86.0	70.7
2G + 8G(1.5)	74.2	89.0	67.4
1G + 8G(4.1) *vs.*	74.3	70.3	45.0
8G + 1G(3.4)	82.8	78.0	43.3
4G + 1G(4.1) *vs.*	90.6	61.4	83.9
2G + 2G(3.2)	97.0	68.8	92.3
1G + 2G(7.3) *vs.*	77.4	77.4	34.5
2G + 1G(6.5)	85.9	85.9	28.8

11. Concluding remarks

In this article, we address survey situations where many alternative auxiliary vectors (**x**-vectors) can be created and considered for use in the calibration estimator $\tilde{Y}_{\text{CAL}}$. For any given **x**-vector, a certain unknown bias remains in $\tilde{Y}_{\text{CAL}}$; we wish by an appropriate choice of **x**-vector to make that bias as small as possible. Hence we examine the bias ratio defined by (4.2) and (4.3). The component Δ_A of the bias ratio was expressed, in (5.8) to (5.10), as product of easily interpreted statistical measures. This led us to suggest several alternative bias indicators, for use in evaluating different **x**-vectors in regard to their capacity to effectively reduce the bias. We studied in particular the indicator H_1 given by (5.12). It functions very well but is geared to a particular study variable y. However, a typical government survey has many study variables, and for practical reasons it is desirable to use the same **x**-vector in estimating all y-totals. A compromise becomes necessary. We argued that

the indicator H_3 in (5.12) suits this purpose; it depends on the $\mathbf{x}_k$ but not on any *y*-data. A topic for further research is to develop other indicators (than H_3) for the "many *y*-variable situation". Another topic for further work is to examine algorithms for stepwise selection of *x*-variables with the indicator H_1, other than the one used in Section 9.

Acknowledgements

The authors are grateful to the referees and to the Associate Editor for comments contributing to an improvement of this paper.

References

Deville, J.-C. (2002). La correction de la nonréponse par calage généralisé. *Actes des Journées de Méthodologie*, I.N.S.E.E., Paris.

Eltinge, J., and Yansaneh, I. (1997). Diagnostics for the formation of nonresponse adjustment cells with an application to income nonresponse in the US Consumer Expenditure Survey. *Survey Methodology*, 23, 33-40.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-98.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.

Rizzo, L., Kalton, G. and Brick, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons, Inc.

Särndal, C.-E., and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, 4, 251-260.

Schouten, B. (2007). A selection strategy for weighting variables under a not-missing-at-random assumption. *Journal of Official Statistics*, 23, 51-68.

Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for the representativeness of survey response. *Survey Methodology*, 35, 101-113.

Thomsen, I., Kleven, Ø., Wang, J.H. and Zhang, L.C. (2006). Coping with deceasing response rates in Statistics Norway. Recommended practice for reducing the effect of nonresponse. Reports 2006/29. Oslo: Statistics Norway.

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 145-155
Statistics Canada, Catalogue No. 12-001-X

Calibration estimation using exponential tilting in sample surveys

Jae Kwang Kim [1]

Abstract

We consider the problem of parameter estimation with auxiliary information, where the auxiliary information takes the form of known moments. Calibration estimation is a typical example of using the moment conditions in sample surveys. Given the parametric form of the original distribution of the sample observations, we use the estimated importance sampling of Henmi, Yoshida and Eguchi (2007) to obtain an improved estimator. If we use the normal density to compute the importance weights, the resulting estimator takes the form of the one-step exponential tilting estimator. The proposed exponential tilting estimator is shown to be asymptotically equivalent to the regression estimator, but it avoids extreme weights and has some computational advantages over the empirical likelihood estimator. Variance estimation is also discussed and results from a limited simulation study are presented.

Key Words: Benchmarking estimator; Empirical likelihood; Instrumental variable calibration; Importance sampling; Regression estimator.

1. Introduction

Consider the problem of estimating $Y = \sum_{i=1}^{N} y_i$ for a finite population of size N. Let A denote the index set of the sample obtained by a probability sampling scheme. In addition to y_i, suppose that we also observe a p-dimensional auxiliary vector $\mathbf{x}_i$ in the sample such that $\mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i$ is known from an external source. We are interested in estimating Y using the auxiliary information $\mathbf{X}$.

The Horvitz-Thompson (HT) estimator of the form

$$\hat{Y}_d = \sum_{i \in A} d_i \, y_i, \tag{1}$$

where $d_i = 1/\pi_i$ is the design weight and π_i is the first order inclusion probability, is unbiased for Y. But, it does not make use of the information given by $\mathbf{X}$. According to Kott (2006), a calibration estimator can be defined as the estimator of the form

$$\hat{Y}_w = \sum_{i \in A} w_i \, y_i$$

where the weights w_i satisfy

$$\sum_{i \in A} w_i \, \mathbf{x}_i = \mathbf{X} \tag{2}$$

and $\hat{Y}_w$ is asymptotically design unbiased (ADU). Calibration estimation has become very popular in survey sampling because it provides consistency across different surveys and often improves the efficiency. (Särndal 2007).

The regression estimator, using the weights

$$w_i = d_i + (\mathbf{X} - \hat{\mathbf{X}}_d)' \left(\sum_{j \in A} d_j \, \mathbf{x}_j \, \mathbf{x}_j' \right)^{-1} d_i \, \mathbf{x}_i, \tag{3}$$

obtained by minimizing

$$\sum_{i \in A} (w_i - d_i)^2 / d_i$$

subject to constraint (2), is asymptotically design unbiased. Note that if an intercept term is included in the column space of X matrix then (2) implies that the population size N is known. If N is unknown, one can require that the sum of the final weights are equal to the sum of the design weights. Thus,

$$\sum_{i \in A} w_i = \hat{N}, \tag{4}$$

where

$$\hat{N} = \begin{cases} N & \text{if } N \text{ is known} \\ \sum_{i \in A} d_i & \text{otherwise,} \end{cases}$$

can be imposed as a constraint in addition to (2), which yields the weights

$$w_i = \frac{\hat{N}}{\hat{N}_d} d_i + \left(\mathbf{X} - \frac{\hat{N}}{\hat{N}_d} \hat{\mathbf{X}}_d \right)' \left\{ \sum_{j \in A} d_j (\mathbf{x}_j - \bar{\mathbf{X}}_d) (\mathbf{x}_j - \bar{\mathbf{X}}_d)' \right\}^{-1} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d), \tag{5}$$

where $\hat{\mathbf{X}}_d = \sum_{i \in A} d_i \, \mathbf{x}_i$, $\hat{N}_d = \sum_{i \in A} d_i$, and $\bar{\mathbf{X}}_d = \hat{\mathbf{X}}_d / \hat{N}_d$. We define the regression estimator to be $\hat{Y}_{\text{reg}} = \sum_{i \in A} w_i \, y_i$ using the weights (5). The regression estimator can be efficient if y_i is linearly related with $\mathbf{x}_i$ (Isaki and Fuller 1982; Fuller 2002), but the weights in the regression estimator can take negative or extremely large values.

1. Jae Kwang Kim, Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A. E-mail: jkim@iastate.edu.

The empirical likelihood (EL) calibration estimator, discussed by Chen and Qin (1993), Chen and Sitter (1999), Wu and Rao (2006), and Kim (2009), is obtained by maximizing the pseudo empirical likelihood

$$\sum_{i\in A} d_i \ln(w_i)$$

subject to constraints (2) and (4). The solution to the optimization problem can be written as

$$w_i = d_i \frac{1}{\lambda_0 + \boldsymbol{\lambda}_1'(\mathbf{x}_i - \mathbf{X}/\hat{N})}, \tag{6}$$

where λ_0 and $\boldsymbol{\lambda}_1$ satisfy constraints (2), (4), and $w_i > 0$ for all i. The EL calibration estimator is asymptotically equivalent to the regression estimator using weights (5) and avoids negative weights if a solution exists, but can result in extremely large weights.

Because the empirical likelihood method requires solving nonlinear equations, the computation can be cumbersome. Furthermore, in some extreme cases, $\bar{\mathbf{X}} = N^{-1}\sum_{i=1}^N \mathbf{x}_i$ does not belong to the convex hull of the sample $\mathbf{x}_i$'s and the solution does not exist. In this extreme situation, the constraint (2) can be relaxed.

Rao and Singh (1997) solved a similar problem by allowing

$$\left|\sum_{i\in A} w_i x_{ij} - X_j\right| \le \delta_j X_j, \quad j = 1, 2, ..., p,$$

for some small tolerance level $\delta_j > 0$ where $X_j = \sum_{i=1}^N x_{ij}$. Note that the choice of $\delta_j = 0$ leads to the exact calibration condition (2). Rao and Singh (1997) chose the tolerance level δ_j using a shrinkage factor in the ridge regression but their approach does not directly apply to the empirical likelihood method and the choice of δ_j is somewhat unclear. Chambers (1996) and Beaumont and Bocci (2008) also discussed a ridge regression estimation in the context of avoiding extreme weights. Breidt, Claeskens and Opsomer (2005) used penalized spline approach to obtain the ridge calibration. Recently, Park and Fuller (2009) developed a method of obtaining the shrinkage factor δ_j using a regression superpopulation model with random components.

Chen, Variyath and Abraham (2008) tackled a similar problem in the context of the empirical likelihood method and proposed a solution by adding an artificial point such that $\bar{\mathbf{X}} = N^{-1}\sum_{i=1}^N \mathbf{x}_i$ would belong to the convex hull of the augmented $\mathbf{x}_i$'s. The proposed estimator in Chen *et al.* (2008) only satisfies the calibration property approximately in the sense that

$$\sum_{i\in A} w_i \mathbf{x}_i - \mathbf{X} = o_p(n^{-1/2}N). \tag{7}$$

This approximate calibration property is attractive because it allows more generality in the choice of weights. In particular, when the dimension of the auxiliary variable $\mathbf{x}$ is large the calibration constraint (2) can be quite restrictive. As can be seen in Section 2, an estimator satisfying the asymptotic calibration property (7) enjoys most of the desirable properties of the empirical likelihood calibration estimator and is computationally efficient.

In this paper, we consider a class of empirical-likelihood-type estimators that satisfy the approximate calibration property (7). In Section 2, the idea of estimated importance sampling of Henmi *et al.* (2007) is discussed and a new estimator using this methodology is proposed. In Section 3, a weight trimming technique to avoid extreme calibration weights is proposed. In Section 4, variance estimation of the proposed estimator is discussed. In Section 5, results from a simulation study are presented. Concluding remarks are made in Section 6.

2. Proposed method

To introduce the proposed method, we first discuss estimated importance sampling introduced by Henmi *et al.* (2007). Suppose that $\mathbf{x}_i$ is observed throughout the population but y_i is observed only in the sample. We assume a superpopulation model for $\mathbf{x}_i$ with density $f(\mathbf{x};\boldsymbol{\eta})$ known up to a parameter $\boldsymbol{\eta} \in \Omega$. The superpopulation model characterized by the density $f(\mathbf{x};\boldsymbol{\eta})$ is a working model in the sense that the model is used to derive a model-assisted estimator (Särndal, Swenson and Wretman 1992).

Let $\hat{\boldsymbol{\eta}}$ be the pseudo maximum likelihood estimator of $\boldsymbol{\eta}$ computed from the sample

$$\hat{\boldsymbol{\eta}} = \arg\max_{\Omega} \sum_{i\in A} d_i \ln\{f(\mathbf{x}_i;\boldsymbol{\eta})\}$$

and let $\boldsymbol{\eta}_{0,N}$ be the maximum likelihood estimator of $\boldsymbol{\eta}$ computed from the population

$$\boldsymbol{\eta}_{0,N} = \arg\max_{\Omega} \sum_{i=1}^N \ln\{f(\mathbf{x}_i;\boldsymbol{\eta})\}.$$

Following Henmi *et al.* (2007), we can construct the following estimated importance weight

$$w_i = d_i \frac{f(\mathbf{x}_i;\boldsymbol{\eta}_{0N})}{f(\mathbf{x}_i;\hat{\boldsymbol{\eta}})}. \tag{8}$$

To discuss the asymptotic properties of the estimator using the weights in (8), assume a sequence of the finite populations and the samples, as in Isaki and Fuller (1982), such that

$$\sum_{i\in A} d_i(\mathbf{x}_i', y_i)'\,(\mathbf{x}_i', y_i) - \sum_{i=1}^{N}(\mathbf{x}_i', y_i)'\,(\mathbf{x}_i', y_i) = O_p(n^{-1/2}N)$$

for all possible A and for each N. The following theorem presents some asymptotic properties of the estimator with the estimated importance weights in (8).

Theorem 1. *Under the regularity conditions given in Appendix A, the estimator* $\hat{Y}_w = \sum_{i\in A} w_i y_i$, *with the* w_i *defined by* (8), *satisfies*

$$\sqrt{n}N^{-1}(\hat{Y}_w - \hat{Y}_l) = o_p(1), \tag{9}$$

where

$$\hat{Y}_l = \hat{Y}_d - \hat{\boldsymbol{\Sigma}}_{sy}'\,\hat{\boldsymbol{\Sigma}}_{ss}^{-1}\,\hat{\mathbf{S}}_{0d}, \tag{10}$$

$\hat{Y}_d$ *is defined in* (1), $\hat{\mathbf{S}}_{0d} = \sum_{i\in A} d_i \mathbf{s}_{i0}, \hat{\boldsymbol{\Sigma}}_{sy} = N^{-1}\sum_{i\in A} d_i\,\mathbf{s}_{i0}\,y_i$, *and* $\hat{\boldsymbol{\Sigma}}_{ss} = N^{-1}\sum_{i\in A} d_i\,\mathbf{s}_{i0}^{\otimes 2}$. *Here,* $\mathbf{s}_{i0} = \partial \ln f(\mathbf{x}_i;\boldsymbol{\eta})/\partial\boldsymbol{\eta}|_{\boldsymbol{\eta}=\boldsymbol{\eta}_{0,N}}$ *and the notation* $B^{\otimes 2}$ *denotes* BB'.

The proof of Theorem 1 is presented in Appendix A. Because $\mathbf{S}_{0N} \equiv \sum_{i=1}^{N}\mathbf{s}_{i0} = \mathbf{0}$, we can write (10) as

$$\hat{Y}_l = \hat{Y}_d + \hat{\boldsymbol{\Sigma}}_{sy}'\hat{\boldsymbol{\Sigma}}_{ss}^{-1}(\mathbf{S}_{0N} - \hat{\mathbf{S}}_{0d}),$$

which is a regression estimator of Y using $\mathbf{s}_i(\boldsymbol{\eta}_{0N})$ as the auxiliary variable. Therefore, under regularity conditions, the proposed estimator using estimated importance sampling is asymptotically unbiased and has asymptotic variance no greater than that of the direct estimator $\hat{Y}_d$. Note that the validity of Theorem 1 does not require that the working model $f(\mathbf{x};\boldsymbol{\eta})$ be true.

If the density of $\mathbf{x}_i$ is a multivariate normal density, then the weights in (8) become

$$w_i = d_i\,\frac{\phi(\mathbf{x}_i;\,\bar{\mathbf{X}}_N,\,\boldsymbol{\Sigma}_{xx,N})}{\phi(\mathbf{x}_i;\,\bar{\mathbf{X}}_d,\,\hat{\boldsymbol{\Sigma}}_{xx,d})}, \tag{11}$$

where $\bar{\mathbf{X}}_d$ is defined after (5), $\hat{\boldsymbol{\Sigma}}_{xx,d} = \sum_{i\in A} d_i(\mathbf{x}_i - \bar{\mathbf{X}}_d)^{\otimes 2}/\hat{N}_d$, $\boldsymbol{\Sigma}_{xx,N} = \sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{X}}_N)^{\otimes 2}/N$, and $\phi(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the density of the multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. If $\boldsymbol{\Sigma}_{xx,N}$ is unknown and only $\bar{\mathbf{X}}_N$ is available, then we can use

$$w_i = d_i\,\frac{\phi(\mathbf{x}_i;\,\bar{\mathbf{X}}_N,\,\hat{\boldsymbol{\Sigma}}_{xx,d})}{\phi(\mathbf{x}_i;\,\bar{\mathbf{X}}_d,\,\hat{\boldsymbol{\Sigma}}_{xx,d})}. \tag{12}$$

Tillé (1998) derived weights similar to those in (12) in the context of conditional inclusion probabilities.

In general, the parametric model for $\mathbf{x}_i$ is unknown. Thus, we consider an approximation for the importance weights in (8) using the Kullback-Leibler information criterion for distance. Let $f(\mathbf{x})$ be a given density for $\mathbf{x}$ and let P_0 be the set of densities that satisfy the calibration constraint. That is,

$$P_0 = \left\{ f_0(\mathbf{x});\ \int f_0(\mathbf{x})\,d\mathbf{x} = 1,\ \int \mathbf{x} f_0(\mathbf{x})\,d\mathbf{x} = \bar{\mathbf{X}}_N \right\}.$$

The optimization problem using Kullback-Leibler distance can be expressed as

$$\min_{f_0\in P_0} \int f_0(\mathbf{x}) \ln\left\{\frac{f_0(\mathbf{x})}{f(\mathbf{x})}\right\} d\mathbf{x}. \tag{13}$$

The solution to (13) is

$$f_0(\mathbf{x}) = f(\mathbf{x})\,\frac{\exp(\hat{\boldsymbol{\lambda}}'\mathbf{x})}{E\{\exp(\hat{\boldsymbol{\lambda}}'\mathbf{x})\}} \tag{14}$$

where $\hat{\boldsymbol{\lambda}}$ satisfies $\int \mathbf{x} f_0(\mathbf{x})\,d\mathbf{x} = \bar{\mathbf{X}}_N$. Thus, the estimated importance weights in (8) using the optimal density in (14) can be written

$$w_i = d_i\,\frac{f_0(\mathbf{x}_i)}{f(\mathbf{x}_i)} = d_i \exp(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1'\mathbf{x}_i) \tag{15}$$

where $\hat{\lambda}_0$ and $\hat{\boldsymbol{\lambda}}_1$ satisfy constraint (2) and (4). The shift from $f(\mathbf{x})$ to $f_0(\mathbf{x})$ in (14) is called exponential tilting. Thus, an estimator using the weight (15) satisfying the calibration constraints (2) and (4) can be called an exponential tilting (ET) calibration estimator. That is, we define the ET calibration estimator as

$$\hat{Y}_{\mathrm{ET}} = \sum_{i\in A} d_i \exp(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1'\mathbf{x}_i)\,y_i, \tag{16}$$

where $\hat{\lambda}_0$ and $\hat{\boldsymbol{\lambda}}_1$ satisfy constraint (2) and (4). Estimators based on exponential tilting have been used in various contexts. For examples, see Efron (1981), Kitamura and Stutzer (1997), and Imbens (2002). When N is known, Folsom (1991) and Deville, Särndal and Sautory (1993) developed the estimator (16) using a very different approach.

To compute λ_0 and $\boldsymbol{\lambda}_1$ in (16), because of the calibration constraints (2) and (4), we need to solve the following estimating equations:

$$\hat{U}_0(\boldsymbol{\lambda}) \equiv \sum_{i\in A} d_i \exp(\lambda_0 + \boldsymbol{\lambda}_1'\,\mathbf{x}_i) - \hat{N} = 0 \tag{17}$$

$$\hat{\mathbf{U}}_1(\boldsymbol{\lambda}) \equiv \sum_{i\in A} d_i \exp(\lambda_0 + \boldsymbol{\lambda}_1'\mathbf{x}_i)\,\mathbf{x}_i - \mathbf{X} = \mathbf{0}, \tag{18}$$

where $\boldsymbol{\lambda}' = (\lambda_0,\ \boldsymbol{\lambda}_1')$. Writing $\hat{\mathbf{U}}' = (\hat{U}_0,\ \hat{\mathbf{U}}_1')$, we can use the Newton-type algorithm of the form

$$\hat{\boldsymbol{\lambda}}_{(t+1)} = \hat{\boldsymbol{\lambda}}_{(t)} - \left\{\frac{\partial}{\partial\boldsymbol{\lambda}'}\hat{\mathbf{U}}(\hat{\boldsymbol{\lambda}}_{(t)})\right\}^{-1}\hat{\mathbf{U}}(\hat{\boldsymbol{\lambda}}_{(t)})$$

and the solution can be written

$$\hat{\boldsymbol{\lambda}}_{1(t+1)} = \hat{\boldsymbol{\lambda}}_{1(t)} + \left\{\sum_{i\in A} w_{i(t)}(\mathbf{x}_i - \bar{\mathbf{X}}_{w(t)})^{\otimes 2}\right\}^{-1}\left(\mathbf{X} - \sum_{i\in A} w_{i(t)}\mathbf{x}_i\right), \tag{19}$$

where $w_{i(t)}= d_i \exp(\hat{\lambda}_{0(t)}+\hat{\boldsymbol{\lambda}}'_{1(t)}\mathbf{x}_i)$ and $\bar{\mathbf{X}}_{w(t)}=\sum_{i\in A} w_{i(t)}\mathbf{x}_i/\sum_{i\in A} w_{i(t)}$, with the initial values $\hat{\boldsymbol{\lambda}}_{1(0)}=\mathbf{0}$. Once $\hat{\boldsymbol{\lambda}}_{1(t)}$ is computed by (19), $\hat{\lambda}_{0(t)}$ is computed by

$$\exp(\hat{\lambda}_{0(t)}) = \frac{\hat{N}}{\sum_{i\in A} d_i \exp(\hat{\boldsymbol{\lambda}}'_{1(t)}\mathbf{x}_i)}. \tag{20}$$

Note that, $w_{i(0)} = d_i\hat{N}/\hat{N}_d$ since $\hat{\boldsymbol{\lambda}}_{1(0)}=\mathbf{0}$. Because $\hat{\mathbf{U}}(\boldsymbol{\lambda})$ is twice continuously differentiable and convex in $\boldsymbol{\lambda}$, the sequence $\hat{\boldsymbol{\lambda}}_{(t)}$ always converges if the solution to $\hat{\mathbf{U}}(\boldsymbol{\lambda})=\mathbf{0}$ exists (Givens and Hoeting 2005). The convergence rate is quadratic in the sense that

$$|\hat{\boldsymbol{\lambda}}_{1(t+1)} - \hat{\boldsymbol{\lambda}}_1| \le C|\hat{\boldsymbol{\lambda}}_{1(t)} - \hat{\boldsymbol{\lambda}}_1|^2$$

for some constant C, where $\hat{\boldsymbol{\lambda}}_1 = \lim_{t\to\infty}\hat{\boldsymbol{\lambda}}_{1(t)}$.

By construction, the t-step exponential tilting (ET) estimator, defined by

$$\hat{Y}_{\mathrm{ET}(t)} = \sum_{i\in A} d_i \exp(\hat{\lambda}_{0(t)} + \hat{\boldsymbol{\lambda}}'_{1(t)}\mathbf{x}_i)\, y_i \tag{21}$$

where $\hat{\lambda}_{0(t)}$ and $\hat{\boldsymbol{\lambda}}_{1(t)}$ are computed by (19) and (20), satisfies the calibration constraint (2) for sufficiently large t. By the recursive form in (19) with $\hat{\boldsymbol{\lambda}}_{1(0)}=\mathbf{0}$, we can write

$$\hat{\boldsymbol{\lambda}}_{1(t)} = \sum_{j=0}^{t-1} (\mathbf{S}_{xx,w(j)})^{-1}(\tilde{\mathbf{X}}_N - \bar{\mathbf{X}}_{w(j)}), \tag{22}$$

where $\tilde{\mathbf{X}}_N = \mathbf{X}/\hat{N}$ and $\mathbf{S}_{xx,w(j)} = \sum_{i\in A} w_{i(t)}(\mathbf{x}_i - \bar{\mathbf{X}}_{w(t)})^{\otimes 2}/\hat{N}$. Thus, the t-step ET estimator (21) can be written as

$$\hat{Y}_{\mathrm{ET}(t)} = \hat{N}\frac{\sum_{i\in A} d_i g_{i(t)} y_i}{\sum_{i\in A} d_i g_{i(t)}},$$

where

$$g_{i(t)} = \prod_{j=0}^{t-1} \frac{\phi\left(\mathbf{x}_i; \tilde{\mathbf{X}}_N, \mathbf{S}_{xx,w(j)}\right)}{\phi\left(\mathbf{x}_i; \bar{\mathbf{X}}_{w(j)}, \mathbf{S}_{xx,w(j)}\right)}.$$

The following theorem presents some asymptotic properties of the exponential tilting estimator.

Theorem 2. The *t-step ET estimator* (21) *based on equations* (19) *and* (20) *satisfies*

$$\sqrt{n}N^{-1}(\hat{Y}_{\mathrm{ET}(t)} - \hat{Y}_{\mathrm{reg}}) = o_p(1), \tag{23}$$

for each $t = 1, 2, \ldots$, *where* $\hat{Y}_{\mathrm{reg}}$ *is the regression estimator using the regression weight in* (5).

The proof of Theorem 2 is presented in Appendix B. Theorem 2 presents the asymptotic equivalence between the t-step ET estimator and the regression estimator. Unlike the regression estimator, the weights of the ET estimator are always positive. For sufficiently large t, the t-step ET estimator satisfies the calibration constraint (2). Deville and Särndal (1992) proved the result (23) for the special case of $t\to\infty$.

Remark 1. *The one-step ET estimator, defined by* $\hat{Y}_{\mathrm{ET}(1)}$, *has a closed-form tilting parameter*

$$\hat{\boldsymbol{\lambda}}_{1(1)} = \left\{\sum_{i\in A} d_i(\mathbf{x}_i - \bar{\mathbf{X}}_d)^{\otimes 2} \Big/ \hat{N}_d\right\}^{-1}(\tilde{\mathbf{X}}_N - \bar{\mathbf{X}}_d), \tag{24}$$

where $\tilde{\mathbf{X}}_N = \mathbf{X}/\hat{N}$ *and* $\bar{\mathbf{X}}_d = \sum_{i\in A} d_i\mathbf{x}_i/\sum_{i\in A} d_i$. *By Theorem* 2, *the one-step ET estimator is asymptotically equivalent to the regression estimator, but the calibration constraint* (2) *is not necessarily satisfied. Using Theorem* 2 *applied to* $\mathbf{x}_i$ *instead of* y_i, *the one-step ET estimator can be shown to satisfy the approximate calibration constraint described in* (7).

Remark 2. *The ET estimator can also be derived by finding the weights that minimize*

$$Q(w) = \sum_{i\in A} w_i \ln\left(\frac{w_i}{d_i}\right) \tag{25}$$

subject to constraints (2) *and* (4). *The objective function* (25) *is often called the minimum discrimination function. The minimum value of* $Q(w)$ *is zero if* (4) *is the only calibration constraint and is monotonically increasing if additional calibration constraints are imposed.*

3. Instrumental-variable calibration

We consider some extension of the proposed method in Section 2 to a more general class of ET calibration estimator using instrumental-variables. Use of instrumental-variable in the calibration estimation has been discussed in Estevao and Särndal (2000) and Kott (2003) in some limited simulations. Let $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i)$ be an instrumental-variable derived from $\mathbf{x}_i$, where the function $\mathbf{z}(\cdot)$ is to be determined. The instrumental-variable exponential tilting (IVET) estimator using the instrumental variable $\mathbf{z}_i$ can be defined as

$$\hat{Y}_{\mathrm{IVET}} = \sum_{i\in A} w_i y_i = \sum_{i\in A} d_i \exp(\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}'_1\mathbf{z}_i)\, y_i, \tag{26}$$

where $\hat{\lambda}_0$ and $\hat{\boldsymbol{\lambda}}_1$ are computed from (2) and (4). Note that the IVET estimator (26) is a class of estimators indexed by $\mathbf{z}_i$. The instrumental-variable approach defined in (26) provides more flexibility in creating the ET estimator. The choice of $\mathbf{z}_i = \mathbf{x}_i$ leads to the standard ET estimator in (16) but some transformation $\mathbf{z}_i = \mathbf{z}(\mathbf{x}_i)$ can make the resulting ET estimator in (26) more attractive in practice. The solution to the calibration equations can be obtained iteratively by

$$\hat{\boldsymbol{\lambda}}_{1(t+1)} = \hat{\boldsymbol{\lambda}}_{1(t)} + \left\{\sum_{i\in A} w_{i(t)}(\mathbf{x}_i - \bar{\mathbf{X}}_{w(t)})(\mathbf{z}_i - \bar{\mathbf{Z}}_{w(t)})'\right\}^{-1} \left(\mathbf{X} - \sum_{i\in A} w_{i(t)}\mathbf{x}_i\right), \tag{27}$$

where $w_{i(t)} = d_i \exp(\hat{\lambda}_{0(t)} + \hat{\boldsymbol{\lambda}}'_{1(t)}\mathbf{z}_i)$ and $\bar{\mathbf{Z}}_{w(t)} = \sum_{i\in A} w_{i(t)}\mathbf{z}_i/\sum_{i\in A} w_{i(t)}$, with equation (20) unchanged and $\hat{\boldsymbol{\lambda}}_{1(0)}=\mathbf{0}$.

The IVET estimator (26) is useful in creating the final weights that have less extreme values. Since the final weight in (26) is a function of $\mathbf{z}_i$, we can make $g_i = w_i/d_i$ bounded by making $\mathbf{z}_i$ bounded. To create bounded $\mathbf{z}_i$, we can use a trimmed version of $\mathbf{x}_i$, noted by $\mathbf{z}_i = (z_{i1}, z_{i2}, ..., z_{ip})$, where

$$z_{ij} = \begin{cases} x_{ij} & \text{if } |x_{ij} - \bar{x}_j| \le C_j S_j \\ \bar{x}_j + C_j S_j & \text{if } x_{ij} > \bar{x}_j + C_j S_j \\ \bar{x}_j - C_j S_j & \text{if } x_{ij} < \bar{x}_j - C_j S_j, \end{cases} \quad (28)$$

$\bar{x}_j = N^{-1}\sum_{i\in A} d_i x_{ij}$, $S_j^2 = N^{-1}\sum_{i\in A} d_i (x_{ij} - \bar{x}_j)^2$, and C_j is a threshold for detecting outliers, for example, $C_j = 3$. Thus, the IVET estimator using the instrumental-variable obtained by trimming $\mathbf{x}_i$ can be used as an alternative approach to weight trimming.

Instead of using the trimmed instrumental variable $\mathbf{z}_i$ in (28), we can consider the following instrumental variable

$$\mathbf{z}_i = \mathbf{x}_i \Phi_i$$

for some symmetric matrix Φ_i such that $\mathbf{z}_i$ is bounded. Some suitable choice of Φ_i can also improve the efficiency of the resulting IVET estimator. To see this, using the same argument from Theorem 2, the instrumental-variable ET estimator (26) using equations (20) and (27) is asymptotically equivalent to

$$\hat{Y}_{\text{IV, reg}} = \tilde{Y}_d + (\mathbf{X} - \tilde{\mathbf{X}}_d)' \hat{\mathbf{B}}_z \quad (29)$$

where

$$(\tilde{\mathbf{X}}_d', \tilde{Y}_d) = \left(\frac{\hat{N}}{\hat{N}_d}\right)(\hat{\mathbf{X}}_d', \hat{Y}_d)$$

and

$$\hat{\mathbf{B}}_z = \left\{\sum_{i\in A} d_i (\mathbf{z}_i - \bar{\mathbf{Z}}_d)(\mathbf{x}_i - \bar{\mathbf{X}}_d)'\right\}^{-1} \sum_{i\in A} d_i (\mathbf{z}_i - \bar{\mathbf{Z}}_d) y_i. \quad (30)$$

The estimator (29) takes the form of a regression estimator and is called the instrumental-variable regression estimator. Thus, under the choice of $\mathbf{z}_i = \Phi_i \mathbf{x}_i$, the instrumental-variable regression estimator can be written as (29) with

$$\hat{\mathbf{B}}_z = \left\{\sum_{i\in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d)\Phi_i (\mathbf{x}_i - \bar{\mathbf{X}}_d)'\right\}^{-1} \sum_{i\in A} d_i (\mathbf{x}_i - \bar{\mathbf{X}}_d)\Phi_i y_i$$

and its variance is minimized for $\Phi_i = V_i^{-1}$ where V_i is the model-variance of y_i given $\mathbf{x}_i$ (Fuller 2009). The model-variance is the variance under the working superpopulation model for the regression of y_i on $\mathbf{x}_i$. Thus, instrumental-variable can be used to improve the efficiency of the resulting calibration estimator, in addition to avoid extreme final weights. Furthermore, the optimal instrumental-variable can be trimmed as in (28) to make the final weightsbounded. Further investigation of the optimal choice of Φ is beyond the scope of this paper and will be a topic of future research.

Remark 3. *Deville and Särndal* (1992) *also considered range-restricted calibration weights of the form*

$$w_i = d_i g_i(\hat{\boldsymbol{\lambda}}) = d_i \frac{L(U-1) + U(1-L)\exp(K\hat{\boldsymbol{\lambda}}'\mathbf{x}_i)}{(U-1) + (1-L)\exp(K\hat{\boldsymbol{\lambda}}'\mathbf{x}_i)}, \quad (31)$$

where $K = (U-L)/\{(1-L)(U-1)\}$, *for some* L *and* U *such that* $0 < L < 1 < U$. *If calibration constraints* (2) *and* (4) *are to be satisfied, then we can use* $\hat{\lambda}_0 + \hat{\boldsymbol{\lambda}}_1' \mathbf{x}_i$ *instead of* $\hat{\boldsymbol{\lambda}}'\mathbf{x}_i$ *in* (31). *The resulting calibration estimator is asymptotically equivalent to the regression estimator using the weights in* (5) *while the IVET estimator is asymptotically equivalent to the instrumental-variable regression estimator* (29). *Computation for obtaining* $\hat{\boldsymbol{\lambda}}$ *is somewhat complicated because* $\partial g_i(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}$ *is not easy to evaluate in* (31). *In the IVET estimator*, *the computation*, *given by* (27), *is straightforward.*

To compare the proposed weight with existing methods, we consider an artificial example of a simple random sample with size $n = 5$ where $x_k = k$, $k = 1, 2, ..., 5$. Calculations are for three population means of x; $\bar{X}_N = 3$, $\bar{X}_N = 4.5$, and $\bar{X}_N = 6$. Table 1 presents the resulting weights for the regression estimator, the empirical likelihood (EL) estimator, the t-step ET estimator (16) with $t = 1$ and $t = 10$, and the t-step instrumental variable exponential tilting (IVET) estimator (26) with $t = 1$ and $t = 10$. For the IVET estimator, the instrumental variable z_i is created by

$$z_i = \begin{cases} 1.5 & \text{if } x_i \le 1.5 \\ x_i & \text{if } x_i \in (1.5,\ 4.5) \\ 4.5 & \text{if } x_i \ge 4.5. \end{cases}$$

The last column of Table 1 presents the estimated mean of X using the respective calibration weights. All the weights are equal to $1/n = 0.2$ for $\bar{X}_N = 3$. The regression estimator is linearly increasing in x_i but has negative weights for the population with $\bar{X}_N = 4.5$ and $\bar{X}_N = 6$. For the population where $\bar{X}_N = 6$, the weights could not be computed for the EL method because $\bar{X}_N$ is outside the range of the sample x_i's. In this extreme case of $\bar{X}_N = 6$, the ET method provides nonnegative weights by sacrificing the calibration constraint and the EL estimator has more extreme weights than the ET estimator or IVET estimator in the sense that the weight for $k = 5$ is the largest among the estimators considered. The weight for the one-step ET estimator is close to that of the regression estimator for large x_i but it is close to that of EL estimator for small x_i. The 10-step ET estimators has better calibration properties in the sense of smaller value of squared error, $(\sum_{k=1}^5 w_k x_k - \bar{X}_N)^2$, than the one-step ET estimator. The ET estimator and the IVET estimator provide almost the same estimates of $\bar{X}_N$ for both t, but the IVET estimator produces less extreme weights than the ET estimator.

Table 1
An example of calibration weights with a sample of size $n = 5$

Method	$\bar{X}_N$	x_i 1	2	3	4	5	$\hat{X}_N$
Reg.	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	-0.100	0.050	0.200	0.035	0.500	4.5
	6.0	-0.400	-0.100	0.200	0.500	0.800	6.0
EL	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.033	0.043	0.063	0.115	0.746	4.5
	6.0	N/A	N/A	N/A	N/A	N/A	N/A
ET ($t = 1$)	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.027	0.057	0.100	0.255	0.540	4.2
	6.0	0.002	0.009	0.039	0.173	0.777	4.7
ET ($t = 10$)	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.009	0.027	0.078	0.227	0.659	4.5
	6.0	0.000	0.000	0.000	0.001	0.999	5.0
IVET ($t = 1$)	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.030	0.047	0.121	0.309	0.493	4.2
	6.0	0.003	0.006	0.041	0.267	0.683	4.6
IVET ($t = 10$)	3.0	0.200	0.200	0.200	0.200	0.200	3.0
	4.5	0.007	0.015	0.066	0.294	0.618	4.5
	6.0	0.000	0.000	0.000	0.087	0.913	4.9

Reg., Regression estimator; EL, empirical likelihood; ET, exponential tilting; IVET, instrumental variable exponential tilting; N/A, Not applicable.

4. Variance estimation

We now discuss variance estimation of the ET calibration estimators of Sections 2 and 3. Because the estimated parameter $(\hat{\lambda}_0, \hat{\boldsymbol{\lambda}}_1')$ in the ET calibration estimator (16) has some sampling variability, variance estimation method should take into account of this sampling variability of these estimated parameters. In this case, variance estimation can be often obtained by a linearization method or by a replication method (Wolter 2007). For the discussion of the linearization method, let the variance of the HT estimator (1) be consistently estimated by

$$\hat{V}(\hat{Y}_d) = \sum_{i\in A}\sum_{j\in A} \Omega_{ij} y_i y_j. \tag{32}$$

The linearization variance estimator for the ET estimator can be obtained by the linearization variance formula for the regression estimator, as in Deville and Särndal (1992), using the asymptotic equivalence between the ET calibration estimator andthe regression estimator, as shown in Theorem 2. Specifically, if the population size N is known, a linearization variance estimator of the IVET estimator in (26) can be written as

$$\hat{V}(\hat{Y}_{\text{IVET}}) = \sum_{i\in A}\sum_{j\in A} \Omega_{ij}\, g_i\, g_j\, \hat{e}_i\, \hat{e}_j \tag{33}$$

where Ω_{ij} are the coefficients of the variance estimator in (32), $g_i = w_i/d_i$ is the weight adjustment factor, and $\hat{e}_i = y_i - \bar{Y}_d - (\mathbf{x}_i - \bar{\mathbf{X}}_d)'\hat{\mathbf{B}}_z$, where $\hat{\mathbf{B}}_z$ is defined in (30). The choice of $\mathbf{z}_i = \mathbf{x}_i$ in (33) gives the linearized variance estimator for the ET estimator in (16). Consistency of the variance estimator (33) can be found in Kim and Park (2010).

For the one-step ET estimator, a replication method can be easily implemented. Let the replication variance estimator be of the form

$$\hat{V}_{\text{rep}} = \sum_{k=1}^{L} c_k (\hat{Y}_d^{(k)} - \hat{Y}_d)^2, \tag{34}$$

where L is the number of replication, c_k is the replication factor associated with replicate k, $\hat{Y}_d^{(k)} = \sum_{i\in A} d_i^{(k)} y_i$, and $d_i^{(k)}$ is the k^{th} replicate of the design weight d_i. For example, the replication variance estimator (34) includes the jackknife and the bootstrap (see Rust and Rao 1996). Assume that the replication variance estimator (34) is a consistent estimator for the variance of $\hat{Y}_d$. The k^{th} replicate of the one-step ET estimator can be computed by

$$\hat{Y}_{\text{ET}(1)}^{(k)} = \sum_{i\in A} d_i^{(k)} \exp(\hat{\lambda}_{0(1)}^{(k)} + \hat{\boldsymbol{\lambda}}_{1(1)}^{(k)'}\mathbf{z}_i)\, y_i \tag{35}$$

where

$$\hat{\boldsymbol{\lambda}}_{1(1)}^{(k)} = \left\{\sum_{i\in A} d_i^{(k)} (\mathbf{x}_i - \bar{\mathbf{X}}_d^{(k)})\,(\mathbf{z}_i - \bar{\mathbf{Z}}_d^{(k)})' \Big/ \hat{N}_d^{(k)}\right\}^{-1} (\mathbf{X}/\hat{N}^{(k)} - \bar{\mathbf{X}}_d^{(k)}),$$

$$\hat{N}^{(k)} = \begin{cases} N & \text{if } \hat{N} = N \\ \hat{N}_d^{(k)} = \sum_{i\in A} d_i^{(k)} & \text{if } \hat{N} = \hat{N}_d, \end{cases}$$

$$(\bar{\mathbf{X}}_d^{(k)},\ \bar{\mathbf{Z}}_d^{(k)}) = \frac{\sum_{i\in A} d_i^{(k)} (\mathbf{x}_i,\ \mathbf{z}_i)}{\sum_{i\in A} d_i^{(k)}},$$

and

$$\exp(\hat{\lambda}_{0(1)}^{(k)}) = \frac{\hat{N}}{\sum_{i\in A} d_i^{(k)} \exp(\mathbf{z}'\hat{\boldsymbol{\lambda}}_{1(1)}^{(k)})}.$$

The replication variance estimator defined by

$$\hat{V}_{\text{rep}} = \sum_{k=1}^{L} c_k (\hat{Y}_{\text{ET}}^{(k)} - \hat{Y}_{\text{ET}})^2, \tag{36}$$

where $\hat{Y}_{\text{ET}}^{(k)}$ is defined in (35), can be used to estimate the variance of the ET calibration estimator in (26).

5. Simulation study

To study the finite sample performance of the proposed estimators, we performed a limited simulation study. In the simulation, two finite populations of size $N = 10{,}000$ were independently generated. In population A, the finite population is generated from an infinite population specified by $x_i \sim \exp(1) + 1$; $y_i = 3 + x_i + x_i e_i$, $e_i \mid x_i \sim N(0, 1)$; $z_i \mid (x_i, y_i) \sim \chi^2(1) + \mid y_i \mid$. In population B, (x_i, e_i, z_i) are the same as in population A but $y_i = (5 - 1/\sqrt{8}) + 1/\sqrt{8}(x_i - 2)^2 + e_i$. The auxiliary variable, x_i, is used for calibration and z_i is the measure of size used for unequal probability sampling. From both of the finite populations generated, $M = 10{,}000$ Monte Carlo samples of size n were independently generated under two sampling schemes described below. The parameter of interest is the population mean of y and we assume that the population size N is known.

The simulation setup can be described as a $2 \times 2 \times 8 \times 2$ factorial design with four factors. The factors are (a) two types of finite populations, (b) Sampling mechanism: simple random sampling and probability proportional to size (z_i) sampling with replacement, (c) Calibration method: no calibration, the regression estimator, the EL method in (6) with $t = 1$ and $t = 10$, the t-step ET method in (21) with $t = 1$ and $t = 10$, and the IVET method (26) with $t = 1$ and $t = 10$, (d) sample size: $n = 100$ and $n = 200$. Since N is assumed to be known, the calibration estimators are computed to satisfy $\sum_{i=1}^{n} w_i(1, x_i) = (1, \bar{X}_N)$ in both populations. For the IVET method (26), the instrumental variable z_i is created using the definitions in (28) with threshold $C = 3$.

Using the Monte Carlo samples generated as above, the biases and the mean squared errors of the eight estimators of the population mean of y, the variable of interest, were computed and are presented in Table 2. The calibration estimators are biased but the bias is small if the regression model holds or the sample size is large. In population A, the linear regression model holds and the regression estimator is efficient in terms of mean squared errors. However, the regression estimator is not efficient in population B because the model used for the regression estimator is not a good fit. The seven calibration estimators show similar performances for the larger sample size. The 10-step IVET estimator performs as well as the regression estimator in population A, and it shows slightly better performance than the other six calibration estimators. In population B, the 10-step IVET estimator performs the best among the calibration estimators considered.

In addition to point estimation, variance estimation was also considered. We considered only the variance estimation for the t-step ET estimators and IVET estimators. The linearization variance estimator in (33) and the replication variance estimator in (36) were computed for each estimator in each sample. In the replication method, the jackknife method was used by deleting one element for each replication. The relative biases of the variance estimators were computed by dividing the Monte Carlo bias of the variance estimator by the Monte Carlo variance. The Monte Carlo relative biases of the linearization variance estimators and the replication variance estimators are presented in Table 3. The theoretical relative bias of the variance estimators is of order $o(1)$, which is consistent with the simulation results in Table 3. The linearization variance estimator slightly underestimates the true variance because it ignores the second order term in the Taylor linearization. The replication variance estimator shows slight positive bias in the simulation. The biases of the variance estimators are generally smaller in absolute values in population A because the linear model holds. In population B, variance estimators for the IVET estimator are less biased than those for the ET estimator because of less extreme weights used by the IVET estimator.

6. Concluding remarks

We have considered the problem of estimating Y with auxiliary information of the form $E\{U(\mathbf{X})\} = 0$ with some known function $U(\cdot)$. The class of the linear estimators of the form $\hat{Y} = \sum_{i \in A} w_i y_i$ with $\sum_{i \in A} w_i \{1, U(\mathbf{x}_i)\} = (\hat{N}, 0)$ and $w_i > 0$ is considered. If the density $f(\mathbf{x}; \boldsymbol{\eta})$ of X is known up to $\boldsymbol{\eta} \in \Omega$, then an efficient estimation can be implemented using the estimated importance weight

$$w_i \propto d_i \frac{f(x_i;\ \boldsymbol{\eta}_{0,N})}{f(x_i;\ \hat{\boldsymbol{\eta}})},$$

where d_i are the initial weights and where $\boldsymbol{\eta}_{0,N}$ and $\hat{\boldsymbol{\eta}}$ are the maximum likelihood estimators of $\boldsymbol{\eta}$ based on the population and the sample, respectively. If the parametric form of $f(\mathbf{x}; \boldsymbol{\eta})$ is unknown, thenthe exponential tilting weights of the form

$$w_{i(\lambda)} \propto \exp\{\boldsymbol{\lambda}' U(\mathbf{x}_i)\}$$

can be used, where $\boldsymbol{\lambda}$ is determined to satisfy

$$\sum_{i \in A} w_{i(\lambda)} U(\mathbf{x}_i) = 0. \tag{37}$$

Table 2
Monte Carlo Biases and Monte Carlo Mean squared errors of the point estimators for the mean of y, based on 10,000 Monte Carlo samples

Population	Sample Size	Estimator	SRS		PPS	
			Bias	MSE	Bias	MSE
A	100	No Calibration	0.00	0.02398	0.00	0.02023
		Regression estimator	0.00	0.01261	0.00	0.01289
		EL estimator ($t = 1$)	0.01	0.01369	0.01	0.01353
		EL estimator ($t = 10$)	0.00	0.01285	0.00	0.01289
		ET estimator ($t = 1$)	0.01	0.01334	0.01	0.01353
		ET estimator ($t = 10$)	0.00	0.01269	0.00	0.01289
		IVET estimator ($t = 1$)	0.01	0.01309	0.01	0.01330
		IVET estimator ($t = 10$)	0.00	0.01263	0.00	0.01289
	200	No Calibration	0.00	0.01069	0.00	0.00925
		Regression estimator	0.00	0.00595	0.00	0.00568
		EL estimator ($t = 1$)	0.01	0.00632	0.01	0.00604
		EL estimator ($t = 10$)	0.00	0.00597	0.00	0.00568
		ET estimator ($t = 1$)	0.00	0.00616	0.01	0.00578
		ET estimator ($t = 10$)	0.00	0.00596	0.00	0.00568
		IVET estimator ($t = 1$)	0.00	0.00605	0.01	0.00574
		IVET estimator ($t = 10$)	0.00	0.00591	0.00	0.00567
B	100	No Calibration	0.00	0.02044	0.00	0.01692
		Regression estimator	-0.01	0.01473	0.00	0.01461
		EL estimator ($t = 1$)	0.01	0.01652	0.01	0.01516
		EL estimator ($t = 10$)	0.00	0.01490	0.01	0.01472
		ET estimator ($t = 1$)	0.00	0.01516	0.01	0.01483
		ET estimator ($t = 10$)	0.00	0.01470	0.00	0.01459
		IVET estimator ($t = 1$)	0.00	0.01497	0.00	0.01458
		IVET estimator ($t = 10$)	0.00	0.01472	0.00	0.01453
	200	No Calibration	0.00	0.00888	0.00	0.00823
		Regression estimator	-0.01	0.00705	0.00	0.00735
		EL estimator ($t = 1$)	0.01	0.00769	0.01	0.00764
		EL estimator ($t = 10$)	0.00	0.00715	0.01	0.00745
		ET estimator ($t = 1$)	0.00	0.00723	0.01	0.00749
		ET estimator ($t = 10$)	0.00	0.00706	0.01	0.00734
		IVET estimator ($t = 1$)	0.00	0.00704	0.00	0.00728
		IVET estimator ($t = 10$)	0.00	0.00699	0.00	0.00725

SRS, simple random sampling; PPS, probability proportional to size sampling; MSE, mean squared error; EL, empirical likelihood; ET, exponential tilting; IVET, instrumental-variable exponential tilting.

Table 3
Monte Carlo Relative Biases of the variance estimators, based on 10,000 Monte Carlo samples

Population	Sample size	Estimator	Linearization		Replication	
			SRS	PPS	SRS	PPS
A	100	ET ($t = 1$)	-7.02	-2.66	10.65	4.11
		ET ($t = 10$)	-4.91	-0.80	5.60	0.67
		IVET ($t = 1$)	-5.28	-3.63	7.67	2.25
		IVET ($t = 10$)	-4.11	-0.87	4.96	0.41
	200	ET ($t = 1$)	-3.97	-0.19	3.65	0.57
		ET ($t = 10$)	-2.93	0.87	2.23	-0.35
		IVET ($t = 1$)	-3.35	-0.10	2.34	0.02
		IVET ($t = 10$)	-2.72	0.78	1.62	-0.53
B	100	ET ($t = 1$)	-7.64	-3.01	10.72	4.50
		ET ($t = 10$)	-5.98	-0.98	7.21	0.74
		IVET ($t = 1$)	-5.77	-2.31	4.53	-0.10
		IVET ($t = 10$)	-5.44	-1.86	5.17	-0.51
	200	ET ($t = 1$)	-2.41	-1.01	5.76	2.53
		ET ($t = 10$)	-1.29	0.18	4.30	1.91
		IVET ($t = 1$)	-1.39	-0.35	2.09	1.04
		IVET ($t = 10$)	-1.15	-0.06	2.04	0.99

SRS, simple random sampling; PPS, probability proportional to size sampling; ET, exponential tilting; IVET, instrumental-variable exponential tilting.

If a solution to (37) exists, it can be expressed as the limit of the form

$$w_{i(t)} \propto \prod_{s=0}^{t-1} \exp\{-\hat{U}'_{(s)}\hat{\Sigma}^{-1}_{aa(s)}U(\mathbf{x}_i)\} \qquad (38)$$

where $\hat{U}_{(s)} = \sum_{i\in A} w_{i(s)}U(\mathbf{x}_i)$, $\hat{\Sigma}_{aa(t)} = \sum_{i\in A} w_{i(t)}\{U(\mathbf{x}_i) - \bar{U}_{(t)}\}^{\otimes 2}$, $\bar{U}_{(t)} = \sum_{i\in A} w_{i(t)}U(\mathbf{x}_i)/\sum_{i\in A} w_{i(t)}$ with the initial weight $w_{i(0)} = d_i(\hat{N}/\hat{N}_d)$. If the solution to condition (37) does not exist, we can still use the weights in (38), but the equality must be relaxed. Instead, approximate equality will be satisfied in (37) in the sense that $\sum_{i\in A} w_{i(t)}U(\mathbf{x}_i)$ converges to zero much faster than $\sum_{i\in A} w_{i(0)}U(\mathbf{x}_i)$ for $t \geq 1$. Approximate equality in (37) is called the approximate calibration condition.

The estimators $\hat{Y}_{(t)} = \sum_{i\in A} w_{i(t)}y_i$ that use the t-step ET weights in (38), including the one-step estimator $\hat{Y}_{(1)}$, are asymptotically equivalent to the regression estimator of the form

$$\hat{Y}_{\text{reg}} = \hat{Y}_{(0)} - \hat{U}'_{(0)}\hat{\Sigma}^{-1}_{aa(0)}\hat{\Sigma}_{ay(0)},$$

where $\hat{Y}_{(0)} = \sum_{i\in A} w_{i(0)}y_i$ and $\hat{\Sigma}_{ay(0)} = \sum_{i\in A} w_{i(0)}\{U(\mathbf{x}_i) - \bar{U}_{(0)}\}y_i$. Unlike the regression estimator, the weights of the proposed method are always nonnegative. Furthermore, using the instrumental variable technique in Section 3, the weights are bounded above. Suitable choice of the instrumental variable also improves the efficiency of the resulting calibration estimator.

The exponential tilting calibration method is asymptotically equivalent to the empirical likelihood calibration method but it is more attractive computationally in the sense that the partial derivatives are not required in the iterative computation. Because the computation is simple, the variance of the proposed estimator can be easily estimated using a replication method, as discussed in Section 4. Further investigation in this direction, including interval estimation, can be a topic of future research.

Acknowledgements

The author wishes to thank Minsun Kim for computational support and two anonymous referees and the associated editor for very helpful comments that greatly improved the quality of the paper. This research was partially supported by a Cooperative Agreement NRCS 68-3A75-4-122 between the US Department of Agriculture Natural Resources Conservation Service and Iowa State University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the USDA Natural Resources Conservation Service.

Appendix

A. Assumptions and proof of Theorem 1

We first assume the following regularity conditions:

[A-1] The density $f(\mathbf{x};\boldsymbol{\eta})$ is twice differentiable with respect to $\boldsymbol{\eta}$ for every $\mathbf{x}$ and satisfy

$$\left|\frac{\partial^2 f(\mathbf{x};\boldsymbol{\eta})}{\partial\eta_i\partial\eta'_j}\right| \leq K(\mathbf{x})$$

for function $K(\mathbf{x})$ such that $E\{K(\mathbf{x})\} < \infty$, in a neighborhood of $\boldsymbol{\eta}_{0,N}$.

[A-2] The pseudo maximum likelihood estimator $\hat{\boldsymbol{\eta}}$ satisfies $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}) = O_p(1)$.

[A-3] The matrix $E\{\mathbf{s}(\boldsymbol{\eta}_{0,N})^{\otimes 2}\}$ exists and is nonsingular, where $\mathbf{s}(\boldsymbol{\eta}_{0,N}) = \partial \ln f(\mathbf{x}_i;\boldsymbol{\eta})/\partial\boldsymbol{\eta}|_{\eta=\eta_{0,N}}$.

To prove Theorem 1, write

$$g_i(\boldsymbol{\eta}) = \frac{f(\mathbf{x}_i;\ \boldsymbol{\eta}_{0,N})}{f(\mathbf{x}_i;\boldsymbol{\eta})},$$

and $w_i(\boldsymbol{\eta}) = d_i g_i(\boldsymbol{\eta})$. The estimated importance weight in (8) can be written $w_i = w_i(\hat{\boldsymbol{\eta}})$. Taking a Taylor expansion of $N^{-1}\sum_{i\in A} d_i\mathbf{s}_i(\hat{\boldsymbol{\eta}}) = 0$ around $\boldsymbol{\eta}_{0,N}$ leads to

$$\mathbf{0} = \frac{1}{N}\sum_{i\in A} d_i\mathbf{s}_i(\boldsymbol{\eta}_{0,N}) + \left\{\frac{\partial}{\partial\boldsymbol{\eta}'}\frac{1}{N}\sum_{i\in A} d_i\mathbf{s}_i(\boldsymbol{\eta}_{0,N})\right\}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}) + o_p(|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0,N}|).$$

Note that the first term on the right side of

$$\frac{1}{N}\frac{\partial}{\partial\eta}\sum_{i\in A} d_i\mathbf{s}_i(\boldsymbol{\eta}) = \frac{1}{N}\sum_{i\in A} d_i\frac{\partial^2 f(\mathbf{x}_i;\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'}{f(\mathbf{x}_i;\boldsymbol{\eta})} - \frac{1}{N}\sum_{i\in A} d_i\left\{\frac{\partial f(\mathbf{x}_i;\boldsymbol{\eta})/\partial\boldsymbol{\eta}}{f(\mathbf{x}_i;\boldsymbol{\eta})}\right\}^{\otimes 2}. \qquad (A1)$$

converges to $\int\{\partial^2 f(\mathbf{x};\boldsymbol{\eta})/\partial\boldsymbol{\eta}\partial\boldsymbol{\eta}'\}d\mathbf{x}$ which equals to zero by the dominated convergence theorem with [A1]. The second term converges to $E\{\mathbf{s}(\boldsymbol{\eta}_{0,N})^{\otimes 2}\}$. Thus, by [A-2],

$$\bar{\mathbf{S}}_{0d} \equiv \frac{1}{N}\sum_{i\in A} d_i\,\mathbf{s}_i(\boldsymbol{\eta}_{0,N}) = O_p(n^{-1/2}) \qquad (A2)$$

and

$$\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_{0N} = \hat{\boldsymbol{\Sigma}}^{-1}_{ss}\bar{\mathbf{S}}_{0d} + o_p(n^{-1/2}). \qquad (A3)$$

Now, taking a Taylor expansion of $N^{-1}\hat{Y}_w = N^{-1}\sum_{i\in A} w_i(\hat{\boldsymbol{\eta}})y_i$ around $\boldsymbol{\eta} = \boldsymbol{\eta}_{0,N}$ leads to

$$\frac{\hat{Y}_w}{N} = \frac{\hat{Y}_d}{N} + \left\{\frac{\partial}{\partial\boldsymbol{\eta}}\frac{1}{N}\sum_{i\in A} w_i(\boldsymbol{\eta}_{0,N})y_i\right\}'(\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_{0,N}) + o_p(|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_{0,N}|) \quad \text{(A4)}$$

by the uniform continuity of $\partial\{\sum_{i\in A} w_i(\boldsymbol{\eta})y_i\}/\partial\boldsymbol{\eta}$ around $\boldsymbol{\eta}_{0,N}$. Now, using

$$\frac{\partial}{\partial\boldsymbol{\eta}} g_i(\boldsymbol{\eta}) = -\frac{f(\mathbf{x}_i;\boldsymbol{\eta})}{f(\mathbf{x}_i;\boldsymbol{\eta})}\times\frac{\partial f(\mathbf{x}_i;\boldsymbol{\eta})/\partial\boldsymbol{\eta}}{f(\mathbf{x}_i;\boldsymbol{\eta})} = -g_i(\boldsymbol{\eta})\times s_i(\boldsymbol{\eta}),$$

where $\mathbf{s}_i(\boldsymbol{\eta}) = \partial \ln f(\mathbf{x}_i;\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, we have

$$\frac{\partial}{\partial\boldsymbol{\eta}}\sum_{i\in A} w_i(\boldsymbol{\eta})y_i = -\sum_{i\in A} w_i(\boldsymbol{\eta})\mathbf{s}_i(\boldsymbol{\eta})y_i.$$

Using $w_i(\boldsymbol{\eta}_{0,N}) = d_i$ and writing $\mathbf{s}_i(\boldsymbol{\eta}_{0,N}) = \mathbf{s}_{i0}$, we have, by (A2),

$$\frac{\partial}{\partial\boldsymbol{\eta}}\frac{1}{N}\sum_{i\in A} w_i(\boldsymbol{\eta}_{0,N})y_i = -\frac{1}{N}\sum_{i\in A} d_i\mathbf{s}_{i0}y_i = -\hat{\boldsymbol{\Sigma}}_{sy} + O_p(n^{-1/2}). \quad \text{(A5)}$$

Using (A5) and (A3) in (A4), result (9) is obtained.

B. Proof of Theorem 2

Write

$$\hat{\theta}(\boldsymbol{\lambda}_1) = \frac{\sum_{i\in A} d_i m_i(\boldsymbol{\lambda}_1)y_i}{\sum_{i\in A} d_i m_i(\boldsymbol{\lambda}_1)},$$

where $m_i(\boldsymbol{\lambda}_1) = \exp(\boldsymbol{\lambda}_1'\mathbf{x}_i)$. Note that $\hat{Y}_{\mathrm{ET}(t)} = \hat{N}\hat{\theta}(\hat{\boldsymbol{\lambda}}_{1(t)})$ and $\hat{\boldsymbol{\lambda}}_{1(t)}$ is defined in (19). By a Taylor expansion of $\hat{\theta}(\hat{\boldsymbol{\lambda}}_{1(t)}) = \hat{N}^{-1}\hat{Y}_{\mathrm{ET}(t)}$ around $\boldsymbol{\lambda}_1 = \mathbf{0}$ and by the continuity of the partial derivatives of $\hat{\theta}(\boldsymbol{\lambda}_1)$, we have

$$\hat{\theta}(\hat{\boldsymbol{\lambda}}_{1(t)}) = \hat{\theta}(\mathbf{0}) + \dot{\theta}(\mathbf{0})'(\hat{\boldsymbol{\lambda}}_{1(t)}-\mathbf{0}) + o_p(|\hat{\boldsymbol{\lambda}}_{1(t)}-\mathbf{0}|), \quad \text{(B1)}$$

where $\dot{\theta}(\boldsymbol{\lambda}) = \partial\hat{\theta}(\boldsymbol{\lambda})/\partial\boldsymbol{\lambda}$. Because $\hat{\boldsymbol{\lambda}}_{1(t)}$ converges in quadratic order and the one-step estimator satisfies $\hat{\boldsymbol{\lambda}}_{1(1)} = O_p(n^{-1/2})$, equation (22) can be written as

$$\hat{\boldsymbol{\lambda}}_{1(t)} = \left\{\hat{N}_d^{-1}\sum_{i\in A} d_i(\mathbf{x}_i-\bar{\mathbf{X}}_d)^{\otimes 2}\right\}^{-1}(\hat{N}^{-1}\mathbf{X}-\bar{\mathbf{X}}_d) + o_p(n^{-1/2}). \quad \text{(B2)}$$

Note that

$$\dot{\theta}(\boldsymbol{\lambda}_1) = \left\{\sum_{i\in A} d_i m_i(\boldsymbol{\lambda}_1)\right\}^{-1}\sum_{i\in A} d_i\dot{m}_i(\boldsymbol{\lambda}_1)\{y_i-\hat{\theta}(\boldsymbol{\lambda}_1)\}$$

where $\dot{m}_i(\boldsymbol{\lambda}_1) = \partial m_i(\boldsymbol{\lambda}_1)/\partial\boldsymbol{\lambda}_1$. Using $m_i(\mathbf{0}) = 1$ and $\dot{m}_i(\mathbf{0}) = \mathbf{x}_i$, we have $\hat{\theta}(\mathbf{0}) = \hat{Y}_d/\hat{N}_d$ and

$$\dot{\theta}(\mathbf{0}) = \hat{N}_d^{-1}\sum_{i\in A} d_i(\mathbf{x}_i-\bar{\mathbf{X}}_d)y_i. \quad \text{(B3)}$$

Therefore, inserting (B2) and (B3) into (B1), we have

$$\hat{\theta}(\hat{\boldsymbol{\lambda}}_{1(t)}) = \frac{\hat{Y}_d}{\hat{N}_d} + \left(\frac{\mathbf{X}}{\hat{N}}-\bar{\mathbf{X}}_d\right)'\left\{\sum_{i\in A} d_i(\mathbf{x}_i-\bar{\mathbf{X}}_d)^{\otimes 2}\right\}^{-1}\sum_{i\in A} d_i(\mathbf{x}_i-\bar{\mathbf{X}}_d)y_i + o_p(n^{-1/2}),$$

which proves (23).

References

Beaumont, J.-F., and Bocci, C. (2008). Another look at ridge calibration. *Metron*, LXVI, 5-20.

Breidt, F.J., Claeskens, G. and Opsomer, J.D. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92, 831-846.

Chambers, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

Chen, J., and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika*, 80, 107-116.

Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica Sinica*, 9, 385-406.

Chen, J., Variyath, A.M. and Abraham, B. (2008). Adjusted empirical likelihood and its properties. *Journal of Computational and Graphical Statistics*, 17, 426-443.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedure in survey sampling, *Journal of the American Statistical Association*, 88, 1013-1020.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9, 139-172.

Estevao, V.M., and Särndal, C.-E. (2000). A functional approach to calibration. *Journal of Official Statistics*, 16, 379-399.

Folsom, R.E. (1991). Exponential and logistic weight adjustment for sampling and nonresponse error reduction. In *Proceedings of the Section on Social Statistics*, American Statistical Association, 197-202.

Fuller, W.A. (2002). Regression estimation for sample surveys. *Survey Methodology*, 28, 5-23.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Givens, G.H., and Hoeting, J.A. (2005). *Computational Statistics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Henmi, M., Yoshida, R. and Eguchi, S. (2007). Importance sampling via the estimated sampler. *Biometrika*, 94, 985-991.

Imbens, G.W. (2002). Generalized method of moments and empirical likelihood. *Journal of Business and Economic Statististics*, 20, 493-506.

Isaki, C., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, 19, 145-157.

Kim, J.K., and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, In press.

Kott, P.S. (2003). A practical use for instrumental-variable calibration. *Journal of Official Statistics*, 19, 265-272.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.

Kitamura, Y., and Stutzer, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica*, 65, 861-874.

Park, M., and Fuller, W.A. (2009). The mixed model for survey regression estimation. *Journal of Statistical Planning and Inference*, 139, 1320-1331.

Rao, J.N.K., and Singh, A. (1997). A ridge shrinkage method for range restricted weight calibration in survey sampling. In *Proceedings of the Section on Survey Research Methods*, American Statist Association, 57-64.

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.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33, 99-119.

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

Tillé, Y. (1998). Estimation in surveys using conditional probabilities: Simple random sampling. *International Statistical Review*, 66, 303-322.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd Ed. New York: Springer-Verlag.

Wu, C., and Rao, J.N.K. (2006). Pseudo empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*, 34, 359-375.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 157-170
Statistics Canada, Catalogue No. 12-001-X

Comparison of survey regression techniques in the context of small area estimation of poverty

Stephen J. Haslett, Marissa C. Isidro and Geoffrey Jones [1]

Abstract

One key to poverty alleviation or eradication in the third world is reliable information on the poor and their location, so that interventions and assistance can be effectively targeted to the neediest people. Small area estimation is one statistical technique that is used to monitor poverty and to decide on aid allocation in pursuit of the Millennium Development Goals. Elbers, Lanjouw and Lanjouw (ELL) (2003) proposed a small area estimation methodology for income-based or expenditure-based poverty measures, which is implemented by the World Bank in its poverty mapping projects via the involvement of the central statistical agencies in many third world countries, including Cambodia, Lao PDR, the Philippines, Thailand and Vietnam, and is incorporated into the World Bank software program PovMap. In this paper, the ELL methodology which consists of first modeling survey data and then applying that model to census information is presented and discussed with strong emphasis on the first phase, *i.e.*, the fitting of regression models and on the estimated standard errors at the second phase. Other regression model fitting procedures such as the General Survey Regression (GSR) (as described in Lohr (1999) Chapter 11) and those used in existing small area estimation techniques: Pseudo-Empirical Best Linear Unbiased Prediction (Pseudo-EBLUP) approach (You and Rao 2002) and Iterative Weighted Estimating Equation (IWEE) method (You, Rao and Kovačević 2003) are presented and compared with the ELL modeling strategy. The most significant difference between the ELL method and the other techniques is in the theoretical underpinning of the ELL model fitting procedure. An example based on the Philippines Family Income and Expenditure Survey is presented to show the differences in both the parameter estimates and their corresponding standard errors, and in the variance components generated from the different methods and the discussion is extended to the effect of these on the estimated accuracy of the final small area estimates themselves. The need for sound estimation of variance components, as well as regression estimates and estimates of their standard errors for small area estimation of poverty is emphasized.

Key Words: Small area models; Nested error regression model; Poverty mapping.

1. Introduction

Poverty is a very complex multidimensional concern: there is no single definition and method of measurement available. In this paper, we adhere to the meaning of poverty that is used by most economists, *i.e.*, households are considered to be in poverty if their income falls below some income threshold called the poverty line. Chambers (2006) described this as income-poverty, and it is the definition adopted by the World Bank in the implementation of their small area poverty mapping projects carried out in conjunction with national statistical agencies and used, for example, for monitoring progress towards the Millennium Development Goals (UN website). Sometimes expenditure-based poverty measures are used instead to assess economic poverty. In public health related contexts, different measures such as standardized weight for age, height for age and weight for height for children (underweight, stunting and wasting, respectively) are used, *e.g.*, in Bangladesh (Haslett and Jones 2004) and Nepal (Haslett and Jones 2006).

Surveys conducted in most third world countries usually allow an acceptable level of precision for reporting poverty statistics at the first and second administrative level or geographical area (*e.g.*, for the Philippines - National and Region respectively). However, for policy makers to properly target assistance and interventions to the neediest communities and households, more disaggregated finer-level poverty statistics are needed. However, survey based poverty statistics at smaller geographical areas or lower administrative level are usually less reliable (have higher standard errors) due to smaller sample sizes, and this is where small area estimation comes into play.

The most common small area estimation methodology used for poverty measures in third world countries proposed by Elbers, Lanjouw and Lanjouw (ELL) (2002, 2003) allows generation of more precise estimates for smaller geographical areas by combining the survey data with information from a recent census. The ELL method consists of two phases: fitting a regression model (or models) to complex survey data and using that model to predict income or expenditure per capita at household level (which is transformed and aggregated to estimate poverty statistics at small area level).

In this paper, we focus specifically on the various algorithms used to fit the phase 1 regression models, and to estimate regression parameter standard errors and variance components from survey data. We emphasise consequences of survey regression modeling decisions rather than the

1. Stephen J. Haslett, Marissa C. Isidro and Geoffrey Jones, Institute of Fundamental Sciences: Statistics, College of Sciences, Massey University, Private Bag 11-222, Palmerston North, New Zealand. E-mail: S.J.Haslett@massey.ac.nz.

entire and rather comprehensive system ELL use to form small area estimates.

The preliminary requirement of the ELL methodology applied to economic measures is to develop an accurate model of per capita income or expenditure of households although this is often used to generate non-linear functions of income or expenditure (*e.g.*, poverty incidence - percentage of households below the poverty line, or poverty gap - sum of relative differences in income or expenditure for households or individuals below the poverty line). The survey-based regression model developed for income or expenditure is critical to accurate poverty statistics, but as we show below the regression model itself is not always the most important element, and other issues such as estimation of variance components deserve emphasis.

Other existing survey-based small area estimation regression techniques - Pseudo-Empirical Best Linear Unbiased Prediction (Pseudo-EBLUP) approach (You and Rao 2002), Iterative Weighted Estimating Equation (IWEE) method (You *et al.* 2003) and the General Survey Regression (GSR) (Skinner, Holt and Smith 1989) method are considered as alternative survey based model-fitting techniques and compared with two variations of the ELL method for fitting regression models to survey data. Our investigation is based on real data from the 2000 Philippine Family Income and Expenditure Survey (FIES), rather than simulated data.

This paper is organized as follows: Section 2 gives relevant background on small area models; the model for income (or expenditure) as presented by Elbers, Lanjouw and Lanjouw is given in Section 3; presented in Section 4 is a summary of the ELL methodology, followed by details on the alternative fitting methods in Section 5, which includes the Pseudo-Empirical Best Linear Unbiased Prediction Approach (5.1), IWEE Method (5.2), and the General Survey Regression Method (5.3). Section 6 discusses differences between the techniques, while Section 7 presents their application to the Philippine FIES 2000 data. This is followed by the conclusion and recommendations (Section 8).

2. Small area models

Ghosh and Rao (1994) classify small area models into two broad categories, area level and unit level models. Area level models refer to sets of models that can be considered when only area-specific auxiliary variables are available. Unit level models, on the other hand, refer to models that can be considered when there are unit-specific auxiliary variables and unit level values of the variable under study can be used. All such models are special cases of a general linear or generalized linear mixed model, and usually involve both fixed and random effects.

For area level models, it is assumed that the population mean $(\bar{Y}_a)$ of the a^{th} small area or some suitable function $\theta_a = g(\bar{Y}_a)$ is related to the area-specific auxiliary variables $\mathbf{x}_a = (x_{a1}, ..., x_{ap})'$ through a linear model

$$\theta_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a \tag{1}$$

where $a = 1, ..., k$, $v_a \sim \text{iid}(0, \sigma_v^2)$, $\boldsymbol{\beta}$ is a vector of regression parameters, c_a are known or estimated positive constants to allow for heteroscedasticity, k is the total number of small areas under study and p is the number of auxiliary variables. It is assumed that a direct design-based estimator, $\hat{\bar{Y}}_a$, of the population mean $\bar{Y}_a$ is available whenever the area sample size $n_a \geq 1$, and that

$$\hat{\theta}_a = \theta_a + e_a \tag{2}$$

where $\hat{\theta}_a = g(\hat{\bar{Y}}_a)$ and the sampling errors e_a are independent $N(0, V_a)$ with known variance V_a. Combining equation (1) and (2) gives the area level linear mixed model:

$$\hat{\theta}_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a + e_a. \tag{3}$$

We note that (3) involves both design-based random variables e_a and model-based random variables v_a (Rao 1999), where design-based variables are due to the sample selection mechanism, and model-based ones to the super-population structure in which the model is embedded.

Area level models have various extensions so they can for example handle correlated sampling errors, spatial dependence of random small area effects, time series and cross-sectional data (see Rao 2003, 1999 and Ghosh and Rao 1994).

The unit level model assumes that the variable of interest Y_{ah} for the h^{th} unit in the a^{th} small area is related to the element-specific auxiliary data $\mathbf{x}_{ah} = (x_{ah1}, ..., x_{ahp})'$ through a nested error regression model:

$$Y_{ah} = \mathbf{x}_{ah}' \boldsymbol{\beta} + v_a + e_{ah} \tag{4}$$

where $a = 1, ..., k$, $h = 1, ..., N_a$, $\boldsymbol{\beta} = (\beta_0, ... \beta_{p-1})$ is $p \times 1$ vector of regression parameters and N_a is the number of population units or households in the a^{th} small area. It is also assumed that the random effects v_a, are iid $N(0, \sigma_v^2)$ and are independent of the unit errors e_{ah} which are assumed to be iid $N(0, \sigma_e^2)$. Extensions that allow errors to be heteroscedastic, with known scaling constant(s) are also possible.

The ELL method uses a unit level model, where the units are households in the case of income or expenditure data, and where the variation is modeled at primary sampling unit, *i.e.*, cluster level and household level. Note that ELL do not include model variation at small area level, only for cluster within small area, and for household within cluster. This is the form of the basic model used for comparisons in this paper since ELL is the standard small area estimation

method for poverty in third world countries. In the real datasets we have studied this additional small area variation has been very small. Despite this empirical evidence however, important questions remain about how best to estimate the small area variance component in the presence of cluster level variation, when there is sample survey weighting, especially where many of the small areas contain only one sampled cluster.

The ELL model has a number of other characteristics not all of which are standard in a statistical sense (see Haslett and Jones 2005, for example). The intention of this paper is not to discuss differences in the available methods generally, but to focus directly on how methods of fitting regression models to survey data differ when the ELL first phase "base structure" of fitting a survey regression model is used. The focus of this paper therefore is on comparison of the available methods of fitting regression models to survey data on income or expenditure using a specified set of regressors, even though ELL can also be (and is) used relatively routinely to find small area estimates for non-linear functions (*e.g.*, poverty incidence, gap or severity) by applying fitted regression models to a census.

The answer to the 'best regression model fitting' question for survey data on which this paper focuses (as with other matters related to the ELL methodology) is particularly important because there are billions of dollars of aid funding that are (or have the potential to be) allocated based on the regression models used as part of small area estimation of poverty.

3. Income/consumption model

Modeling per capita income or expenditure of households instead of poverty measures themselves (such as poverty incidence and gap) is one of the distinctive features of the ELL method. As mentioned in the previous section, the ELL method involves fitting the income or expenditure model to the survey data and applying it to the census data prior to the generation of the small area estimates of poverty measures. The income/expenditure model is as follows:

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + u_{bh} \tag{5}$$

where $b = 1, ..., M$, $h = 1, ..., N_b$; Y_{bh} is the log-transformed per capita income or expenditure of the h^{th} unit or household in the b^{th} cluster, M is the total number of clusters in the population and N_b is the total number of households in the b^{th} cluster in the population. $\mathbf{x}_{bh}$ is a set of the auxiliary variables available in both the survey and the census, which generally need to be contemporaneous; u_{bh} is the random error term representing that part of Y_{bh} that cannot be explained by $\mathbf{x}_{bh}$. Income and expenditure data almost invariably have a skewed distribution, hence a transformation (usually logarithmic) is applied to make the data more symmetrical.

The households for which data on per capita income or expenditure is collected are seldom independent, but have natural groupings or clusters, often defined administratively. Households that are close to each other or in the same cluster, tend to be similar in many respects. In the survey data, the clusters are usually also the primary sampling units (PSUs) for the sample survey design. To account for the clustering of households, the random error term u_{bh} in the regression model is usually assumed to have the following specification:

$$u_{bh} = v_b + e_{bh} \tag{6}$$

where v and e are independent of each other and uncorrelated with $\mathbf{x}_{bh}$, v_b is the error term held in common by the b^{th} group or cluster (*e.g.*, barangay for the Philippines) and e_{bh} is the household level error within the cluster. The importance of each term is measured by their respective variances or variance components, σ_v^2 and σ_e^2. There are various procedures for estimating these variances. This important topic is covered in the sections that follow.

Model (5) can be written as

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + v_b + e_{bh} \tag{7}$$

which is similar in form to the unit level model or nested error regression model mentioned in the previous section. However while the form of the model is similar, the group being referred to is different, *e.g.*, Y_{ah} refers to the h^{th} household in the a^{th} small area, while Y_{bh} refers to the h^{th} household in the b^{th} cluster. Clusters, based on the survey design, will typically be much smaller than the areas for which small area estimates are sought, and generally (unlike almost all the small areas) not all clusters are sampled. For example in the Philippines, estimates are sought at the municipal level which is composed of barangays or clusters.

4. The ELL methodology

In the ELL methodology, the estimate of the regression parameter $\boldsymbol{\beta}$ is given, in Elbers *et al.* (2002, page 11 footnote 8) and in the POVMAP software Zhao (2006) developed for the ELL method, as

$$\hat{\boldsymbol{\beta}}_{\text{ELL}} = \left(\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{X}_b \right)^{-1} \left(\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{y}_b \right) \tag{8}$$

and the corresponding variance-covariance matrix as

$$\mathbf{V}(\hat{\boldsymbol{\beta}}_{\text{ELL}}) = \mathbf{D} \left[\left(\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{W}_b \mathbf{X}_b \right)^{-1} \right] \mathbf{D} \tag{9}$$

where $\mathbf{V}_b = (\sigma_e^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{1}_{n_b} \mathbf{1}'_{n_b})$, (σ_υ^2) is the cluster level variance, while (σ_e^2) is the household level variance, $\mathbf{I}_{n_b}$ is an identity matrix, $\mathbf{1}'_{n_b} = (1...1)$ is a constant vector, $\mathbf{D} = (\sum_{b=1}^{m} \mathbf{X}'_b \mathbf{W}_b \mathbf{V}_b^{-1} \mathbf{X}_b)^{-1}$, $\mathbf{X}_b = (\mathbf{x}_{b1}, ..., \mathbf{x}_{bn_b})'$; $\mathbf{y}_b = (y_{b1}, ..., y_{bn_b})'$; $\mathbf{W}_b$ is a diagonal matrix of sampling weights; m is the number of clusters in the sample and n_b is the number of households in each sampled cluster. Equation (8) assumes $\mathbf{V}_b$ is known. In practice we need to estimate σ_e^2 and σ_v^2 to get the estimator $\hat{\mathbf{V}}_b$. We note that the variance expression in (9) is derived under a vaguely specified model assumed for the sample (see Elbers *et al.* 2002). Under the ELL method, fitting the income/expenditure model (7) involves obtaining the initial estimate of $\boldsymbol{\beta}$ through weighted least squares (WLS) method and using the residuals of the initial model to estimate the covariance matrix $\mathbf{V}_b$ needed to obtain $\hat{\boldsymbol{\beta}}_{\mathrm{ELL}}$. The estimate of the cluster level (σ_v^2) and household level (σ_e^2) variances, are derived by Elbers *et al.* (2002) as follows:

$$\hat{\sigma}_v^2 = \max\left(\frac{\sum_b w_b (u_{b.} - u_{..})^2}{\sum_b w_b (1 - w_b)} - \frac{\sum_b w_b (1 - w_b) \tau_b^2}{\sum_b w_b (1 - w_b)}; 0 \right) \quad (10)$$

where $\tau_b^2 = \sum_h (e_{bh} - e_{b.})^2 / (n_b (n_b - 1)); w_b = \sum_h w_{bh} / \sum_b \sum_h w_{bh}$, is the by-cluster transformed sampling weights which sum to one across clusters and w_{bh} is the re-scaled sampling weights which sum to the total sample size. Here $u_{b.} = \sum_h u_{bh}$ and $u_{..} = \sum_b \sum_h u_{bh}$ (which is equal to zero) where u_{bh} is as defined in equation (6).

There are two ways suggested by Elbers *et al.* (2002) to generate the estimate of the household level variance component: "direct" computation which is denoted by $(\hat{\sigma}_e^2)$ or the heteroscedasticity model-based $(\hat{\sigma}_{e,bh}^2)$. Direct computation involves using the difference between the estimated mean square error from the initial WLS regression and the computed estimate of σ_v^2, while the heteroscedasticity model-based computation uses a logistic-type link function to bound the variance as follows:

$$\sigma_{e,bh}^2(z_{bh}, \boldsymbol{\alpha}, A, B) = \left[\frac{A \exp(z'_{bh} \boldsymbol{\alpha}) + B}{1 + \exp(z'_{bh} \boldsymbol{\alpha})} \right] \quad (11)$$

where A and B are the upper and lower bounds respectively, estimated with the parameter vector $\boldsymbol{\alpha}$ using a standard pseudomaximum likelihood procedure (Elbers *et al.* 2003), and where z_{bh} are auxiliary variables. Elbers *et al.* claim that imposing a minimum bound of zero and a maximum bound of $A^* = (1.05) \max\{e_{bh}^2\}$ in general yields similar estimates of the parameters $\boldsymbol{\alpha}$. These restrictions allow one to estimate the simpler form

$$\ln\left[\frac{e_{bh}^2}{A^* - e_{bh}^2} \right] = z'_{bh} \boldsymbol{\alpha} + r_{bh} \quad (12)$$

where r_{bh} is an error term and the other variables are as defined earlier. In most of the World Bank poverty mapping projects, slight modifications are usually made, for example, adding a constant δ to e_{bh}^2 in model (11).

By using model (12), and employing the delta method, $\hat{\sigma}_{e,bh}^2$ is computed as:

$$\hat{\sigma}_{e,bh}^2 = \left[\frac{A^* C_{bh}}{1 + C_{bh}} \right] + \frac{1}{2} \hat{\sigma}_r^2 \left[\frac{A^* C_{bh} (1 - C_{bh})}{(1 + C_{bh})^3} \right] \quad (13)$$

where $C_{bh} = \exp\{z'_{bh} \hat{\boldsymbol{\alpha}}\}$, and $\hat{\sigma}_r^2$ is the estimated variance of the residuals under model (12). If the household level variance component is based on a heteroscedastic model, then, $\mathbf{V}_b = (\sigma_{e,bh}^2 \mathbf{I}_{n_b} + \sigma_v^2 \mathbf{1}_{n_b} \mathbf{1}'_{n_b})$. Heteroscedasticity modeling is conducted on the assumption that variation at the household level depends on some covariates.

As discussed in more detail in the appendix, the way in which the weight matrix $\mathbf{W}_b$ enters the calculation in equation (9) above leads to an asymmetric estimated covariance matrix. A rather better approach based on 'pseudomaximum likelihood' is outlined by Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998) and involves splitting $\mathbf{X}'_b \mathbf{V}_b^{-1} \mathbf{X}_b$ into separate sums of squares and cross-product terms, and weighting each appropriately - if we write $\mathbf{V}_b^{-1} = c \mathbf{I}_{n_b} + d \mathbf{1}_{n_b} \mathbf{1}'_{n_b}$ then the appropriate weighting is $c \mathbf{X}'_b \mathbf{W}_b \mathbf{X}_b + d \mathbf{X}'_b \mathbf{W}_b \mathbf{1}_{n_b} \mathbf{1}'_{n_b} \mathbf{W}_b \mathbf{X}_b$.

Since the ELL version, $\mathbf{W}_b \mathbf{V}_b^{-1}$, is not generally symmetric, neither is $\mathbf{D}$ in equation (9). As a consequence the supposed covariance matrix of $\hat{\boldsymbol{\beta}}_{\mathrm{ELL}}$, $\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathrm{ELL}})$, is also not symmetric. The POVMAP software attempts to solve this problem by taking the average of their $\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathrm{ELL}})$ and its transpose, thereby forcing the matrix to be symmetric.

Note again that under the ELL method, the regression fit to the survey data and the estimation of variance components is only the first phase. The consequent phase involves prediction at household level based on the entire census data and aggregation to small area level.

The survey fitting methods (derivation of the estimate of $\boldsymbol{\beta}$ and its corresponding variance-covariance matrix) of three alternative regression procedures to ELL are presented in the following sections.

5. Alternative fitting methods

5.1 The pseudo-empirical best linear unbiased prediction approach

You and Rao (2002) proposed an estimator of the small area mean by deriving an estimator of $\boldsymbol{\beta}$ based on the unit level model (4). The process of deriving the estimator of $\boldsymbol{\beta}$ starts with the computation of the best linear unbiased predictor (BLUP) of v_a given the parameters $\boldsymbol{\beta}$, σ_e^2 and

σ_v^2 from the aggregated (survey-weighted) area level model:

$$\bar{Y}_{aw} = \bar{\mathbf{x}}'_{aw}\boldsymbol{\beta} + v_a + \bar{e}_{aw} \tag{14}$$

which proceeds as follows:

$$\hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2) = \gamma_{aw}(\bar{y}_{aw} - \bar{\mathbf{x}}'_{aw}\boldsymbol{\beta}) \tag{15}$$

where $\bar{\mathbf{x}}_{aw} = \sum_{h=1}^{n_a} w_{ah}\mathbf{x}_{ah}$, $\bar{y}_{aw} = \sum_{h=1}^{n_a} w_{ah}y_{ah}$, $\gamma_{aw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\,\delta_a^2)$, $w_{ah} = \tilde{w}_{ah}/\sum_{h=1}^{n_a}\tilde{w}_{ah}$, $\delta_a^2 = \sum_{h=1}^{n_a} w_{ah}^2$, and $\tilde{w}_{ah}$ are the unit level survey weights; then solving for the survey-weighted estimating equation for $\boldsymbol{\beta}$:

$$\sum_{a=1}^{k}\sum_{h=1}^{n_a} \tilde{w}_{ah}\mathbf{x}_{ah}[y_{ah} - \mathbf{x}'_{ah}\boldsymbol{\beta} - \hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2)] = 0 \tag{16}$$

from which the estimator of $\boldsymbol{\beta}$ is obtained as

$$\hat{\boldsymbol{\beta}}_w = \left\{\sum_{a=1}^{k}\sum_{h=1}^{n_a} \mathbf{x}_{ah}\mathbf{z}'_{ah}\right\}^{-1}\left\{\sum_{a=1}^{k}\sum_{h=1}^{n_a} \mathbf{z}_{ah}y_{ah}\right\} \tag{17}$$

where $\mathbf{z}_{ah} = \tilde{w}_{ah}(\mathbf{x}_{ah} - \gamma_{aw}\bar{\mathbf{x}}_{ah})$. The corresponding covariance matrix is then as follows:

$$\begin{aligned}\boldsymbol{\Phi}_w = {} & \sigma_e^2\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a} \mathbf{x}_{ah}\mathbf{z}'_{ah}\right)^{-1} \\ & \left(\sum_{a=1}^{k}\sum_{h=1}^{n_a} \mathbf{z}_{ah}\mathbf{z}'_{ah}\right)\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a} \mathbf{x}_{ah}\mathbf{z}'_{ah}\right)^{-1} \\ & + \sigma_v^2\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a} \mathbf{x}_{ah}\mathbf{z}'_{ah}\right)^{-1} \\ & \left\{\sum_{a=1}^{k}\left(\sum_{h=1}^{n_a} \mathbf{z}_{ah}\right)\left(\sum_{a=1}^{n_a} \mathbf{z}_{ah}\right)'\right\}\left\{\left(\sum_{a=1}^{k}\sum_{h=1}^{n_a} \mathbf{x}_{ah}\mathbf{z}'_{ah}\right)^{-1}\right\}'. \end{aligned} \tag{18}$$

The variance components are estimated using Henderson's Method 3 (Henderson 1953), to generate unbiased estimates even in the presence of correlated elements in the model. The estimators of the variance components are as follows:

$$\hat{\sigma}_{eH}^2 = (n - k - p + 1)^{-1}\sum_{a=1}^{k}\sum_{h=1}^{n_a} \hat{\varepsilon}_{ah}^2 \tag{19}$$

where $\{\hat{\varepsilon}_{ah}^2\}$ are residuals from the OLS regression of $(y_{ah} - \bar{y}_a)$ on $\{x_{ah1} - \bar{x}_{a.1}, ..., x_{ahp} - \bar{x}_{a.p}\}$ and $(\bar{y}_a, \bar{x}_{a.1}, ..., \bar{x}_{a.p})$ are the sample means in the a^{th} group.

$$\hat{\sigma}_{vH}^2 = n_*^{-1}\left[\sum_{a=1}^{k}\sum_{h=1}^{n_a} \hat{u}_{ah}^2 - (n - p)\hat{\sigma}_{eH}^2\right] \tag{20}$$

where $n_* = n - \mathrm{tr}[(\mathbf{X}'\mathbf{X})^{-1}\sum_{a=1}^{k} n_a^2\,\bar{\mathbf{x}}_a\,\bar{\mathbf{x}}'_a]$ with $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_k)$, and the $\{\hat{u}_{ah}\}$ are the residuals from the OLS regression of y_{ah} on $\{x_{ah1}, ..., x_{ahp}\}$. For the model (7), the subscript a is replaced by b.

However, the Henderson's estimators above do not account for the sampling weights. To address this, an estimation technique has been proposed by You *et al.* (2003) which extends the Pseudo-EBLUP method by incorporating the weights in the estimation of the variance components. This is described in the next section.

5.2 The iterative weighted estimating equation method

The estimator proposed by You *et al.* (2003) is similar to the Pseudo-EBLUP estimator, except that it incorporates the sampling weights in the computation of the variance components, and it generates the parameter estimate $\boldsymbol{\beta}$ and the variance components by using an iterative weighted estimating equation (IWEE) approach. The authors derived the estimator of σ_e^2 and σ_v^2 as follows:

$$\begin{aligned}\hat{\sigma}_{ew}^{2(t)} &= \frac{\sum_{a=1}^{k}\sum_{h=1}^{n_a} \tilde{w}_{ah}[y_{ah} - \bar{y}_{aw} - (\mathbf{x}_{ah} - \bar{\mathbf{x}}_{aw})'\hat{\boldsymbol{\beta}}^{(t-1)}]^2}{\sum_{a=1}^{k}\left[(1 - \delta_a^2)\sum_{h=1}^{n_a} \tilde{w}_{ah}\right]} \\ &\equiv \tilde{\sigma}_{ew}^{2(t)}(\boldsymbol{\beta}) \end{aligned} \tag{21}$$

and

$$\begin{aligned}\hat{\sigma}_{vw}^{2(t)} &= \frac{1}{k}\sum_{a=1}^{k} \tilde{v}_{aw}^2 + \frac{\tilde{\sigma}_{vw}^{2(t-1)}}{k}\sum_{a=1}^{k}(\gamma_{aw} - 1)^2 + \frac{\tilde{\sigma}_{ew}^{2(t)}}{k}\sum_{a=1}^{k}\delta_a^2\gamma_{aw}^2 \\ &\equiv \tilde{\sigma}_{vw}^{2(t)}(\tilde{v}_w, \sigma_e^2, \sigma_v^2). \end{aligned} \tag{22}$$

The survey weighted estimates of $\boldsymbol{\beta}$, σ_e^2, σ_v^2 are obtained simultaneously by following iterative updating steps, t in the equation above stands for the t^{th} iteration. Since the variance components σ_v^2 and σ_e^2 are unknown, initial estimates for the iterative steps are generated by Henderson's method. Again, as for Pseudo-EBLUP, for the ELL regression model formulation (7), the subscript a is replaced by b.

This approach is similar to the probability-weighted iterative generalized least squares (PIWGLS) method proposed by Pfeffermann *et al.* (1998) for fitting multilevel models where the estimation process considered the unequal selection probabilities at each stage of sampling and involves iterating between the parameter $\boldsymbol{\beta}$ and the variance components until convergence. A model-based approach is also proposed by Pfeffermann, Moura and Silva (2006), which involves deriving the hierarchical model for given sample data as a function of the population model and the selection probabilities, and then fitting the sample model using Bayesian approach by use of Markov Chain Monte Carlo algorithm.

5.3 General survey regression method

Another approach to generate the estimator of the parameter β and its variance is the design-based methodology for fitting regression models (Lohr 1999). This

technique is currently used in the Stata, Sudaan, and WesVar package, for example. The estimator of $\boldsymbol{\beta}$ given below is the sample weighted regression estimator for a model with homoscedastic variance structure and uncorrelated observations in the population.

$$\hat{\boldsymbol{\beta}}_S = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \qquad (23)$$

This estimator is not derived under the model specified by (7) even under the homoscedastic variances for household errors. The linearized/robust variance estimate for $\hat{\boldsymbol{\beta}}_S$ is based on the design-based variance estimator for a total, given as,

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_S) = \mathbf{D}\left\{\frac{m}{m-1}\sum_{b=1}^{m}\left(\sum_{h=1}^{n_b} w_{bh}\mathbf{d}_{bh}\right)'\left(\sum_{h=1}^{n_b} w_{bh}\mathbf{d}_{bh}\right)\right\}\mathbf{D} \qquad (24)$$

where $\mathbf{d}_{bh} = \hat{e}_{bh}\mathbf{x}_{bh}$; $\hat{e}_{bh}$ is the residual from WLS regression; $\mathbf{x}_{bh}$ is a vector of the independent variables; w_{bh} is a sampling weight; $\mathbf{D} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$; and $\mathbf{W}$ is a diagonal matrix of the sampling weights.

The General Survey Regression method differs from the other techniques in the computation of the estimates, and generates the estimates without computing the variance components, σ_v^2 and σ_e^2. As shown above, the equations for the estimator of the parameter $\boldsymbol{\beta}$ and its corresponding estimated covariance matrix only involve the sampling weights matrix $\mathbf{W}$. The estimated covariance matrix in (24) is often referred to as a sandwich estimator.

6. Comparison of the model fitting techniques

The ELL methodology is claimed to be a weighted GLS estimation procedure. However, as pointed out earlier, the sampling weights are not properly incorporated in the estimation process and this leads to non-interpretability of the elements in some matrices involved in the estimation, as well as asymmetry in the estimated covariance matrix. For the ELL method of estimating the variance components, the weights are accounted for only at the cluster level. The two ways (direct computation and heteroscedasticity model-based) that ELL use for generating the household level variance component do not incorporate the sampling weights. For direct computation, the household level variance component is determined from the residual of the survey-weighted (WLS) regression conducted at the preliminary step and the weighted estimate of the cluster level component. The heteroscedasticity based computation is based on modeling the square of the residuals from the WLS regression.

While the ELL methodology follows a GLS-like estimation procedure, the pseudo-EBLUP and IWEE method follow the Generalized Estimating Equation (GEE) procedure (Liang and Zeger 1986) using an exchangeable working correlation matrix, *i.e.*, all the off-diagonal elements of the correlation matrix within clusters are equal, and in Pseudo-EBLUP and IWEE are equal to $\sigma_v^2/(\sigma_v^2 + \sigma_e^2)$. An exchangeable or equicorrelated working correlation matrix is one of the common working correlation matrices presented in the paper of Horton and Lipsitz (1999) when reviewing different software for fitting GEE regression models.

The two procedures, Pseudo-EBLUP and IWEE, both incorporate the sampling weights in the estimation of the parameter $\boldsymbol{\beta}$ and the corresponding standard error, although the Pseudo-EBLUP method uses Henderson's method in the estimation of the variance components. While Henderson's method generates unweighted estimates of the variance components, the IWEE method incorporates the sampling weights iteratively from estimation of variance components for computation of standard error of the estimate of the regression parameter.

There is a very limited published literature on the application to real data sets of the Pseudo-EBLUP and IWEE methods. Those that there are consider the clusters as the small area, and often use the data in Battese, Harter and Fuller (1988), whose data set contains information on hectares of corn and soybeans per segment for counties in North Central Iowa and assumes simple random sampling within areas or clusters. An exception is the recent paper by Militino, Ugarte, Goicoa and Gonzalez-Audicana (2006), which applies Pseudo-EBLUP to estimating the total area occupied by olive trees in Navarra, Spain, where (as in Battese *et al.*) the units are self weighting. Generally for poverty estimation, Pseudo-EBLUP and IWEE techniques must be applied in more complex situations, since sampling clusters and small areas are not identical and the sample is not self weighting. In the example in the next section, the clusters (barangay) are different from the small areas (municipalities), the clusters are sub-units of the small area and the sampling scheme is not self weighting.

The GSR method is one of the least complicated estimation procedures as it employs a weighted least squares procedure using the sandwich estimator for estimating the variance of the estimator of the regression parameter. As mentioned earlier, this method differs from the other techniques in that the estimate of the regression parameters and their corresponding standard errors are generated without computing the variance components.

Based on the discussion above, for all the techniques considered, the survey-based estimation procedure for the parameter $\boldsymbol{\beta}$ and its corresponding standard error are theoretically sound given their assumptions, except for the ELL method where there are some inconsistencies in the estimation of parameters $\boldsymbol{\beta}$ and the covariance of $\hat{\boldsymbol{\beta}}$.

7. Application to real data

In this section, the four different regression techniques (one of which contains two variants of ELL) are compared using the Philippine 2000 Family Income and Expenditure Survey (FIES). The FIES data is a nationwide survey undertaken by the Philippines National Statistics Office (NSO) every three years. The survey gathers details on family income and expenditure as well as information affecting income and expenditure. Selected households are interviewed in two separate operations, each covering a half-year period, in order to allow for seasonal patterns in income and expenditure. For FIES 2000 the interviews were conducted in July 2000, for the period 01 January to 30 June and January 2001 for the period 01 July to 31 December. The sample design for FIES used a multi-stage stratified random sampling technique. Barangays are the primary sampling units (PSUs) and are stratified into urban and rural within each province and selected using systematic sampling with probability proportional to size. Large barangays are further divided into enumeration areas and subjected to further sampling before the final stage in which households are systematically sampled from the 1995 Population Census List of Households. Interview non-response was only 3.4 percent, with 39,615 of the sample households being successfully interviewed in both survey visits. Deterministic imputation was done to address item non-response, *i.e.*, entry for a particular missing item is deduced from other items in the questionnaire.

The auxiliary variables used in this paper are adopted from the variables included in the model formulated by Haslett and Jones (2005) that was fitted without using POVMAP for the small area poverty mapping project in the Philippines. The auxiliary variables included both household characteristics and municipal means (in which the household data used have the same value for every sampled household in a given municipality, *i.e.*, small area). These auxiliary variables are not only derived from the FIES data but also from the Philippine 2000 Labor Force Survey (LFS) and Census of Population and Housing (CPH). The LFS collects socioeconomic characteristics of the population over 15 years old. It is conducted on a quarterly basis by the NSO by personal interview, using previous week as reference period. Being part of the Integrated Survey of Households (NSCB 2000), the July 2000 and January 2001 surveys used the same sample of households as the 2000 FIES. Thus the two data sets can be merged to form a richer set of auxiliary variables. Additional auxiliary variables were also taken from the 2000 CPH in the form of municipal means. Census variables in both the short and long form were averaged at municipal level to create new data sets that could be merged with the set of auxiliary variables from FIES and LFS.

Presented in Tables 1, 2, and 3 are the computed estimates of the parameter $(\boldsymbol{\beta})$ and the corresponding standard errors as well as the estimates of the variance components at the national, regional and provincial levels, respectively. Table 2 is one of the regional models of the 16 models fitted at the regional level (there are 16 regions in the Philippines in the year 2000). Similarly, Table 3 shows one of the provincial models of the 20 models formulated for 20 selected provinces. To standardize comparison, exactly the same set of predictor variables are used for all the different model fitting techniques. (There are five sets of parameter estimates, although there are only four basic methods considered, because ELL is used both with and without heteroscedasticity.) Note that in practice when ELL is applied, the survey data is often subdivided and separate models fitted to each subsample, *e.g.*, to each regionally-based stratum as the 16 regions in the Philippines or even provincial level models. This can lead to overfitted models and downwardly biased standard errors for small area estimates. For the analysis here, a single model (or the national level model) has been fitted. In practice intermediate models with some but not all possible regional effects seem to work best. See for example Haslett and Jones (2005).

To assess the differences of the estimates generated from the different techniques, an informal comparison of the "significance" of the different estimates of β is conducted by subtracting from the estimate by one method the mean of the other methods' estimates, then dividing by the standard error of the one method. At the national level (Table 1), estimates of the regression coefficients generated from the different methods are significantly different from each other for a number of the independent variables. GSR tends to generate estimates of the regression coefficients for the majority of the variables that are significantly different from the other methods. As pointed out earlier, the GSR estimator is the sample weighted regression estimator for a model with homoscedastic variance structure and uncorrelated observations in the population and hence this estimator is not derived under the model specified by (7). However, it is the most conservative as it generates the highest standard error for all the household level characteristics. On the other hand, the IWEE method has the highest estimated standard error for all the municipal means. The ELL_H (ELL with heteroscedasticity) method can be considered to be the least conservative since it produces the lowest standard errors for all the estimated regression coefficients of the household level characteristics as well as for the municipal means, except for two variables where GSR generated the smallest estimates. As to the estimates of the variance components, the ELL method generates the smallest estimated cluster level variance, which is about 92% of the Pseudo-EBLUP method and 86% of the IWEE method. As to the household level variance, the IWEE method generates the smallest estimate.

Table 1
National level estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household (mean = 0.1576633) **Based from the ELL results

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.11867	0.00181	-0.12034	0.00165	-0.11875	0.00183	-0.11888	0.00180	-0.11405	0.00216
famsizesqc	0.00937	0.00039	0.00981	0.00036	0.00938	0.00039	0.00939	0.00038	0.00898	0.00044
type_mult	0.03876	0.01697	0.03703	0.01588	0.03699	0.01717	0.03466	0.01692	0.11460	0.02194
per_kids	-0.20342	0.01476	-0.20818	0.01322	-0.20293	0.01491	-0.20216	0.01467	-0.22864	0.01617
roof_light	-0.06314	0.01291	-0.05808	0.01056	-0.06263	0.01306	-0.06175	0.01287	-0.09251	0.01413
per_61up	-0.09402	0.01420	-0.08331	0.01371	-0.09392	0.01435	-0.09389	0.01412	-0.09705	0.01698
roof_strong	0.05882	0.01135	0.05633	0.00962	0.05944	0.01148	0.06030	0.01132	0.03118	0.01293
wall_light	-0.05459	0.01182	-0.04979	0.00975	-0.05426	0.01195	-0.05392	0.01178	-0.06286	0.01353
wall_salvaged	-0.10814	0.02505	-0.11327	0.02058	-0.10748	0.02533	-0.10607	0.02495	-0.15702	0.02925
wall_strong	0.14248	0.01051	0.12964	0.00910	0.14274	0.01063	0.14319	0.01047	0.12662	0.01284
fa_xs	-0.17052	0.00941	-0.16756	0.00782	-0.17144	0.00952	-0.17236	0.00939	-0.14213	0.01110
fa_s	-0.08368	0.00861	-0.08242	0.00725	-0.08403	0.00871	-0.08454	0.00857	-0.06667	0.00964
fa_l	0.09016	0.00908	0.08478	0.00792	0.09065	0.00918	0.09106	0.00904	0.07848	0.01047
fa_xl	0.16959	0.01104	0.15404	0.00992	0.17034	0.01117	0.17121	0.01100	0.14300	0.01334
fa_xxl	0.27072	0.01144	0.24485	0.01094	0.27172	0.01157	0.27274	0.01140	0.23913	0.01457
fa_xxxl	0.36190	0.01371	0.31369	0.01286	0.36270	0.01387	0.36382	0.01367	0.32123	0.02025
all_eled	0.19084	0.01535	0.20497	0.01307	0.19031	0.01551	0.18964	0.01527	0.21344	0.01831
all_hsed	0.42325	0.01250	0.43771	0.01083	0.42192	0.01263	0.42024	0.01244	0.48180	0.01475
all_coed	1.21591	0.01371	1.29368	0.01379	1.21324	0.01386	1.20935	0.01366	1.35022	0.01827
dom_help	0.60207	0.01629	0.61218	0.01886	0.60035	0.01645	0.59733	0.01620	0.70307	0.02656
head_male	-0.05878	0.00988	-0.04581	0.00932	-0.05862	0.00998	-0.05819	0.00982	-0.07410	0.01173
no_spouse	-0.09367	0.00987	-0.07376	0.00917	-0.09361	0.00997	-0.09351	0.00981	-0.09599	0.01123
hou_9600	0.28537	0.07654	0.25643	0.07375	0.28871	0.07911	0.28783	0.08066	0.31956	0.07941
hea_rel_mus	0.09058	0.02645	0.10859	0.02507	0.09753	0.02728	0.09731	0.02782	0.10196	0.02737
Per_eng	0.17273	0.06529	0.14561	0.06298	0.17782	0.06754	0.17799	0.06887	0.17076	0.06407
Hou_coelpg	0.37463	0.04348	0.39784	0.04210	0.37934	0.04494	0.37792	0.04581	0.42682	0.03711
Hou_own_ref	0.17716	0.10497	0.18342	0.10178	0.17189	0.10843	0.17329	0.11055	0.13791	0.09766
Hou_own_tel	1.39287	0.13356	1.42109	0.12987	1.38551	0.13723	1.38974	0.13989	1.23506	0.13019
Per_wor_prh	0.46957	0.15484	0.40302	0.14926	0.47517	0.16006	0.47208	0.16317	0.50814	0.15210
Per_ind_52	-0.76245	0.21708	-0.78120	0.21073	-0.76326	0.22410	-0.76307	0.22849	-0.73294	0.21214
const	9.54013	0.05525	9.54456	0.05290	9.53566	0.05698	9.53594	0.05791	9.52622	0.05613
Variance Components Estimate	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH** level	Cluster** level
	0.18461	0.04741	NA*	0.04741	0.18820	0.05172	0.18185	0.05498	0.18461	0.04741

Table 2
Regional level estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household (mean = 0.18930) **Based from the ELL results

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.12327	0.00760	-0.12934	0.00689	-0.12377	0.00752	-0.12380	0.00749	-0.11786	0.00997
famsizesqc	0.01096	0.00164	0.01190	0.00147	0.01101	0.00163	0.01102	0.00162	0.01030	0.00195
dom_help	0.81037	0.08873	0.75624	0.10986	0.80727	0.08784	0.80708	0.08751	0.84490	0.08911
wall_light	-0.06808	0.04289	-0.06390	0.03743	-0.06020	0.04272	-0.05973	0.04257	-0.14472	0.04226
wall_strong	0.13761	0.03745	0.15212	0.03469	0.14514	0.03737	0.14560	0.03725	0.06116	0.04249
fa_xs	-0.22074	0.04910	-0.22368	0.04518	-0.22723	0.04875	-0.22761	0.04858	-0.14856	0.05665
fa_s	-0.13540	0.03840	-0.12255	0.03344	-0.13775	0.03805	-0.13789	0.03791	-0.11059	0.04538
fa_l	0.09484	0.03709	0.08894	0.03429	0.09590	0.03676	0.09597	0.03663	0.08529	0.04122
fa_xl	0.16627	0.04315	0.15519	0.04072	0.16938	0.04284	0.16958	0.04269	0.13698	0.04897
fa_xxl	0.33706	0.04545	0.31196	0.04829	0.34173	0.04516	0.34201	0.04500	0.29156	0.05148
fa_xxxl	0.33103	0.06185	0.30377	0.06029	0.33762	0.06134	0.33801	0.06111	0.26052	0.06635
all_hsed	0.33987	0.05253	0.35591	0.04783	0.33807	0.05209	0.33796	0.05189	0.35776	0.04843
all_coed	1.21824	0.05734	1.24762	0.05842	1.20787	0.05692	1.20726	0.05671	1.32979	0.06227
per_kids	-0.24699	0.06440	-0.24047	0.05846	-0.24439	0.06371	-0.24424	0.06347	-0.27423	0.07050
per_61up	-0.14609	0.06126	-0.15938	0.05787	-0.14703	0.06063	-0.14708	0.06040	-0.13525	0.07124
hou_9600	1.13985	0.49103	1.27035	0.47888	1.14320	0.52137	1.14357	0.52172	1.07509	0.51937
Hou_own_ref	1.45233	0.24550	1.51020	0.23864	1.44986	0.26072	1.44985	0.26089	1.44779	0.23585
const	9.36877	0.20322	9.32363	0.19660	9.36597	0.21502	9.36569	0.21512	9.41385	0.21430
Variance Components Estimate	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH** level	Cluster** level
	0.19544	0.03073	NA*	0.03073	0.19052	0.03728	0.18902	0.03748	0.19544	0.03073

Table 3
Provincial level estimates of regression parameters with the standard errors and the variance components for the four techniques. *Different value for each household (mean = 0.23749) **Based from the ELL results

Explanatory	ELL(no hetero)		ELL(w/ hetero)		Pseudo-EBLUP		IWEE		GSR	
Variables	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error	Beta	Std. Error
famsize	-0.1450	0.0175	-0.1489	0.0156	-0.1452	0.0179	-0.1449	0.0171	-0.1413	0.0097
famsizesqc	0.0090	0.0063	0.0124	0.0067	0.0091	0.0065	0.0090	0.0062	0.0085	0.0055
fa_xs	-0.4549	0.1126	-0.3816	0.1010	-0.4552	0.1149	-0.4546	0.1095	-0.4479	0.0718
fa_s	-0.2550	0.0976	-0.2653	0.0794	-0.2545	0.0995	-0.2555	0.0951	-0.2693	0.1198
wall_light	-0.2055	0.0945	-0.1474	0.0778	-0.2057	0.0965	-0.2058	0.0919	-0.2063	0.1070
all_hsed	0.4007	0.1643	0.3531	0.1448	0.4015	0.1673	0.4006	0.1601	0.3891	0.1585
all_coed	1.5411	0.1677	1.8202	0.1769	1.5429	0.1709	1.5429	0.1635	1.5439	0.2326
Hou_own_tel	3.4373	1.0270	3.2630	1.0582	3.4265	1.0622	3.4274	0.9871	3.4392	0.5733
Per_wor_prh	-1.1075	1.1933	-1.5801	1.2008	-1.1049	1.2327	-1.1056	1.1483	-1.1150	0.8729
const	10.0976	0.1480	10.0798	0.1279	10.0988	0.1517	10.0981	0.1435	10.0872	0.1373
Variance Components Estimate	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH level	Cluster level	HH** level	Cluster** level
	0.25753	0.01871	NA*	0.25753	0.26682	0.02079	0.24498	0.01671	0.25753	0.01871

At the regional level, estimates of the regression coefficients are generally similar for all the different estimation methods, except that the GSR and/or ELL_H methods generated estimates for a few variables which were significantly different from the other methods. Similar to the national level estimated standard errors, GSR also tends to be the most conservative method for the majority of the regional level models - it generated the highest estimated standard errors for most of the regression coefficients of the household characteristics. IWEE has the highest estimated standard error for most of the coefficients of the municipal means. The ELL_H method produces the lowest standard errors for the majority of the regression coefficients of the household characteristics and municipal means. The ELL method tends to generate the smallest estimated cluster level variance with ratios to Pseudo-EBLUP and IWEE ranging from around 82% to 100%. The IWEE method still has the smallest household level variance.

Similar to the regional level estimates, the regression coefficients' estimates at the provincial level are similar except for some discrepancies from the GSR and ELL_H estimates. For the estimated standard errors of the regression coefficients, the ELL_H still produces the lowest estimates for the majority of the coefficients of the household characteristics; however, the GSR method (instead of the ELL_H method) now produces the lowest estimated standard error for the majority of the municipal means. The ELL method still tends to generate the smallest estimated cluster level variance for most provinces with the smallest ratio to Pseudo-EBLUP about 53% and to IWEE about 48%. For a number of provinces, IWEE tends to generate the smallest estimated cluster level variance. For the household level variance, IWEE still generated the smallest estimate. Generally, estimates of the cluster level variance tend to be more variable at the provincial level which is due to smaller sample sizes.

For small area estimates of poverty, after the regression model is applied to census data, estimated standard errors in the regression are only one part of the small area estimates' standard errors. There is also variation at the cluster level in (7) that needs to be considered (to different degrees depending on the level of aggregation used to construct the small areas) and there is variation at household level too. These additional sources of variation can be assessed via the estimated variance components. As shown above, regardless of the level (national, regional and provincial) at which the model is formulated, the IWEE method generates the smallest household level variance, while the ELL method generates the smallest cluster level variance. Since the cluster level variation usually makes a much larger contribution to the estimated standard error at the small area level, ELL is again the least conservative. We note that the household level variance under the ELL method with heteroscedasticity model varies from one unit to another, hence, the mean value is reported, and that the estimated R^2 for the heteroscedasticity model is negligible, $R^2 =$ 0.03 even at the national level, so that in terms of regression model fit at least it may offer few advantages for this data set. In our experience with applying the ELL method we have found that heteroscedasticity modeling is unnecessary.

Returning to the regression (*i.e.*, the estimates generated for $\boldsymbol{\beta}$ and the estimated standard error for the different techniques), IWEE is the method that best incorporates the sampling weights from the computation of the variance components necessary for the generation of small area estimates and their estimated standard errors. In terms of implementation, the GSR method would generally be the simplest option as it is available for example in packages such as Stata, Sudaan or WesVar. The ELL method combines sampling weights and covariance structure in a way that is non-standard in that it uses an estimate of $\mathbf{W}_b\mathbf{V}_b^{-1}$ in (8) and (9) to produce an asymmetric estimated covariance matrix for the estimates of $\boldsymbol{\beta}$ and for estimating $\boldsymbol{\beta}$ itself. For estimating $\boldsymbol{\beta}$ this would be acceptable if the asymmetric matrix were a generalized inverse of the correct covariance matrix. It is however clearly not acceptable as an

estimated covariance matrix, a problem ELL attempt to circumvent (*e.g.*, in the World Bank's POVMAP software) by averaging each of the relevant pairs of off-diagonal elements to meet the necessary condition that a covariance matrix be symmetric.

Generally in the ELL method of poverty estimation only variables matching in terms of average and standard deviation in both survey and census plus census averages can be used. This is because, after the regression model has been fitted to the survey data, in the second phase it is applied to the census data as a predictor at household level, *i.e.*, the regression equation (however it has been estimated) is used to find predicted values of per capita income or expenditure for each census household, generated via

$$\hat{Y}_{bh} = \mathbf{x}'_{bh}\hat{\boldsymbol{\beta}} + \hat{v}_b + \hat{e}_{bh} \qquad (25)$$

using imputed values of v_b and e_{bh} (based for example on bootstrap sampling from their survey estimates). Here $\mathbf{x}_{bh}$ are auxiliary variables from the census. Poverty indices are typically based on non-linear functions of log-income or log-expenditure, so the predictions from (25) are transformed appropriately before averaging over each small area. Note that in practice v_b can be estimated for the sampled clusters, but the sample and census codes usually do not match so these cannot be identified in the census, and it is the bootstrap (by selecting from the sampled barangays, *i.e.*, PSUs) that provides imputed values for all barangays; a parallel comment applies to $\hat{e}_{bh}$ for households within clusters. The general benefit of using census data in this way (as ELL does) is that the predictor variables can be used for all census households (of which there are many) not just those in the survey, thereby increasing accuracy of the small area estimates (conditional on the model being correct). Note that the estimates in (25) remain unbiased even if v_b and e_{bh} are not included in the prediction itself, but the variance estimate for small area a needs to be computed based on equation (25) so that it incorporates the necessary additional variation at cluster and household levels.

In poverty estimation, we are interested in area-level summaries of non-linear functions of $\hat{Y}_{bh}$, for example, whether it is below the poverty line (poverty incidence) and poverty gap rather than the regression fitting per se. It is instructive here to examine the effects of model uncertainty on area mean estimates

$$\bar{y}_a = \bar{\mathbf{x}}'_a \hat{\boldsymbol{\beta}} \qquad (26)$$

where $\bar{\mathbf{x}}_a$ is the population (*i.e.*, census) mean for area a of the covariates including the constant 1, after the regression model has been applied to the census data as in phase 2 of ELL. By similarly averaging (7) to get the true mean $\bar{Y}_a$, subtracting from (26), and applying the variance operator, we get the prediction error variance equation:

$$V(\bar{y}_a - \bar{Y}_a) = \bar{\mathbf{X}}_a \boldsymbol{\Phi}_w \bar{\mathbf{X}}'_a + \frac{1}{N_a^2}\sum_{b=1}^{m} N_b^2\, \sigma_v^2 + \frac{1}{N_a}\sigma_e^2 \qquad (27)$$

where N_a is the population size at a particular level of aggregation, N_b is the population size in each cluster, $\boldsymbol{\Phi}_w$ is the variance-covariance matrix of the regression coefficient estimates, and (σ_v^2, σ_e^2) are the cluster and household level variance components, respectively. Note that estimating this prediction error variance requires estimates of the variance components, but any bias caused by uncertainty in these would be a second order effect (see Prasad and Rao 1990).

Based on (27), the extent of the influence of the survey based regression model and other variance components (cluster and household level) on the accuracy of the final small area estimates can be compared for any fitting technique and/or levels of aggregation. Generally, it is either the regression model (via the estimate of the regression parameters) or the cluster effect that dominates the estimated accuracy of the computed small area estimate. Using the national level model in Table 1 and the survey data (instead of the census) auxiliary variables to estimate the first term in (27), shows that the extent to which the regression model effect contributes to small area estimate variance increases markedly as household data are more aggregated - about 0.25% at the municipal level, 20% at the provincial level and 70% at the regional level. In other words, the more aggregated the data into larger areas, the greater the dominance of the regression model parameter uncertainty, regardless of the regression fitting method. This is as expected because even at high levels of aggregation, the contribution to the overall variance from the model effect depends on the average covariate values, not on the population size. This is the reason that, at the most aggregated regional level, small area techniques usually offer little improvement over direct estimates. This is also why it is important (as this paper has done) to examine in detail the regression fitting procedures applied in small area estimation of third world poverty.

The effect of cluster level variation is different: at lower levels of aggregation (*e.g.*, municipality) the computed variance of the small area estimates are dominated by the cluster component of variance or cluster level effect, *i.e.*, for small areas (other than regional estimates) the variance component, not the regression model, has the greatest impact on the value of the standard error of the small area estimates. Consequently, the accuracy of estimates of variance components especially at cluster level can be crucial to accurate estimation of standard error of small area estimates at the aggregation level at which they are most useful (for example at municipal level in the Philippines). Again, this is why the method used for phase 1 fitting for variance components as discussed in this paper, are critical to small area estimation of poverty.

Presented in Tables 4-6 are Kruskal-Wallis (KW) tests (Siegel 1956) for the various fitting methods conducted on the estimated variances at the municipal (Table 4), provincial (Table 5) and regional (Tables 6) levels. In Table 4 significant differences exist among the variance estimates generated by the various small area techniques, as shown by the p-values of the Kruskal-Wallis statistics. Multiple comparison of mean ranks shows the Pseudo-EBLUP and IWEE methods have variance estimates at cluster level that are significantly higher than the other methods, but not significantly different from each other (although for the IWEE method the Z-value for the difference from average rank is in general rather higher than all the others).

The ELL method and the GSR method generate signifycantly lower and similar variance component estimates. This is principally because we used the ELL variance components estimation technique in generating variance components for the GSR method (because GSR does not usually estimate variance components), although the residuals we used were not identical for the two regression fitting methods. As expected, at the municipal level for which small area estimates were used in practice, the cluster effect (rather than regression coefficient uncertainty) is generally the dominant part of the small area variance estimates. Since the ELL and GSR methods have similar cluster level variance, their corresponding variance estimates at small area also tend to be similar. Explicitly, observe from Table 4 that the ranking of the variance estimates generally conforms with the ranking of the cluster effects.

In poverty estimation, estimates at higher levels of aggregation, such as those in Table 5 and 6, are generally carried out for comparison with direct survey estimates at these more aggregated levels, even though they are not particularly useful for aid allocation. The results do however, support those indicated for lower level of aggregation. In Table 5 and Table 6, the estimated variances for the poverty estimates generated by the different techniques are not significantly different from each other at the provincial and regional level, an effect that is partially due to the small number of provinces and even smaller number of regions. The variances and hence the standard errors may not be significantly different from each other, but it is worth noting that the GSR method tends to generate the smallest estimated standard error for the regression model and in turn the smallest variance estimate for poverty at the regional level, even though GSR generates higher standard errors for the individual regression coefficients (corresponding to the diagonal elements only in the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$). As expected, at an even higher level of aggregation for all methods, the relative effect of the regression component is more pronounced.

The general conclusion is that, whether fitting survey data alone or using survey based regression parameter estimates in conjunction with census data, it is crucial not only to find a suitable model (*i.e.*, set of regressors) based on an adequate sample size, but also to get sound estimates of the regression parameters and their standard errors under this model as well as good estimates of the variance components at all relevant levels of aggregation. Usually the relevant levels of aggregation are determined via the survey design, rather than simply through the level at which small area estimates are sought, although the number of levels need not be limited to two (*e.g.*, to cluster-level and household-level).

Survey data, whether used for poverty estimation or in other context, also introduces problems involving survey weights that can be important not only for regression parameter estimation (and their estimated standard errors) but also for estimating variance components. Incorporating survey weights into regression models with correlated data introduces problems because it is the population correlation as it applies to the weighted survey data that needs to be properly modeled, so that weighting correlation matrices using matrix multiplication (as ELL do) is not technically adequate (see Appendix).

For the Philippine data and for the specified list of regressors, regardless of which of the four methods are used, parameter estimates were very similar, which suggests that the more important issue is possible underestimation of standard errors of parameter estimates and of variance components particularly at cluster level. ELL is the least conservative in that it gave the lowest estimates of both variance measures, and in this respect (as with its use of asymmetric estimated covariance matrices) some caution may be warranted with the regression and variance component aspects of the ELL technique. GSR gave similar estimates of standard errors for the small area estimates to ELL when using the same technique for variance components, despite having higher standard errors (and using a sound covariance matrix) for regression parameters. This is because when there is less aggregation, the level at which most small area estimates are actually used, variance components dominate.

The Pseudo-EBLUP and IWEE methods incorporate survey weights correctly (given a suitable choice of pseudo-likelihood and hence GEE) and gave larger (*i.e.*, more conservative) estimates of cluster level variance components. This suggests that these two methods and particularly IWEE are among the best of the currently available methods, not necessarily for estimating regression equations (where availability of standard software may give GSR an advantage), but for estimating the crucial variance components.

Table 4
Kruskal-Wallis test for estimated variances at the municipal level (N = 1,243)

SAE	Cluster Effect			Beta Effect			Variance		
Techniques	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.002843	2,961.2(a)	-3.22	0.0002311	3,067.3(ab)	-0.89	0.00318	2,963.4(a)	-3.18
ELL(w/ hetero)	0.002843	2,961.2(a)	-3.22	0.0002128	2,802.0(c)	-6.72	0.00316	2,930.8(a)	-3.89
Pseudo-EBLUP	0.003094	3,229.4(b)	2.67	0.0002449	3,257.5(ad)	3.28	0.00346	3,241.3(b)	2.93
IWEE	0.003294	3,426.9(b)	7.01	0.0002529	3,364.5(d)	5.64	0.00366	3,441.3(b)	7.32
GSR(Stata)	0.002843	2,961.2(a)	-3.22	0.0002311	3,048.7(b)	-1.3	0.00317	2,963.1(a)	-3.18
Overall		3,108			3,108			3,108	
KW Statistic	H = 69.92	(P = 0.000)		H = 72.19	(P = 0.000)		H = 78.06	(P = 0.000)	

Table 5
Kruskal-Wallis test for estimated variances at the provincial level (N = 83)

SAE	Cluster Effect			Beta Effect			Variance		
Techniques	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.0002518	200.3	-0.65	0.0001162	207.7	-0.03	0.00039	202.3	-0.48
ELL(w/ hetero)	0.0002518	200.3	-0.65	0.0001095	190.1	-1.52	0.00038	196.3	-0.99
Pseudo-EBLUP	0.000274	214.9	0.59	0.0001239	224.2	1.37	0.00042	217.1	0.78
IWEE	0.0002916	224.2	1.38	0.0001287	234.1	2.22	0.00045	227.8	1.68
GSR (Stata)	0.0002517	200.3	-0.65	0.00010	184	-2.04	0.00037	196.4	-0.98
Overall		208			208			208	
KW Statistic	H = 2.82	(P = 0.589)		H = 10.61	(P = 0.031)		H = 4.48	(P = 0.344)	

Table 6
Kruskal-Wallis test for estimated variances at the regional level (N = 16)

SAE	Cluster Effect			Beta Effect			Variance		
Techniques	Median	Mean Rank	Z	Median	Mean Rank	Z	Median	Mean Rank	Z
ELL(no hetero)	0.000050	38.2	-0.45	0.000077	40.9	0.08	0.00013	39.3	-0.23
ELL(w/ hetero)	0.000050	38.2	-0.45	0.000073	35.1	-1.05	0.00012	37	-0.67
Pseudo-EBLUP	0.000055	42.6	0.4	0.000082	46.9	1.23	0.00014	44	0.67
IWEE	0.000058	45.3	0.93	0.000085	50.1	1.85	0.00015	46.6	1.17
GSR(Stata)	0.000050	38.2	-0.45	0.000070	29.6	-2.1	0.00013	35.6	-0.94
Overall		40.5			40.5			40.5	
KW Statistic	H = 1.30	(P = 0.861)		H = 8.36	(P = 0.079)		H = 2.58	(P = 0.630)	

Of course, such considerations (while central) need to be predicated by adequate data cleaning, sound matching of possible regressor variables (in terms of mean, variance, and meaning) between survey and census where census data is also being used. Also needed are the proper, time consuming consideration of a wide range of possible regressor variables and recognition of the limits placed on subdividing survey data by small sample sizes, since all estimated standard errors for both regression parameter and small area estimates (whatever method is used for fitting the variance component estimate) are conditional on the regression model being correct.

8. Conclusion and recommendation

There is a great need for sound poverty statistics in order to effectively monitor interventions and assistance to various impoverished localities. Small area estimation techniques are one methodology that is being used to provide such statistics. In this sense the issues raised in this paper concerning the accuracy of the small area estimates are not simply an academic issue but are central to the Millennium Development Goals and to aid allocation in what is a multi-billion dollar industry.

In this paper, we have considered four estimation techniques for fitting regression models using survey data and related them to small area poverty estimation. We have shown that although differences in estimates are insufficient to invalidate the published national studies, the most frequently implemented survey data fitting technique, ELL with heteroscedasticity, recommended by the World Bank, has some limitations since (like its homoscedastic version) it lacks sound theoretical underpinning. Replacing the survey fitting part of the ELL method is recommended. For the other methodologies considered (the Pseudo-EBLUP, IWEE, and the GSR method), all have valid theoretical basis mathematically and the results generated can be clearly interpreted once the assumptions have been checked. The different methodologies when applied to complex weighted survey data from the Philippines indicate that for variance component estimation from survey data and hence for small area estimation at a fine level, Pseudo-EBLUP and particularly IWEE are likely to be better than the GSR or the ELL methods, although GSR is sound and easy to use because it is available in off-the-shelf software.

We have also shown that at the level where small area estimation is actually used for aid allocation, the variance estimate of the small area tends to be dominated by the cluster level variance rather than by the accuracy of the regression parameter estimates. Hence, it is particularly important that the cluster-level component of variance (and, if fitted as recommended, any small area level variance component) is properly estimated. It is also important that the regression model used in the generation of small area estimates (including choice of suitable regressors) is appropriate. Essentially, at lower levels of aggregation it is the variance components that dominate the standard error of the small area estimates, so that the estimation of the variance components is critical whatever the choice of aggregation level. Sound survey-based regression method, good choice of regression variables, and care with sample size (especially if separate regression models are fitted to subsets of survey data), also remain central to sound small area estimation of third world poverty.

Acknowledgements

The authors would like thank the referees and the Associate Editor for their careful reading of the manuscript and for their helpful suggestions.

Appendix

In footnote 8 of the Elbers *et al.* (2002) World Bank working paper and implicitly in Elbers *et al.* (2003) in *Econometrica*, the covariance of the error process is denoted $\mathbf{\Omega}$ and it is stated that $\mathbf{W\Omega}^{-1} = \mathbf{P}^T\mathbf{P}$ where $\mathbf{W}$ is 'a weighting matrix of expansion factors'. In the notation of Section 4 above, $\mathbf{W}$ is block diagonal with or diagonal with diagonal blocks $\mathbf{W}_b$, and $\mathbf{\Omega}$ is block diagonal with diagonal blocks $\mathbf{V}_b$.

However, either $\mathbf{W}$ and $\mathbf{\Omega}$ (or $\mathbf{\Omega}^{-1}$) are non-conformable (with weighting factors in $\mathbf{W}$ at cluster level and the observations and hence $\mathbf{\Omega}^{-1}$ at individual level), or if conformable $\mathbf{W\Omega}^{-1}$ is generally asymmetric (even if W is diagonal) unless $\mathbf{W}$ is a simple multiple of the identity matrix, *i.e.*, $\mathbf{W} = \sigma^2\mathbf{I}$.

Hence, $\mathbf{W\Omega}^{-1}$ does not equal $\mathbf{P}^T\mathbf{P}$ as has been claimed since $\mathbf{P}^T\mathbf{P}$ is symmetric in general and $\mathbf{W\Omega}^{-1}$ is not. Making $\mathbf{W\Omega}^{-1}$ symmetric by adding it to its transpose and dividing by two, as is done in the World Bank PovMap software, is not a technically adequate solution to this problem. (Note that even in the simple case where $\mathbf{W}$ and $\mathbf{\Omega}^{-1}$ are conformable, and $\mathbf{W}$ is diagonal but not all diagonal elements are equal, $\mathbf{W\Omega}^{-1}$ is not diagonal because it has every element of row i of $\mathbf{\Omega}^{-1}$ multiplied by w_i (where w_i is the i^{th} diagonal element of $\mathbf{W}$) but the i^{th} column does *not* have every element multiplied by an identical weight.)

Putting this issue of symmetry to one side, and using $\mathbf{P}^T\mathbf{P}$ in place of $\mathbf{W\Omega}^{-1}$, ELL seem to be claiming that comparing their 'sample survey adjusted weighted GLS estimator' to the 'unadjusted GLS' estimator implies that instead of using $\mathbf{\Omega}^{-1}$ as the underlying metric (*i.e.*, the inverse of the relevant covariance matrix), a weighted version namely $\mathbf{W\Omega}^{-1}\mathbf{W}^T$ should be used. This creates no asymmetry issue in itself (provided $\mathbf{P}^T\mathbf{P}$ were used in place of $\mathbf{W\Omega}^{-1}$). However, even if $\mathbf{W}$ were diagonal and $\mathbf{P}^T\mathbf{P}$ used, the weight matrix $\mathbf{W}$ cannot use even unequal diagonal weights corresponding to the sampled units, *i.e.*, w_i say, because the ij^{th} element of $\mathbf{\Omega}^{-1}$ (unlike the ij^{th} element of $\mathbf{\Omega}$) does *not* correspond to the i^{th} and j^{th} unit in the sample (or in the population), so it is rather unclear what $\mathbf{W}$ is or how $\mathbf{W}$ can be sensibly defined as 'a weighting matrix of expansion factors'.

This argument still applies when $\mathbf{V}_b$ is replaced by its estimator $\hat{\mathbf{V}}_b$ which uses estimates in place of σ_e^2 and σ_v^2.

References

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association*, 83, 28-36.

Chambers, R. (2006). What is poverty? Who asks? Who answers? *Poverty in Focus*, UNDP, December 2006, 3-4.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355-364.

Elbers, C., Lanjouw, J. and Lanjouw, P. (2002). *Micro-level Estimation of Welfare.* Research Working Paper 2911, World Bank, Development Research Group, Washington, D.C.

Ghosh, M., and Rao J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Haslett, S., and Jones, G. (2004). *Local Estimation of Poverty and Malnutrition in Bangladesh*, Bangladesh Bureau of Statistics and United Nations World Food Programme.

Haslett, S., and Jones, G. (2005). *Local Estimation of Poverty in the Philippines*, Philippine National Statistics Co-ordination Board/World Bank Report. http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Local_Estimation_of_Poverty_Philippines.pdf.

Haslett, S., and Jones, G. (2005). Small area estimation using surveys and censuses: Some practical and statistical issues. *Statistics in Transition*, 7, 541-556.

Haslett, S., and Jones, G. (2006). *Small Area Estimation of Poverty, Caloric Intake and Malnutrition in Nepal.* Published: Nepal Central Bureau of Statistics/World Food Programme, United Nations/World Bank, September 2006, 184pp, ISBN 999337018-5.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

Horton, N.J., and Lipsitz, S.R. (1999). Review of software to fit Generalized Estimating Equation regression models. *The American Statistician*, 53, 160-169.

Liang, K.L., and Zeger, S. (1986). Longitudinal data analysis using Generalized Linear Models. *Biometrika*, 73, 13-22.

Lohr, S.L. (1999). *Sampling*: *Design and Analysis.* Duxbury Press, Brooks/Cole Publishing Company.

Militino, A.F., Ugarte, M.D., Goicoa, T. and Gonzalez-Audicana, M. (2006). Using small area models to estimate the total area occupied by olive trees. *Journal of Agricultural*, *Biological*, *and Environmental Statistics*, 11, 450-461.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society* B, 60, 23-40.

Pfeffermann, D., Moura, F.A. and Silva, P.L. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 949-959.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.

Rao, J.N.K. (2003). *Small Area Estimation*, Wiley Series in Survey Methodology. Wiley-Interscience, John Wiley & Sons, Inc.

NSCB (2000). *Profile of Censuses and Surveys*. National Statistical Coordination Board, Philippines.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences.* McGraw-Hill Series in Psychology. New York: McGraw-Hill.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Chichester: John Wiley & Sons.

UN website. *http://www.un.org/milleniumgoals/*.

You, Y., and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. and Kovačević, M. (2003). Estimating fixed effects and variance components in a random intercept model using survey data. *Proceedings*: *Symposium 2003*, *Challenges in Survey Taking for the Next Decade*. Statistics Canada.

Zhao, Q. (2006). User manual for PovMap, The World Bank. http://siteresources.worldbank.org/INTPGI/Resources/342674-1092157888460/Zhao_ManualPovMap.pdf.

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 171-180
Statistics Canada, Catalogue No. 12-001-X

Small area estimation of the number of firms' recruits by using multivariate models for count data

Maria Rosaria Ferrante and Carlo Trivisano [1]

Abstract

The number of people recruited by firms in Local Labour Market Areas provides an important indicator of the reorganisation of the local productive processes. In Italy, this parameter can be estimated using the information collected in the Excelsior survey, although it does not provide reliable estimates for the domains of interest. In this paper we propose a multivariate small area estimation approach for count data based on the Multivariate Poisson-Log Normal distribution. This approach will be used to estimate the number of firm recruits both replacing departing employees and filling new positions. In the small area estimation framework, it is customary to assume that sampling variances and covariances are known. However, both they and the direct point estimates suffer from instability. Due to the rare nature of the phenomenon we are analysing, counts in some domains are equal to zero, and this produces estimates of sampling error covariances equal to zero. To account for the extra variability due to the estimated sampling covariance matrix, and to deal with the problem of unreasonable estimated variances and covariances in some domains, we propose an "integrated" approach where we jointly model the parameters of interest and the sampling error covariance matrices. We suggest a solution based again on the Poisson-Log Normal distribution to smooth variances and covariances. The results we obtain are encouraging: the proposed small area estimation model shows a better fit when compared to the Multivariate Normal-Normal (MNN) small area model, and it allows for a non-negligible increase in efficiency.

Key Words: Multivariate Poisson-Log Normal distribution; Zero counts; Generalized Variance Function; Hierarchical Bayesian models.

1. Introduction

The number of people recruited by firms for a certain period can be taken as a key indicator of ongoing changes in the economic system. To highlight the dynamic of the demand for local labour, we consider the number of people recruited by firms in Local Labour Market Areas (LLMAs), these last grouped according to i) productive specialization, ii) firms' size classes and iii) industrial sector. Domains are defined by cross-classifying these three variables. In order to emphasise the signals of the reorganisation of the productive process, we focus on the numbers of "recruits replacing employees leaving the firm (substitute recruits – SR)" and "recruits filling new positions (new recruits – NR)". In Italy, information about firms' recruits is collected by the Excelsior Survey co-sponsored by the Union of Italian Chambers of Commerce (UNIONCAMERE), the Ministry of Labour and the European Union. Unfortunately, this survey does not provide reliable estimates of firms' recruits for each of these domains due to small domain sample size. As a consequence, a small area estimation (SAE) technique has to be adopted in order to obtain estimates with an acceptable degree of variability.

In this paper, we propose a SAE approach for the estimation of counts. Due to data constraints, we adopt an aggregated area-level model.

Since we aim at estimating SR and NR, we adopt a multivariate SAE model that borrows strength not only from areas but also from the correlations between the NR and SR true values. In order to estimate the median income of different sized groups of families, Fay (1987) proposed a multivariate regression model in an Empirical Bayes context. Multivariate SAE approaches have also been developed by Ghosh, Nangia and Kim (1996) and Datta, Fay and Ghosh (1991), Datta, Ghosh, Nangia and Natarajan (1996) and Datta, Lahiri, Maiti and Lu (1999) for continuous data in the hierarchical cross-section time series model framework. Fabrizi, Ferrante and Pacei (2005, 2008) adopted multivariate area level models to estimate a vector of continuous poverty parameters. As in the univariate Fay-Herriot model (Fay and Herriot 1979), all of the papers mentioned above assume the use of small area normal sampling and linking models.

Since the sampling correlations between SR and NR estimators are mainly negative, we propose a SAE model based on the Multivariate Poisson-Log Normal (MPLN) distribution. Unlike other multivariate distributions for counts proposed in the literature, this particular distribution allows for unconstrained (that is, both positive and negative) correlations (Aitchison and Ho 1989).

We also deal with the instability of estimators of sampling error variances and covariances. An approximately unbiased estimate of the variance of direct estimators is

1. Maria Rosaria Ferrante, Department of Statistics - University of Bologna, Italy. E-mail: maria.ferrante@unibo.it; Carlo Trivisano, Department of Statistics - University of Bologna, Italy. E-mail: carlo.trivisano@unibo.it.

usually available in SAE. However, in area-level models it is customary to assume that the sampling variance is known and equal to its estimate (Rao 2003; page 76). This assumption is commonly stated and largely accepted in the case of large samples, whereas both the variance estimator and direct point estimators suffer from instability in the case of small samples. As a partial solution, sampling variance estimates are often smoothed through the generalized variance functions (GVF) approach (Wolter 1985). In You, Rao and Gambino (2003), sampling variances and covariances were smoothed over areas and times. In order to consider the extra variability associated with the estimated sampling variances, Arora and Lahiri (1997) proposed an integrated Hierarchical Bayes (HB) smoothing approach for continuous data. See You and Chapman (2006), Liu, Lahiri and Kalton (2007) and You (2008) for different extensions of Arora and Lahiri (1997).

Due to the rarity of recruits in certain domains, a further problem arises that is linked to the instability of sampling error variances and covariances estimators. When direct estimates of SR or NR (or both) are equal to zero, estimated sampling error variances and covariances are also equal to zero. Note that observing estimated variances equal to zero does not necessarily imply that the estimates have a high degree of accuracy. This problem was encountered in previous small area estimation problems (*e.g.*, Elazar 2004; Chattopadhyay, Lahiri, Larsen and Reimnitz 1999). Chen (2001) proposed a unit level hierarchical modeling to handle the problem. Moreover, some studies (Cohen 2000) use the logarithmic transformation of the mean (or total) direct estimates of the count data in order to adopt a linear SAE model, simply discarding the estimates equal to zero. Although this solution overcomes the "zero variance" problem, it also leads to biased estimates and neglects a portion of the sample.

In order to deal with the instability of variances and covariances estimators as well as the problem of estimated sampling variances equal to zero, we suggest an "integrated" approach in the spirit of that proposed by Arora and Lahiri (1997), Liu *et al.* (2007) and You (2008). Within an HB framework, we jointly model the parameters of interest and the sampling error covariance matrices by adopting a smoothing covariance solution based once again on the Poisson-Log Normal distribution.

The layout of this paper is as follows. The data set employed is described in section 2, while section 3 presents direct domain estimation and its associated sampling error variances and covariances. In section 4, we describe the multivariate SAE model we propose for estimating counts as well as the solution we suggest for overcoming the instability of sampling error variances and covariances estimators in the presence of zero counts. Section 5 reports the results obtained by measuring the performance of the adopted SAE model. Details on the Poisson-Log Normal distribution are given in the Appendix.

2. The excelsior survey

The Excelsior Survey is one of the most complete Italian statistical sources for labour demand data, providing estimates of the number of people recruited by Italian firms. Each year, a stratified simple random sample of about 100,000 firms with at least one employee is contacted and asked about the number of people it plans to hire in the short term. The factors used for stratification are the firm's industrial sector and size class. The allocation of the sample in the strata satisfies a constraint on the maximum estimated standard error corresponding to a 95% significance level (Baldi, Bellisai, Fivizzani and Sorrentino 2007). By focusing on local geographical details, the survey is designed to produce reliable estimates for the administrative provinces (NUTS3, following the "Nomenclature of Units for Territorial Statistics" reported in http://europa.eu.int/comm/eurostat/ramon/nuts). This geographical unit, singled out on the basis of administrative criteria, does not appear to be the best choice when analysing the dynamics of the local labour demand. In order to shed some light on the signals of the reorganization of the local productive process, a better territorial subdivision would be LLMAs (following the OECD definition). LLMAs are groups of municipalities sharing the same labour market conditions (for the location of LLMAs in Italy, see Sforzi 1991). In Italy, following the strategy proposed by Sforzi and Lorenzini (2002) and adopted by the Italian Statistical Institute (ISTAT), certain LLMAs are labelled "industrial districts" (IDs). IDs are geographically defined productive systems characterized by a dominant specialization. In the 1990s, these were considered to be the main stimulus for the growth of the Italian economic system (Becattini 1992).

Estimating the number of substitute and new recruits in firms operating within/outside of IDs can help us verify whether IDs are still a source of dynamism for the Italian economy as a whole. In order to refer to types of ID, we group them according to their productive specialization. Similarly, LLMAs not labelled as IDs can be classified according to their economic vocation (LLMAs can be characterized by a specific manufacturing activity, tourist area, city, *etc*.). Moreover, the comparison between ID and non-ID firms makes economic sense if the industrial sector and size of the firms are also taken into account. Finally, as already noted, domains of interest are defined by cross-classifying: i) groups of LLMAs obtained according to their productive specialization, ii) firm's industrial sector and iii) firm's size.

This paper focuses on the manufacturing sector characterising the IDs' economic activity. The analysis is limited to two Italian regions containing a large quantity of IDs, namely Tuscany and Emilia-Romagna, and to firms with fewer than 100 employees (as censuses are taken for the other size classes). The target population consists of 54,089 firms employing a total of 809,059 people.

3. Direct estimates

Table 1 provides details of the categories defining the 208 domains of interest. Note that the number of domains is less than that expected due to the absence of a number of domains within the population. The domains are unplanned since they are formed grouping LLMAs contained in the same planned stratum. For the sake of simplicity, in the following we avoid using the stratum subscription wherever possible.

Let θ_{i1} and θ_{i2} be the true number of NR and SR for domain i ($i = 1, ..., 208$), respectively. We shall first define a direct estimator of θ_{ij} ($i = 1, ..., 208;\ j = 1, 2$). Let y_{ijl} be the response of the l^{th} unit related to the j^{th} variable in the i^{th} domain ($l = 1, ..., n_i$, where n_i is the sample size in domain i; $i = 1, ..., 208;\ j = 1, 2$). As design based (direct) estimator we use a ratio domain estimator defined as $\hat{\theta}_{ij} = \sum_{l=1}^{n_i} y_{ijl} / (n_i / N_i)\ N_i / \hat{N}_i$, where N_i and n_i are respectively the population size and the sampling size referred to domain i, and $\hat{N}_i = n_i / n_{t\ni i}\ N_{t\ni i}$, where $N_{t\ni i}$ and $n_{t\ni i}$ are respectively the population size and the sampling size of the stratum t containing the domain i (Särndal, Swensson and Wretman 1992; page 391).

Since we are estimating the number of occurrences of rare events, in 50 of the 208 domains, direct estimates of NR and/or of SR are equal to zero, that is, $\hat{\theta}_{i1} = 0$ and/or $\hat{\theta}_{i2} = 0$. Zero point estimates imply that $\hat{V}(\hat{\theta}_{i1}) = 0$ and/or $\hat{V}(\hat{\theta}_{i2}) = 0$, where $\hat{V}(\hat{\theta}_{i1})$ and $\hat{V}(\hat{\theta}_{i2})$ are the standard design-based variance estimates of $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$, respectively. This result gives a false impression of high accuracy, whereas the exact opposite is more likely to be true in a small area context. Moreover, design based estimates of NR and/or of SR equals to zero produce $\hat{\text{COV}}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$, where $\hat{\text{COV}}(\hat{\theta}_{i1}, \hat{\theta}_{i2}) = 0$ denotes the standard design-based estimate of the design-based covariance between $\hat{\theta}_{i1}$ and $\hat{\theta}_{i2}$. As a result, covariances also need to be smoothed in a multivariate SAE model.

We hereafter refer to the set of the 50 small areas having one or both zero estimated variances and zero covariances as the "Zero Count" (ZC) set. The complementary set of 158 domains, where $\hat{V}(\hat{\theta}_{i1}) > 0$ and $\hat{V}(\hat{\theta}_{i2}) > 0$, is named the "Non Zero Count" (NZC) set.

Considering the data generating process and the nature of the outcome variables, we expect mainly negative correlations between θ_{i1} and θ_{i2}. Briefly, we need a suitable distribution for both smoothing covariance matrices and modeling small area parameters that allows for an unrestricted covariance matrix, that is, for both positive and negative correlations.

Table 1
Variables defining domains of interest

LLMAs grouped by productive specialization	Firm size [(b)]	Industrial sector[(a)]
Industrial district[(a,c)]	1-9	1 Food, beverages and tobacco
Food, beverages and tobacco	10-49	2 Textiles and clothing
Textiles and clothing	50-99	3 Paper products, printing and publishing
Paper products, printing and publishing	≥ 100	4 Machinery
Machinery		5 Chemicals and basic metals
Jewellery, musical instruments, games, *etc*.		6 Leather and footwear
Leather and footwear		7 Wood, furniture and household equipment
Wood, furniture and household equipment		8 Jewellery, musical instruments, games, *etc*.
LLMAs not defined as district [(c)]		9 Builders, contractors
Non-specialised manufacturing		10 Other manufacturing
Non-specialized, excluding manufacturing		
Tourist		
Cities		

(a) As defined by the 2-digit ATECO 91-ISIC 3 level classification and by Sforzi (1991).
(b) Defined according to the number of employees.
(c) Defined in accordance with Istat (1997).

4. An integrated multivariate small area model for count data

Multivariate count data can have a non-trivial correlation structure. In general, the modeling of this structure significantly affects the estimators' efficiency and the computation of correct standard errors. A number of multivariate models for count data have been proposed in the literature, such as the Multivariate Poisson, Multivariate Negative Binomial and Multivariate Poisson-Gamma Mixture models (for a review of such models, see Winkelmann 2003). Unfortunately, these distributions are not suitable for modeling our data since they are based on the hypothesis that correlation is the result of an individual factor that does not vary across outcomes, thus implying a covariance structure restricted to non-negative correlations. In the bivariate case, a more flexible covariance structure is provided by the Latent Poisson Normal distribution (van Ophem 1999); however, any extensions to higher dimensional multivariate data appear impractical.

Aitchison and Ho (1989) proposed a *d*-variate distribution that allows for an unrestricted covariance structure, the Multivariate Poisson-Log Normal distribution (MPLN). No closed form exists for this distribution, but it can be represented as a simple mixture allowing for parameter estimation in an MCMC approach (Chib and Winkelmann 2001). Details of the MPLN distribution are provided in the Appendix.

4.1 Smoothing sampling covariance matrices

As previously mentioned, the instability of standard errors in SAE is usually dealt with using a GVF approach. In this section, we present a GVF model with a regression function inspired by the MPLN distribution.

Let $\mathbf{y}_{il} = [y_{i1l},\ y_{i2l}]'$ be the vector of the two outcome variables referring to the l^{th} unit in the i^{th} domain. Let $\mathbf{y}_{il} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \perp \mathbf{y}_{il'} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i$ and $\mathbf{y}_{il} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i \sim \text{PLN}_2(\boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i)$, $\forall i,\ \forall l$. Under these hypotheses, the moments leading up to the second order can be expressed as follows:

$$E(y_{ijl} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \exp(\lambda_{ij} + \sigma_{i,jj}/2) = \zeta_{ij}$$

$$V(y_{ijl} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \zeta_{ij} + \zeta_{ij}^2[\exp(\sigma_{i,jj}) - 1]$$

$$\text{COV}(y_{ijl}, y_{ihl} \mid \boldsymbol{\lambda}_i, \boldsymbol{\Sigma}_i) = \zeta_{i1}\zeta_{i2}[\exp(\sigma_{i,jh}) - 1],\ j \neq h$$

where $\sigma_{i,jh}$ denotes the (j, h), $j, h = 1, 2,$ element of $\boldsymbol{\Sigma}_i$.

To deal with the problem of smoothing covariance matrices, Otto and Bell (1995), suggested an approach based on a Wishart distributional assumption; specifically, they used smoothed estimates in a small area Normal-Normal model. In the same spirit, we propose a Bayesian approach using the following GVF strategy. Under simple random sampling, let us assume that the sampling covariance matrix in domain i, $\mathbf{C}_i$ follows a Wishart distribution with $n_i - 1$ degrees of freedom:

$$\mathbf{C}_i \mid n_i, \boldsymbol{\Gamma}_i \sim W_2(n_i - 1, \boldsymbol{\Gamma}_i)$$

where $\boldsymbol{\Gamma}_i = E(\mathbf{C}_i \mid n_i, \boldsymbol{\Gamma}_i)$, $i = 1, 2, ..., 158,$ and elements (j, h) of $\mathbf{C}_i$ are defined as $C_{i,jh} = n_i^{-1}\sum_{i=1}^{n_i}(y_{ijl} - \bar{y}_{ij})(y_{ijh} - \bar{y}_{ih})$, where $\bar{y}_{ij} = n_i^{-1}\sum_{i=1}^{n_i} y_{ijl}$.

If ζ_{ij} parameters are known, then $E(\mathbf{C}_i \mid n_i, \boldsymbol{\Gamma}_i)$ only depends on elements of the $\boldsymbol{\Sigma}_i$ matrix. We propose to estimate ζ_{ij} using the design based estimator $\hat{\zeta}_{ij} = N_i^{-1}\hat{\theta}_{ij}$. Thus, we can express each element of the $\boldsymbol{\Gamma}_i$ matrix as a function of estimates $\hat{\zeta}_{ij}$ and of the elements of the $\boldsymbol{\Sigma}_i$ matrix:

$$\Gamma_{i,11} = \hat{\zeta}_{i1} + \hat{\zeta}_{i1}^2(\exp(\sigma_{i,11}) - 1)$$

$$\Gamma_{i,22} = \hat{\zeta}_{i2} + \hat{\zeta}_{i2}^2(\exp(\sigma_{i,22}) - 1)$$

$$\Gamma_{i,12} = \hat{\zeta}_{i1}\hat{\zeta}_{i2}(\exp(\sigma_{i,12}) - 1)$$

where $\sigma_{i,11} = \bar{\boldsymbol{\sigma}}_{11}'\mathbf{Z}_i$, $\sigma_{i,22} = \bar{\boldsymbol{\sigma}}_{22}'\mathbf{Z}_i$, $\sigma_{i,12} = \bar{\boldsymbol{\sigma}}_{12}'\mathbf{Z}_i$, being $\mathbf{Z}_i$ is a 3×1 vector of dummy variables identifying the firm's size class in the domain *i*, and

$$\bar{\boldsymbol{\sigma}}_{11} = \begin{pmatrix} \bar{\sigma}_{1,11} \\ \bar{\sigma}_{2,11} \\ \bar{\sigma}_{3,11} \end{pmatrix}, \bar{\boldsymbol{\sigma}}_{22} = \begin{pmatrix} \bar{\sigma}_{1,22} \\ \bar{\sigma}_{2,22} \\ \bar{\sigma}_{3,22} \end{pmatrix}, \bar{\boldsymbol{\sigma}}_{12} = \begin{pmatrix} \bar{\sigma}_{1,12} \\ \bar{\sigma}_{2,12} \\ \bar{\sigma}_{3,12} \end{pmatrix}$$

that is, we assume that parameters $\boldsymbol{\Sigma}_i$ are equal for domains belonging to the same firm size class.

We estimate $\bar{\boldsymbol{\sigma}}_{11}, \bar{\boldsymbol{\sigma}}_{22}, \bar{\boldsymbol{\sigma}}_{12}$ parameters on NZC data. Since we are following a Bayesian approach, prior specifications for $\bar{\sigma}_{k,jj}$ and $\bar{\sigma}_{k,12}$ $k = 1, 2, 3$ are needed. We use the following prior specifications: $\bar{\sigma}_{k,11}^{1/2} \sim U^+$, $\bar{\sigma}_{k,22}^{1/2} \sim U^+$, $\bar{\rho}_k \sim U(-1,1)$, where $\bar{\sigma}_{k,12} = \bar{\rho}_k(\bar{\sigma}_{k,11}\bar{\sigma}_{k,22})^{1/2}$ and U^+ denotes a uniform distribution over a subset of R^+ with a large but finite length. In section 4.3, we show how these estimates can be used to integrate the SAE model with a model for sampling error covariance matrices.

4.2 A Multivariate Normal-Poisson-Log Normal small area model

In this section, we propose a multivariate SAE model based on the MPLN distribution in order to jointly estimate SR and NR using the NZC set.

Let $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})^T$ be the vector of the two parameters of interest for the i^{th} domain in the set of NZC data $(i = 1, ..., 158)$, and let $\hat{\boldsymbol{\theta}}_i$ be the corresponding vector of direct estimates. The SAE model consists of two separate models. The first model is a sampling model:

$$\hat{\boldsymbol{\theta}}_i \mid \boldsymbol{\theta}_i \sim \text{ind}\ N_2(\boldsymbol{\theta}_i \mid \boldsymbol{\Psi}_i), \quad i = 1, ..., 158. \qquad (1)$$

As in Lahiri and Rao (1995), we justify the normality assumption in (1) using the central limit argument. It is standard practice to assume that sampling error covariance matrices $\boldsymbol{\Psi}_i$ are known, and a GVF method is generally used to estimate $\boldsymbol{\Psi}_i$. Here, as a smoothed estimation of $\boldsymbol{\Psi}_i$ we adopt $\hat{\boldsymbol{\Psi}}_i = E(\boldsymbol{\Gamma}_i | \mathbf{C}_i, n_i) K_i$, where $K_i = N_i (N_{t\ni i}/n_{t\ni i} - 1)$. From this point on we will refer to $\hat{\boldsymbol{\Psi}}_i$ as Smoothed Sampling Error Covariance matrix (SMSEC).

The second component of the SAE model is a linking model that relates $\boldsymbol{\theta}_i$ to area specific auxiliary data:

$$\boldsymbol{\theta}_i \sim \text{ind PLN}_2(\boldsymbol{\eta}_i, \boldsymbol{\Sigma}_v), \quad i = 1, ..., 158,$$

where (2)

$$\boldsymbol{\eta}_i = \boldsymbol{\alpha} + \boldsymbol{\gamma}\mathbf{Z}_i + \boldsymbol{\beta}\mathbf{Z}_i x_i$$

$\mathbf{Z}_i$ is a 3×1 vector of dummy variables identifying the firm's size class in the domain i and $x_i = \log(x_i^*)$, where x_i^* is the number of employees in the domain i.

At the end, $\boldsymbol{\Sigma}_v$ is the covariance matrix related to the area-specific random effects:

$$\boldsymbol{\Sigma}_v = \begin{pmatrix} \sigma_{v,11} & \sigma_{v,12} \\ \sigma_{v,21} & \sigma_{v,22} \end{pmatrix}$$

and

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} 0 & \gamma_{12} & \gamma_{13} \\ 0 & \gamma_{22} & \gamma_{23} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{pmatrix}.$$

From here on, we refer to this small area model as "Multivariate Normal-Poisson-Log Normal" (MNPLN).

We adopt a fully hierarchical Bayesian approach. In this framework, relatively complex (*e.g.*, multivariate) models can be implemented easily; in addition, posterior distributions can be approximated using MCMC algorithms. Computing small area multivariate estimates, and estimates of their MSE in particular, can be difficult within a frequentist approach. The specification of priors for the described model is as follows:

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \sim N_2(\mathbf{0}, a\mathbf{I}_2),$$

$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \sim N_2(\mathbf{0}, g_{k'}\mathbf{I}_2) \qquad k' = 2, 3,$$

$$\begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix} \sim N_2(\mathbf{0}, b_k\mathbf{I}_2) \qquad k = 1, 2, 3,$$

$$\boldsymbol{\Sigma}_v^{-1} \sim W(s, \mathbf{I}_2),$$

$$\begin{pmatrix} \gamma_{1k'} \\ \gamma_{2k'} \end{pmatrix} \perp \begin{pmatrix} \beta_{1k} \\ \beta_{2k} \end{pmatrix},$$

where $s = 3$ and a, $g_{k'}$, b_k are large compared with the scale of the data. This is to reflect the lack of prior information about model parameters, thus defining diffuse but proper specification of priors. The posterior means $\hat{\boldsymbol{\theta}}_i^{\text{HB}} = E(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Psi}}_i)$ are taken as estimators of the area parameters, while the posterior variance $V(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Psi}}_i)$ is used as a measure of uncertainty.

For the sake of comparison, we take the standard Multivariate Normal-Normal (MNN) model as a benchmark, where the sampling model is defined as in (1) and the linking model is defined as follows:

$$\boldsymbol{\theta}_i \sim \text{ind } N_2(\boldsymbol{\mu}_i^*, \boldsymbol{\Sigma}_v^*), \tag{3}$$

where $\boldsymbol{\mu}_i^* = \boldsymbol{\alpha}^* + \boldsymbol{\gamma}^*\mathbf{Z}_i + \boldsymbol{\beta}^*\mathbf{Z}_i x_i^*$. Parameters $\boldsymbol{\alpha}^*$, $\boldsymbol{\gamma}^*$, $\boldsymbol{\beta}^*$ and their prior distributions are defined as $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ in the previous model.

4.3 An integrated MNPLN small area model

In order to account for the extra variability due to the estimated covariance matrices of sampling errors, as well as to overcome the zero variances and covariances problem, we suggest a solution in the spirit of that proposed by Arora and Lahiri (1997), Liu *et al.* (2007) and You (2008). We integrate the model for sampling error covariance matrices of section 4.1 into SAE models (1) and (2). Thus, we here refer to the whole set of 208 domains.

In this context, the small area sampling model is formulated as usual, that is, $\hat{\boldsymbol{\theta}}_i | \boldsymbol{\theta}_i \sim \text{ind } N_2(\boldsymbol{\theta}_i, \boldsymbol{\Psi}_i^*)$, $i = 1, ..., 208$. Under the hypotheses regarding $\mathbf{y}_{il}$ formulated in section 4.1, assuming that the $\boldsymbol{\Sigma}_i$s are known and assuming that $\theta_{ij} = N_i\zeta_{ij}$, the elements of the sampling error covariance matrix $\boldsymbol{\Psi}_i^*$ can be expressed as follows:

$$\Psi_{i,jj}^* = K_i[\theta_{ij}/N_i + \theta_{ij}^2/N_i^2 (\exp(\hat{\bar{\boldsymbol{\sigma}}}_{jj}'\mathbf{Z}_i) - 1)] \tag{4}$$

$$\Psi_{i,12}^* = K_i[N_i^{-2}\theta_{i1}\theta_{i2} (\exp(\hat{\bar{\boldsymbol{\sigma}}}_{12}'\mathbf{Z}_i) - 1)] \tag{5}$$

where $\hat{\bar{\boldsymbol{\sigma}}}_{jj}'$ $j = 1, 2$ and $\hat{\bar{\boldsymbol{\sigma}}}_{12}'$ are posterior means of parameters $\bar{\boldsymbol{\sigma}}_{jj}$ and $\bar{\boldsymbol{\sigma}}_{12}$, respectively, computed using the model of section 4.1.

Since the sampling error covariance matrices are expressed as a function of the $\boldsymbol{\theta}_i$ parameters, here they can be considered Model Based Sampling Error Covariances (MBSEC). The posterior means $\hat{\boldsymbol{\theta}}_i^{\text{HB}} = E(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i)$ are taken as estimators of $\boldsymbol{\theta}_i'$s, while the posterior variance $V(\boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_i)$ is used as a measure of uncertainty.

We note that the MNN model cannot be implemented following the integrated approach described above. In fact, (3) does not ensure the positivity of $\boldsymbol{\theta}_i$ nor of the diagonal elements of $\boldsymbol{\Psi}_i$ as a result.

5. Data analysis

In section 5.1, we compare the MNPLN model with the benchmark MNN model and their univariate counterparts. We assume SMSEC for both models; we thus refer to the two strategies as MNPLN-SMSEC and MNN-SMSEC from here on. Since these models do not allow us to deal with the zero count problem, we refer this analysis to the NZC set. In section 5.2, we compare the SAE integrated strategy based on the MNPLN model and MBSEC (MNPLN-MBSEC), which we presented in Section 4.3, with the strategy based on the MNPLN-SMSEC. We limit the analysis to the NZC set in order to evaluate the two strategies under the same conditions. Finally, in section 5.3 we evaluate the overall performance of the proposed SAE model MNPLN-MBSEC for the whole data set (NZC+ZC).

Posterior distributions of parameters were obtained for all models, using Monte Carlo integration via the Gibbs sampling algorithm. We used the MCMC software WinBUGS (Spiegelhalter, Thomas, Best and Gilks 1995) to run three parallel chains (each with 25,000 runs), the starting point being drawn from an over-dispersed distribution. WinBUGS codes are available at the URL http://www2.stat.unibo.it/trivisano/. The convergence of the Gibbs sampler was monitored by visual inspection of the chains' plots and of autocorrelation diagrams, and by means of the potential scale reduction factor proposed by Gelman and Rubin (1992). Although all models displayed fast convergence, we discarded the first 5,000 iterations from each chain. In multivariate models, the fairly strong autocorrelation of chains is reduced by thinning the chain (1 out of every 3 values has been considered for posterior summaries). See Rao (2003, pages 228-232) for details.

The performances of the small area models discussed in sections 4.2 and 4.3 are compared using various measures. In order to choose among competing models, we computed the Deviance Information Criterion (DIC). The DIC is a model selection criterion according to which a model's performance is evaluated as the sum of a measure of fit (the posterior mean of the deviance $\bar{D}$) and a measure of complexity obtained as the difference between $\bar{D}$ and the deviance evaluated at the parameters' posterior mean. In this way, a model is preferred if it displays a lower DIC value (Spiegelhalter, Best, Carlin and Van der Linde 2002).

In order to verify the strength of the multivariate approach to SAE, we use as a benchmark the univariate versions of models discussed in sections 4.2 and 4.3, defined as follows. For all models, we set $\sigma_{v,12} = 0$ in $\boldsymbol{\Sigma}_v$, and we assume $\sigma_{v,11} \perp \sigma_{v,22}$, $\sigma^{1/2}_{v,jj} \sim U(0, U^+)$, $j = 1, 2$. For SMSEC models, we set $\boldsymbol{\Psi}_i = \text{diag}(\hat{\boldsymbol{\Psi}}_i)$, while for MBSEC models we set $\sigma_{1,12} = 0$ in (5). In addition, a new set of estimates for parameters $\bar{\boldsymbol{\sigma}}_{11}$ and $\bar{\boldsymbol{\sigma}}_{22}$ is obtained by setting $\bar{\rho}_k = 0$ in the model of section 4.1.

Table 2 reports the DIC results for the whole set of small area models.

Table 2
Model comparison using DIC statistic

Model	Data set	DIC
MNN-SMSEC	NZC	2,742.2
(univariate version)	NZC	2,745.4
MNPLN-SMSEC	NZC	2,656.9
(univariate version)	NZC	2,661.0
MNPLN-MBSEC	NZC	2,623.6
(univariate version)	NZC	2,638.1
MNPLN-MBSEC	NZC+ZC	3,202.7
(univariate version)	NZC+ZC	3,214.3

All the multivariate models considered perform better in terms of DIC than their univariate counterparts (Table 2). In addition, for all multivariate models we find that posterior credibility intervals of $\rho_v = \sigma_{v,12}/\sqrt{\sigma_{v,11}\sigma_{v,22}}$ do not contain zero. We thus focus on multivariate models in the following paragraphs.

We checked the adequacy of the specified multivariate models using posterior predictive checks. Simulated values of a suitable discrepancy measure are generated from the posterior predictive distribution and are then compared with the values of the same measure computed from observed data. Let $\hat{\boldsymbol{\theta}}_{\text{obs}}$ and $\hat{\boldsymbol{\theta}}_{\text{new}}$ denote the observed and generated data, respectively. The posterior predictive *p*-value is defined as $p = P\{d(\hat{\boldsymbol{\theta}}_{\text{new}}, \boldsymbol{\theta}) > d(\hat{\boldsymbol{\theta}}_{\text{obs}}, \boldsymbol{\theta}) \mid \hat{\boldsymbol{\theta}}_{\text{obs}}\}$. We consider a discrepancy measure proposed in Datta *et al.* (1999), which is defined as

$$d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \sum_{i=1}^{N} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)' \, \boldsymbol{\Psi}^{-1} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i). \qquad (6)$$

Computing the *p*-value is straightforward using the MCMC output. Extreme values of the probability *p* indicate a given model's lack of fit. Following Rao (2003, page 245-246) and You and Rao (2002), we computed two statistics that are useful in order to assess model fit at the individual domain level. The first statistic, $p^*_{ij} = P(\hat{\theta}_{ij,\text{new}} < \hat{\theta}_{ij,\text{obs}} | \hat{\boldsymbol{\theta}}_{\text{obs}})$, provides information about the degree of consistent overestimation or underestimation of $\hat{\theta}_{ij,\text{obs}}$.

The second statistics is defined as

$$d^*_{ij} = [E(\hat{\theta}_{ij} | \hat{\boldsymbol{\theta}}_{\text{obs}}) - \hat{\theta}_{ij,\text{obs}}] \Big/ \sqrt{V(\hat{\theta}_{ij} | \hat{\boldsymbol{\theta}}_{\text{obs}})},$$

where expectation and variance are under the posterior predictive distribution. Table 3 summarizes results relative to p, p_{ij}^* and d_{ij}^*.

To further check the consistency of the data, we calculated direct and model-based estimates of ${}_A\theta_{sj}, s = 1, ..., 10,$ that is, the total number of NR and SR for the ten domains identified by classifying firms only according to the industrial sector. Let $w_{is} = 1$ if the number of recruits in the domain i refers to the industrial sector s and $w_{is} = 0$; otherwise, then

$$ {}_A\theta_{sj} = \sum_i \theta_{ij} w_{is}. \quad (7)$$

At this level of aggregation, direct estimates can be considered accurate. Consequently, given two sets of model-based estimates referring to these large domains, we prefer the one that agrees with the direct estimates. Domains identified by industrial sectors are planned in the Excelsior Survey; each industrial sector is stratified according to firm size. Therefore, direct estimates ${}_A\hat{\theta}_{sj}$ for each industrial sector are calculated using the standard Horwitz-Thompson estimator. Aggregated model-based estimates are computed based on the MCMC output. For models referring to NZC data, we aggregated following (7) at each MCMC step $t, t = 1, ..., T,$ with samples ${}^t\theta_{ij}^*$ and ${}^t\theta_{ij}^{**}$ generated respectively from the posterior distribution of θ_{ij} for domains belonging to the NZC set and from the predictive distribution of θ_{ij} for domains belonging to the ZC set. The HB estimator is defined as ${}_A\hat{\theta}_{sj}^{\mathrm{HB}} = T^{-1}\sum_{t=1}^{T}(\sum_{i\in NZC} {}^t\theta_{ij}^* w_{is} + \sum_{i\in ZC} {}^t\theta_{ij}^{**} w_{is})$. Otherwise, for the model on NZC+ZC data, we aggregated following (7) MCMC samples from the posterior distributions of θ_{ij}. In this case, the HB estimator is defined as ${}_A\hat{\theta}_{sj}^{\mathrm{HB}} = T^{-1}\sum_{t=1}^{T}(\sum_{i\in NZC} {}^t\theta_{ij}^* w_{is})$. Table 4 reports summaries of ${}_A\hat{\theta}_{sj}$ and ${}_A\hat{\theta}_{sj}^{\mathrm{HB}}$.

For all the multivariate models, we examined the following variants of the prior distributions: independent non-informative flat prior distributions were used for the elements of vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*,$ and $\boldsymbol{\gamma}^*; \sigma_{v,jj}^{1/2} \sim U^+$, $j = 1, 2$, $\rho_v \sim U(-1,1)$, $\sigma_{v,12} = \rho_v (\sigma_{v,12}\, \sigma_{v,12})^{1/2}$. We do the same for the elements of matrix $\boldsymbol{\Sigma}^*$ in the MNN model. We did not find any relevant changes in the posterior distributions of parameters of interest.

5.1 Comparing the MNPLN-SMSEC and MNN-SMSEC models on the NZC set

We find that the MNPLN-SMSEC model largely outperforms the MNN-SMSEC one in terms of DIC (Table 2). This last model shows a lack of fit as it displays a p-value equal to 0.034 (Table 3), whereas a value of 0.65 suggests the adequacy of the MNPLN-SMSEC model. This finding is confirmed when p_{ij}^* and d_{ij}^* measures (Table 3) for the two models are compared. For the MNN-SMSEC model, p_{ij}^* ranges over domains from 0.000 to 0.995 for NR $(j = 1)$ and from 0.003 to 0.993 for SR $(j = 2)$, respectively, indicating overestimation and underestimation in some domains. In addition, summaries of the standardized residuals d_{ij}^* indicate that there are predicted values outside two standard deviations of the corresponding observed values. The same measures for the MNPLN-SMSEC model indicate an adequate fit.

We also find that the MNPLN-SMSEC model outperforms the MNN-SMSEC models when performances are evaluated with reference to estimates for large domains (Table 4). In fact, credibility intervals for the MNN-SMSEC only cover 2 aggregated direct estimates for NR and 4 for SR, while credibility intervals under the MNPLN-SMSEC cover 6 aggregated direct estimates for NR and 6 for SR.

Table 3
Posterior predictive checks; summaries of p_{ij}^* and d_{ij}^* calculated with respect to i

Model	Data set	p		p_{i1}^*	p_{i2}^*	d_{i1}^*	d_{i2}^*
MNN-SMSEC	NZC	0.034	min	0.000	0.003	-3.764	-2.867
			median	0.591	0.616	0.257	0.295
			max	0.995	0.993	2.656	-2.515
MNPLN-SMSEC	NZC	0.65	min	0.154	0.129	-0.965	-1.165
			median	0.535	0.561	0.124	0.149
			max	0.891	0.912	1.216	1.286
MNPLN-MBSEC	NZC	0.78	min	0.090	0.134	-1.085	-0.983
			median	0.515	0.519	-0.084	-0.085
			max	0.916	0.914	1.401	1.787
MNPLN-MBSEC	NZC+ZC	0.79	min	0.072	0.111	-1.164	-0.945
			median	0.506	0.523	-0.076	-0.094
			max	0.903	0.913	1.301	1.778

Table 4
Direct and HB estimates for industrial sectors; in italic HB estimates whose credibility intervals cover direct estimates

	Direct estimates		HB estimates											
			MNN-SMSEC (NZC)			MNPLN-SMSEC (NZC)			MNPLN-MBSEC (NZC)			MNPLN-MBSEC (NZC+ZC)		
s	${}_A\hat{\theta}_{s1}$	se(${}_A\hat{\theta}_{s1}$)	${}_A\hat{\theta}^{HB}_{s1}$	95% cred int.		${}_A\hat{\theta}^{HB}_{s1}$	95% cred int.		${}_A\hat{\theta}^{HB}_{s1}$	95% cred int.		${}_A\hat{\theta}^{HB}_{s1}$	95% cred int.	
1	1,702.0	41.3	1,077.0	964.3	1,201.0	1,266.0	1,055.0	1,509.0	*1,649.0*	*1,434.0*	*1,906.0*	*1,630.0*	*1,406.0*	*1,899.0*
2	1,758.8	41.9	1,936.0	1,793.0	2,091.0	*1,720.0*	*1,441.0*	*2,011.0*	*1,975.0*	*1,665.0*	*2,347.0*	*1,908.0*	*1,598.0*	*2,291.0*
3	725.0	26.9	557.8	460.6	662.7	534.6	435.8	642.3	*696.6*	*573.3*	*842.3*	*682.8*	*575.5*	*811.8*
4	373.9	19.3	202.7	123.0	294.8	192.1	129.1	277.0	*370.0*	*291.1*	*471.4*	*319.8*	*252.1*	*408.3*
5	142.4	11.9	*158.2*	*66.5*	*258.2*	*146.0*	*98.4*	*205.7*	235.6	164.3	326.9	*149.7*	*108.3*	*205.0*
6	5,624.1	75.0	4,134.0	3,800.0	4,484.0	*5,235.0*	*4,814.0*	*5,670.0*	*5,537.0*	*5,136.0*	*5,963.0*	*5,594.0*	*5,187.0*	*6,029.0*
7	887.7	29.8	659.9	549.1	783.7	629.6	526.4	743.4	*872.7*	*761.7*	*1,003.0*	*844.6*	*732.3*	*980.3*
8	223.9	15.0	*263.3*	*188.2*	*340.6*	*260.6*	*182.8*	*351.3*	362.0	262.8	494.1	*288.7*	*203.1*	*410.8*
9	661.5	25.7	893.7	790.3	999.4	*777.6*	*624.7*	*948.7*	931.0	754.8	1,150.0	*803.3*	*638.7*	*1,017.0*
10	1,792.6	42.3	1,460.0	1,334.0	1,598.0	*1,579.0*	*1,381.0*	*1,798.0*	*1,847.0*	*1,650.0*	*2,074.0*	*1,813.0*	*1,610.0*	*2,053.0*
	${}_A\hat{\theta}_{s2}$	se(${}_A\hat{\theta}_{s2}$)	${}_A\hat{\theta}^{HB}_{s2}$	95% cred int.		${}_A\hat{\theta}^{HB}_{s2}$	95% cred int.		${}_A\hat{\theta}^{HB}_{s2}$	95% cred int.		${}_A\hat{\theta}^{HB}_{s2}$	95% cred int.	
1	942.7	300.2	482.0	428.5	531.3	503.7	413.3	600.4	*832.6*	*706.4*	*987.6*	*817.8*	*686.0*	*980.0*
2	920.0	135.7	*883.9*	*798.7*	*967.4*	*849.8*	*694.8*	*1,022.0*	*949.8*	*778.9*	*1,161.0*	*922.3*	*747.6*	*1,167.0*
3	253.2	35.6	*249.2*	*209.2*	*292.1*	*254.1*	*202.1*	*309.9*	338.8	269.2	423.1	*284.7*	*226.2*	*354.5*
4	150.5	36.0	84.4	53.3	120.4	84.7	56.8	119.2	*160.6*	*116.7*	*218.0*	*131.5*	*97.0*	*179.6*
5	39.8	16.6	*66.7*	*31.2*	*104.2*	*62.0*	*37.3*	*89.3*	116.3	74.3	173.0	*60.9*	*38.4*	*90.5*
6	2,304.0	131.5	1,869.0	1,692.0	2,054.0	2,070.0	1,856.0	2,282.0	*2,273.0*	*2,060.0*	*2,508.0*	*2,297.0*	*2,079.0*	*2,542.0*
7	532.7	105.8	293.0	247.7	345.6	299.0	245.9	357.2	*471.5*	*402.8*	*553.2*	*443.3*	*377.2*	*538.3*
8	80.8	32.3	115.7	85.7	143.5	*100.5*	*67.7*	*140.3*	*139.5*	*76.7*	*210.4*	*98.0*	*58.5*	*156.9*
9	362.7	66.3	*407.0*	*358.6*	*453.0*	*361.0*	*285.8*	*438.8*	*432.1*	*335.4*	*552.9*	*360.4*	*274.7*	*476.2*
10	856.3	70.7	661.1	598.1	722.6	714.4	614.0	824.7	*855.4*	*740.5*	*984.6*	*832.7*	*719.8*	*964.5*

5.2 Comparing the MNPLN-SMSEC and MNPLN-MBSEC models on the NZC set

Values of p, p^*_{ij} and d^*_{ij} are approximately comparable for the MNPLN-SMSEC and MNPLN-MBSEC models (Table 3). Likewise, model-based estimates produced by MNPLN-SMSEC assume values very close to those obtained using MNPLN-MBSEC; in fact, the correlation between the posterior means of θ_{i1} under the two models is equal to 0.98, while the same measure referring to θ_{i2} is equal to 0.94. The same results arise for the correlation between posterior standard errors, which are 0.92 and 0.94, respectively. Performances of the MNPLN-MBSEC model in terms of agreement with direct estimates of large domains (Table 4) are slightly better than those of the MNPLN-SMSEC model: respectively, 7 direct estimates of *NR* and 8 of SR are covered by the credibility interval calculated under this model.

Given these results, we conclude that the fit of the MNPLN-MBSEC model is adequate.

5.3 Evaluating the performances of MNPLN-MBSEC models on the NZC+ZC set

We observe that the performances of the MNPLN-MBSEC model on the whole dataset in terms of p, p^*_{ij} and d^*_{ij} measures are satisfactory and comparable with those of the same model on the NZC data set (Table 3). Obviously, DIC values for the two models cannot be compared as the two models are estimated on different data sets.

As can be seen in Table 4, all the credibility intervals calculated using this model cover direct estimates referring to large domains; in other words, the agreement of HB estimates with direct estimates is very satisfactory. This result can be explained by noting that zero counts are more probable in small domains, which are characterized by a small number of employees (the covariate in all models). Therefore, estimating models on NZC data can lead to biased estimates of parameter $\boldsymbol{\beta}$. We conclude that integrating a sampling covariance model into the MNPLN small area model leads to an appreciable increase in the reliability

of small area estimates. To describe the efficiency gain of the HB estimates, we computed on the NZC set the average percent CV reduction (You 2008), defined as the average of the difference of the direct CV and HB CV (the ratio of the square root of the posterior variance and the posterior mean) relative to direct CV. The average CV reduction is 23.1% for NR and 29.1% for SR.

Acknowledgements

The authors would like to thank the Editor, Associate Editor and Referee for their helpful comments and suggestions. The research leading to this paper was partially supported by Miur-PRIN 2003/2003133249 and Miur-Prin 2008/2008CEFF37-001.

Appendix

The Multivariate Poisson-Log Normal distribution

Let $\mathbf{y} = (y_1, y_2, ..., y_j, ..., y_d)$ be a d-dimensional vector of counts, and suppose that $y_j|\tau_j \sim \text{Po}(\tau_j)$, with $y_j|\tau_j \perp y_{j'}|\tau_{j'}\ (j \neq j')$. Let the vector of parameters $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_j, ..., \tau_d)$ follow a multivariate Log Normal, that is, $\boldsymbol{\tau}|\boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \text{LN}_d(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$, where $\boldsymbol{\lambda} = E(\log \boldsymbol{\tau})$ and $\boldsymbol{\Sigma} = \text{COV}(\log \boldsymbol{\tau})$. Then the marginal distribution of $\mathbf{y}$ is a Multivariate Poisson-Log Normal (MPLN) distribution, which is a log normal mixture of d independent $\text{Po}(\tau_j)$, that is, $\mathbf{y}|\boldsymbol{\lambda}, \boldsymbol{\Sigma} \sim \text{PLN}_d(\boldsymbol{\lambda}, \boldsymbol{\Sigma})$. By denoting the (j, h), $j, h = 1, 2, ..., d$ element of $\boldsymbol{\Sigma}$ as σ_{jh}, marginal moments can be obtained easily through conditional expectation results and the standard properties of the Poisson and Log Normal distributions:

$$E(y_j|\boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \exp(\lambda_j + \sigma_{jj}/2) = \zeta_j$$

$$V(y_j|\boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \zeta_j + \zeta_j^2\left[\exp(\sigma_{jj}) - 1\right]$$

$$\text{COV}(y_j, y_h|\boldsymbol{\lambda}, \boldsymbol{\Sigma}) = \zeta_j\zeta_h[\exp(\sigma_{jh}) - 1],\ j \neq h.$$

Note that the MPLN model allows for overdispersion provided that $\sigma_{jj} > 0$, thus leading to $V(y_j|\boldsymbol{\lambda}, \boldsymbol{\Sigma}) > E(y_j|\boldsymbol{\lambda}, \boldsymbol{\Sigma})$. Moreover, the correlation structure of counts is unrestricted, since $\text{COV}(y_j, y_h|\boldsymbol{\lambda}, \boldsymbol{\Sigma})$ can be either positive or negative depending on the sign of σ_{jh}. Aitchison and Ho (1989), as well as Good and Pirog-Good (1989), studied a bivariate MPLN distribution, albeit exclusively in cases without covariates. However, the same model can easily be extended to take covariates into consideration (Chib and Winkelmann 2001).

References

Aitchison, J., and Ho, C.H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, 76, 643-653.

Arora, V., and Lahiri, P. (1997). On the superiority of the Bayesian method over the BLUP in small area estimation problems. *Statistica Sinica*, 7, 1053-1063.

Baldi, C., Bellisai, D., Fivizzani, S. and Sorrentino, M. (2007). Production of job vacancy statistics: Coverage. Contributi Istat, Istituto Nazionale di Statistica.

Becattini, G. (1992). The Marshallian industrial district as a socio-economic notion. In *Industrial Districts and International Co-operation in Italy*, (Eds., F. Pyke, G. Becattini and W. Sengenberger). Internation Labor Office, Geneva.

Chattopadhyay, M., Lahiri, P., Larsen, M. and Reimnitz, J. (1999). Composite estimation of drug prevalence for sub-state areas. *Survey Methodology*, 25, 81-86.

Chen, S. (2001). Empirical best prediction and hierarchical Bayes methods in small area estimation. Ph.D. Dissertation, Department of Mathematics and Statistics, University of Nebraska, Lincoln.

Chib, S., and Winkelmann, R. (2001). Markov chain Monte Carlo analysis of correlated count data. *Journal of Business & Economic Statistics*, 19, 428-435.

Cohen, M.L. (2000). Evaluation of Census Bureau's small-area poverty estimates. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 62-68.

Datta, G.S., Fay, R.E. and Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small area estimation. *Proceedings of Bureau of the Census 1991 Annual Research Conference*, U. S. Bureau of the Census, Washington, DC, 63-79.

Datta, G.S., Ghosh, M., Nangia, N. and Natarajan, K. (1996). Estimation of median income of four-person families: A Bayesian approach. In *Bayesian Analysis in Statistics and Econometrics*, (Eds., D.A. Berry, K.M. Chaloner and J.M. Geweke). New York: John Wiley & Sons, Inc., 129-140.

Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 488, 1074-1082.

Elazar, D. (2004). Small area estimation of disability in Australia. *Statistics in Transition*, 6, 5, 667-684.

Fabrizi, E., Ferrante, M.R. and Pacei, S. (2005). Estimation of poverty indicators at sub-national level using multivariate small area models. *Statistics in Transition*, 7, 3, 587-608.

Fabrizi, E., Ferrante, M.R. and Pacei, S. (2008). Measuring sub-national income poverty by using a small area multivariate approach. *Review of Income and Wealth*, 54, 4, 597-615.

Fay, R.E. (1987). Application of multivariate regression to small domain estimation. In *Small Area Statistics*, (Eds., R. Platek, J.N.K. Rao, C.-E. Särndal and M.P. Singh). New York: John Wiley & Sons, Inc., 91-102.

Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.

Ghosh, M., Nangia, N. and Kim, D. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Good, D.H., and Pirog-Good, M.A. (1989). Models for bivariate count data with an application to teenage delinquency and paternity. *Sociological Methods and Research*, 17, 4, 409-431.

Istat (1997). I sistemi locali del lavoro 1991. *Argomenti*, Roma 1997, 10.

Lahiri, P., and Rao, J.N.K. (1995). Robust estimation of mean square error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.

Liu, B., Lahiri, P. and Kalton, G. (2007). Hierarchical Bayes modeling of survey weighted small area proportions. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 3181-3186.

Otto, M.C., and Bell, W.R. (1995). Sampling error modelling of poverty and income statistics for states. *Proceedings of the Section on Government Statistics*, American Statistical Association, 160-165.

Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey: John Wiley & Sons, Inc.

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

Sforzi, F. (1991). I distretti industriali marshalliani nell'economia italiana. In *Distretti industriali e cooperazione fra imprese in Italia*, (Eds., F. Pyke, G. Becattini and W. Sengenberger). Quaderni di Studi e Informazioni, 34.

Sforzi, F., and Lorenzini, F. (2002). I distretti industriali. In *Ministero delle Attività Produttive-IPI*, *L'esperienza italiana dei distretti industriali*, Roma, IPI.

Spiegelhalter, D.J., Best, N., Carlin, B.P. and Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, Series B, 64, 583-639.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1995). *BUGS*: *Bayesian Inference Using Gibbs Sampling*. Version 0.50, Medical Research Council Biostatistics Unit, Cambridge.

Van Ophem, H. (1999). A general method to estimate correlated discrete random variables. *Econometric Theory*, 15, 228-237.

Winkelmann, R. (2003). *Econometric Analysis of Count Data*. Springer, Berlin.

Wolter, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

You, Y. (2008). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 1, 19-27.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 181-192
Statistics Canada, Catalogue No. 12-001-X

Linearization variance estimation for generalized raking estimators in the presence of nonresponse

Julia D'Arrigo and Chris Skinner [1]

Abstract

Alternative forms of linearization variance estimators for generalized raking estimators are defined via different choices of the weights applied (a) to residuals and (b) to the estimated regression coefficients used in calculating the residuals. Some theory is presented for three forms of generalized raking estimator, the classical raking ratio estimator, the 'maximum likelihood' raking estimator and the generalized regression estimator, and for associated linearization variance estimators. A simulation study is undertaken, based upon a labour force survey and an income and expenditure survey. Properties of the estimators are assessed with respect to both sampling and nonresponse. The study displays little difference between the properties of the alternative raking estimators for a given sampling scheme and nonresponse model. Amongst the variance estimators, the approach which weights residuals by the design weight can be severely biased in the presence of nonresponse. The approach which weights residuals by the calibrated weight tends to display much less bias. Varying the choice of the weights used to construct the regression coefficients has little impact.

Key Words: Calibration; Nonresponse; Raking; Variance estimation; Weight.

1. Introduction

Survey weighting is widely used to adjust for nonresponse bias. Generalized raking estimation (Deville, Särndal and Sautory 1993) provides a class of weighting methods which may be used when population totals of auxiliary variables are available. These methods can, in principle, remove (large-sample) nonresponse bias when the probability of nonresponse is related to the values of the auxiliary variables via a generalized linear model.

This paper presents some theory for linearization variance estimation for such methods in the presence of nonresponse. It also reports a simulation study of the properties of alternative raking estimators and associated variance estimators in settings designed to mimic two European surveys conducted by national statistical institutes. We consider three forms of raking estimator: the classical raking ratio estimator, the 'maximum likelihood' raking estimator (Brackstone and Rao 1979; Fuller 2002) and the generalized regression estimator (GREG). The first estimator has been used in practice in the British Labour Force Survey (LFS), the first survey upon which our simulation study is based. A version of the second estimator has been used in practice in the German Survey of Income and Expenditure (SIE), the second survey upon which our simulation study is based. The GREG estimator is widely used in many surveys, in particular in the context of nonresponse (Särndal and Lundström 2005).

A number of weighting methods, which do not fall into the class of generalized raking methods considered here, have also been proposed. See Särndal and Lundström (2005) for a historical account and Kott (2006) and Chang and Kott (2008) for some recent developments where the auxiliary variables for which population-level information is available may differ from those variables which are used as covariates in the generalized linear model for the probability of nonresponse.

The primary focus of this paper is on variance estimation and specifically on linearization methods, for which there exist a number of slightly different forms of variance estimator in the literature. In our simulation study we shall compare the properties of alternative raking estimators and associated variance estimators with respect to the effects of both sampling and nonresponse. A previous simulation study by Stukel, Hidiroglou and Särndal (1996) found little difference between two forms of linearization estimator with respect to sampling. However, there are reasons why nonresponse may lead to greater differences. Conditions for unbiasedness of raking estimation methods under nonresponse models vary between estimation methods (*e.g.*, Kalton and Maligalig 1991; Kalton and Flores-Cervantes 2003) and the choice of variance estimator may be more important in the presence of nonresponse (*e.g.*, Fuller 2002, Section 8).

The paper is structured as follows. The generalized raking estimators are defined in section 2 and, after introducing an asymptotic framework, the bias of these estimators is considered in section 3. Linearization variance estimators are defined in section 4. The simulation study is presented in section 5, the results are discussed in section 6 and some concluding remarks are given in section 7.

2. Generalized raking estimation

We consider the class of weighted estimators of a population total $T_y = \sum_U y_i$, which may be expressed as

1. Julia D'Arrigo and Chris Skinner, University of Southampton. E-mail: C.J.Skinner@soton.ac.uk.

$\hat{T}_y = \sum_s w_i y_i$, where y_i is the value of a survey variable for a unit i in a sample s from a population U and w_i is the *survey weight* which may depend on the sample but not on the choice of survey variable. We suppose here that the sample s consists of the set of respondents remaining after sampling and possible unit nonresponse. Generalized raking is a form of weighted estimation which may be employed when auxiliary population information is available in the form of a vector $T_x = \sum_U x_i$ of population totals of values x_i of a vector of auxiliary variables, where x_i is known for all units in s. Following Deville and Särndal (1992), the weights w_i are said to be *calibrated* if they satisfy the *calibration equations* $\sum_s w_i x_i = T_x$. The vector T_x is referred to as the vector of *calibration totals*. The class of generalized raking weights w_i is obtained by minimising the objective function:

$$\sum_s d_i G\ (w_i / d_i), \tag{2.1}$$

subject to the weights w_i being calibrated, where $G(.)$ is a specified objective function which meets certain criteria (see Deville *et al.* 1993) and d_i is an initial weight. We shall take this to be the design weight, *i.e.*, $d_i = \pi_i^{-1}$, where π_i is the probability that unit i is sampled. Deville and Särndal (1992) show that (subject to $G(.)$ obeying certain conditions), the solution of the above constrained optimisation problem may be expressed as:

$$w_i = d_i\ F\ (x_i' \hat{\lambda}), \tag{2.2}$$

where $F\ (u) = g^{-1}(u)$ denotes the inverse function of $g(u) = dG(u)/du$ and $\hat{\lambda}$ is the Lagrange multiplier which solves the calibration equations:

$$\sum_s d_i\ F\ (x_i' \hat{\lambda})\ x_i = T_x. \tag{2.3}$$

Deville and Särndal (1992) discuss various choices of the $G(.)$ function and associated $F(.)$ function. We consider the following three choices:

linear:

$$G_L(u) = (1/2)(u-1)^2,\ F_L(u) = 1 + u;$$

multiplicative (*raking ratio*):

$$G_M(u) = u \log(u) - u + 1,\ F_M(u) = \exp(u);$$

maximum likelihood raking:

$$G_{ML}(u) = u - 1 - \log(u),\ F_{ML}(u) = (1-u)^{-1}.$$

See also Deville *et al.* (1993) and Fuller (2009, section 2.9) regarding the above terminology for these functions. With the linear choice of $G(.)$, the optimisation problem has a closed form solution and the generalized raking estimator becomes $\hat{T}_y = \hat{T}_{yd} + (T_x - \hat{T}_{xd})' \hat{B}_s$, the *generalised regression estimator* (GREG), where $\hat{T}_{yd} = \sum_s d_i y_i$, $\hat{T}_{xd} = \sum_s d_i x_i$ and

$$\hat{B}_s = \left(\sum_s d_i x_i x_i' \right)^{-1} \sum_s d_i x_i y_i. \tag{2.4}$$

With the multiplicative choice of $G(.)$, the calibrated estimator of T_y is the classical raking ratio estimator (Brackstone and Rao 1979) when T_x contains the population counts in the categories of two or more categorical auxiliary variables. For example, in the context of the Britain Labour Force Survey, x_i denotes the vector of indicator variables of three categorical auxiliary variables: $x_i = (\delta_{1..i}, ..., \delta_{A..i}, \delta_{.1.i}, ..., \delta_{.B.i}, \delta_{..1i}, ..., \delta_{..Ci})'$, where $\delta_{a..i} = 1$ if unit i is in category a of the first auxiliary variable and 0 otherwise, $\delta_{.b.i} = 1$ if unit i is in category b of the second auxiliary variable and 0 otherwise and so on. The population total T_x of this vector thus contains the population counts in each of the (marginal) categories of each of the three auxiliary variables. The construction of the weights for classical raking ratio estimation has traditionally involved the use of iterative proportional fitting (Brackstone and Rao 1979). Ireland and Kullback (1968) demonstrate that this method converges to a solution of the above optimisation problem.

The function $G_{ML}(u)$ leads to an alternative 'maximum likelihood' version of raking adjustment, when x_i takes the same form, denoting indicator variables of categorical auxiliary variables. In this case, the objective function in (2.1) may be interpreted as a quantity which is proportional to minus a log likelihood in the case of simple random sampling with replacement (Brackstone and Rao 1979; Fuller 2002).

3. Asymptotic framework and nonresponse bias

We now consider the asymptotic properties of $\hat{T}_y$ with respect to both the sampling design and the nonresponse mechanism. We assume that the latter is such that each unit in the population responds, if sampled, with probability q_i, where this probability is not dependent on the choice of the sample and different units respond independently. We consider an asymptotic framework defined in terms of sequences of finite populations and associated probability sampling designs and response mechanisms (Fuller 2009, section 1.3), with orders of magnitude terms expressed in terms of $n = \sum_U \pi_i q_i$, the expected number of responding units, and N, the population size. We assume there exist positive constants K_1, K_2 and K_3 such that $K_1 < nN^{-1} d_i < K_2$ and $K_3 < q_i$ for all i.

We shall suppose that Horvitz-Thompson estimators of means are consistent for the corresponding finite population

means and that central limit theorems hold (as expressed formally in the conditions of Theorem 1.3.9 of Fuller 2009). In particular, we assume that the sequences and the function $F(.)$ are such that there is a unique solution λ of

$$\sum_U q_i F(x_i'\lambda)\, x_i = T_x, \qquad (3.1)$$

with

$$\hat{\lambda} = \lambda + O_p(n^{-0.5}), \qquad (3.2)$$

and that

$$\hat{T}_y = \sum_U q_i F(x_i'\lambda) y_i + O_p(Nn^{-0.5}). \qquad (3.3)$$

Deville and Särndal (1992) show that $\lambda = 0$ under certain assumptions (their Result 2). However, their assumptions apply just to the distribution induced by the sampling design and include the requirement that $N^{-1}(\hat{T}_{xd} - T_x) \to 0$ in probability. In the case of nonreponse, however, this requirement will often be implausible (*c.f.* Fuller 2002, page 15) and we do not require that λ be the zero vector.

A key assumption which we shall make is:

Condition C: there exists a vector α such that $F(x_i'\alpha) = q_i^{-1}$.

If condition C holds then α solves (3.1) and so $\lambda = \alpha$. It follows from (3.3) that $\hat{T}_y$ is consistent for T_y for any choice of variable y if this condition holds. Thus, we may view condition C as a sufficient condition for the absence of (asymptotic) nonresponse bias. This property of Condition C has been discussed by Fuller, Loughlin and Baker (1994), Fuller (2009, page 284) and Särndal and Lundström (2005, Proposition 9.2) for the case when F is linear. Fuller (2002, page 15), Kott (2006) and Chang and Kott (2008) also consider estimating response probabilities using general models of the form $q_i^{-1} = F(x_i'\alpha)$.

To illustrate what might happen if condition C does not hold, suppose that x_i is just a scalar with $x_i \equiv 1$. Then the unique solution of (3.1) is $\lambda = g(N/\sum_U q_i)$ and $p\lim(\hat{T}_y) = N(\sum_U q_i y_i)/(\sum_U q_i)$. Hence, the asymptotic nonresponse bias will only disappear for those survey variables which are 'uncorrelated' with the response probabilities q_i.

4. Linearization variance estimation

We now proceed to consider the asymptotic variance of $\hat{T}_y$ and its estimation. As in the previous section, the variance is defined with respect to the joint distribution induced by both sampling and nonresponse.

Note first that in general (and in particular for $G_M(.)$ and $G_{ML}(.)$), iteration is needed to solve the calibration equations. There does exist a literature (see Deville *et al.* 1993) which seeks to estimate the variance of $\hat{T}_y$ after a finite number of iterations. We follow instead the approach of Deville *et al.* (1993) and, for example, Binder and Théberge (1988) by approximating the variance of $\hat{T}_y$ by the variance of the 'converged' estimator, *i.e.*, the hypothetical estimator arising from an infinite number of iterations, represented by $\mathrm{var}(\sum_s w_i y_i)$, where the w_i are the 'converged' weights which solve the constrained optimisation problem in section 2.

A linearization variance estimator is obtained by approximating $\mathrm{var}(\sum_s w_i y_i)$ by $\mathrm{var}(\sum_s d_i z_i)$ for a 'linearized variable' z_i (Deville 1999). We now seek to construct this variable using a large sample argument. We first obtain an expression for $\hat{\lambda}$. A Taylor expansion of the left side of the calibration equations in (2.3) gives

$$\sum_s d_i F(x_i'\hat{\lambda}) x_i = \sum_s d_i F_i x_i + \sum_s d_i f(x_i'\lambda^*) x_i x_i'(\hat{\lambda} - \lambda),$$

where $F_i = F(x_i'\lambda)$, λ^* is between $\hat{\lambda}$ and λ and $f(u) = dF(u)/du$ is assumed to exist. Assuming also continuity of $f(.)$, the existence of $\lim_{N\to\infty} N^{-1}\sum_U q_i f_i x_i x_i'$ and using (3.2), we have

$$N^{-1}\sum_s d_i F(x_i'\hat{\lambda}) x_i = N^{-1}\sum_s d_i F_i x_i + N^{-1}\sum_s d_i f_i x_i x_i'(\hat{\lambda} - \lambda) + o_p(n^{-0.5}), \qquad (4.1)$$

where $f_i = f(x_i'\lambda)$. Then, assuming $\lim_{N\to\infty} N^{-1}\sum_U q_i f_i x_i x_i'$ is non-singular and using (2.3), we obtain

$$\hat{\lambda} - \lambda = \left[\sum_s d_i f_i x_i x_i\right]^{-1}\left[T_x - \sum_s d_i F_i x_i\right] + o_p(n^{-0.5}). \qquad (4.2)$$

See Fuller (2009, proof of Theorem 1.3.9) for formal details of how (4.1) and (4.2) may be derived and the underlying regularity conditions. Note that to ensure $\lim_{N\to\infty} N^{-1}\sum_U q_i f_i x_i x_i'$ is non-singular may require dropping redundant variables from x_i and possibly (as in Deville and Särndal 1992) modifying the estimator for samples with small probability that result in singularity of this matrix.

A similar argument involving the Taylor expansion of w_i in (2.2) about λ gives:

$$w_i = d_i[F_i + f_i x_i'(\hat{\lambda} - \lambda)] + o_p(Nn^{-1.5}). \qquad (4.3)$$

Then, assuming the existence of necessary population moments so that the remainder term in (4.3) holds uniformly across i (Fuller 2009, Corollary 2.7.1.1.), we have

$$\hat{T}_y \equiv \sum_s w_i y_i = \sum_s d_i\left[F_i + f_i x_i'(\hat{\lambda} - \lambda)\right] y_i + o_p(Nn^{-0.5}) \qquad (4.4)$$

and hence from (4.2) and (4.4):

$$\hat{T}_y = \sum_s d_i F_i y_i + B\left[T_x - \sum_s d_i F_i x_i\right] + o_p(Nn^{-0.5}), \quad (4.5)$$

where

$$B = \left[\sum_s d_i f_i y_i x_i'\right]\left[\sum_s d_i f_i x_i x_i'\right]^{-1}. \quad (4.6)$$

Note that $F_i = f_i = 1$ under the assumptions of Deville and Särndal (1992) (since in this case $\lambda = 0$ and it follows from the assumptions about $G(.)$ that $F(0) = f(0) = 1$). Hence, under these assumptions, expression (4.5) corresponds to Result 5 of Deville and Särndal (1992), *i.e.*, the generalized raking estimator is asymptotically equivalent to the GREG estimator. Therefore, the asymptotic variance of $\hat{T}_y$ is the same as that of $\sum_s d_i z_i$, where z_i is the linearized variable:

$$z_i = F_i(y_i - \beta x_i), \quad (4.7)$$

and it is assumed that B converges to a finite limit matrix β. An alternative derivation of this expression is given by Demnati and Rao (2004, section 3.4).

For the purpose of linearization variance estimation, $\hat{T}_y$ is treated as the linear estimator $\sum_s d_i \hat{z}_i$, where

$$\hat{z}_i = \hat{F}_i(y_i - \hat{B}\, x_i) \quad (4.8)$$

is treated as a fixed variable.

A number of choices of $\hat{F}_i$ and $\hat{B}$ have been discussed in the literature. Starting with $\hat{F}_i$, the natural choice implied by the above argument is $\hat{F}_i = F(x_i'\hat{\lambda})$. A simpler choice, however, would be to take $\hat{F}_i = 1$. Deville and Särndal (1992) note that, in their classical theory with $\lambda = 0$, these choices are asymptotically equivalent but they express a preference for the choice $\hat{F}_i = F(x_i'\hat{\lambda})$. In our setting with nonresponse and with $\lambda = 0$ not necessarily holding, the second choice seems preferable and this is emphasized by Fuller (2002, page 15). Note that these two choices imply that $\sum_s d_i \hat{z}_i$ either takes the form $\sum w_i(y_i - \hat{B}\, x_i)$ when $\hat{F}_i = F(x_i'\hat{\lambda})$ or $\sum d_i(y_i - \hat{B}\, x_i)$ when $\hat{F}_i = 1$. We shall therefore refer to these choices as either w_i*-weighted residuals* or d_i *- weighted residuals*.

Regarding $\hat{B}$, it follows from our argument on the choices of $\hat{F}_i$ that f_i in (4.2) should be replaced by $\hat{f}_i = f(x_i'\hat{\lambda})$, giving:

(i) $\hat{B} = [\sum_s d_i \hat{f}_i y_i x_i'][\sum_s d_i \hat{f}_i x_i x_i']^{-1}$, as also proposed by Demnati and Rao (2004).

Other choices are

(ii) $\hat{B} = \hat{B}_s$, as in (2.4), as proposed by Deville *et al.* (1993).

(iii) $\hat{B} = [\sum_s w_i y_i x_i'][\sum_s w_i x_i x_i']^{-1}$, as proposed by Deville and Särndal (1992, equation 3.4), which might be more practical to compute than $\hat{B}_s$ for users of survey data files which include the w_i weights but not the d_i weights.

The extent to which these choices differ depends on the choice of $G(.)$ function. For the linear case $f(u) = 1$ so that the estimators in (i) and (ii) are identical. In the case of classical raking adjustment, $f(u) = F(u) = \exp(u)$ so that $\hat{f}_i = \hat{F}_i$ and $d_i \hat{f}_i = w_i$ and the estimators (i) and (iii) are identical. For the 'maximum likelihood' raking estimator we have $F(u) = (1-u)^{-1}$ and $f(u) = (1-u)^{-2}$ so that $d_i \hat{f}_i = w_i^2/d_i$ and the three variance estimators are all distinct.

Having determined the form of $\hat{z}_i$ in (4.8), the linearization variance estimator for $\hat{T}_y$ is obtained by estimating the variance of the linear estimator $\sum_s d_i \hat{z}_i$, treating d_i and $\hat{z}_i$ as fixed. In the case of a stratified multistage sampling design, assuming "with replacement" sampling of primary sampling units (PSUs) within strata, a standard estimator of the variance (*e.g.*, Stukel *et al.* 1996) is:

$$\hat{V}(\hat{T}_y) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{j=1}^{n_h} (z_{hj} - \bar{z}_h)^2 \quad (4.9)$$

where $z_{hj} = \sum_k d_{hjk} \hat{z}_{hjk}$, $\bar{z}_h = \sum_j z_{hj}/n_h$ and $\hat{z}_{hjk}$ is the value of the variable defined in (4.8) for the k^{th} individual within the j^{th} selected PSU in stratum h. This estimator remains appropriate in the presence of nonresponse if individual response in each PSU is independent of response in all other PSUs and if at least one individual is observed in each selected PSU (Fuller *et al.* 1994, page 78).

5. Simulation studies

In order to compare the performance of the weighted estimators and their corresponding variance estimators, two simulation studies were undertaken by constructing artificial populations using data from the British Labour Force Survey (LFS) and the German Sample Survey of Income and Expenditure (SIE). In each case, $R = 1{,}000$ samples were generated from these populations by first sampling, in a way designed to mimic the real sampling scheme after some simplification, and then removing nonresponding cases according to two nonresponse models. The first assumes multiplicative nonresponse which, from Condition C in section 3, might be expected to lead to least bias for the raking ratio method. The second model assumed additive nonresponse, which might be expected to lead to least bias for the GREG estimator.

For each of the R samples, point estimates of parameters were calculated using the different generalized raking methods presented in section 2 and variance estimates were calculated using the different linearization methods presented in section 4. The properties of the estimators were then summarised.

5.1 Study based on the British Labour Force Survey

The first study was based upon data from the March-May 1998 quarter of the British LFS, a survey of persons living in private households in Britain, designed to provide information on the British labour market and carried out by the Office for National Statistics (ONS). The sample of approximately 58,000 households was treated as an artificial population. Repeated samples were drawn from this population in a way intended to mimic the design used for the LFS (ONS 1998, Section 3). Each sample consisted of 1,211 households selected by stratified simple random sampling with proportional allocation across 19 strata, defined by region of residence. These regions were designed to mimic interviewer areas which defined strata in the LFS. In the LFS all individuals in a sampled household are interviewed if possible. In this simulation study, all the respondents in a sample household were retained, except those aged under 16, who are not relevant for the estimates of interest.

The following two nonresponse models, based upon results of a study of Foster (1998), were used to determine whether sampled individuals responded.

Multiplicative Nonresponse Model:

$$\begin{aligned} q_i^{-1} = 1.15 \ &\times \ 1.17 \text{ (if London)} \\ &\times \ 1.13 \text{ (if aged under 35)} \\ &\times \ 1.1 \text{ (if female)} \end{aligned}$$

Additive Nonresponse Model:

$$\begin{aligned} q_i^{-1} = 1.15 \ &+ \ 0.20 \text{ (if London)} \\ &+ \ 0.15 \text{ (if aged under 35)} \\ &+ \ 0.10 \text{ (if female)} \end{aligned}$$

where q_i is the response probability defined at the beginning of section 3 and the form of the model is chosen to satisfy Condition C.

Three parameters of interest are defined for the artificial population: the total number of persons unemployed, employed or inactive in the workforce. Weights were constructed for responding individuals, with calibration totals consisting of population counts in the categories of three categorical auxiliary variables and with Horvitz-Thompson initial weights d_i, as in section 2. The choice of auxiliary variables was designed to mimic those used in the LFS. However, because of the reduced scale of our artificial population and the consequent smaller numbers of individuals within strata, we simplified the LFS calibration variables to the following three categorical factors, defining 83 control totals:

- area of residence with 23 categories;
- a cross-classification of sex by 10 age groups (consisting of single years for those between 16 and 24 and a separate age group for 25 or older) with 20 categories;
- a cross-classification of region (Northern England; London and South East; Midlands and East Anglia; Scotland) by sex by age in 15-year age groups (16-29, 30-44, 45-59, 60-75 and 75 or older) with 40 categories.

5.2 Study based on the German sample Survey of Income and Expenditure

Our second study is based on the 1998 German Survey of Income and Expenditure (SIE), a national household survey conducted every 5 years by the Federal Statistical Office, to provide information about the economic and social situation of households, especially regarding the distribution of income and expenditure (Muennich and Schulrle 2003). We used data from a synthetic population of 64,326 households, created to represent 20% of all households from the Bremen region, excluding those with a monthly household net income of DM 35,000 or above (DM denotes the currency of German marks). A quota sampling design was employed for this survey and we have not attempted to mimic this design. Instead, our simulation study employs simple random sampling together with nonresponse. Repeated simple random samples of 1,340 households were drawn from the artificial population, representing a sampling fraction of about 1/48. Nonresponse models were constructed using the results of studies of similar surveys in Great Britain: the Family Expenditure Survey and the National Food Survey (Foster 1998). For each selected sample, the subset of responding households was determined by the following nonresponse models:

Multiplicative Model:

$$\begin{aligned} q_i^{-1} = 1.44 \ &\times \ 1.09 \text{ (if self-employed)} \\ &\times \ 1.03 \text{ (if unemployed)} \\ &\times \ 0.97 \text{ (if employed)} \\ &\times \ 1.16 \text{ (if no children in the household).} \end{aligned}$$

Additive Model:

$$\begin{aligned} q_i^{-1} = 1.44 \ &+ \ 0.13 \text{ (if self-employed)} \\ &+ \ 0.04 \text{ (if unemployed)} \\ &- \ 0.04 \text{ (if employed)} \\ &+ \ 0.23 \text{ (if no children in the household).} \end{aligned}$$

The parameters of interest are the total household net income per quarter and the total household expenditure per quarter, computed from the finite artificial population.

As for the LFS study, each sampled household was assigned a weight. In the actual SIE the weights are constructed using essentially the maximum likelihood raking method by adjusting the sample data simultaneously to the marginal

distributions of several characteristics, such as household type, social economic status of the reference person, household net income class and region (land). We try to mimic this adjustment, as far as possible, in our study. However, as for the LFS, because of the problem of strata with small numbers of households we simplify the SIE calibration variables to the following three categorical factors:

- household type with 7 categories
  - mother/father alone + 1child,
  - mother/father alone + 2 or more children,
  - couple with 1 child – spouse employed,
  - couple with 1 child – spouse unemployed,
  - couple with 2 or more children – spouse employed,
  - couple with 2 or more children – spouse unemployed,
  - other.
- social status of the reference person with 5 categories
  - self-employed,
  - civil servant or military,
  - employee,
  - worker,
  - unemployed, pensioner, student or other.
- household net income per quarter with 3 categories
  - 0-5,000 DM,
  - 5-7,000 DM,
  - 7-35,000 DM.

6. Results

6.1 Properties of point estimators

Table 6.1 presents the properties of the point estimators of total unemployed in the LFS study for different calibration methods and alternative assumptions about nonresponse. The properties are assessed following usual practice in simulation studies. For example, the bias in Table 6.1 is obtained from $\hat{B}(\hat{T}_y) = \hat{E}(\hat{T}_y) - T_y$, where $\hat{E}(\hat{T}_y) = 1/R \sum_{r=1}^{R} \hat{T}_{y_r}$, $\hat{T}_{y_r}$ is the value of $\hat{T}_y$ for sample r and R is the number of simulated samples. We observe from this table that the standard error remains virtually constant across alternative raking methods for a given nonresponse model. Nonresponse leads to an increase in the standard error across all estimators as expected (since the sample size is reduced). The table does show evidence of nonresponse bias, which is of a similar order for each of the raking methods. We do not find that this bias is least when the estimator matches the nonresponse model (*i.e.*, the GREG estimator for additive response and the raking estimator for multiplicative response) as we might have expected. Perhaps this is because the covariates used in the nonresponse models (*e.g.*, the aged 35+ variable) are not all included in the calibrating variables. Nevertheless, the nonresponse bias is small in the sense that the root mean square error is very similar to the standard error in each case. Under nonresponse, the GREG calibration method generates some negative weights whereas this is avoided by the two raking methods, as expected. A greater number of very large weights are observed, however, for the 'maximum likelihood' raking estimator.

Corresponding results for the SIE data are presented in Table 6.2. The pattern of results is broadly similar, although there is now no evidence of significant nonresponse bias (*i.e.*, the observed bias could be explained by simulation variation). The standard errors and root mean square errors also remain virtually constant across weighting methods for a given nonresponse model.

Table 6.1
Simulation properties of point estimators of total unemployed using data from LFS with R = 1,000

Nonresponse Model/Point Estimator	Bias (simulation standard error)	Standard Error	Root Mean Square Error	Number of Negative Weights[1]	Number of Very Large Weights[1, 2]
Complete Response:					
GREG	7.6 (14.3)	452.8	452.8	0	0
Classical Raking	8.3 (14.3)	452.8	452.9	0	0
'ML' Raking	9.0 (14.3)	453.3	453.4	0	1
Multiplicative nonresponse:					
GREG	-45.6 (15.8)	498.3	500.3	4	1
Classical Raking	-42.1 (15.8)	498.8	500.6	0	2
'ML' Raking	-39.7 (15.8)	499.4	501.0	0	7
Additive nonresponse:					
GREG	-37.3 (15.7)	497.4	498.8	5	1
Classical Raking	-34.7 (15.7)	497.5	498.7	0	3
'ML' Raking	-32.4 (15.8)	498.1	499.1	0	7

[1] the number of such weights across all sample units and all 1000 samples.
[2] the number of weights more than 10 times the corresponding design weight.

Table 6.2
Simulation properties of point estimators of total income using data from SIE with R = 1,000

Nonresponse Model/Point Estimator	Bias (simulation standard error)	Standard Error	Root Mean Square Error	Number of Negatives Weights	Number of Very Large Weights
Complete Response:					
GREG	-172.2 (331.3)	10,477.3	10,478.7	0	0
Classical Raking	-170.6 (331.5)	10,484.1	10,485.8	0	0
'ML' Raking	-169.8 (331.8)	10,491.5	10,492.9	0	0
Multiplicative nonresponse:					
GREG	-495.7 (429.7)	13,586.8	13,595.8	0	0
Classical Raking	-493.8 (429.6)	13,584.6	13,593.5	0	0
'ML' Raking	-463.5 (429.5)	13,582.8	13,590.7	0	0
Additive nonresponse:					
GREG	-473.2 (430.5)	13,614.8	13,623.0	0	0
Classical Raking	-469.4 (430.5)	13,612.9	13,621.0	0	0
'ML' Raking	-439.5 (430.5)	13,613.5	13,620.6	0	0

6.2 Properties of variance estimators

The properties of the different estimators of the variances of the point estimators of the total unemployed from the LFS are shown in the Table 6.3 (the 'standard error estimate' in the table refers to the square root of the variance estimate). We make a number of observations:

- weighting the residuals by w_i rather than by d_i reduces the bias and root mean squared error of the standard error estimator. The bias arising from the use of d_i weighted residuals in the case of nonresponse is particularly important (as noted by Fuller 2002) but there are also non-negligible reductions of bias even in the complete response case.
- The choice of weight used in $\hat{B}$ for the calculation of residuals seems to have little impact.
- For a given nonresponse setting and choice of weighting the residuals, there is little difference in the results for the different choices of point estimator.

The results in Table 6.3 are extended in Table 6.4 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider two additional parameters: total numbers employed and inactive. We see again that the relative bias arising from using d_i weighted residuals can be substantial in the presence of nonresponse, over 20% in several cases, and that this is reduced using the w_i weighted residuals. Again, little change is observed in the percent relative bias of the standard error estimators when different choices of weights are used in the calculation of $\hat{B}$ for the residuals.

Corresponding results for the SIE data when estimating total income are shown in Table 6.5. Again, the pattern of results is broadly similar to that for the LFS data in Table 6.3. For the complete response case, the use of w_i weighted residuals rather than d_i weighted residuals leads to modest improvement in bias and RMSE of the standard error estimators. For the nonresponse cases the improvements are considerable. Little change in the standard error estimators is observed when modifying the choice of weight used to compute the estimated regression coefficients. The results in Table 6.5 are extended in Table 6.6 to consider relative bias of the standard error estimators, rather than their absolute bias, and to consider one additional parameter: total expenditure per quarter. We see again that the relative bias arising from using d_i weighted residuals can be substantial in the presence of nonresponse, over 35% in all cases, and that this is reduced using the w_i weighted residuals, for which the relative bias never exceeds about 3%.

Table 6.3
Properties of variance estimators when estimating total unemployed from the LFS (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Mean of Standard Error Estimator	Bias of SE Estimator (simulation s.e.)	RMSE of SE Estimator	Coverage[2] of Confidence Interval (%)
Complete Response:						
GREG	*d*	*d*	433.9	-18.8 (0.9)	33.4	93.5
	d	*w*	434.3	-18.5 (0.9)	33.3	93.5
	w	*d*	442.8	-10.0 (1.0)	31.9	93.8
	w	*w*	441.9	-10.8 (1.0)	32.0	93.7
Classical Raking	*d*	*d*	433.9	-18.8 (0.9)	33.4	93.5
	d	*w*	434.2	-18.5 (0.9)	33.3	93.5
	w	*d*	443.0	-9.8 (1.0)	32.0	93.8
	w	*w*	442.0	-10.7 (1.0)	32.0	93.8
'ML' Raking	*d*	*d*	433.9	-19.4 (0.9)	33.7	93.5
	d	*w*	434.3	-19.1 (0.9)	33.6	93.5
	d	*df*	435.4	-17.9 (0.9)	33.0	93.5
	w	*d*	443.7	-9.6 (1.0)	32.5	93.7
	w	*w*	442.3	-11.1 (1.0)	32.4	93.7
	w	*df*	441.6	-11.8 (1.0)	32.3	93.7
Multiplicative nonresponse:						
GREG	*d*	*d*	385.7	-112.6 (0.9)	116.0	85.8
	d	*w*	386.1	-112.1 (0.9)	115.5	85.8
	w	*d*	489.5	-8.8 (1.2)	39.2	94.2
	w	*w*	487.8	-10.4 (1.2)	39.2	94.2
Classical Raking	*d*	*d*	385.7	-113.1 (0.9)	116.5	85.7
	d	*w*	386.1	-112.7 (0.9)	116.1	85.7
	w	*d*	490.3	-8.5 (1.2)	39.6	94.3
	w	*w*	488.4	-10.4 (1.2)	39.5	94.1
'ML' Raking	*d*	*d*	385.7	-113.7 (0.9)	117.1	85.4
	d	*w*	386.2	-113.2 (0.9)	116.6	85.6
	d	*df*	387.8	-111.6 (0.9)	115.0	85.8
	w	*d*	491.9	-7.5 (1.3)	40.4	94.2
	w	*w*	488.9	-10.5 (1.2)	39.9	94.0
	w	*df*	487.5	-11.9 (1.2)	39.8	94.0
Additive nonresponse:						
GREG	*d*	*d*	386.5	-110.9 (0.9)	114.4	86.0
	d	*w*	387.0	-110.5 (0.9)	113.9	86.0
	w	*d*	489.3	-8.2 (1.2)	39.0	94.6
	w	*w*	487.6	-9.8 (1.2)	39.0	94.6
Classical Raking	*d*	*d*	386.5	-111.0 (0.9)	114.4	85.8
	d	*w*	387.0	-110.6 (0.9)	114.0	85.8
	w	*d*	490.1	-7.4 (1.2)	39.2	94.7
	w	*w*	488.1	-9.4 (1.2)	39.1	94.6
'ML' Raking	*d*	*d*	386.5	-111.6 (0.9)	115.0	85.6
	d	*w*	387.0	-111.1 (0.9)	114.6	85.6
	d	*df*	388.6	-109.5 (0.9)	113.0	85.9
	w	*d*	491.6	-6.5 (1.3)	40.0	94.7
	w	*w*	488.6	-9.5 (1.2)	39.5	94.6
	w	*df*	487.3	-10.8 (1.2)	39.4	94.6

[1] see text following equation (4.8), where choices *df*, *d* and *w* correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.
[2] percentage of 95% normal-theory confidence intervals containing true value.

Table 6.4
Relative bias (%) of standard error estimators of unemployed, employed and inactive totals from LFS (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Relative Bias of Standard Error Estimator		
			Unemployed	Employed	Inactive
Complete Response:					
GREG	*d*	*d*	-4.2	-3.4	0.5
	d	*w*	-4.1	-3.3	0.6
	w	*d*	-2.2	-2.2	1.9
	w	*w*	-2.4	-2.3	1.7
Classical Raking	*d*	*d*	-4.2	-3.3	0.7
	d	*w*	-4.1	-3.2	0.8
	w	*d*	-2.2	-2.1	2.1
	w	*w*	-2.4	-2.2	1.9
'ML' Raking	*d*	*d*	-4.3	-3.3	0.7
	d	*w*	-4.2	-3.3	0.8
	d	*df*	-4.0	-3.1	1.1
	w	*d*	-2.1	-2.0	2.3
	w	*w*	-2.4	-2.2	1.9
	w	*df*	-2.6	-2.3	1.8
Multiplicative nonresponse:					
GREG	*d*	*d*	-22.6	-22.3	-18.2
	d	*w*	-22.5	-22.2	-18.1
	w	*d*	-1.8	-3.3	1.8
	w	*w*	-2.1	-3.5	1.5
Classical Raking	*d*	*d*	-22.7	-30.6	-18.4
	d	*w*	-22.6	-30.5	-18.3
	w	*d*	-1.7	-13.5	1.7
	w	*w*	-2.1	-13.7	1.3
'ML' Raking	*d*	*d*	-22.8	-22.0	-18.4
	d	*w*	-22.7	-21.9	-18.3
	d	*df*	-22.3	-21.7	-17.9
	w	*d*	-1.5	-2.7	1.9
	w	*w*	-2.1	-3.1	1.3
	w	*df*	-2.4	-3.3	1.1
Additive nonresponse:					
GREG	*d*	*d*	-22.3	-21.8	-18.5
	d	*w*	-22.2	-21.7	-18.4
	w	*d*	-1.6	-2.9	1.1
	w	*w*	-2.0	-3.1	0.8
Classical Raking	*d*	*d*	-22.3	-30.2	-18.0
	d	*w*	-22.2	-30.1	-17.9
	w	*d*	-1.5	-13.3	1.8
	w	*w*	-1.9	-13.5	1.4
'ML' Raking	*d*	*d*	-22.4	-21.6	-18.0
	d	*w*	-22.3	-21.5	-17.9
	d	*df*	-22.0	-21.3	-17.6
	w	*d*	-1.3	-2.4	2.0
	w	*w*	-1.9	-2.8	1.5
	w	*df*	-2.2	-3.0	1.3

[1] see text following equation (4.8), where *df*, *d* and *w* correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.

Table 6.5
Properties of variance estimators when estimating total income from the SIE (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Mean of Standard Error Estimator	Bias of SE Estimator (s.e.)	RMSE of SE Estimator	Coverage[2] of Confidence Interval (%)
Complete Response:						
GREG	*d*	*d*	10,338.8	-138.5 (6.9)	259.0	93.8
	d	*w*	10,339.2	-138.2 (6.9)	258.8	93.8
	w	*d*	10,377.9	-99.5 (6.9)	240.0	94.1
	w	*w*	10,376.8	-100.5 (6.9)	240.3	94.1
Classical Raking	*d*	*d*	10,338.8	-145.3 (6.9)	262.7	93.8
	d	*w*	10,339.2	-144.9 (6.9)	262.5	93.8
	w	*d*	10,370.0	-106.1 (6.9)	243.1	94.0
	w	*w*	10,376.9	-107.2 (6.9)	243.5	94.0
'ML' Raking	*d*	*d*	10,338.8	-152.7 (6.9)	266.9	93.9
	d	*w*	10,339.2	-152.4 (6.9)	266.7	93.9
	d	*df*	10,340.3	-151.3 (6.9)	266.1	94.0
	w	*d*	10,378.3	-113.2 (6.9)	246.5	94.0
	w	*w*	10,377.1	-114.4 (6.9)	247.0	94.0
	w	*df*	10,376.7	-114.8 (6.9)	247.2	94.0
Multiplicative nonresponse:						
GREG	*d*	*d*	8,104.7	-5,482.1 (7.4)	5,487.1	75.8
	d	*w*	8,105.5	-5,481.3 (7.4)	5,486.3	75.8
	w	*d*	13,214.5	-372.3 (12.8)	549.7	94.5
	w	*w*	13,210.9	-375.9 (12.8)	551.7	94.5
Classical Raking	*d*	*d*	8,104.7	-5,479.8 (7.4)	5,484.9	75.8
	d	*w*	8,105.5	-5,479.1 (7.4)	5,484.1	75.8
	w	*d*	13,214.1	-370.4 (12.8)	549.4	94.5
	w	*w*	13,210.4	-374.2 (12.8)	551.5	94.5
'ML' Raking	*d*	*d*	8,104.7	-5,478.1 (7.4)	5,483.1	75.8
	d	*w*	8,105.5	-5,477.3 (7.4)	5,482.3	75.8
	d	*df*	8,108.1	-5,474.7 (7.4)	5,479.7	75.9
	w	*d*	13,215.2	-367.6 (12.9)	549.4	94.5
	w	*w*	13,210.6	-372.2 (12.9)	551.6	94.5
	w	*df*	13,208.9	-373.9 (12.9)	552.3	94.5
Additive nonresponse:						
GREG	*d*	*d*	8,106.3	-5,508.5 (7.4)	5,513.5	75.6
	d	*w*	8,107.1	-5,507.7 (7.4)	5,512.7	75.6
	w	*d*	13,207.9	-407.0 (12.8)	573.8	94.3
	w	*w*	13,204.3	-410.5 (12.8)	575.9	94.3
Classical Raking	*d*	*d*	8,106.3	-5,506.6 (7.4)	5,511.6	75.7
	d	*w*	8,107.1	-5,505.9 (7.4)	5,510.9	75.7
	w	*d*	13,207.7	-405.3 (12.8)	573.6	94.1
	w	*w*	13,203.9	-409.0 (12.8)	575.8	94.1
'ML' Raking	*d*	*d*	8,106.3	-5,507.2 (7.4)	5,512.2	75.9
	d	*w*	8,107.1	-5,506.4 (7.4)	5,511.4	75.9
	d	*df*	8,109.7	-5,503.8 (7.4)	5,508.8	75.9
	w	*d*	13,208.9	-404.6 (12.9)	574.8	94.1
	w	*w*	13,204.2	-409.2 (12.9)	577.3	94.1
	w	*df*	13,202.5	-411.0 (12.9)	578.1	94.1

[1] see text following equation (4.8), where choices *df*, *d* and *w* correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.
[2] percentage of 95% normal-theory confidence intervals containing true value.

Table 6.6
Relative bias (%) of variance estimators of expenditure and income totals from SIE (R = 1,000)

Weighting Method	*w*- or *d*-weighted residuals[1]	weight used for $\hat{B}$ in residual[1]	Relative Bias of Standard Error Estimator	
			Expenditure	Income
Complete Response:				
GREG	*d*	*d*	0.7	-1.3
	d	*w*	0.7	-1.3
	w	*d*	1.3	-1.0
	w	*w*	1.3	-1.0
Classical Raking	*d*	*d*	0.7	-1.4
	d	*w*	0.7	-1.4
	w	*d*	1.2	-1.0
	w	*w*	1.2	-1.0
'ML' Raking	*d*	*d*	0.6	-1.5
	d	*w*	0.6	-1.5
	d	*df*	0.6	-1.4
	w	*d*	1.2	-1.1
	w	*w*	1.2	-1.1
	w	*df*	1.2	-1.1
Multiplicative nonresponse:				
GREG	*d*	*d*	-38.2	-40.4
	d	*w*	-38.2	-40.3
	w	*d*	-0.3	-2.7
	w	*w*	-0.3	-2.8
Classical Raking	*d*	*d*	-38.2	-40.3
	d	*w*	-38.2	-40.3
	w	*d*	-0.3	-2.7
	w	*w*	-0.3	-2.8
'ML' Raking	*d*	*d*	-38.2	-40.3
	d	*w*	-38.2	-40.3
	d	*df*	-38.2	-40.3
	w	*d*	-0.3	-2.7
	w	*w*	-0.3	-2.7
	w	*df*	-0.4	-2.8
Additive nonresponse:				
GREG	*d*	*d*	-38.1	-40.5
	d	*w*	-38.1	-40.5
	w	*d*	-0.2	-3.0
	w	*w*	-0.2	-3.0
Classical Raking	*d*	*d*	-38.1	-40.5
	d	*w*	-38.1	-40.5
	w	*d*	-0.2	-3.0
	w	*w*	-0.2	-3.0
'ML' Raking	*d*	*d*	-38.2	-40.5
	d	*w*	-38.2	-40.5
	d	*df*	-38.1	-40.4
	w	*d*	-0.2	-3.0
	w	*w*	-0.3	-3.0
	w	*df*	-0.3	-3.0

[1] see text following equation (4.8), where *df*, *d* and *w* correspond to $\hat{B}$ in (i), (ii) and (iii) respectively.

7. Conclusions

The simulation study showed little difference between the bias or variance properties of the three calibration estimators considered: the GREG estimator, the classical raking estimator and the maximum likelihood raking estimator. Some small differences in the distribution of extreme weights were observed: the maximum likelihood raking estimator had the most very large weights and the GREG estimator was the only one with a few negative weights.

Amongst the variance estimators, the main finding was the contrast between the approach which weights residuals by the design weight and that which weights them by the calibrated weight. It was found that the latter variance estimator always had smaller bias and that this effect was very marked in the presence of nonresponse, when the former estimator could be severely biased. The bias of the latter estimator was generally small and the coverage level of the associated confidence intervals was generally close to the nominal coverage.

Alternative ways of weighting the observations in constructing the regression coefficients, when calculating the residuals in the linearization variance estimator, were considered but little effect was observed and there was no evidence that this choice is important in practice.

In general, the findings for the categorical variables in the British Labour Force Survey were remarkably similar to the findings for the continuous variables in the German Income and Expenditure survey.

Acknowledgements

Comments from two referees helped improve this paper significantly. We are grateful to the Office for National Statistics for making the Labour Force Survey data available and to Ralf Münnich and colleagues on the DACSEIS project (http://www.dacseis.de/) for providing the synthetic population based on the German Survey of Income and Expenditure. This research was supported by the Economic and Social Research Council.

References

Binder, D.A., and Théberge, A. (1988). Estimating the variance of raking ratio estimators. *Canadian Journal of Statistics*, 16, Supp. 47-55.

Brackstone, G.J., and Rao, J.N.K. (1979). An investigation of raking ratio estimators. *Sankhyã*, Series C, 41, 97-114.

Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30, 17-34.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-82.

Deville, J.-C., Särndal, C.-E. and Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-20.

Foster, K. (1998). Evaluating nonresponse on household surveys. *GSS Methodology Series*, 8, Office for National Statistics. London.

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

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken: Wiley.

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

Ireland, C.T., and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika*, 55, 179-188.

Kalton, G., and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, 19, 81-97.

Kalton, G., and Maligalig, D.S. (1991). A comparison of methods for weighting adjustment for nonresponse. *Proceedings of the US Bureau of the Census 1991 Annual Research Conference*, 409-428.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.

Muennich, R., and Schulrle, J. (2003). Monte Carlo simulation study of European surveys, Workpackage 3, Deliverables 3.1 and 3.2. DACSEIS project. Available at http://www.uni-trier.de/index.php?id=29730.

Office for National Statistics (1998). *Labour Force Survey User Guide, Volume* 1: *Background and Methodology*. London.

Särndal, C.-E., and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons, Chichester, England.

Stukel, D.M., Hidiroglou, M.A. and Särndal, C.-E. (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 117-125.

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 193-201
Statistics Canada, Catalogue No. 12-001-X

Linearization variance estimators for model parameters from complex survey data

Abdellatif Demnati and J.N.K. Rao [1]

Abstract

Taylor linearization methods are often used to obtain variance estimators for calibration estimators of totals and nonlinear finite population (or census) parameters, such as ratios, regression and correlation coefficients, which can be expressed as smooth functions of totals. Taylor linearization is generally applicable to any sampling design, but it can lead to multiple variance estimators that are asymptotically design unbiased under repeated sampling. The choice among the variance estimators requires other considerations such as (i) approximate unbiasedness for the model variance of the estimator under an assumed model, and (ii) validity under a conditional repeated sampling framework. Demnati and Rao (2004) proposed a unified approach to deriving Taylor linearization variance estimators that leads directly to a unique variance estimator that satisfies the above considerations for general designs. When analyzing survey data, finite populations are often assumed to be generated from super-population models, and analytical inferences on model parameters are of interest. If the sampling fractions are small, then the sampling variance captures almost the entire variation generated by the design and model random processes. However, when the sampling fractions are not negligible, the model variance should be taken into account in order to construct valid inferences on model parameters under the combined process of generating the finite population from the assumed super-population model and the selection of the sample according to the specified sampling design. In this paper, we obtain an estimator of the total variance, using the Demnati-Rao approach, when the characteristics of interest are assumed to be random variables generated from a super-population model. We illustrate the method using ratio estimators and estimators defined as solutions to calibration weighted estimating equations. Simulation results on the performance of the proposed variance estimator for model parameters are also presented.

Key Words: Calibration; Ratio estimators; Total variance; Logistic regression; Weighted estimating equations.

1. Introduction

In survey sampling, estimation of a finite population total $Y = \sum_{k=1}^{N} y_k \equiv Y(y)$ is often of interest, where N is the size of the finite population. For a general sampling design with positive inclusion probabilities π_k, a custumary design unbiased estimator of the total Y is given by $\hat{Y} = \sum_{i \in s} y_i/\pi_i \equiv \sum_{k=1}^{N} d_k(s) y_k$, where s is a sample, $d_k(s) = a_k(s)/\pi_k$ are the design weights with $a_k(s) = 1$ if $k \in s$ and $a_k(s) = 0$ otherwise. We use operator notation and write $\hat{Y}(z) = \sum_{k=1}^{N} d_k(s) z_k$ so that $\hat{Y} = \hat{Y}(y)$. Henceforth, all the sums are considered on the whole population and hence write $\sum_{k=1}^{N} y_k = \sum y_k$ and $\hat{Y}(z) = \sum d_k(s) z_k$, to simplify the notation. Again, using the operator notation, we denote an unbiased estimator of the variance of $\hat{Y}(z)$ as a quadratic function, $\vartheta(z)$, in the z_k's.

More complex estimators of a total Y based on known population auxiliary information, such as ratio and regression estimators, and estimators of more complex parameters obtained as solutions to sample weighted estimating equations, such as estimators of "census" logistic regression coefficients, are also often used in practice. Estimators that can be expressed as a general functional $T(\hat{M})$ have also been studied, where $\hat{M}$ denotes a measure that allocates the weight $d_k(s)$ to y_k; for example, $T(\hat{M}) = \int x d\hat{M}(x) = \sum d_k(s) y_k$ if the population parameter is the total $T(M) = \int x dM(x) = Y$, where the measure M allocates a unit mass to each y_k (Deville 1999). Large-sample estimation of the variance of such complex estimators, $\hat{\theta}$ say, has received considerable attention in the literature. In particular, Taylor linearization methods of estimating the variance of $\hat{\theta}$ are generally applicable to any sampling design that permits an unbiased variance estimator $\vartheta(z)$ of $\hat{Y}(z)$. Binder (1983) studied estimators $\hat{\theta}$ that are solutions to weighted estimating equations and applied Taylor linearization to obtain a variance estimator that can be expressed as $\vartheta(\tilde{z})$, where the linearized variable $\tilde{z}_k$ depends on unknown parameters, and $\tilde{z}_k$ is replaced by an estimator z_k that may be based on the substitution method. Deville (1999) derived a Taylor linearization variance estimator of the functional $T(\hat{M})$ as $\vartheta(\tilde{z})$, where $\tilde{z}_k = I_T(M; y_k)$ denotes the influence function of T at y_k, and then replaced $\tilde{z}_k$ by the sample estimator $z_{k1} = I_T(\hat{M}; y_k)$. For example, when $\hat{\theta}$ is the ratio estimator $(\hat{Y}/\hat{X})X = \hat{R}X$ of the total Y, where $\hat{X} = \hat{Y}(x)$ and $X = Y(x)$ is the known total of an auxiliary variable x, we get $\tilde{z}_k = y_k - Rx_k$ and $z_{k1} = y_k - \hat{R}x_k$. However, $z_k = (X/\hat{X})(y_k - \hat{R}x_k)$ is also a candidate to estimate $\tilde{z}_k$ and the resulting $\vartheta(z)$ is often preferred over $\vartheta(z_1)$; see Demnati and Rao (2004). Thus the choice of an

1. Abdellatif Demnati, Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. E-mail: Abdellatif.Demnati@statcan.gc.ca; J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6. E-mail: JRao@math.carleton.ca.

estimator of $\tilde{z}_k$ is somewhat arbitrary under Deville's approach.

Demnati and Rao (2004) studied general estimators that can be expressed as smooth functions of the weights $\boldsymbol{d}(s) = \{d_1(s), ..., d_N(s)\}^T$, say $\hat{\theta} = f(\boldsymbol{d}(s))$, and obtained a Taylor linearization variance estimator directly as $\vartheta(z)$ with known linearized variables $z_k = \partial f(b)/\partial b_k |_{\boldsymbol{b}=\boldsymbol{d}(s)}$ without estimating $\tilde{z}_k$ first and then replacing it by an estimator. For example, in the case of the ratio estimator their method automatically leads to z_k given above. This method can be applied to a variety of estimators including estimators of "census" logistic regression parameters based on calibration weights (Demnati and Rao 2004). Previous work on direct variance estimation includes Binder (1996).

When analyzing survey data, the population values y_k, $k = 1, ..., N$, are often assumed to be generated from a super-population model, and the user is often interested in making inferences on the model parameters. Let θ_N be a "census" parameter, *i.e.*, an estimator of a model parameter θ when the population y_k -values are all known, and let $\hat{\theta}$ be a design-unbiased estimator of θ_N, the "census" parameter. Suppose that $\hat{\theta}$ is design-model unbiased for θ, *i.e.*, $E_m E_p(\hat{\theta}) = \theta$, where E_m and E_p respectively denote the expectations with respect to the design and the model. Then the total variance of $\hat{\theta}$ is $V(\hat{\theta}) = E_m E_p(\hat{\theta} - \theta)^2$ which can be decomposed as

$$V(\hat{\theta}) = E_m V_p(\hat{\theta}) + V_m(\theta_N), \qquad (1.1)$$

where $V_p(\hat{\theta}) = E_p(\hat{\theta} - \theta_N)^2$ is the design variance of $\hat{\theta}$ and $V_m(\theta_N)$ is the model variance of θ_N. It follows from (1.1) that the total variance may be estimated using a design-based estimator of $V_p(\hat{\theta})$ if the last term $V_m(\theta_N)$ is negligible relative to $E_m V_p(\hat{\theta})$. In that case, the distinction between θ_N and θ can be ignored (Skinner, Holt and Smith 1989, page 14). On the other hand, it is necessary to estimate the total variance $V(\hat{\theta})$ when the model variance $V_m(\theta_N)$ is not negligable relative to $E_m V_p(\hat{\theta})$. This requires consideration of the joint design and model random processes. Molina, Smith and Sugden (2001) argued that the combined process of generation of the finite population and selection of the sample should be the basis for analytical inferences on model parameters. Rubin-Bleuer and Şchiopu-Kratina (2005) have provided a mathematical framework for joint model and design-based inference. However, a broadly applicable method is needed for the estimation of total variance. The main purpose of this paper is to provide such a method, by extending the Demnati-Rao approach for finite population parameters.

In Section 2, we consider the case of a scalar parameter θ and present linearization variance estimators by expanding the Demnati and Rao (2004) approach. The method is illustrated for the special case of a ratio estimator of a super-population mean θ. Results of Section 2 are extended in Section 3 to estimators of a vector parameter $\boldsymbol{\theta}$ obtained as solutions to weighted estimating equations, and the method is illustrated for the special case of parameters of a logistic regression model. Simulation results are also presented.

2. Scalar model parameter

2.1 Point estimators

Consider a finite population U of N elements, and let $d_k(s) = a_k(s)/\pi_k$ be the design weights attached to the population element k, where $a_k(s) = 1$ if element k is in the sample s and $a_k(s) = 0$ otherwise, and π_k is the inclusion probability associated with k. We consider estimators $\hat{\theta}$ of a scalar parameter θ that can be expressed as functions of random variables under the design and the assumed model. In particular, $\hat{\theta} = f(\boldsymbol{A}_d)$, where $\boldsymbol{A}_d$ is a $(p+1) \times N$ matrix with columns $\boldsymbol{d}_k = (d_k h_{1k}, d_k h_{2k}, ..., d_k h_{(p+1)k})^T \equiv (d_{1k}, ..., d_{(p+1)k})^T$ where $d_k = d_k(s)$ is random under the design, $h_{1k} = 1$, and h_{ik} $(i = 2, ..., p+1)$ are random under the model.

For example, consider the ratio model with fixed covariates x_k:

$$E_m(y_k) = \beta x_k,\ V_m(y_k) = \sigma^2 x_k,\ \mathrm{Cov}_m(y_k,\ y_t) = 0,\quad k \neq t,\ k, t = 1, ..., N, \qquad (2.1)$$

where E_m, V_m, and Cov_m denote model expectation, model variance, and model covariance respectively and $\sigma^2 > 0$. Suppose that we are interested in estimating the super-population mean $\theta = E_m(\bar{Y}) = N^{-1} \sum E_m(y_k) = \beta \bar{X}$ where $\bar{Y}$ is the finite population mean of y. In this case, a ratio estimator of θ is given by

$$\hat{\theta} = \bar{X}(\hat{Y}/\hat{X}) \equiv \bar{X}\hat{R}, \qquad (2.2)$$

where $\hat{Y} = \sum d_k(s) y_k$ and $\hat{X} = \sum d_k(s) x_k$ are the design-unbised estimators of the totals Y and X, and $\bar{X}$ is the know population mean of x. We can write the ratio estimator (2.2) in the form $\hat{\theta} = \bar{X}(\sum d_{2k})/\sum d_{1k} x_k$, where $d_{1k} = d_k(s)$ and $d_{2k} = d_k(s) y_k$. This is a special case of $f(\boldsymbol{A}_d)$ with $p = 1$ and $h_{2k} = y_k$.

Let E_p be the design expectation and $E = E_m E_p$ be the total expectation. Then, we have $E(d_{1k}) = E_m(1) = 1 \equiv \mu_{1k}$ and $E(d_{ik}) = E_m(g_{ik}) \equiv \mu_{ik}$, $i = 2, ..., p+1$, noting that $E_p(d_k(s)) = 1$. We assume that $f(\boldsymbol{A}_\mu) = \theta$, where $\boldsymbol{A}_\mu$ is a $(p+1) \times N$ matrix with columns $\boldsymbol{\mu}_k = (\mu_{1k}, \mu_{2k}, ..., \mu_{(p+1)k})^T$. Hence, $\hat{\theta}$ is asymptotically pm -unbised for θ. In the special case of the ratio estimator, we have $f(\boldsymbol{A}_\mu) = \beta \bar{X} = \theta$, noting that $\mu_{1k} = 1$ and $\mu_{2k} = \beta x_k$.

2.2 Linearization variance estimator

We first derive an estimator of the total variance of a linear estimator $\hat{U} = \sum \boldsymbol{u}_k^T \boldsymbol{d}_k$, where $\boldsymbol{u}_k$ is a vector of constants. The total variance of $\hat{U}$ may be decomposed as

$$V(\hat{U}) = E_m V_p(\hat{U}) + V_m E_p(\hat{U}) \equiv I + II, \quad (2.3)$$

where V_p and V_m denote design variance and model variance respectively. A design-unbiased estimator of the component I of the total variance (2.3) is obtained by estimating the design variance $V_p(\hat{U})$ for fixed $\boldsymbol{h}_k = (h_{1k}, ..., h_{(p+1)k})^T$. Now, noting that $\hat{U} = \sum b_k d_k(s)$ is the standard Narain-Horvitz-Thompson (NHT) estimator of the total $U = \sum b_k$ when $b_k = \boldsymbol{u}_k^T \boldsymbol{h}_k$ are fixed conditionally, we can use either the Sen-Yates-Grandy (SYG) variance estimator for fixed sample size designs or the Horvitz-Thompson (HT) variance estimator for arbitrary designs. The SYG estimator is given by

$$\begin{aligned} \text{est}(I) &= \vartheta_{\text{SYG}}(\hat{U}) \\ &= \sum\sum_{k<t} d_{kt}(s) \frac{(\pi_k \pi_t - \pi_{kt})}{\pi_k \pi_t} (b_k - b_t)^2, \end{aligned} \quad (2.4)$$

where $d_{kt}(s) = \{a_k(s)\, a_t(s)\}/\pi_{kt}$ and π_{kt} is the inclusion probability for units k and t $(k \neq t)$. The HT variance estimator is given by

$$\text{est}(I) = \vartheta_{\text{HT}}(\hat{U}) = \sum\sum d_{kt}(s) \frac{(\pi_{kt} - \pi_k \pi_t)}{\pi_k \pi_t} b_k b_t, \quad (2.5)$$

where $d_{kk}(s) = d_k(s)$. For the special case of stratified random sampling (2.4) and (2.5) are identical.

Turning to the component II of the total variance (2.3), we have $V_m E_p(\hat{U}) = V_m(\sum \boldsymbol{u}_k^T \boldsymbol{h}_k) = \sum\sum \boldsymbol{u}_k^T \text{Cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\, \boldsymbol{u}_t$ and a pm-unbiased estimator is therefore given by

$$\text{est}(II) = \sum\sum d_{kt}(s)\, \boldsymbol{u}_k^T \text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\, \boldsymbol{u}_t, \quad (2.6)$$

after replacing $\text{Cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)$ by an estimator $\text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)$. The estimator of total variance (2.3) is now given by $\text{est}(I) + \text{est}(II)$. We denote it, in operator notation, as $\vartheta(\boldsymbol{u})$.

We now turn to the estimation of total variance of $\hat{\theta}$. Following Demnati and Rao (2004), a Taylor expansion of $\hat{\theta} - \theta$ may be written as

$$\hat{\theta} - \theta \approx \sum \tilde{\boldsymbol{z}}_k^T (\boldsymbol{d}_k - \boldsymbol{\mu}_k) \quad (2.7)$$

where $\tilde{\boldsymbol{z}}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_\mu}$ and $\boldsymbol{A}_b$ is a $(p+1) \times N$ matrix with k^{th} column $\boldsymbol{b}_k$, a vector of arbitrary real numbers. The approximation (2.7) is valid for any $\hat{\theta}$ that can be expressed as a smooth function of estimated totals. Following Demnati and Rao (2004), a linearization estimator of the total variance is now given by

$$\vartheta_{\text{DR}}(\hat{\theta}) = \vartheta(\boldsymbol{z}), \quad (2.8)$$

which is obtained from $\vartheta(\boldsymbol{u})$ by replacing $\boldsymbol{u}_k$ by the "linearized variable" $\boldsymbol{z}_k = \partial f(\boldsymbol{A}_b)/\partial \boldsymbol{b}_k \,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$. A rigorous theoretical justification of (2.8) follows along the lines of Deville (1999).

2.3 Special case of ratio estimator

For the ratio estimator $\hat{\theta} = \bar{X}\hat{R}$ of the model parameter $\theta = \beta \bar{X}$, $\boldsymbol{z}_k$ reduces to

$$\boldsymbol{z}_k = (\bar{X}/\hat{X})(-\hat{R}\, x_k, 1)^T = (z_{1k}, z_{2k})^T. \quad (2.9)$$

Further, b_k in (2.4) or (2.5) is replaced by

$$\begin{aligned} \boldsymbol{z}_k^T \boldsymbol{h}_k &= z_{1k} + z_{2k} y_k \\ &= (\bar{X}/\hat{X})(y_k - \hat{R}\, x_k) \equiv (\bar{X}/\hat{X})\, e_k, \end{aligned}$$

using (2.9). Also, replacing $\boldsymbol{u}_k$ by $\boldsymbol{z}_k$ in (2.6) we get

$$\boldsymbol{z}_k^T \text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\, \boldsymbol{z}_t = z_{2k} z_{2t} \text{cov}_m(y_k, y_t).$$

Under the ratio model (2.1) with unspecified model variance $V_m(y_k) = \sigma_k^2$, $k = 1, ..., N$, we can estimate $\sigma_k^2 = E_m(y_k - \beta x_k)^2$ by $(y_k - \hat{R} x_k)^2$ and letting $\text{cov}_m(y_k, y_t) = 0$, for $k \neq t$.

We now study the special case of simple random sampling without replacement. In this case, both (2.4) and (2.5) reduce to

$$\text{est}(I) = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{1}{n}\left(1 - \frac{n}{N}\right) s_e^2, \quad (2.10)$$

where $s_e^2 = \sum a_k(s)\, e_k^2/(n-1)$, and (2.6) reduces to

$$\text{est}(II) = \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{(n-1)}{nN} s_e^2. \quad (2.11)$$

Hence, using (2.10) and (2.11), the variance estimator (2.8) reduces to

$$\begin{aligned} \vartheta_{\text{DR}}(\hat{\theta}) &= \text{est}(I) + \text{est}(II) \\ &= \left(\frac{\bar{X}}{\bar{x}}\right)^2 \frac{1}{n} \frac{N-1}{N} s_e^2. \end{aligned} \quad (2.12)$$

It is interesting to note that the "g-weight" $\bar{X}/\bar{x}$ appears automatically in $\vartheta_{\text{DR}}(\hat{\theta})$, given by (2.12), and that the finite population correction $1 - n/N$ is absent in $\vartheta_{\text{DR}}(\hat{\theta})$ unlike in $est(I)$ given by (2.10).

In the customary approach to the estimation of total variance (see *e.g.*, Korn and Graubard 1998) $V(\hat{\theta})$ is first written as

$$V(\hat{\theta}) = E_m V_p(\hat{\theta}) + V_m E_p(\hat{\theta})$$

$$\approx E_m V_p(\hat{\theta}) + V_m(\bar{Y})$$

$$= E_m V_p(\hat{\theta}) + N^{-2} \sum E_m (y_k - \beta\, x_k)^2, \quad (2.13)$$

under the ratio model with unspecified σ_k^2, $k = 1, ..., N$. The first term $E_m V_p(\hat{\theta})$ in (2.13) is then estimated by a design-consistent estimator of $V_p(\hat{\theta})$, typically by (2.10) without the *g*-factor $(\bar{X}/\bar{x})^2$. The second term is estimated by $N^{-2} \sum d_k(s)(y_k - \hat{R}\, x_k)^2 = (nN)^{-1}(n-1)s_e^2$. The sum of the two estimated terms then equals (2.12) without the *g*-factor. We denote this customary variance estimator by $\vartheta_{\text{cus}}(\hat{\theta})$. On the other hand, if (2.10) with the *g*-factor is used to estimate $V_p(\hat{\theta})$, the sum of this estimated term and the previous estimator of the second term leads to a "hybrid" variance estimator

$$\vartheta_{\text{mix}}(\hat{\theta}) = \text{est}(I) + (nN)^{-1}(n-1)s_e^2,$$

where the *g*-term is absent in the last term. It is clear from the above results that the choice of estimator of total variance under the customary approach is not unique, unlike under the proposed approach.

If the parameter of interest is $\beta = \theta/\bar{X}$ instead of θ, then $\hat{\beta} = \hat{\theta}/\bar{X} = \hat{R}$ and $\vartheta_{\text{DR}}(\hat{\beta})$ under simple random sampling is give by

$$\vartheta_{\text{DR}}(\hat{\beta}) = \bar{X}^{-2}\vartheta_{\text{DR}}(\hat{\theta}) = \bar{x}^{-2}\frac{1}{n}\frac{N-1}{N}s_e^2. \quad (2.14)$$

The customary approach leads to the same variance estimator, (2.14).

2.4 Simulation study

We conducted a small simulation study to examine the performances of different variance estimators, both un-conditionally and conditionally on $\hat{X}$. We first generated $R = 2{,}000$ finite populations $\{y_1, ..., y_N\}$ each of size $N =$ 393, from the ratio model

$$y_k = 2\, x_k + x_k^{1/2}\varepsilon_k, \quad (2.15)$$

with independent values ε_k generated from $N(0, 1)$, where the fixed x_k are the "number of beds" for the Hospitals population studied in Valliant, Dorfman and Royall (2000, page 424-427). One simple random sample of specified size n is drawn from each generated population. Our parameter of interest is $\theta = \beta\bar{X}$, where $\beta = 2$.

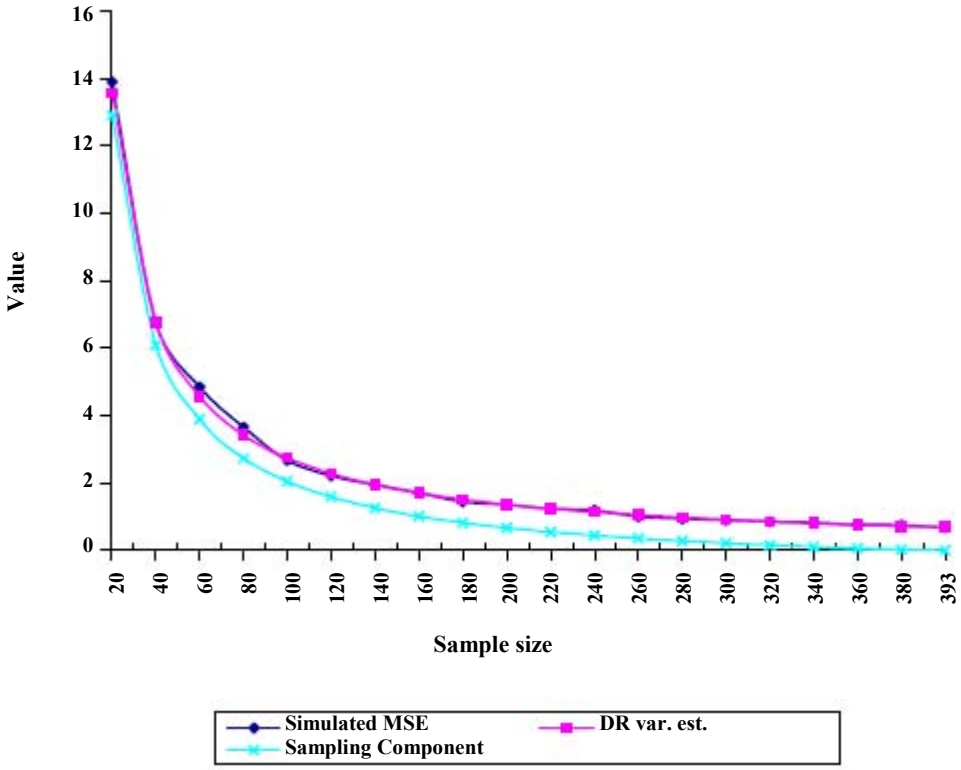


Figure 1 **Averages of variance estimates for selected sample sizes compared to estimated MSE of the ratio estimator. $\bar{\vartheta}_{\text{DR}}$ = DR var. est., $\bar{\vartheta}_s$ = Sampling component: ratio model**

Simulated total MSE of the ratio estimator $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$ is calculated as $M(\hat{\theta}) = R^{-1}\sum_{r=1}^{2,000}(\hat{\theta}_r - \theta)^2$, where $\hat{\theta}_r$ is the value of $\hat{\theta}$ for the r^{th} simulated sample and $(\bar{y}, \bar{x})$ are the sample means. We calculated the total variance estimate $\vartheta_{\text{DR}}(\hat{\theta})$, and its components $\vartheta_s = \text{est}(I)$ and $\vartheta_m = \text{est}(II)$ from each simulated sample r and their averages $\bar{\vartheta}_{\text{DR}}$, $\bar{\vartheta}_s$, and $\bar{\vartheta}_m$ over r. Figure 1 gives a plot of the average of variance estimates, $\bar{\vartheta}_{\text{DR}}$ and $\bar{\vartheta}_s$, and the simulated total MSE for $n = 20, 40, ..., 380, 393$. In the case of $n = N$, $\bar{\vartheta}_s = 0$. It is seen from Figure 1, that ϑ_{DR} is approxiamatly unbiased, whereas ϑ_s leads to severe underestimation as the sample size, n, increases.

We also examined the conditional performance of the variance estimators under simple random sampling given $\bar{x}$, by conducting another simulation study for inference on θ, using model (2.15). The study is similar to the study of Royall and Cumberland (1981) for inference on the finite population mean $\theta_N = \bar{Y}$ from a fixed population $\{y_1, ..., y_N\}$. We generated $R = 20{,}000$ finite populations $\{y_1, ..., y_N\}$, each of size $N = 393$ from (2.15) using the number of beds as x_k, and from each population we then selected one simple random sample of size $n = 100$. We arranged the 20,000 samples in ascending order of $\bar{x}$-values and then grouped them into 20 groups each of size 1,000 such that the first group, G_1, contained 1,000 samples with the smallest $\bar{x}$-values, the next group, G_2, contained the next 1,000 smallest $\bar{x}$-values, and so on to get $G_1, ..., G_{20}$. For each of the 20 groups so formed, we calculated the average values of the ratio estimates $\hat{\theta} = \bar{X}(\bar{y}/\bar{x})$ and the mean estimates $\bar{y}$, and the resulting conditional relative bias (CRB) in estimating $\theta = 2\bar{X}$; see Figure 2. It is clear from Figure 2 that $\bar{y}$ is conditionally biased unlike $\hat{\theta}$: negative CRB (-14%) for G_1 increasing to positive CRB (+14%) for G_{20}. Note that both $\bar{y}$ and $\hat{\theta}$ are unconditionally unbiased for θ. The conditional bias of $\hat{\theta}$ and $\bar{y}$ in estimating the model parameter θ is similar to the conditional bias in estimating the "census" parameter $\theta_N = \bar{Y}$, as observed by Royall and Cumberland (1981).

We also calculated the conditional MSE of $\hat{\theta}$ and the associated CRB of the variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix} based on the average values of ϑ_{DR}, ϑ_{cus} and ϑ_{mix} in each group; see Figure 3. It is evident from Figure 3 that CRB of ϑ_{cus} ranges from -28% to 20% across the groups whereas ϑ_{DR} exhibits no such trend and its CRB is less than 5% in absolute value except for G_6 and G_{20}. Also, the CRB of ϑ_{mix} is largely negative and below that of ϑ_{DR} for the first half of the groups and above for the second half, but ϑ_{mix} exhibits no visible trends unlike ϑ_{cus}.

Figure 4 reports the conditional coverage rates (CCR) of normal theory confidence intervals based on ϑ_{DR}, ϑ_{cus}, ϑ_{mix} and ϑ_s (ignoring the component ϑ_m) for nominal level of 95%. As expected, the use of ϑ_s leads to severe undercoverage because the sampling fraction, 100/393, is significant. On the other hand, CCR associated with ϑ_{DR} is closer to nominal level across groups, while ϑ_{cus} exhibits a trend across groups with CCR ranging from 91% to 97%. Further, CCR associated with ϑ_{mix} is slightly below that of ϑ_{DR} for the first half of the groups but ϑ_{mix} and ϑ_{DR} perform similarly.

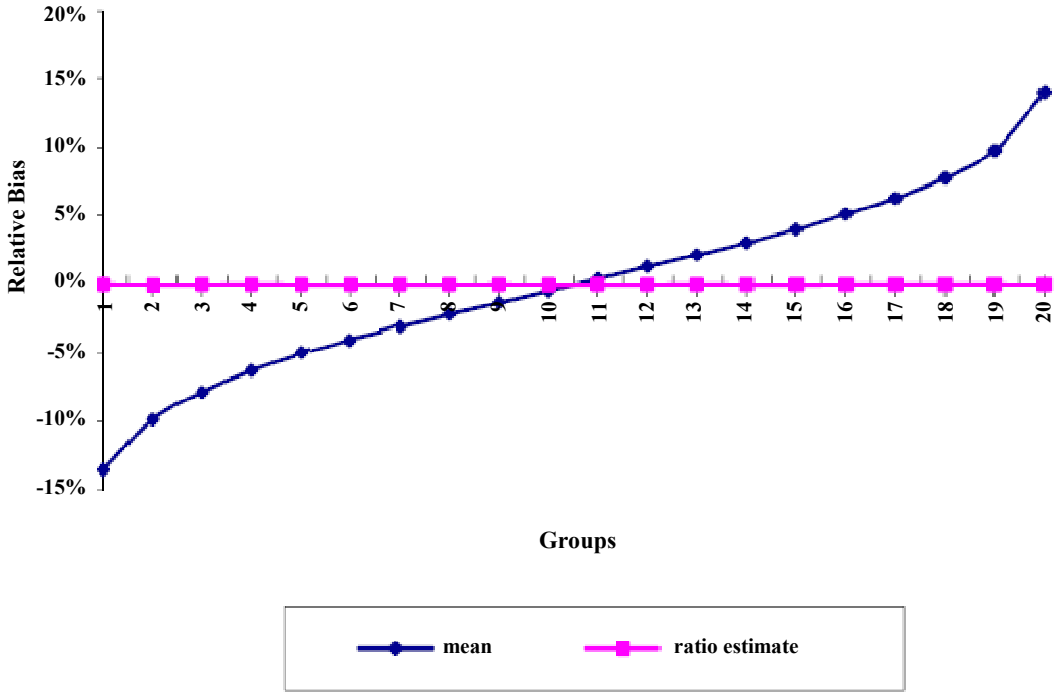


Figure 2 Conditional relative bias of the expansion and ratio estimators: ratio model

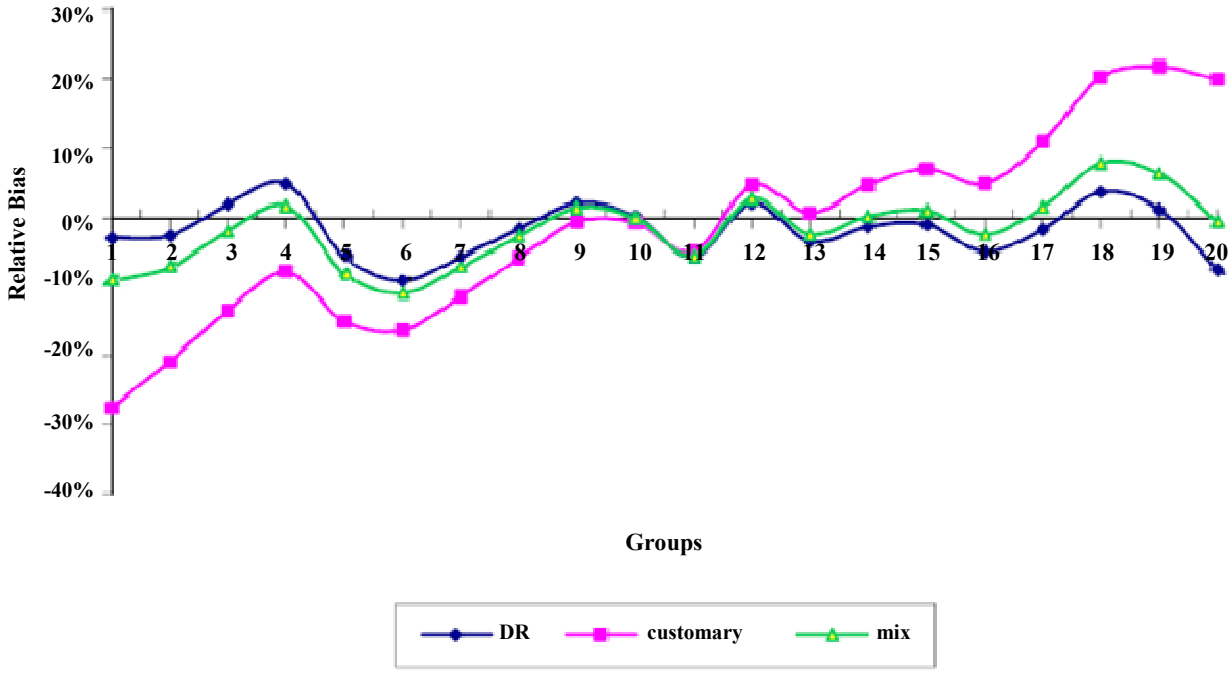


Figure 3 Conditional relative bias of variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix}: ratio model

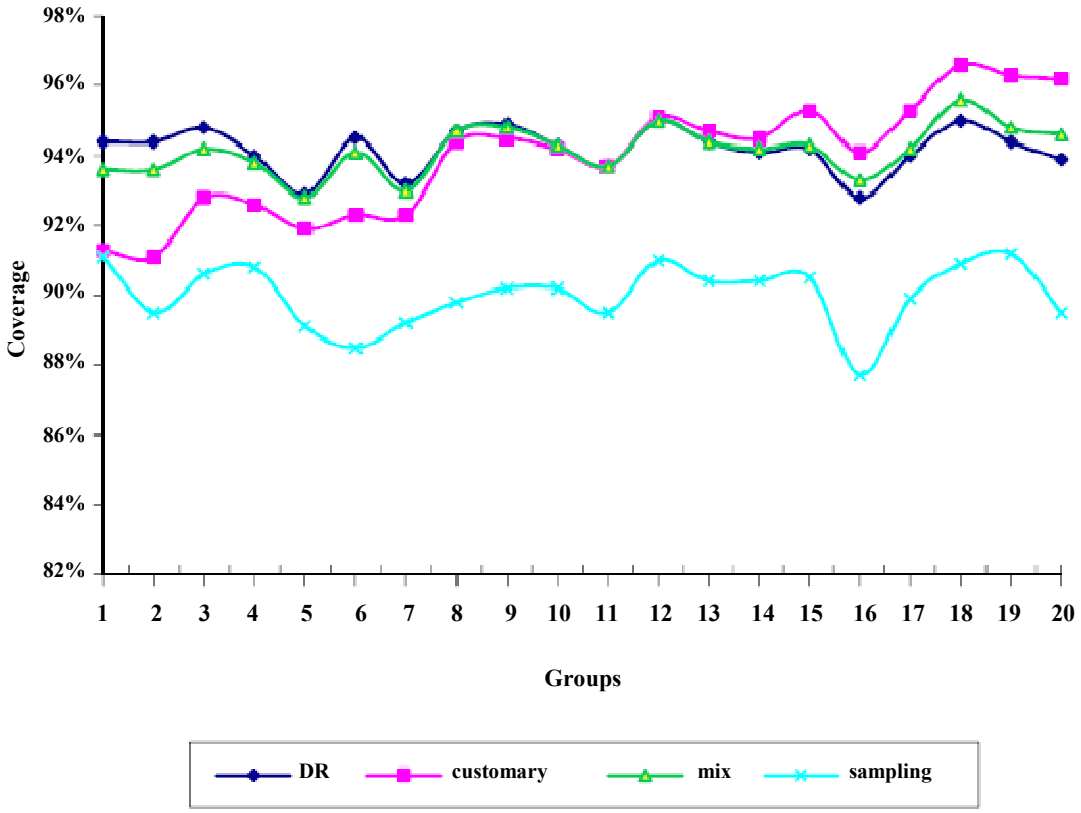


Figure 4 Conditional coverage rates of normal theory confidence intervals based on ϑ_{DR}, ϑ_{cus}, ϑ_{mix} and ϑ_{s} for nominal level of 95%: ratio model

3. Calibration weighted estimating equations

3.1 Estimators of model parameters

Suppose that the super-population model on the responses y_k is specified by a generalized linear model (McCullagh and Nelder 1989) with mean $E_m(y_k) = \mu_k(\boldsymbol{\theta}) = h(\boldsymbol{x}_k^T\boldsymbol{\theta})$, where $\boldsymbol{x}_k$ is a $p\times 1$ vector of explanatory variables, $\boldsymbol{\theta}$ is the p-vector of model parameters and $h(.)$ is a "link" function. For example, $h(a) = a$ gives a linear regression model and $h(a) = e^a/(1+e^a)$ gives a logistic regression model for binary responses y_k.

We define census estimating equations (CEE) , based on estimating functions $\boldsymbol{l}_k(\boldsymbol{\theta})$, as $\boldsymbol{l}(\boldsymbol{\theta}) = \sum \boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}$ with $E_m\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}$, and the solution to CEE gives the census parameter vector $\boldsymbol{\theta}_N$. For example, $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{x}_k(y_k - \boldsymbol{\mu}_k(\boldsymbol{\theta}))$ for linear and logistic regression models. We use generalized regression (GREG) weights $w_k(s) = d_k(s)g_k(\boldsymbol{d}(s))$, where the "$g$-weights" are given by

$$g_k(\boldsymbol{d}(s)) = 1 + (\boldsymbol{T} - \hat{\boldsymbol{T}})^T\left[\sum d_k(s)c_k\boldsymbol{t}_k\boldsymbol{t}_k^T\right]^{-1}c_k\boldsymbol{t}_k,$$

for specified c_k, where $\hat{\boldsymbol{T}} = \sum d_k(s)\boldsymbol{t}_k$ is the HT estimator of the known total $\boldsymbol{T}$ of a $q\times 1$ vector of calibration variables $\boldsymbol{t}_k$ and $\boldsymbol{d}(s)$ is the $N\times 1$ vector of the weights $d_k(s)$. The GREG weights, $w_k(s)$, have the calibration property $\sum w_k(s)\boldsymbol{t}_k = \boldsymbol{T}$ and lead to efficient estimators $\tilde{Y} = \sum w_k(s)y_k$ of totals $Y = \sum y_k$, when y_k and t_k are linearly related (Särndal, Swensson and Wretman 1989, chapter 6).

We use the calibration weights, $w_k(s)$, to estimate the CEE. The calibration weighted estimating equations are given by

$$\tilde{\boldsymbol{l}}(\boldsymbol{\theta}) = \sum w_k(s)\boldsymbol{l}_k(\boldsymbol{\theta}) = \sum d_k(s)g_k(\boldsymbol{d}(s))\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{0}. \quad (3.1)$$

The solution to (3.1), obtained by the Newton-Raphson-type iterative method, gives the calibration-weighted estimator $\tilde{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}}$ is approximately design-model unbiased for $\boldsymbol{\theta}$, *i.e.*, $E(\tilde{\boldsymbol{\theta}}) \approx \boldsymbol{\theta}$. It follows from (3.1) that $\tilde{\boldsymbol{\theta}}$ is of the form $\boldsymbol{f}(\boldsymbol{A}_d)$ with $\boldsymbol{d}_k = (d_k(s), d_k(s)\boldsymbol{l}_k^T(\boldsymbol{\theta}))^T$, where $\boldsymbol{f}(\boldsymbol{A}_d)$ is a $p\times 1$ vector and $\boldsymbol{A}_d$ is a $(p+1)\times N$ matrix with k^{th} column $\boldsymbol{d}_k$. Here we have $h_{1k} = 1$ and $(h_{2k}, ..., h_{(p+1)k}) = \boldsymbol{l}_k(\boldsymbol{\theta})$.

3.2 Linearized variance estimators

We first extend the result on variance estimation for the scalar case $\hat{U} = \sum \boldsymbol{b}_k^T\boldsymbol{d}_k$ (Section 2.2) to the vector case $\hat{\boldsymbol{U}} = \sum \boldsymbol{U}_k\boldsymbol{d}_k = \sum \boldsymbol{b}_k^T d_k(s)$, where $\boldsymbol{b}_k = \boldsymbol{U}_k\boldsymbol{h}_k$ is a p-vector and $\boldsymbol{U}_k$ is a $p\times(p+1)$ matrix with rows $\boldsymbol{u}_{jk}^T$, $j = 1, ..., p$. In this case, the SYG variance estimator (2.4) is changed to

$$\begin{aligned}\mathbf{est}(I) &= \vartheta_{\text{SYG}}(\hat{U})\\ &= \sum\sum_{k<t} d_{kt}(s)\frac{(\pi_k\pi_t - \pi_{kt})}{\pi_k\pi_t}(\boldsymbol{b}_k - \boldsymbol{b}_t)(\boldsymbol{b}_k - \boldsymbol{b}_t)^T. \quad (3.2)\end{aligned}$$

Similarly, the H-T variance estimator (2.5) is changed to

$$\mathbf{est}(I) = \vartheta_{\text{HT}}(\hat{U}) = \sum\sum d_{kt}(s)\frac{(\pi_{kt} - \pi_k\pi_t)}{\pi_k\pi_t}\boldsymbol{b}_k\boldsymbol{b}_t^T. \quad (3.3)$$

Turning to the component II of the total variance of $\hat{\boldsymbol{U}}$, (2.6) is changed to

$$\mathbf{est}(II) = \sum\sum d_{kt}(s)\boldsymbol{U}_k\,\text{cov}_m(\boldsymbol{h}_k, \boldsymbol{h}_t)\boldsymbol{U}_t^T. \quad (3.4)$$

The total variance of $\hat{\boldsymbol{U}}$ is estimated by the sum of (3.2) and (3.4) for fixed sample size designs or by the sum of (3.3) and (3.4) for arbitrary designs.

A linearization variance estimator of the total variance of $\tilde{\boldsymbol{\theta}}$ is obtained from the estimated total variance estimator of $\hat{\boldsymbol{U}}$ by replacing $\boldsymbol{U}_k$ by the linearized variable $\boldsymbol{Z}_k = \partial\boldsymbol{f}(\boldsymbol{A}_b)/\partial\boldsymbol{b}_k\,|_{\boldsymbol{A}_b = \boldsymbol{A}_d}$. Following the implicit differentiation method of Demnati and Rao (2004), $\boldsymbol{Z}_k$ reduces to

$$\boldsymbol{Z}_k = [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}g_k(\boldsymbol{d}(s))(-\hat{\boldsymbol{B}}_l^T\boldsymbol{t}_k, \boldsymbol{I}_p),$$

with

$$\hat{\boldsymbol{B}}_l = \left[\sum d_k(s)c_k\boldsymbol{t}_k\boldsymbol{t}_k^T\right]^{-1}\sum d_k(s)c_k\boldsymbol{t}_k\boldsymbol{l}_k^T(\tilde{\boldsymbol{\theta}}),$$

$$\tilde{\boldsymbol{J}}(\boldsymbol{\theta}) = -\sum d_k(s)g_k(\boldsymbol{d}(s))(\partial\boldsymbol{l}_k(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T),$$

and $\boldsymbol{I}_p$ is the $p\times p$ identity matrix.

After some simplification, the first component $\mathbf{est}(I)$ is given by (3.2) or (3.3) with $\boldsymbol{b}_k$ changed to

$$\boldsymbol{Z}_k\boldsymbol{h}_k = [\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}e_k(\tilde{\boldsymbol{\theta}})g_k(\boldsymbol{d}(s)), \quad (3.5)$$

where

$$\boldsymbol{e}_k(\tilde{\boldsymbol{\theta}}) = \boldsymbol{l}_k(\tilde{\boldsymbol{\theta}}) - \hat{\boldsymbol{B}}_l^T\boldsymbol{t}_k.$$

Similarly, the second component $\mathbf{est}(II)$ simplifies to

$$\mathbf{est}(II) =$$

$$[\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}\sum d_k(s)g_k^2(\boldsymbol{d}(s))\boldsymbol{l}_k(\tilde{\boldsymbol{\theta}})\boldsymbol{l}_k^T(\tilde{\boldsymbol{\theta}})[\tilde{\boldsymbol{J}}(\tilde{\boldsymbol{\theta}})]^{-1}, \quad (3.6)$$

if $\text{Cov}_m[\boldsymbol{l}_k(\tilde{\boldsymbol{\theta}})\boldsymbol{l}_t^T(\tilde{\boldsymbol{\theta}})] = \boldsymbol{0}$ for $k \neq t$.

The total variance estimator of $\tilde{\boldsymbol{\theta}}$ is now estimated by

$$\vartheta_{\text{DR}}(\tilde{\boldsymbol{\theta}}) = \mathbf{est}(I) + \mathbf{est}(II). \quad (3.7)$$

This variance estimator of $\tilde{\boldsymbol{\theta}}$ automatically takes account of the g-weights as in Section 2.

A customary variance estimator of $\tilde{\boldsymbol{\theta}}$, $\vartheta_{\text{cus}}(\tilde{\boldsymbol{\theta}})$, is obtained from (3.7) by ignoring the g-weights in (3.5) and (3.6). Similarly, a hybrid variance estimator, $\vartheta_{\text{mix}}(\tilde{\boldsymbol{\theta}})$, is

obtained from (3.7) by retaining the g-weights in $\mathbf{est}(I)$ and ignoring them in $\mathbf{est}(II)$.

3.3 Simulation study

We conducted a simulation study to compare the relative performances of the three variance estimators ϑ_{DR}, ϑ_{cus}, and ϑ_{mix}, for the special case of a logistic regression model:

$$E_m(y_k) = \mu_k(\boldsymbol{\theta}) = \exp(\boldsymbol{x}_k^T\boldsymbol{\theta})/\{1+\exp(\boldsymbol{x}_k^T\boldsymbol{\theta})\} \qquad (3.8)$$

$$V_m(y_k) = \mu_k(\boldsymbol{\theta})(1-\mu_k(\boldsymbol{\theta})),\ \mathrm{Cov}_m(y_k, y_t) = 0,\ k \neq t.$$

In this case, we have $\boldsymbol{l}_k(\boldsymbol{\theta}) = \boldsymbol{x}_k(y_k - \mu_k(\boldsymbol{\theta}))$, and

$$\tilde{\boldsymbol{J}}(\boldsymbol{\theta}) = \sum d_k(s) g_k(\boldsymbol{d}(s)) \boldsymbol{x}_k \boldsymbol{x}_k^T \mu_k(\boldsymbol{\theta})(1-\mu_k(\boldsymbol{\theta})).$$

For the simulation study, we set $\boldsymbol{x}_k = (1,\ x_k)^T$, where the x_k denote the number of beds for the Hospitals population of size $N = 393$ studied in Section 2.2. We implemented post-stratification by dividing the population into two classes with $N_1 = 171$ hospitals k having $x_k < 350$ in class 1 and $N_2 = 122$ hospitals k with $x_k \geq 350$ in class 2. Here, $g_k(\boldsymbol{d}(s)) = N_h/\hat{N}_h$, $h = 1, 2$, if k belongs to class h, where $\hat{N}_h = \sum d_k(s) t_{hk}$ is the design-weight estimator of N_h, and $\boldsymbol{t}_k = (t_{1k}, t_{2k})^T$ is the vector of class indicator variables t_{hk}.

We generated $R = 40{,}000$ finite populations $\{y_1, ..., y_N\}$, each of size $N = 393$, assuming the logistic regression model (3.8) with $\boldsymbol{\theta} = (\theta_0, \theta_1)^T = (-1,\ 0.005)^T$. The parameter of interest is $\theta_1 = 0.005$. From each generated population, we selected one simple random sample of size $n = 150$, and then obtained the calibration-weighted estimated $\tilde{\theta}_1$ and associated variance estimators $\mathrm{est}(I) = \vartheta_s(\tilde{\theta}_1)$, $\vartheta_{DR}(\tilde{\theta}_1)$, $\vartheta_{cus}(\tilde{\theta}_1)$ and $\vartheta_{mix}(\tilde{\theta}_1)$ from each sample r. We obtained the averages of the estimates and the variance estimates as $\mathrm{av}(\hat{\theta}_1) \approx 0.00514$, $\mathrm{av}(\vartheta_{DR}) \approx 0.0989$, $\mathrm{av}(\vartheta_{cus}) \approx 0.0987$, $\mathrm{av}(\vartheta_{mix}) \approx 0.0988$, and $\mathrm{av}(\vartheta_s) \approx 0.0613$. Also, the estimated total MSE of $\hat{\theta}_1$ is equal to 0.0998. Hence, unconditionally the estimator $\tilde{\theta}_1$ is approximately unbiased for θ_1, and the bias of the three variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix} is negligible. On the other hand ignoring the second component and using only the first component, $\mathrm{est}(I) = \vartheta_s(\tilde{\theta}_1)$, leads to severe underestimation, as expected.

We also examined the conditional performances of the three variance estimators along the line of Section 2.2. We arranged the 40,000 samples in ascending order of the sample size, n_1, in class 1, and then grouped the samples into twenty groups, each of size 2,000, such that the first group, G_1, contained the 2,000 samples with the smallest n_1-values, the second group, G_2, contained the 2,000 samples with the next smallest n_1-values, and so on to get twenty groups, $G_1, ..., G_{20}$.

We calculated the conditional MSE of $\tilde{\theta}_1$ and the associated conditional relative bias (CRB) of the variance estimators ϑ_{DR}, ϑ_{cus} and ϑ_{mix} based on the average values of ϑ_{DR}, ϑ_{cus} and ϑ_{mix} in each group; see Figure 5. We can see from Figure 5 that CRB of ϑ_{cus} ranges from 20% to -20% across the groups, whereas ϑ_{DR} exibits no such trend and its CRB is less than 5% in absulate value except for two groups. Also, the CRB of ϑ_{mix} exhibits a trend but less pronounced than ϑ_{cus}. Figure 6 reports the conditional coverage rates (CCR) of normal theory intervals based on ϑ_{DR}, ϑ_{cus} and ϑ_{mix} for nominal level of 95%. We can see from Figure 6 that ϑ_{cus} exhibits a trend across groups with CCR ranging from 97% to 92%, whereas CCR associated with ϑ_{DR} is close to the nominal level across groups. Further, CCR associated with ϑ_{mix} is slighthy above that of ϑ_{DR} for the first half of the groups and slighty below for the remaing groups.

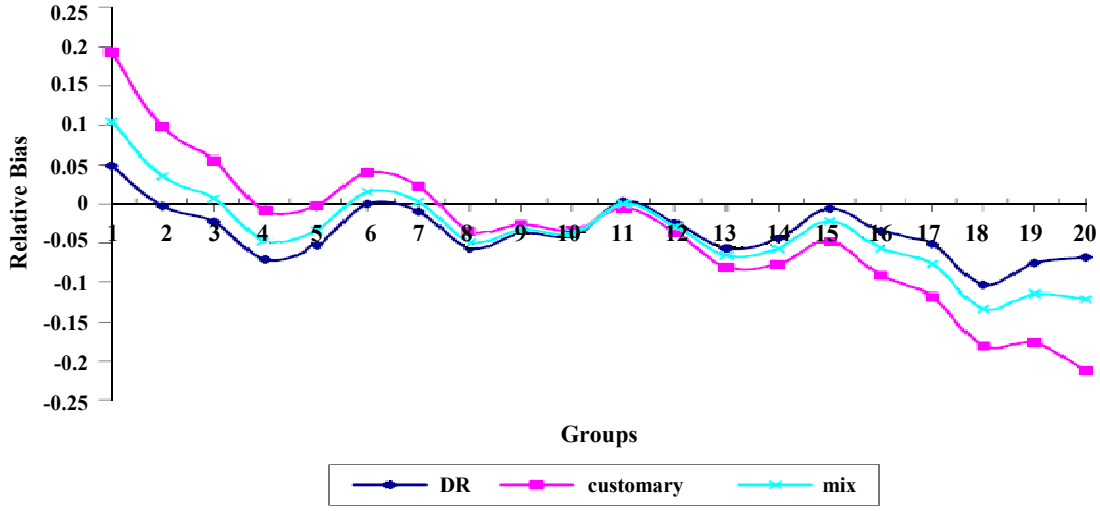


Figure 5 Conditional relative bias of variance estimators: logistic regression

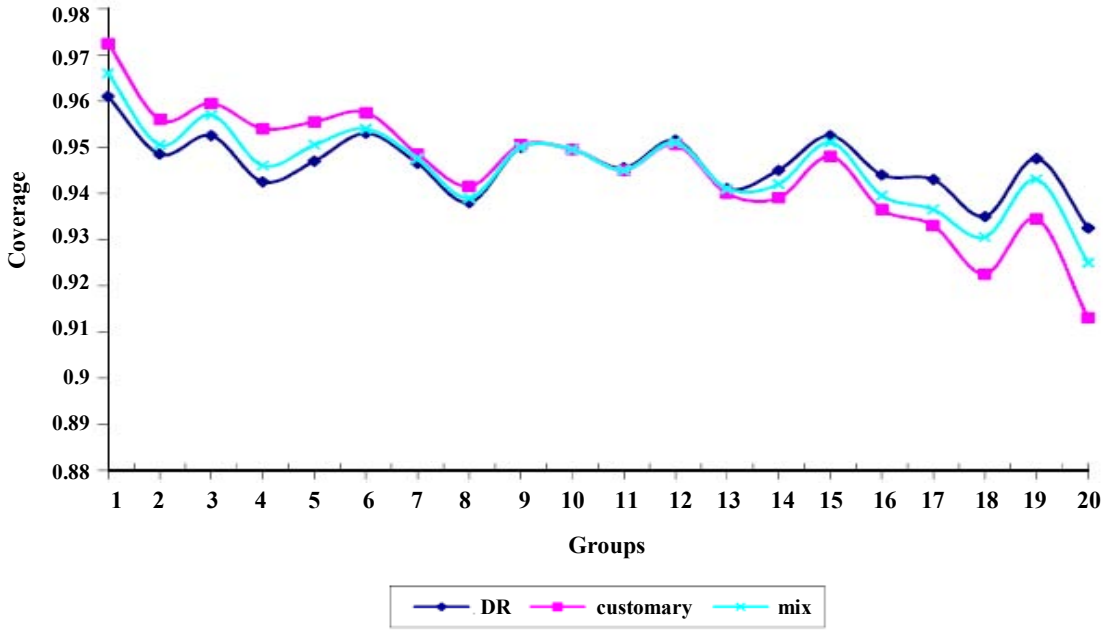


Figure 6 Conditional coverage rates of normal theory confidence intervals for nominal level of 95%: logistic regression

Concluding remarks

We have studied the estimation of total variance of estimators of model parameters under an assumed super-population model. Our approach leads directly to a linearization variance estimator which is shown to perform well under a conditional framework when calibration weights are used for estimation. We are currently investigating extensions of our method to estimation of total variance under imputation for item nonresponse and integration of two independent surveys.

Acknowledgements

We thank two referees for constructive comments and suggestions. J.N.K. Rao's work was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada.

References

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data (with discussion). *Survey Methodology*, 30, 17-34.

Deville, J.-C. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25, 193-203.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*, New York: John Wiley & Sons, Inc.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Ed. Chapman & Hall, London.

Molina, E.A., Smith, T.M.F. and Sugden, R.A. (2001). Modeling overdispersion for complex survey data. *International Statistical Review*, 69, 373-384.

Royall, R.M., and Cumberland, W.G. (1981). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, 76, 66-77.

Rubin-Bleuer, S., and Şchiopu-Kratina, I. (2005). On the two-phase framework for joint model and design-based inference. *Annals of Statistics*, 33, 2789-2810.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527-537.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*, New York: John Willey & Sons, Inc.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite population sampling and inference*: *A prediction approach*, New York: John Wiley & Sons, Inc.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 203-215
Statistics Canada, Catalogue No. 12-001-X

Statistical foundations of cell-phone surveys

Kirk M. Wolter, Phil Smith and Stephen J. Blumberg [1]

Abstract

The size of the cell-phone-only population in the USA has increased rapidly in recent years and, correspondingly, researchers have begun to experiment with sampling and interviewing of cell-phone subscribers. We discuss statistical issues involved in the sampling design and estimation phases of cell-phone studies. This work is presented primarily in the context of a nonoverlapping dual-frame survey in which one frame and sample are employed for the landline population and a second frame and sample are employed for the cell-phone-only population. Additional considerations necessary for overlapping dual-frame surveys (where the cell-phone frame and sample include some of the landline population) are also discussed. We illustrate the methods using the design of the National Immunization Survey (NIS), which monitors the vaccination rates of children age 19-35 months and teens age 13-17 years. The NIS is a nationwide telephone survey, followed by a provider record check, conducted by the Centers for Disease Control and Prevention.

Key Words: Cell-phone study; Random digit dialing; Dual-frame survey; Network sampling; Indirect sampling; Linking rules; Weighting of survey data; National Immunization Survey.

1. Introduction

The number of persons with cell phones in the USA has increased rapidly in recent years, and the percent of adults living in households with cell phones is expected to soon exceed the percent living in households with landlines (CTIA 2008; Blumberg and Luke 2008; Arthur 2007; Ehlen and Ehlen 2007). Correspondingly, survey researchers have begun to experiment with the sampling and interviewing of cell-phone subscribers (Lavrakas, Shuttles, Steeh and Fienberg 2007). This article is about the issues of statistical design and estimation that arise in cell-phone surveys. It emphasizes theoretically rigorous but practical solutions to the emergent problems survey researchers are facing in cell-phone surveys today.

Standard telephone surveys driven by random-digit-dialing (RDD) sampling only cover the population of households that have at least one working landline telephone actually used for voice communications. In an RDD survey, one assumes that the landline telephone is a household appliance and that all persons in the population are attached to one and only one household. Thus, one can sample people indirectly by sampling their telephone numbers and proceed from there to use reasonably standard and well-known methods of estimation.

The cell-phone survey brings a paradigm shift and new challenges. Most people think of the cell phone as a personal appliance, not a household device. Some people do share a cell phone, including 10-20 percent of cell-phone-only adults (Carley-Baxter, Peytchev and Lynberg 2008), but many do not, and thus it cannot be assumed that all residents of a household can be reached through the same cell-phone line. Some residents of a household can be reached through more than one cell-phone line. Some residents can be reached only by a cell-phone line while others can be reached through both cell and landline telephones. Thus, in the cell-phone survey, the household may no longer provide the same unifying organization that it does in standard telephone surveys.

To address the growing risk of bias (due to under-coverage) in telephone surveys, one can consider dual-frame telephone survey designs that include both an RDD sample of landline telephones and a sample of cell-phone lines. The telephone numbers on the two sampling frames are non-overlapping, but the corresponding people and households that may be the objects of the survey are partially overlapping.

A rigorous theory of estimation for such telephone survey designs has been lacking, although some initial descriptions of weighting have been advanced by Brick, Dipko, Presser, Tucker and Yuan (2006), Brick, Edwards and Lee (2007), and Frankel, Battaglia, Link and Mokdad (2007). In this article, we provide a general theory of unbiased estimation for population totals in the context of dual-frame telephone survey designs and derive the corresponding survey weights. We show what information must be collected in the survey itself to enable the calculation of the sampling weights.

To introduce ideas, we let A signify the portion of the overall population of interest accessible through the landline sampling frame, let B denote the portion accessible through the cell-phone sampling frame, and let C denote the portion not accessible through either frame (the *phoneless population* and other relatively small components of the

1. Kirk M. Wolter, NORC and the University of Chicago. E-mail: wolter-kirk@norc.org; Phil Smith, National Center for Immunization and Respiratory Diseases; Stephen J. Blumberg, National Center for Health Statistics.

total population). We let *a* be the subpopulation in *A* not accessible through cell-phone lines (the *landline-only population*), let *b* be the subpopulation in *B* not accessible through landlines (the *cell-phone-only population*), and let *ab* be the subpopulation accessible through both landlines and cell-phone lines (the *mixed population*). We will sharpen this notation in succeeding sections.

Whether or not a unit in the population of interest is accessible through landlines or cell-phone lines is itself a complex matter. Throughout this article, when we say that a unit is accessible through landlines, we shall mean that there is both physical access to one or more landlines (usually residential landlines only) and a respondent would actually answer the landline if it rang for voice communications. Many adults today maintain a landline telephone strictly for computer communications and utilize a cell phone for all voice communications. By our definition, such adults are not considered to have landline access and instead are considered to be in the cell-phone-only population. Similarly, when we say that a unit is accessible through cell-phone lines, we shall mean that there is both physical access to a cell phone and intent to answer the cell phone if it rang. All other units in the population of interest that are not accessible through either landlines or cell-phone lines are considered phoneless. Current evidence suggests, although no one knows for sure, that about 20 to 30 percent of adults are domain *b*, 5 to 10 percent are in domain *C*, and the balance are spread across domains *a* and *ab*.

What we know so far from the cell-phone surveys we and others have conducted is that the data collection is relatively expensive, with average-interviewer-hours-per-completed case running around three times the average for standard RDD surveys. The higher cost is brought, in part, by the legal requirement (in the US, the Telephone Consumer Protection Act) of manually dialing the selected cell-phones. Response rates are somewhat lower than those achieved in RDD surveys. Interview length may be problematic, with some respondents less willing to submit to a lengthy interview by cell phone than by landline phone. Privacy issues may constrain the cell-phone interview, if the respondent is not in a private place at the time of the interview. The cell-phone user's propensity to respond may vary monotonically with his or her level of use of the cell phone, with the heavy user more willing to answer the phone than the lighter or occasional user. Most breakoffs occur during the opening seconds of the interview attempt. Because cell-phone surveys are relatively new, people are not used to being called and the interviewer has mere seconds to sell the survey. On the other hand, we find many cell-phone respondents to be quite cooperative once their attention has been held through the survey's introductory script.

Due to all of these circumstances in the environment, we currently view the cell-phone sample as a relatively small supplementary sample, with the main sample continuing to be a larger RDD sample of landlines. The cell-phone sample is intended to round out the coverage of the population of interest. In the future, as the environment matures and if costs come down, it may be possible to shift towards a more balanced approach with similarly sized landline and cell-phone samples, or even to a state where the cell-phone sample begins to dominate and the landline sample is used as a supplement to round out coverage.

In Section 2, we introduce the topic of *networks* of *sampling units*, *reporting units*, and *estimation units* and show how cell-phone surveys equate to a sampling of networks. Section 3 introduces various key concepts that will be needed as we discuss survey estimation, among them being the idea of a *link* (or edge) between the *nodes* (or vertices) in the network. Section 4 describes the duality that exists between the populations corresponding to the different types of nodes. Our approach will remind some readers of Lavallée's (2007) methods for indirect sampling. The heart of the paper is Section 5, which sets forth unbiased estimators of population totals for cell-phone surveys and for corresponding dual-frame telephone survey designs. Section 6 gives an example, illustrating implications of the new methods of estimation for an existing telephone survey regarding the vaccination coverage of young children and teenagers. We close in Section 7 with a brief summary.

Throughout the article, we emphasize the development of rigorous but practical design and estimation procedures for population *B*. The methods of RDD surveys, *i.e.*, the methods for population *A*, are well known and, to a degree, have been used for decades; for a recent review of these methods see Wolter, Chowdhury and Kelly (2008).

2. Networks of units and the response protocol

In general, at least three types of units arise in the context of a cell-phone survey, as follows:

- Sampling units (SU)
- Reporting units (RU)
- Estimation units (EU).

The SU is the unit of sampling in the survey. In actual practice, telephone numbers may be sampled directly from cell-phone frames, or they may be sampled in stages, with perhaps exchanges or banks of numbers serving as the primary sampling units and numbers themselves being selected in one or more stages of subsampling within the primary units. To keep the discussion simple, in this article we will present the telephone number itself as the SU.

The actual target of the survey interview and the unit of analysis is what we shall call the EU. Some surveys focus on the collection and analysis of data on households or families, in which case the household or family is the EU. Other surveys focus on person level data, where the eligible persons may be children under age 18, adults age 18+, or some demographic segment of the population, such as Hispanic females aged 0-34. Still other surveys focus on both household- and person-level data, in which case the survey involves at least two types of EUs and two levels of analysis.

The adult is the respondent or RU in telephone surveys. The EU may or may not have the capacity to respond directly for itself, and instead an RU responds on its behalf. If the EU is an adult, then the same adult or even a different adult may serve as the corresponding RU. If the EU is a household, family, consumer unit, or child, then one or more adults may serve as the corresponding RU. The response protocol, specified by the survey methodologist, actually determines which RUs are permitted to respond for which EUs. In a typical survey, one respondent adult (or RU) would be contacted by telephone and interviewed for each SU selected into the sample.

SUs, RUs, and EUs may bear different relationships to one another in a cell-phone survey. Figure 1 gives nine networks that illustrate some of the types of relationships that are possible. In the first network, one SU is linked to one RU, which in turn responds for one EU. This arrangement could occur if one adult uses one telephone line, and the adult in turn reports for the household or for him or herself or for one child. In the second network, one SU is linked to two RUs, each of which can respond for the EU. This arrangement would occur, for example, if two adults shared the same telephone line and each was permitted by survey protocol to respond for the household. The fifth network could occur if two adults each had their own telephone line not shared with the other adult, while each adult in the pair is allowed by survey protocol to respond for each of two children.

More complicated networks are possible and surely must exist in the world. For example, the eighth network shows an arrangement of three adults sharing two telephone lines. The first of the lines is shared by all three adults, while the second line is only used by the third adult. The first of the adults is permitted by survey protocol to respond for two EUs, such as the adult's biological children; the second adult is not permitted to respond for any EUs; and the third adult is permitted to respond only for a third EU that is not reportable by the first two adults.

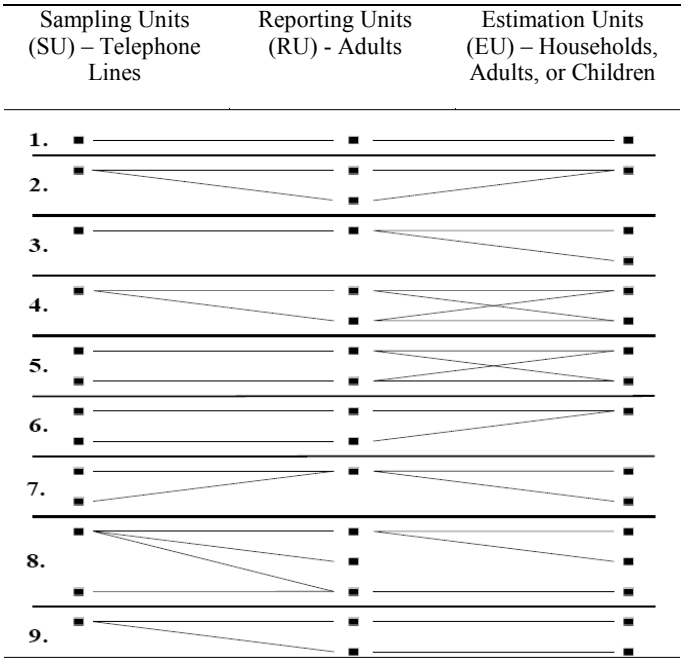


Figure 1 Examples of networks in a cell-phone survey

3. Links between units in the network

A *link* is a salient relationship between two nodes in the network. In the context of Figure 1, the links are represented by the line segments that join the different nodes. To provide a foundation for survey estimation, we need to explore links between (i) RUs and SUs, (ii) EUs and RUs, (iii) and EUs and SUs.

3.1 Link of RU and SU

Two concepts are central to creating a link between an RU and an SU, namely, the concepts of (a) an *Active Personal Cell Number* (*APCN*) and (b) *usual access* to the cell-phone line.

An APCN is a telephone line that is in service at the time of the cell-phone survey and can ring through to an eligible adult who uses the cell phone, at least partially, for personal matters. In other words, an APCN meets three tests:

- It is in service
- It connects to an eligible adult respondent
- It is not used exclusively for business purposes.

We say that a given adult has usual access to a given APCN if and only if the individual has

- Regular,
- Substantial, and
- Ongoing use of the cell-phone line.

Each APCN has one or more regular adult users, and each individual user has usual access to one or more cell phones. In many cases, there is a unique one-to-one relationship between the cell-phone line and the adult user. In some cases, there is a one-to-many relationship between the cell-phone line and its users.

We treat a given SU and a given RU as linked if and only if the SU is an APCN and the RU has usual access to the SU. A cell-phone survey must work with and recognize the links that exist between the population of SUs and the population of RUs.

3.2 Link of EU and RU

A given EU is linked to one or more RUs via natural relationships that exist in the world, such as those created by family or place of residence. For example, an adult respondent may respond to the survey interview on behalf of his or her household, family, or consumer unit. He or she may respond for him or herself, for a dependent child under age 18, or for his or her own parent or sibling.

All surveys require a response protocol that defines which adult respondents are to respond for which EUs. The protocol is selected by the survey methodologist in light of feasibility, cost, and accuracy-of-reporting concerns. It is this protocol that establishes the links between EUs and RUs.

3.3 Link of EU and SU

The foregoing links between RUs and SUs and between EUs and RUs determine the links between EUs and SUs. We say a given EU is linked to a given SU if and only if the EU is linked to at least one RU that in turn is linked to the SU.

Some notation will become useful in our work in the following sections. Let j denote a given EU in the population of interest and let i be a given SU in the population. Then define the indicator or link variables

$$\ell_{ij} = 1, \quad \text{if the } j^{\text{th}} \text{ EU is linked to the } i^{\text{th}} \text{ SU}$$
$$= 0, \quad \text{otherwise.}$$

4. Duality between the populations of SUs and EUs

To begin the process of determining an unbiased estimation procedure for cell-phone surveys, we establish that a duality exists between the population of SUs or cell phones (henceforth denoted by U^{SB}) and the population of EUs that are linked to cell phones (denoted by U^{EB}). The goal of a cell-phone survey is to make inferences concerning U^{EB}, but we will soon see that this goal is equivalent to making certain inferences concerning U^{SB} (in this notation, the first superscript designates the type of unit while the superscript B refers to the cell-phone sampling frame. Later we will use the superscript A to signify the landline sampling frame).

In the EU domain, a population total of interest is given by

$$Y^{\text{EB}} = \sum_{j \in U^{\text{EB}}} Y_j,$$

where the Y-variable on the right-hand side is a questionnaire item or other recoded or derived variable attached to the units in the population U^{EB}. Similarly, in the SU domain, a population total is defined by

$$X^{\text{SB}} = \sum_{i \in U^{\text{SB}}} X_i,$$

where the X-variable on the right-hand side is any fixed characteristic attached to the units in the population U^{SB}.

While the interest of the survey analyst centers on the total from the population of EUs (and on other parameters of this population), one can obtain a corresponding parameter in the SU domain by writing

$$Y^{\text{EB}} = \sum_{j \in U^{\text{EB}}} Y_j = \sum_{j \in U^{\text{EB}}} \sum_{i \in U^{\text{SB}}} \frac{Y_j \ell_{ij}}{\sum_{i' \in U^{\text{SB}}} \ell_{i'j}} = \sum_{i \in U^{\text{SB}}} X_i = X^{\text{SB}}, \quad (1)$$

where the X-variable is now defined specifically by

$$X_i = \sum_{j \in U^{\text{EB}}} \frac{Y_j \ell_{ij}}{\sum_{i' \in U^{\text{SB}}} \ell_{i'j}}. \quad (2)$$

From (1), one can see the correspondence between estimation in the SU domain and estimation in the EU domain. The total X^{SB}, with X_i defined as in (2), is equivalent to the total of interest Y^{EB}, and thus the problem of estimation of Y^{EB} is equivalent to the problem of estimation of X^{SB}.

We note that (2) arises in substantially the same form in the theory of indirect sampling. See Lavallée (2007), Theorem 4.1. In indirect sampling, SUs are linked to naturally defined clusters of EUs; if a given SU is selected into the sample, the survey data are collected for all EUs in the linked clusters. The analogy here is that the clusters are defined by the RUs that respond to the cell-phone interview attempt, and survey data are collected from the respondent for all EUs to which he or she is linked. The current situation is such that the cluster is defined by the SU-RU pair. An identifiability problem arises in this regard that does not occur in general in indirect sampling, and we elaborate on this matter in Section 5.5.

In (2), we effectively allocate an equal share of Y_j to each SU i to which it is linked. We could, alternatively, achieve the same ends by allocating Y_j to its linked SUs in proportion to some other known measure of the intensity of

the relationship between j and i. Although one could conceive of an optimal allocation of Y_j to its linked SUs, as in Deville and Lavallée (2006), such an allocation may be difficult to execute or may not be of great import in large scale practical settings.

5. Estimation

As mentioned in the introduction, some EUs will be linked exclusively to cell phones, some will be linked exclusively to landlines, and some will be linked to both landlines and cell phones. Phoneless EUs, if any, will not be linked to cell phones or to landlines. To provide notation for this environment, let U^{E} be the overall population of EUs of interest, and let U^{S} be the overall population of SUs. Let U^{EA} be the elements of U^{E} that are linked to landlines, let U^{EB} be the elements that are linked to cell-phone lines, let U^{Ea} be the elements that are linked only to landlines, let U^{Eb} be the elements that are linked only to cell-phone lines, let U^{Eab} be the elements that are linked to both landlines and cell-phone lines, and let U^{EC} be the elements that are phoneless. Note that $U^{\mathrm{E}} = U^{\mathrm{EA}} \cup U^{\mathrm{EB}} \cup U^{\mathrm{EC}}$, $U^{\mathrm{EA}} = U^{\mathrm{Ea}} \cup U^{\mathrm{Eab}}$, and $U^{\mathrm{EB}} = U^{\mathrm{Eab}} \cup U^{\mathrm{Eb}}$, where U^{Ea}, U^{Eab}, and U^{Eb} are disjoint sets. Also, let U^{SA} be the population of landlines, such that $U^{\mathrm{S}} = U^{\mathrm{SA}} \cup U^{\mathrm{SB}}$. Landlines and cell-phone lines reflect disjoint subsets of the overall population of SUs.

In the following Sections 5.1 and 5.2, we discuss unbiased estimation for the subpopulation, say $U^{\mathrm{ET}} = U^{\mathrm{EA}} \cup U^{\mathrm{EB}}$, that is linked to at least one telephone of any kind. We use the super-script T to designate this telephone subpopulation. Subsequently, in Section 5.4, we briefly discuss coverage of the phoneless population.

For EUs in U^{E}, define the indicator variables

δ_j = 1, if none of the RUs linked to j have access to landline service, while at least one of these RUs has usual access to cell-phone service

= 0, otherwise

ϕ_j = 1, if none of the RUs linked to j have usual access to cell-telephone service, while at least one of these RUs has access to landline service

= 0, otherwise.

The δ-variable is an indicator of cell-phone-only status and the ϕ-variable is an indicator of landline-only status.

Then the population total of interest may be decomposed as

$$Y^{\mathrm{ET}} = Y^{\mathrm{EA}} + Y^{\mathrm{Eb}}, \tag{3}$$

where

$$Y^{\mathrm{Eb}} = \sum_{j \in U^{\mathrm{ET}}} \delta_j Y_j$$

is the total of the cell-phone-only domain, and

$$Y^{\mathrm{EA}} = \sum_{j \in U^{\mathrm{ET}}} (1 - \delta_j)\, Y_j$$

is the total of the complement of this domain, including EUs that are linked exclusively to landlines and mixed EUs that are linked to both landlines and cell phones. The total of EUs may also be written as

$$Y^{\mathrm{ET}} = Y^{\mathrm{Ea}} + Y^{\mathrm{Eab}} + Y^{\mathrm{Eb}}, \tag{4}$$

where

$$Y^{\mathrm{Ea}} = \sum_{j \in U^{\mathrm{ET}}} \phi_j Y_j$$

is the total of the landline-only population, and

$$Y^{\mathrm{Eab}} = \sum_{j \in U^{\mathrm{ET}}} (1 - \delta_j)\,(1 - \phi_j)\, Y_j$$

is the total of the mixed population that has a combination of landline and cell-phone access. Finally, the population total may be written as

$$Y^{\mathrm{ET}} = Y^{\mathrm{Ea}} + Y^{\mathrm{EB}}, \tag{5}$$

where

$$Y^{\mathrm{EB}} = \sum_{j \in U^{\mathrm{ET}}} (1 - \phi_j) Y_j$$

is the total of the complement (in the telephone population) of the landline-only population.

We view (3) and, to some extent, (4) as the decompositions of current practical interest and importance in telephone surveys in the USA and, in what follows, we present methods of estimation for each. Because of the current high relative cost of cell-phone interviews, surveys based on decomposition (5) would not be cost effective. It would almost always be better to represent the domain U^{Eab} using a sample of landlines than using a sample of cell phones. If the relative cost of cell-phone interviewing shifts downward in the future, decomposition (5) could become economically viable. It may also be viable for surveys in other countries where the cost structure is more favorable to cell-phone interviews.

5.1 Case of nonoverlapping domains

In this section, we will use a sample of cell-phone lines for purposes of estimation for the cell-phone-only population U^{Eb} and a sample of landlines for estimation for the entire landline population U^{EA}. We observe that it is not

possible to directly select a sample of cell-phone-only lines, because cell-phone-only status is not available on the sampling frame but rather is determined in the survey screening interview. To operationalize this design, one would screen-out cell-phone respondents who classify themselves in the mixed domain and terminate the interview, continuing the interview only for cell-phone-only respondents.

Let s^{SB} denote a probability sample of SUs (cell-phone lines) selected from the population U^{SB}, and let $\{W_i^{SB}\}$ denote the set of base sampling weights such that

$$\hat{X}^{SB} = \sum_{i \in s^{SB}} W_i^{SB} X_i$$

is an unbiased estimator of the population total X^{SB}, where X_i is a characteristic of the i^{th} unit in the population. Assuming simple random sampling without replacement within strata, the base weights are of the form

$$W_i^{SB} = N_h / n_h, \tag{6}$$

where h signifies the sampling stratum in which the i^{th} SU is selected, N_h is the number of SUs on the sampling frame in stratum h, and n_h is the sample size in stratum h. Typically, the cell-phone sampling frame would include all telephone numbers within the exchanges assigned by the telephone system to cell phones. Simple random sampling would be the most common method of sample selection from such exchanges. There is little information available on the cell-phone sampling frame to enable stratification of the sample, except for the coarse geographic information embodied within the area code.

Let s^{EB} be the corresponding sample of EUs, *i.e.*, $s^{EB} = \{j \in U^{EB} \mid j$ is linked to at least one SU i in $s^{SB}\}$. We will use this sample to estimate the domain total of EUs that are linked only to a cell phone, Y^{Eb}. From (1) and (2), we can readily see that the unbiased estimator of the domain total is given by

$$\hat{Y}^{Eb} = \sum_{i \in s^{SB}} W_i^{SB} \left\{ \sum_{j \in U^{EB}} \delta_j Y_j \ell_{ij} \Big/ \sum_{i' \in U^{SB}} \ell_{i'j} \right\}$$

$$= \sum_{j \in s^{EB}} \delta_j Y_j W_j^{EB}, \tag{7}$$

where the EU level sampling weights are defined by

$$W_j^{EB} = \sum_{i \in s^{SB}} W_i^{SB} \ell_{ij} \Big/ \sum_{i' \in U^{SB}} \ell_{i'j} \quad \text{for } j \in s^{EB}. \tag{8}$$

Again, see Lavallée (2007) for expression of these weights in the context of indirect sampling.

Before leaving domain *b*, we observe in passing that it is possible to subsample the EUs and collect the survey information only for the subsample instead of enumerating all EUs linked to the sample RUs. If the statistician would choose some form of subsampling, perhaps to control sample size or cost, then an additional weighting factor would appear in the weights in (8). Such subsampling is referred to as two-stage indirect sampling in Lavallée (2007, Section 5.1).

Turning to domain *A*, let s^{SA} denote a standard RDD sample of landline telephones, let s^{EA} be the implied sample of EUs, *i.e.*, $s^{EA} = \{j \in U^{EA} \mid j$ is linked to at least one SU i in $s^{SA}\}$, and let

$$\hat{Y}^{EA} = \sum_{j \in s^{EA}} W_j^{EA} Y_j \tag{9}$$

be the standard unbiased estimator of the population total. For brevity, we shall not derive the standard sampling weights here; for more information about these weights, see Wolter *et al.* (2008).

From (7) and (9), the unbiased estimator of the population total of the EUs is given by

$$\hat{Y}^{ET} = \hat{Y}^{EA} + \hat{Y}^{Eb} \tag{10}$$

and the weights needed to support this estimator are $\{W_j^{EA}\}$ and $\{W_j^{EB}\}$.

5.2 Case of overlapping domains

We now proceed with estimation starting from the decomposition (4). This means that in the cell-phone sample we will interview not only the cell-phone-only population, but also the mixed population (*i.e.*, those that use both landline and cell telephones). The estimator of the population total of interest is now of the form

$$\hat{Y}^{ET} = \hat{Y}^{Ea} + \hat{Y}^{Eab} + \hat{Y}^{Eb}, \tag{11}$$

where

$$\hat{Y}^{Ea} = \sum_{j \in s^{EA}} W_j^{EA} \phi_j Y_j$$

is the estimator for the landline-only domain derived from the landline sample, $\hat{Y}^{Eb}$ is defined in (7) and is the estimator for the cell-phone-only domain derived from the cell-phone sample, and $\hat{Y}^{Eab}$ is an estimator of the mixed domain obtained from both samples. The estimator of the mixed domain is

$$\hat{Y}^{Eab} = \lambda \sum_{j \in s^{EA}} W_j^{EA} (1 - \phi_j) Y_j$$

$$+ (1 - \lambda) \sum_{j \in s^{EB}} W_j^{EB} (1 - \delta_j) Y_j. \tag{12}$$

The weights need to support estimator (11) are $\{W_j^{EA}\}$ and $\{W_j^{EB}\}$.

See Hartley (1962) for discussion of the mixing parameter λ in a dual-frame survey, focusing on considerations

of sampling variability. Turning to considerations of bias, Brick *et al.* (2006) report that the propensity to respond to a cell-phone survey may be positively related to the frequency of use of the cell phone. Thus, the two pieces on the right side of (12) may be subject to a differential nonresponse bias not removed by the standard weighting-class methods. In the mixed population, infrequent users of the cell phone may be less likely to respond if surveyed in the cell-phone sample than if surveyed in the landline sample. If these adults would be substantially different from other adults in the mixed population with respect to the key characteristics under study in the survey, then (12) and also (11) could be subject to a nonreponse bias.

5.3 Variance estimation

To make inferences from the sample to the overall population, we require an estimator of the variance of the estimated total. First, consider the case of nonoverlapping domains. By working in the SU population, we can employ methods of variance estimation appropriate to the survey design. From (7), the estimated total for the cell-phone only domain may be written by

$$\hat{Y}^{\mathrm{Eb}} = \sum_{i \in s^{\mathrm{SB}}} W_i^{\mathrm{SB}} X_i,$$

where

$$X_i = \sum_{j \in U^{\mathrm{EB}}} \delta_j Y_j \ell_{ij} \Big/ \sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j}. \qquad (13)$$

Assuming simple random sampling, the unbiased estimator of the variance of the estimated total is given by

$$v(\hat{Y}^{\mathrm{Eb}}) = \sum_{h=1}^{L} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{1}{n_h} s_{xh}^2,$$

where

$$s_{xh}^2 = \frac{1}{n_h - 1} \sum_{i \in s_h^{\mathrm{SB}}} \left(X_i - \frac{1}{n_h} \sum_{i' \in s_h^{\mathrm{SB}}} X_{i'} \right)^2.$$

If we would ignore the finite population correction factor, which would be possible in almost any real telephone survey, the variance estimator becomes

$$v(\hat{Y}^{\mathrm{Eb}}) = \sum_{h=1}^{L} \frac{n_h}{n_h - 1} \sum_{i \in s_h^{\mathrm{SB}}} \left(W_i^{\mathrm{SB}} X_i - \frac{1}{n_h} \sum_{i' \in s_h^{\mathrm{SB}}} W_{i'}^{\mathrm{SB}} X_{i'} \right)^2. \qquad (14)$$

Now let $v(\hat{Y}^{\mathrm{EA}})$ be an estimator of the variance of $\hat{Y}^{\mathrm{EA}}$ for the RDD sample of landlines. Such estimators are well known and we do not review them here; see for example, Wolter *et al.* (2008). Because sampling is independent in the landline and cell-phone sampling frames, the unbiased estimator of the variance of the estimated total for the entire telephone population becomes

$$v(\hat{Y}^{\mathrm{ET}}) = v(\hat{Y}^{\mathrm{EA}}) + v(\hat{Y}^{\mathrm{Eb}}). \qquad (15)$$

To facilitate the following developments, we let $\hat{V}^{\mathrm{EB}}[\delta Y]$ be another symbol to represent the estimator of variance in (14). This notation will emphasize the fact that the estimator of variance is based on the X_i variable in (13) defined in terms of the characteristic $\delta_j Y_j$, which is the characteristic of interest for cell-phone-only EUs. Also, let the symbol $\hat{V}^{\mathrm{EA}}[Y]$ be the estimator $v(\hat{Y}^{\mathrm{EA}})$ defined in terms of the characteristic Y_j. With this notation, (15) becomes $v(\hat{Y}^{\mathrm{ET}}) = \hat{V}^{\mathrm{EA}}[Y] + \hat{V}^{\mathrm{EB}}[\delta Y]$.

Second, consider variance estimation for the case of overlapping domains. The estimator of the total of the telephone population is now $\hat{Y}^{\mathrm{ET}}$ in (11). For fixed λ, the unbiased estimator of variance is clearly seen from the work done in (14) and (15). It is

$$v(\hat{Y}^{\mathrm{ET}}) = \hat{V}^{\mathrm{EA}}[\phi Y + \lambda(1-\phi)Y] + \hat{V}^{\mathrm{EB}}[\delta Y + (1-\lambda)(1-\delta)Y]. \qquad (16)$$

The first term on the right side of (16) is the variance estimator for the RDD sample of landlines applied to the composite characteristic $\phi_j Y_j + \lambda(1-\phi_j)Y_j$, which is the characteristic for landline-only EUs plus a λ-portion of the characteristic for mixed EUs. The second term on the right side of (16) is the variance estimator for the cell-phone sample applied to the composite characteristic $\delta_j Y_j + (1-\lambda)(1-\delta_j)Y_j$, which is the characteristic for cell-phone-only EUs plus a $(1-\lambda)$-portion of the characteristic for mixed EUs.

Estimators of covariance matrices can be built up from expressions like (15) and (16), facilitating statistical inference concerning other population parameters of interest.

5.4 Adjustments of the sampling weights

The sampling weights may be adjusted because of nonresponse or a planned calibration to known control totals.

Thus far, we have not addressed the various types of missing data that may occur in a cell-phone survey. We will focus on deriving adjustments for missing data that arise during the cell-phone interviews, assuming that standard adjustments for missingness in the landline sample have already been incorporated in the $\{W_j^{\mathrm{EA}}\}$ weights.

Missing data can arise due to three factors: (i) *nonresolution* of the SU; (ii) an *incomplete screening interview* of the RU; and (iii) an *incomplete main interview* of the RU. In this article, we adopt the convention that the resolution step refers to the classification of the SU as an ACPN or something else, such as a disconnected line or a dedicated business line; nonresolved SUs and SUs resolved as

non-ACPNs do not continue with the interview. The screening step refers to a brief preliminary interview intended to ascertain telephone status and to determine any demographic or other eligibility characteristics of any EUs linked to the RU; RUs for which the screening interview is incomplete or for which the screening interview is complete but no eligible EUs are linked to the RU do not continue with the interview. If the survey protocol calls for including only cell-phone-only EUs, as in Section 5.1, then the interview would terminate at this point for any mixed EUs. On the other hand, if the survey protocol calls for including both cell-phone-only and mixed EUs, as in Section 5.2, then the interview would continue for all such EUs. The interview step refers to the collection of the main survey items that form the substance of the survey for each of the eligible EUs linked to the RU. The survey methodologist must institute a definition of what constitutes a completed interview. In particular, the methodologist must decide whether *breakoffs* (an interview attempt that is completed for some but not all of the eligible EUs linked to the RU) are to be treated as a completed interview or not. Some other authors may organize the steps in the survey response process somewhat differently than the convention adopted here.

Adjustments to the sampling weights can be made for nonresolution and screener nonresponse, assuming a missing-at-random model for the response mechanism. These two adjustments must be made at the SU level. Let $\{s_\alpha^{\mathrm{SB}}\}$ be a partition of the cell-phone sample into user-specified weighting cells α, and let the base sampling weights from (6) now be denoted by W_{1i}^{SB}, where the subscript 1 has been added simply to signify the first step in a multi-step adjustment process. Telephone area codes, rate centers, and census environmental variables at the county or area code level can be used to form the weighting cells; otherwise, little covariate information is available concerning cell-phone numbers. The cell-specific resolution completion rates are defined by

$$R_{1\alpha} = \frac{\sum_{i' \in s_\alpha^{\mathrm{SB}}} r_{1i'} W_{1i'}^{\mathrm{SB}}}{\sum_{i' \in s_\alpha^{\mathrm{SB}}} W_{1i'}^{\mathrm{SB}}},$$

where r_{1i} is a resolution indicator variable (= 1, if resolved, = 0, if not resolved), and the nonresolution adjusted weights are $W_{2i}^{\mathrm{SB}} = r_{1i} W_{1i}^{\mathrm{SB}} / R_{1\alpha}$ for $i \in s_\alpha^{\mathrm{SB}}$.

Let e_{1i} be an indicator of whether i is a resolved APCN (= 1, if resolved APCN, = 0, otherwise), and let $\{s_\beta^{\mathrm{SB}}\}_{\beta=1}^{B}$ be a partition of the cell-phone sample into user-specified weighting cells, which could be the same as or different than the foregoing partition. Then, the cell-specific screener completion rates are

$$R_{2\beta} = \frac{\sum_{i' \in s_\beta^{\mathrm{SB}}} r_{2i'} e_{1i'} W_{2i'}^{\mathrm{SB}}}{\sum_{i' \in s_\beta^{\mathrm{SB}}} e_{1i'} W_{2i'}^{\mathrm{SB}}},$$

where r_{2i} is a screener indicator variable (= 1, if screener completed, = 0, if screener not completed), and the screener-nonresponse adjusted weights are $W_{3i}^{\mathrm{SB}} = r_{2i} e_{1i} W_{2i}^{\mathrm{SB}} / R_{2\beta}$ for $i \in s_\beta^{\mathrm{SB}}$. Note that the appropriate sum of the weights is preserved at each step of the adjustment process.

Next, an adjustment to the sampling weights must be made for interview nonresponse. Depending on how breakoffs are classified by the survey methodologist, there may be two cases to consider: (i) the RU completes or fails to complete the interview for all of its linked and eligible EUs en masse, or (ii) the RU selectively completes or fails to complete the interview on an EU by EU basis. If breakoffs would be classified as incomplete interviews, then only Case i would apply. Let e_{2i} be an indicator of whether the RU is screened and is linked to at least one EU that is eligible for the interview (= 1, if screened and eligible, = 0, otherwise), and let r_{3i} be the interview indicator variable (= 1, if the interview is complete, = 0, otherwise).

For Case i, the weight adjustment can be made at the SU level and is given by $W_{4i}^{\mathrm{SB}} = r_{3i} e_{2i} W_{3i}^{\mathrm{SB}} / R_{3\gamma}$ for $i \in s_\gamma^{\mathrm{SB}}$, where $R_{3\gamma}$ is the weighted interview completion rate computed within user-specified weighting cells γ. Again, options for constructing weighting cells are limited in a cell-phone survey; they may be specified in terms of the information available at the previous weighting steps or any information collected in the screening interview. The weighted interview completion rate is

$$R_{3\gamma} = \frac{\sum_{i' \in s_\gamma^{\mathrm{SB}}} r_{3i'} e_{2i'} W_{3i'}^{\mathrm{SB}}}{\sum_{i' \in s_\gamma^{\mathrm{SB}}} e_{2i'} W_{3i'}^{\mathrm{SB}}}.$$

The estimated total for the cell-phone-only domain may now be expressed by

$$\hat{Y}^{\mathrm{Eb}} = \sum_{j \in s^{\mathrm{EB}}} \delta_j Y_j W_{4j}^{\mathrm{EB}}, \qquad (17)$$

where

$$W_{4j}^{\mathrm{EB}} = \sum_{i \in s^{\mathrm{SB}}} W_{4i}^{\mathrm{SB}} \ell_{ij} \Big/ \sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j}$$

and s^{EB} is the set of eligible EUs reported in the screening interviews. The weight is zero for any eligible EUs in s^{EB} for which the RU failed to complete the main interview. The estimated total for the mixed domain, if called for by the survey protocol, is defined similarly by

$$\hat{Y}^{\mathrm{Eab}} = \lambda \sum_{j \in s^{\mathrm{EA}}} W_j^{\mathrm{EA}} (1 - \varphi_j) Y_j + (1 - \lambda) \sum_{j \in s^{\mathrm{EB}}} W_{4j}^{\mathrm{EB}} (1 - \delta_j) Y_j.$$

For Case ii, the noninterview adjustment must be made at the EU level. The EUs are treated as spawned cases and a decision is made for each one as to whether it has a completed interview or not. The estimated total for the cell-phone-only domain is (17), where the weight is now defined by

$$W_{4j}^{\mathrm{EB}} = r_{3j}\, e_{2j}\, W_{3j}^{\mathrm{EB}} / R_{3\gamma} \quad \text{for} \quad j \in s^{\mathrm{EB}},$$

$$W_{3j}^{\mathrm{EB}} = \sum_{i \in s_3^{\mathrm{SB}}} W_{3i}^{\mathrm{SB}} \ell_{ij} / \sum_{i' \in U^{\mathrm{SB}}} \ell_{i'j},$$

and

$$R_{3\gamma} = \frac{\sum_{j' \in s_\gamma^{\mathrm{EB}}} r_{3j'} W_{3j'}^{\mathrm{EB}}}{\sum_{j' \in s_\gamma^{\mathrm{EB}}} W_{3j'}^{\mathrm{EB}}}.$$

Here, the weighting cells, γ, are defined in terms of characteristics of the EUs as determined from the screening interview and other sources.

For either Case i or ii, to facilitate computations, take W_{4j}^{EA} to be defined and equal to zero for EUs in the cell-phone sample, and take W_{4j}^{EB} to be equal to zero for EUs in the landline sample. If the survey protocol is as in Section 5.1, then we conclude that the survey weights for estimating the population total of interest are defined by

$$W_j = W_{4j}^{\mathrm{EA}} + W_{4j}^{\mathrm{EB}}\, \delta_j \tag{18}$$

for $j \in s^{\mathrm{ET}}$, where $s^{\mathrm{ET}} \in s^{\mathrm{EA}} \cup s^{\mathrm{EB}}$. Otherwise, if the survey protocol is as in Section 5.2, then we conclude that the survey weights are defined by

$$W_j = W_{4j}^{\mathrm{EA}} \{\phi_j + \lambda (1 - \phi_j)\}$$

$$+ W_{4j}^{\mathrm{EB}} \{\delta_j + (1 - \lambda)(1 - \delta_j)\} \tag{19}$$

for $j \in s^{\mathrm{ET}}$.

The nonresponse-adjusted weights from (18) or (19) may be calibrated (Deville and Särndal 1992) to external control totals within socio-economic or geographic cells for the population of EUs, using poststratification, raking, or GREG (generalized regression estimation) techniques. If accurate sources are available, control totals may be established and calibration may be conducted separately for domains *A* and *b* or for domains *a*, *ab*, and *b*. If control totals are not available by telephone status, then calibration must use control totals for the entire population regardless of telephone status.

To illustrate these ideas, we briefly examine the GREG estimator. Let us suppose that we have available a $1 \times p$ auxiliary variable $\mathbf{Z}_j$ for the observed, eligible EUs for which the control totals $\mathbf{Z}^{\mathrm{ET}} = \sum_{j \in U^{\mathrm{ET}}} \mathbf{Z}_j$ are known. For example, the *z*-variable may arise from a fully saturated model in terms of explanatory variables age, race, and sex. Let s_4^{ET} be the set of EUs with a completed main interview and let $n_4^{\mathrm{ET}} = \#(s_4^{\mathrm{ET}})$ be the number of eligible EUs reported in the completed interviews obtained within the consolidated telephone sample. Stack the *y*-values, *z*-values, and weights into the matrices $\mathbf{Y} = (Y_1, ..., Y_{n_4^{\mathrm{ET}}})', \mathbf{Z} = (\mathbf{Z}_1', ..., \mathbf{Z}_{n_4^{\mathrm{ET}}}')'$, and $\mathbf{W} = \mathrm{diag}(W_1, ..., W_{n_4^{\mathrm{ET}}})'$. Then the GREG estimator (Cassel, Särndal, and Wretman 1976) of the total of the telephone population of interest takes the familiar form

$$\tilde{Y}^{\mathrm{ET}} = \hat{Y}^{\mathrm{ET}} + (\mathbf{Z}^{\mathrm{ET}} - \hat{\mathbf{Z}}^{\mathrm{ET}})\, \hat{\beta} = \sum_{j \in s_4^{\mathrm{ET}}} W_j\, g_j\, Y_j,$$

where the estimated coefficients are given by $\hat{\beta} = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{W}\mathbf{Y}$, $\hat{Y}^{\mathrm{ET}} = \sum_{j \in s_4^{\mathrm{ET}}} W_j\, Y_j$, $\hat{\mathbf{Z}}^{\mathrm{ET}} = \sum_{j \in s_4^{\mathrm{ET}}} W_j\, \mathbf{Z}_j$, and $g_j = 1 + (\mathbf{Z}^{\mathrm{ET}} - \hat{\mathbf{Z}}^{\mathrm{ET}})\mathbf{Z}_j'$. Lavallée (2007, Chapter 7) derives the Taylor series estimator of the variance of the GREG estimator in an indirect sampling context. Also see Wolter (2007, Chapter 6) for estimation of the variance of the GREG estimator.

Before leaving the topic of calibration, we note that we have largely left aside the small phoneless population, which fundamentally is impossible to sample in a telephone survey. Yet, in all likelihood, the overall population total $Y^{\mathrm{E}} = Y^{\mathrm{ET}} + Y^{\mathrm{EC}}$ will be the parameter of interest, not the total of the telephone population Y^{ET}, and the known control totals used in calibration may be totals for the overall population $\mathbf{Z}^{\mathrm{E}} = \mathbf{Z}^{\mathrm{ET}} + \mathbf{Z}^{\mathrm{EC}}$, not totals for the telephone population $\mathbf{Z}^{\mathrm{ET}}$. To include the phoneless population, we may consider use of a revised GREG estimator with $g_j = 1 + (\mathbf{Z}^{\mathrm{E}} - \hat{\mathbf{Z}}^{\mathrm{ET}})\mathbf{Z}_j'$. This revision takes the same model for the phoneless population as for the telephone population. See Keeter (1995) and Chowdhury, Montgomery and Smith (2008) for other considerations in the calibration of weights for the phoneless population.

5.5 Identifiability assumptions

The foregoing theory assumes fundamentally that if SU *i* is selected into the sample of cell-phone lines, then X_i defined in (2) is observable in the cell-phone interview. Yet the 9th network (and also the 8th) in Figure 1 illustrates a potential problem for the theory. For this network, two RUs are linked to one SU, and in turn each RU is linked to only one EU. To continue this illustration, we suppose that these two EUs are not linked to any other RUs in the population. At the time of the survey interview, only one of the RUs will typically be reached and interviewed (unless the survey protocol would specifically mandate that an interview be

attempted with each RU linked to the selected SU). The respondent RU will report for its linked EU, but by the very nature of this network, the respondent cannot report for the EU that is linked to the companion RU who shares the sample cell-phone line. Thus, there is at least one EU that is linked to the SU that cannot be observed, *i.e.*, data cannot be collected in the cell-phone interview. Thus, we say X_i is *not identifiable*. The situation regarding the reportability of the two EUs would be reversed if the cell-phone interview attempt would have rung through to the companion RU.

To maintain the unbiasedness of the estimator of the population total, the X_i must be identifiable for every respondent SU selected into the sample of cell-phone lines. We need to make one of two assumptions. First, we could assume the problem away by acting as if networks like numbers 8 and 9 either do not exist or are trivial in number.

Secondly, the more realistic case would be to assume an extra randomization step, namely, that the interview call attempt to the given SU has reached a randomly selected RU linked to the SU. This randomization could be viewed as conceptual (that is, occurring naturally and not directed by the survey methodologist). To be formal and rigorous, one would need to collect information on the number of RUs linked to the SU and the probability that the cell-phone call attempt would ring through to the respondent RU. The probability would be approximated by the respondent's self-report of his or her share of use of the cell phone. If only one RU is linked to the SU, then this probability is 1.0 and clearly this simple value would not need to be collected in the interview once it is reported that there is only one RU. If two or more RUs are linked to the SU, then the probability or share to be collected is denoted by τ_{ik} for RUs indexed by *k*, where $\sum_{k\in U_i^{\mathrm{RB}}} \tau_{ik} = 1$ and U_i^{RB} is the set of RUs that are linked to the i^{th} SU. With this additional information in hand, an unbiased estimator of

$$X_i = \sum_{j\in U^{\mathrm{EB}}} \frac{\delta_j Y_j \ell_{ij}}{\sum_{i'\in U^{\mathrm{SB}}} \ell_{i'j}}$$

is given by

$$\hat{X}_i = \sum_{j\in U^{\mathrm{EB}}} \frac{1}{\sum_{i'\in U^{\mathrm{SB}}} \ell_{i'j}} \sum_{k\in U_i^{\mathrm{RB}}} \alpha_{ik} \frac{\delta_j Y_j \ell_{ij} \ell_{ikj}}{\tau_{ik} \sum_{k'\in U_i^{\mathrm{RB}}} \ell_{ik'j}}, \quad (20)$$

where α_{ik} is an indicator variable signifying whether the k^{th} RU was the realized respondent or not for the i^{th} SU in s^{SB} and

$$\begin{aligned} \ell_{ikj} &= 1, \quad \text{if SU } i \text{ is linked to RU } k \text{ which in turn is linked to EU } j \\ &= 0, \quad \text{otherwise.} \end{aligned}$$

The data are now identified and one can plug (20) into (7), giving the revised estimator

$$\hat{Y}^{\mathrm{Eb}} = \sum_{j\in s^{\mathrm{EB}}} \delta_j Y_j W_{0j}^{\mathrm{EB}} \quad (21)$$

with revised weights

$$W_{0j}^{\mathrm{EB}} = \sum_{i\in s^{\mathrm{SB}}} W_i^{\mathrm{SB}} \frac{1}{\sum_{i'\in U^{\mathrm{SB}}} \ell_{i'j}} \sum_{k\in U_i^{\mathrm{RB}}} \alpha_{ik} \frac{\ell_{ij}\ell_{ikj}}{\tau_{ik} \sum_{k'\in U_i^{\mathrm{RB}}} \ell_{ik'j}}. \quad (22)$$

As an approximation, one could take the RUs to be equal users of the cell phone, in which case τ_{ik} would simply be the reciprocal of the number of RUs linked to the SU *i* for all RUs *k*. Adjustments for nonresponse and calibration to control totals would proceed as before.

Alternatively, the survey methodologist could call for a real randomization step, which would require that the interviewer make a roster of the RUs linked to the SU and select one at random, or a pseudo randomization step using the last birthday method. Such methods are probably not feasible at this time, due to the difficulty of gaining cooperation in cell-phone interviews.

5.6 Implications for data collection

Certain information must be collected in the survey interview in order to support the calculation of the estimators discussed here.

To support the use of δ_j, the cell-phone survey must collect information to establish whether any of the RUs linked to the EU have access to a landline telephone. The respondent RU must report this information both for himself or herself and for other RUs that may be linked to the EU.

To support the use of ϕ_j, the landline survey must collect information to establish whether any of the RUs linked to the EU have regular access to a cell phone. The respondent RU must report this information both for himself or herself and for other RUs that may be linked to the EU. This report may be quite straightforward in the event that the response protocol only links EUs to RUs within the same household. For more complicated response protocols, the report could be difficult to obtain.

To support the use of $\sum_{i'\in U^{\mathrm{SB}}} \ell_{i'j}$ in calculating the survey weights, the survey must collect information to establish how many SUs in the population are linked to the reported EU *j*. The respondent RU must be able to report the number of cell phones, including their own, that ring to an RU who is linked to the given EU.

If the estimator given in (21) and (22) would be used in order to identify all of the EUs, then additional information must be collected in the interview. The respondent RU must know and report the number of RUs, including themselves, that are linked to both the selected SU and the reported EU.

The respondent RU must also know and report their share of use of the cell phone on which the interview is completed or be able to say that use is approximately equal.

6. Example: The National Immunization Survey (NIS)

We illustrate the information that must be collected in the survey interview using the NIS, a survey of parents of children age 19-35 months and of teens age 13-17 years sponsored by the Centers for Disease Control and Prevention (CDC) for the purpose of monitoring vaccination coverage rates (*i.e.*, the proportion of children who are up-to-date with respect to the recommended vaccination schedule) in the USA. Data collection in the NIS occurs in two phases: an RDD telephone survey of households with landline telephones that have children or teens in the eligible age range, followed by a survey mailed to the vaccination providers of the age-eligible children. The sampling frame for the telephone survey phase of the NIS consists of all landline telephone numbers in 1+ banks in the USA. Cellular telephone numbers in dedicated cellular banks are currently not included in the NIS sampling frame. When a household with an age-eligible child is identified in the telephone survey, the interview is conducted with the adult in the household who is identified as the most knowledgeable about the vaccination status of the child (nearly always the mother or father). During the telephone interview, data are collected for each age-eligible child in the household, including the demographic characteristics of the child, demographic characteristics of the child's mother, and socio-economic characteristics of the child's household. At the end of the telephone interview, consent is asked to contact the child's vaccination providers. If consent is given, all vaccination providers named by the telephone interview respondent are contacted by mail to obtain the child's provider-reported vaccination history, which is used in statistical analysis to evaluate vaccination status. Smith, Hoaglin, Battaglia, Khare and Barker (2005) provide a detailed description of the statistical methods used by the NIS.

Because of the growth of the cell-phone-only population, the proportion of the NIS target population that is covered by the landline sampling frame has decreased in recent years. Using data from the National Health Interview Survey, Khare, Singleton, Wouhib and Jain (2008) estimate that about 18 percent of eligible children and 10 percent of eligible teens may be missing from the NIS sampling frame. To address the increase in cell-phone-only households in the NIS target population, cell-phone interviews could be added to the NIS.

For the NIS, the telephone number is the SU, the knowledgeable mother or father is the RU, and the age-eligible child is the EU. For the landline RDD or A sample, the parent is a resident of the household to which the sample landline number is assigned, while for the cell-phone or B sample, the parent has regular access to the cell phone to which the sample telephone number is assigned. Children are not subsampled in the NIS, but rather the knowledgeable parent reports for all of their age-eligible children who live in their home (but not for any children who may live elsewhere). These elements of the survey protocol establish the links between RUs and SUs and between EUs and RUs.

One comprehensive NIS design is to conduct estimation by way of nonoverlapping domains and decomposition (3). That is, the A sample is used to represent all children linked to a landline household and the B sample is used to represent all children linked to a cell-phone-only parent. We considered and rejected decompositions (4) and (5) due to considerations of cost and the potential for differential nonresponse bias in estimation for the mixed population.

To implement the estimator in (10), we determine whether the A-sample child is landline-only through use of the following three questionnaire items:

A1. Next I have some questions about cell phones in your household. In total, how many working cell phones do you and your household members have available for personal use? Please don't count cell phones that are used exclusively for business purposes.

A2. How many [of these] cell phones do [LIST ALL ELIGIBLE CHILDREN]'s parents and guardians usually use?

A3. Of all the telephone calls that you and your family receive, are nearly all received on cell phones, nearly all received on regular phones, or some received on cell phones and some received on regular phones? (IF ASKED ABOUT INCLUDING BUSINESS CALLS: Please do not include any business-related calls in your answer).

For the cell-phone or B sample, we establish whether the child is cell-phone-only using the following two questions.

B1. Do you have a landline in your household? (INTERVIEWER PROBE IF YES: Please do not include modem only lines, fax only lines, lines used just for a home security system, beepers, pagers, or the cell phone).

B2. Thinking just about the landline home phone, not your cell phone, if that telephone rang and someone was home, under normal circumstances how likely is it that it would be answered? Would you say

extremely likely, somewhat likely, somewhat unlikely, or not at all likely?

We would use Question B2, due to Cantor, Brownlee, Zukin and Boyle (2008), to determine whether the landline is actually used for voice communications and thus whether the respondent is in the *ab* or *b* domain.

Also for the B sample, to determine the number of cell phones in the population that are linked to a given age-eligible child, we would use the following two questions:

B3. Next, I have some questions about cell phones in your household. In total, how many working cell phones do you and your household members have available for personal use? Please do not count cell phones that are used exclusively for business purposes, and please include the number we called.

B4. How many of these cell phones do [LIST CHILDREN]'s parents and guardians usually use? Please include the number we called.

Responses to questions A1-A3 and B1-B4 permit the calculation of survey weights and implementation of the unbiased estimator of the population total given in (10).

7. Summary

In this article, we used some theory of indirect sampling and network sampling to demonstrate a statistical framework for the design and analysis of cell-phone surveys. We exhibited an unbiased estimator of the population total with respect to estimation units linked to sampling units. By implication, this theory gives a means of constructing estimators of other population parameters that can be expressed as functions of totals. We illustrated the issues using the NIS, a telephone survey about young children and teens.

Information from the survey interviews is needed to classify estimation units into the cell-phone-only domain, the landline-only domain, or the mixed domain. Reporting error could result in misclassifications and undermine the unbiasedness of the estimator, as could survey nonresponse in the cell-phone and landline interviews.

Acknowledgements

The authors thank associate editor for helpful comments.

References

Arthur, A. (2007). The birth of a cellular nation. *The Source*. Mediamark Research Inc. Available from: http://www.mediamark.com/mri/TheSource/sorc2007_09.htm, 3.

Blumberg, S.J., and Luke, J.W. (2008). Wireless substitution: Early release of estimates from the National Health Interview Survey. National Center for Health Statistics. Available from: http://www.cdc.gov/nchs/nhis.htm.

Brick, J.M., Dipko, S., Presser, S., Tucker, C. and Yuan, Y. (2006). Nonresponse bias in a dual frame sample of cell and landline numbers. *Public Opinion Quarterly*, 70, 780-793.

Brick, J.M., Edwards, W.S. and Lee, S. (2007). Sampling telephone numbers and adults, interview length, and weighting in the California Health Interview Survey cell phone pilot study. *Public Opinion Quarterly*, 71, 793-813.

Cantor, J., Brownlee, S., Zukin, C. and Boyle, J. (2008). Do We Need to Worry About Wireless Substitution in Public Opinion Polls about Health Reform. Presentation at the AcademyHealth 25th Annual Research Meeting, Washington, DC.

Carley-Baxter, L., Peytchev, A. and Lynberg, M. (2008). Comparison of cell phone and landline surveys: A design perspective. Paper presented at the annual meeting of the American Association for Public Opinion Research, New Orleans, LA.

Cassel, C.-M., Särndal, C.-E. and Wretman, J.H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63, 615-620.

Chowdhury, S., Montgomery, R. and Smith, P.J. (2008). Adjustment for noncoverage of nonlandline telephone households in and RDD Survey. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexanderia, VA.

CTIA (2008). Wireless Quick Facts. Available from http://www.ctia.org/advocacy/research/index.cfm/AID/10323.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, 32, 165-176.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Ehlen, J., and Ehlen, P. (2007). Cellular-only substitution in the United States as lifestyle adoption: Implications for telephone survey coverage. *Public Opinion Quarterly*, 71, 717-733.

Frankel, M., Battaglia, M., Link, M. and Mokdad, A. (2007). Integrating cell phone numbers into Random Digit-Dialed (RDD) landline surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, (Alexandria, VA), 3793-3800.

Hartley, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section*, American Statistical Association, 203-206.

Keeter, S. (1995). Estimating non-coverage bias from a phone survey. *Public Opinion Quarterly*, 59, 196-217.

Khare, M., Singleton, J.A., Wouhib, A. and Jain, N. (2008). Assessment of Potential Bias in the National Immunization Survey (NIS) from the Increasing Prevalence of Households Without Landline Telephones. Presented at the National Immunization Conference, Centers for Disease Control and Prevention.

Lavallée, P. (2007). *Indirect Sampling*. New York: Springer Science+Business Media, LLC.

Lavrakas, P.J., Shuttles, C.D., Steeh, C. and Fienberg, H. (2007). The state of surveying cell phone numbers in the United States: 2007 and Beyond. *Public Opinion Quarterly*, 71, 840-854.

Smith, P.J., Hoaglin, D.C., Battaglia, M.P., Khare, M. and Barker, L.E. (2005). Statistical methodology of the National Immunization Survey, 1994-2002. National Center for Health Statistics, Hyattsville, MD. *Vital and Health Statistics*, Series 2, 138.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, *Second Edition*. New York: Springer-Verlag.

Wolter, K.M., Chowdhury, S. and Kelly, J. (2008). Design, conduct, and analysis of random digit dialing surveys. In *Handbook of Statistics*: *Sample Surveys*, *Theory*, *Methods and Inference*, (Eds., D. Pfeffermann and C.R. Rao), Elsevier, Oxford, UK.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 217-222
Statistics Canada, Catalogue No. 12-001-X

Collecting data for poverty and vulnerability assessment in remote areas in Sub-Saharan Africa

Rudolf Witt, Diemuth E. Pemsl and Hermann Waibel [1]

Abstract

Data collection for poverty assessments in Africa is time consuming, expensive and can be subject to numerous constraints. In this paper we present a procedure to collect data from poor households involved in small-scale inland fisheries as well as agricultural activities. A sampling scheme has been developed that captures the heterogeneity in ecological conditions and the seasonality of livelihood options. Sampling includes a three point panel survey of 300 households. The respondents belong to four different ethnic groups randomly chosen from three strata, each representing a different ecological zone. In the first part of the paper some background information is given on the objectives of the research, the study site and survey design, which were guiding the data collection process. The second part of the paper discusses the typical constraints that are hampering empirical work in Sub-Saharan Africa, and shows how different challenges have been resolved. These lessons could guide researchers in designing appropriate socio-economic surveys in comparable settings.

Key Words: Socio-economic household surveys; Survey design; Data collection challenges; Sub-Saharan Africa.

1. Introduction

To collect economic data in small-scale fisheries in Sub-Saharan Africa (SSA) is challenging, as patterns and constraints of resource use vary considerably, *i.e.*, spatially, seasonally and over time. This requires careful planning of the collection of data that is needed for meaningful poverty and vulnerability assessment. Although small-scale fisheries (SSF) can generate significant profits and make considerable contributions to poverty alleviation and food security, little information exists about their actual contribution to livelihoods and household economics in Sub-Saharan Africa (FAO 2005, 2006). The key constraints for empirical studies in this field are difficulties associated with data collection, such as remoteness and inaccessibility especially during the rainy season. High variability of natural resource conditions, and thus production, cause additional requirements for survey design. For preparation and implementation of a survey in SSA, researchers can draw upon similar studies in other parts of the world concerning survey methodology, questionnaire design, and interview procedure, *e.g.*, the World Bank's Living Standard Measurement Survey (LSMS) questionnaire. However, many peculiarities of rural communities in SSA require an adapted and elaborated approach.

Some of these peculiarities are of an ecological nature, such as seasonal changes in access to resources and markets, which are directly affecting patterns and constraints of resource use. Others pertain to the economic side of household behavior, since income-generating activities of rural households in SSA compose complex portfolios. Particularly households in fishery-dependent communities have adopted a flexible and strongly seasonal matrix of diversified activities (Béné, Neiland, Jolley, Ovie, Sule, Ladu, Mindjimba, Belal, Tiotsop, Baba, Dara, Zakara and Quensiere 2003a; Béné, Neiland, Jolley, Ladu, Ovie, Sule, Baba, Belal, Mindjimba, Tiotsop, Dara, Zakara and Quensiere 2003b; Béné, Mindjimba, Belal, Jolley and Neiland 2003c; Neiland, Jaffry and Kudasi 2000, Neiland, Madaka and Béné 2005; Sarch 1997). The local populations are alternatively or simultaneously fishers, herders, and farmers, and each piece of land is potentially a fishing ground, a grazing area and a cultivated field, depending on the flood cycle (Béné *et al.* 2003a, page 20). Due to high vulnerability of the ecological and economic system to shocks, such as flood, drought and pest outbreaks which result in year to year variation in fish stocks and in high crop losses, households have diversified their activities portfolio, thus spreading the risk of income losses. Capturing the dynamic interplay of the different livelihood elements is a special challenge in conducting socio-economic household surveys. Other constraints for data collection are culturally determined, for example tensions between different ethnic groups, the existence of a multitude of languages and patois spoken in the study region, or some peculiarities of the Muslim-African culture.

The data required for poverty and vulnerability assessment demand an appropriate survey methodology, for data quality to meet the requirements of a robust econometric analysis. Data needs for economic poverty assessment and the evaluation of SSF's contribution to poverty and vulnerability alleviation are substantial. Detailed information on

1. Rudolf Witt and Hermann Waibel, Institute of Development and Agricultural Economics, Faculty of Economics and Management, Leibniz University Hannover, Königsworther Platz 1, 30167 Hannover, Germany. E-mail: witt@ifgb.uni-hannover.de; Diemuth E. Pemsl, Economist, The WorldFish Centre, Penang, Malaysia.

household income, including different income sources such as agricultural production, fishing, livestock rearing, off-farm work *etc*., is necessary. Also, data on the stock and value of productive and convertible assets, as well as on the distribution of consumption expenditures need to be elicited. In addition, information on control variables, such as ecological, economic or social shocks that have occurred in the past, subjective risk assessments, debts and liabilities, household composition, and others, is required.

This paper presents the collection procedure of quantitative household data from poor households in the Logone floodplain, a major inland fisheries region in Northern Cameroon. The objective of collecting household level panel data in 2007-2008 was to assess the role of small-scale fisheries (SSF) in mitigating risk through portfolio diversification, thus contributing to reducing vulnerability to poverty. In this paper, we emphasize the requirements of the general methodological approach for sampling and survey design. Due to the complex nature of the SSF sector outlined above, a procedure for sampling and data collection is required that allows the assessment of poverty and vulnerability of SSF households. Particularly, the survey design needs to account for the high variation in income generating activities over time as a result of the high variability of access to natural resources and resulting adjustments in a household's food security situation, consumption, income and assets.

2. Study site and sampling procedure

The study site is the Logone floodplain in the Far-North province of Cameroon. The floodplain covers about 8,000 km^2 and is part of the bigger Logone-Chari subsystem in the Lake Chad Basin, which supplies 95% of Lake Chad's total riverine inputs and has a basin area of approximately 650,000 km^2 (UNEP 2004). Within this vast area a representative region was defined in collaboration with national experts and other key informants, while considering the accessibility and logistic feasibility of the study. The study area covers about 2,400 km^2, spreading from the Maga Lake in the south to Ivyé village in the north, where the Logomatya joins the Logone River. This area is relatively densely populated and is characterized by rich fish stocks and intensive fishing, fish processing and fish trading.

The livelihoods of the rural population in this area are particularly exposed to harsh climatic conditions, such as limited and erratic rainfall, which result in a large variation of production outcomes from year to year (In this respect, the study area is representative for many similar rural settings, particularly in the Sudano-Sahelian zone of Sub-Saharan Africa.) and thus considerable income risk. However, the impact is different between the sub-regions of the study area. Based on Neyman (1938), as cited in Rao (2005), a stratified random sampling procedure was therefore considered most effective. To draw a representative sample of households in the study area while accounting for different production conditions (such as access to fish resources), a stratification of the study site into different agroecological zones was undertaken. It was assumed that under different ecological and production conditions the role of fisheries in terms of income generation would differ. This procedure allowed capturing the whole continuum of fishing intensity (from specialized/full-time fishermen to purely agriculture/livestock rearing oriented households).

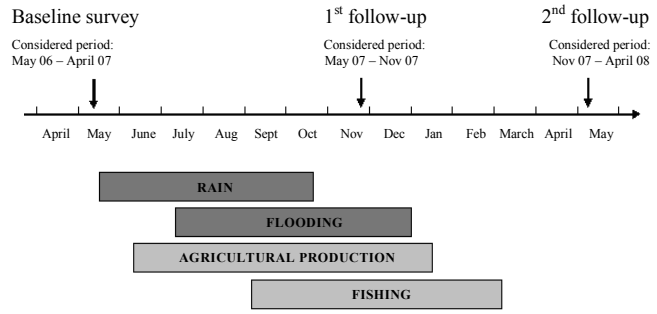
In a second step, a complete list of villages in the study area (N = 88) was compiled. These villages served as the primary sampling unit. Following the recommendations of local fisheries experts, 14 villages were selected proportional to the total number of villages per zone. The average village size in the floodplain (study area) is about 45 households, with a range of 15 to 100 households. Within villages every second household was chosen randomly from household lists established by the village headman. Hence, a sample size of 300 households was chosen proportional to the size of the village populations, which equates to a sampling ratio of 7% of the total population (estimated at 20,000 by the Ministry of Livestock, Fisheries and Animal Industries, MINEPIA).

All selected villages were visited before commencing the household level survey with the aim to establish contacts between the researcher and the village headmen and conduct focus group discussions (FGDs) with the village leaders. The objective of the FGDs was twofold. First, some general information was collected such as the village size, infrastructure, and access to fish resources and markets. Second, complete household lists for every selected village were compiled, since no official statistical information existed. For this study, a household was defined as an economically independent unit consisting of the household head, one or more spouse(s), children and other directly dependent members, living in the household or having migrated to other locations. Household size varies from two (*i.e.*, normally husband and spouse) to more than 15. Large households are common for Northern Cameroon, since due to widespread polygamy household heads often live together with up to four wives. Mostly, households do not live separately from other kin households, but usually form a clan, living together in a larger compound. However, within the compound, households are independent from each other. During the visits, special attention was paid to list the names of individual household heads and not only those of the compound/clan leaders. The additional information collected during the FGDs was necessary to get a first understanding of the livelihood options and constraints in

the study area, which proved to be helpful for the development of the household questionnaire. In the last step, the compiled household lists were used for a weighted random sampling of the 300 sample households.

3. Survey design

Seasonality is an important characteristic of the livelihood conditions in the Logone floodplain. Therefore, in order to capture seasonal variation, the survey was designed to yield a two-period panel data set (2006 – 2007), with an additional third survey six months after conducting the baseline survey (see Figure 1). The baseline survey was accomplished right at the end of the dry season, when income-generating activities are extremely limited, and the financial resources, generated during the rainy season in 2006, are being used up. The period covered in the baseline survey was May 2006 to April 2007, constituting a stock check of average income flows, consumption expenditures, and an asset inventory. The first follow-up survey captured the busy time of the year, where expenditures rise due to investments (*e.g.*, purchase of new fishing nets and other productive assets), and variable production costs in agriculture and fishing. Finally, the second follow-up survey covered the second half of the year, giving account of the economic household activities in this period. This approach was chosen to improve the accuracy of data on livelihood activities by reducing the recall period, and to make sure to capture seasonal variation in income and consumption.



Source: own illustration

Figure 1 Livelihood options in the study area and design of the survey

Before the start of each survey, enumerator training workshops of 3 to 4 days were conducted, including pre-testing of the questionnaire in order to detect weaknesses and the necessity to eliminate, rephrase or add additional questions. The baseline pre-test was carried out in two villages of zone 1 and 2, in order to test the suitability of the questionnaire for different livelihood conditions. The baseline study was completed within 3 weeks in May 2007 by four enumerators, working in a team, and accompanied and directly supervised by the first author. This procedure gave the opportunity for immediate cross-checking for missing information, and also enabled the researcher to observe and reinforce interview techniques and immediately discuss problems or questions.

Due to the relative remoteness of the villages and difficulties of access, careful logistical planning was necessary. The field trips often covered several days, and it was inevitable to spend the nights in the villages. Hence, the survey procedure adopted was as follows: the whole team arrived in a village, presenting itself to the village chief, who had been previously informed about the arrival date of the team during the FGD visit. The chief then called the heads of the selected households to a central meeting place, usually under a tree in front of the chief's house. After the interview, which normally took about one hour, the respondent was given a small present as a compensation for his time (a package of sugar and a bag of tea), and the next household head was called to sit down. Working in a group enabled the team to finish a village in about one or two days and proceed to the next one. That course of action strongly motivated and encouraged the enumerators for security and psychological reasons. The interview time, and hence the time planned to be spent per village, was held flexible, so that careful cross-checking for consistency and plausibility of responses was ensured. Hence, during the enumerator training workshops and throughout the data collection process, special emphasis was placed on the ultimate primacy of data quality.

4. Data collection challenges and lessons learnt

This section describes some challenges and constraints in data collection, which have been encountered during this study, but which are not limited to the study region. Similar settings are found in many wetlands and floodplains in SSA, and the lessons learnt in this study may prove helpful for comparable data collection endeavors.

Seasonality

When collecting data in rural fisheries-dependent communities in SSA, the seasonal nature of the livelihood systems and the ecological constraints need to be taken into consideration. Very often, villages are spatially marginalized and access is extremely difficult during certain periods of the year. For example, in the Logone floodplain in North Cameroon, access to the villages is very restricted during several weeks twice a year due to the annual flood cycle. At the beginning of the flooding season, and during the deflooding period, access is not possible, neither by vehicle, nor by boat. Hence, the placing of the survey periods need to be adapted to these conditions. For example,

although it would have been more reasonable to place a follow-up survey at the end of the production cycle in January, thus better capturing agricultural production and fishing harvests, this procedure proved to be unfeasible. From mid December to end of February access to the sampled villages was not possible at all. The research team decided for a compromise, collecting data in December, even if this falls in the midst of the harvesting season. The missed data on yields and income was then recollected during the second follow-up. Similar problems arise in other major inland fisheries such as the Hadejia-Nguru Wetlands in Nigeria or the Lower Shire river basin in Malawi.

Defining time periods

For recall surveys and particularly for panel surveys (*i.e.*, the research team is repeatedly revisiting the same households) it is important to assure a common understanding of the time period that is considered in the questionnaire. Different notions of the time span may result in biased information concerning income or consumption flows and can flaw the results and conclusions drawn from the study. In order to assure a common understanding of the requested time period, the respective cultural understanding of time needs to be taken into account. We found that in the Logone floodplain, people do not think in time units such as weeks or months. Hence, questions, such as: "How much did you spend on food items in the last 6 months?" were not appropriate. In this case, it proved instrumental to refer to certain region-wide acknowledged social events or celebrations. For example, the survey in November coincided with the Tabaski festivities, so that it was easy for the respondents to delimit the time period considered in the second follow-up survey.

Selection of enumerators and their cultural competence

Perhaps the most important factor in empirical work is the choice of the enumerators. To achieve good data quality, enumerators must not only provide the needed skills and knowledge, but also dispose over additional soft skills, such as mastering of languages, social competence, and the will to work under severe conditions.

The lack of sufficiently educated interviewer personnel in the Far-North Province in Cameroon presented a serious constraint. For this study, a team of five MINEPIA staff, who work as government officials in the survey area, was recruited as enumerators. While respondents can have reservations to provide information to government officers, the more important factor was that the survey team represented the two ethnic groups of the study area. Also, enumerators spoke the languages of the region, they were familiar with the local peculiarities, and used to the conditions in the field. In addition, respondents' willingness to provide information was actually encouraged in expectations of a follow-up governmental support.

Another advantage of the selected enumerators was awareness and sensitivity towards ethnic tensions. Enumerators were careful not to take sides with either one of the involved parties, and avoided offensive statements. This was especially important with regard to multiple visits of villages and respondents during the follow-up surveys. Any disaccord between respondents and enumerators would have resulted in significant attrition and the need to drop entire villages from the sample.

Certain cultural or religious norms also demanded tactfulness and respect. For example, in a number of villages only men could be interviewed since women in that African-Muslim culture are not allowed to meet or talk to men other than direct family members. In cases where the household head was not present at the time of the visit, it was not possible to interview the spouse (or any other woman in the household) instead. An adult male household member had to be chosen to provide the required information. For the same reason, interviews could not take place in the house of the respondents. For the sake of compliance to these cultural norms, the interview procedure had to be adapted. Instead of visiting the chosen households one by one, all sampled household representatives in each village were called to a central meeting place by the village chief (usually in front of the chief's house). If the household head was not present, another adult member of the household (usually male) was interviewed. The enumerators then seated themselves at a distance of about three to five meters from each other, calling the respective respondent to be interviewed in private, while the others were waiting for their turn.

Sample attrition

A particular challenge of panel surveys in general is to maintain the size of the sample over time (Jäckle and Lynn 2008, Laaksonen 2007). Attrition can be high due to several reasons. For example, in some cases the household head has died, the whole household has moved away, or the respondents lose interest to participate especially if no or not enough incentives are provided. The loss of willingness to participate in a follow-up survey caused a problem during the second visit. Due to budget constraints the survey team decided not to compensate the participants for their time at the second visit. For the baseline survey, each respondent had received a box of sugar and a package of tea which turned out to be a strong extrinsic incentive. When households learned that no remuneration had been foreseen at the second visit, 69 households (23% of the total sample) announced that they were "too busy" to participate. Considering this reaction, compensation was again offered at the

third survey, so that most of the lost households could be regained. They were even willing to respond to both questionnaires (1st and 2nd follow-up). Thus the missing data could be completed during the last survey round albeit at the cost of lower reliability due to memory bias. Such respondent behavior is consistent with findings by Jäckle and Lynn (2008), who report significant positive effects of continued incentive payments on attrition, bias and item non-response. At the end of the survey period, 14 households (4.7%) have been lost due to permanent migration or other reasons, and hence were removed from the sample.

5. Summary and conclusions

Data collection for poverty analysis in SSA is a challenging endeavor. Often, cultural, ecological and economic constraints push researchers to put up with a compromise between data quality and feasibility of the study. On the other hand, collection of such data is important because little is known about poverty and vulnerability of marginalized groups such as fisheries communities in remote areas of SSA. In this paper, we present the approach that has been taken in the course of a study on poverty and vulnerability in the Logone floodplain, which is a major fishing area in Northern Cameroon. We identify typical constraints that are often hampering empirical work in SSA, and show how different challenges can be overcome by an adequate survey design, sampling and careful application of the survey instrument. Major constraints encountered were the difficulties to access the target population, limitations in finding qualified enumerators and high demand for cultural sensitivity of the research team.

Of eminent importance is a close collaboration with local authorities and experts in the respective field of research, as well as a good understanding of and compliance with local cultural norms and values. Learning from the local population and empathizing with it's peculiar ways of living before starting the survey per se has been found to be a key success factor for working in that region. Summing up, it can be concluded that despite a number of difficulties, quantitative data collection in rural Sub-Saharan Africa is a task that can be completed with satisfying results. An appropriate survey design and interview procedure developed in collaboration with local staff and experts can assure adequate data quality for economic poverty and vulnerability analysis.

Acknowledgements

We thank the German Federal Ministry of Economic Cooperation and Development (BMZ) for the financial support through the project on Food security and poverty alleviation through improved valuation and governance of river fisheries in Africa which was coordinated by the WorldFish Center. The authors also thank two anonymous referees for their very thorough review of the paper and their extremely valuable comments. We also want to thank the editor for the excellent guidance of the review process and additional useful comments on the paper. The views expressed in this paper are not necessarily those of the donor agency nor of our project partner the WorldFish Center.

References

Béné, C., Neiland, A., Jolley, T., Ovie, S., Sule, O., Ladu, B., Mindjimba, K., Belal, E., Tiotsop, F., Baba, M., Dara, L., Zakara, A. and Quensiere, J. (2003a). Inland fisheries, poverty, and rural livelihoods in the Lake Chad Basin. *Journal of Asian and African Studies*, 38, 1, 17-51.

Béné, C., Neiland, A., Jolley, T., Ladu, B., Ovie, S., Sule, O., Baba, M., Belal, E., Mindjimba, K., Tiotsop, F., Dara, L., Zakara, A. and Quensiere, J. (2003b). Natural-resource institutions and property rights in inland African fisheries - The case of the Lake Chad Basin region. *International Journal of Social Economics*, 30, 3, 275-301.

Béné, C., Mindjimba, K., Belal, E., Jolley, T. and Neiland, A. (2003c). Inland fisheries, tenure systems and livelihood diversification in Africa: The case of the Yaéré floodplains in Lake Chad Basin. *African Studies*, 62, 2, 187-212.

FAO (2005). Technical guidelines for responsible fisheries Nr 10: Increasing the contribution of small-scale fisheries to poverty alleviation and food security. FAO, Rome.

FAO (2006). FAO's Activities on Small-scale Fisheries: An Overview. Advisory Committee on Fisheries Research (ACFR), Sixth Session, Rome, 17-30 October 2006.

Jäckle, A., and Lynn, P. (2008). Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias. *Survey Methodology*, 34, 1, 105-117.

Laaksonen, S. (2007). Weighting for twophase surveyed data. *Survey Methodology*, 33, 2, 121-130.

Neiland, A.E., Madaka, S.P. and Béné, C. (2005). Traditional Management Systems, poverty and change in the Arid Zone Fisheries of Northern Nigeria. *Journal of Agrarian Change*, 5, 117-48.

Neiland, A.E., Jaffry, S. and Kudasi, D.K. (2000). *Fishing Income, Poverty and Fisheries Management in North-East Nigeria.* Fisheries of North East Nigeria and the Lake Chad Basin, Volume I - A compilation of research project reports in two volumes 1993-2004, compiled by Dr. Arthur E. Neiland, 291-319.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.

Rao, J.N.K. (2005). Interplay Between Sample Survey Theory and Practice: An Appraisal. *Survey Methodology*, 31, 2, 117-138.

Sarch, M.-T. (1997). Fishing and Farming in Lake Chad: Implications for Fisheries Development. *Development Policy Review*, 15, 141-57.

UNEP (2004). Map of the Lake Chad Basin. In *Lake Chad Basin*, *GIWA Regional assessment 43*, (Eds., M.P. Fortnam and J.A. Oguntola), University of Kalmar, Kalmar, Sweden.

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 223-227
Statistics Canada, Catalogue No. 12-001-X

Respondent differences and length of data collection in the Behavioral Risk Factor Surveillance System

Mohamed G. Qayad, Pranesh Chowdhury, Shaohua Hu and Lina Balluz [1]

Abstract

The current economic downturn in the US could challenge costly strategies in survey operations. In the Behavioral Risk Factor Surveillance System (BRFSS), ending the monthly data collection at 31 days could be a less costly alternative. However, this could potentially exclude a portion of interviews completed after 31 days (late responders) whose respondent characteristics could be different in many respects from those who completed the survey within 31 days (early responders). We examined whether there are differences between the early and late responders in demographics, health-care coverage, general health status, health risk behaviors, and chronic disease conditions or illnesses. We used 2007 BRFSS data, where a representative sample of the noninstitutionalized adult U.S. population was selected using a random digit dialing method. Late responders were significantly more likely to be male; to report race/ethnicity as Hispanic; to have annual income higher than $50,000; to be younger than 45 years of age; to have less than high school education; to have health-care coverage; to be significantly more likely to report good health; and to be significantly less likely to report hypertension, diabetes, or being obese. The observed differences between early and late responders on survey estimates may hardly influence national and state-level estimates. As the proportion of late responders may increase in the future, its impact on surveillance estimates should be examined before excluding from the analysis. Analysis on late responders only should combine several years of data to produce reliable estimates.

Key Words: BRFSS; Responders; Differences; Length of data collection.

1. Introduction

The Behavioral Risk Factor Surveillance System (BRFSS) is a state-based household telephone survey in the United States (U.S.) and its territories which monitors health risk behaviors and chronic disease conditions for the adult noninstitutionalized population (Centers for Disease Control and Prevention [CDC] 2009a, BRFSS Turning Information into Public Health, http://www.cdc.gov/brfss/about.htm). It is the largest telephone survey in the world and is implemented by the 50 states, the District of Columbia, and U.S. territories, in collaboration with the CDC. The survey is conducted continuously throughout the year.

CDC dispenses the samples (phone numbers) to states quarterly. At the state level, the samples are divided into 12 monthly lists for operational purposes. Trained interviewers call each sampled telephone number. After each call to a sampled telephone number, a disposition code is assigned. States and their contractors are required to give final dispositions to their monthly released samples within that month. Over 90% of the monthly samples and completed interviews receive final dispositions within 31 days. States continue to complete their remaining samples afterwards (Qayad, Balluz and Garvin 2009).

Because of economic downturns, states and survey organizations may face budget cuts that could adversely affect their survey operations. Such unforeseen circumstances warrant searching for alternative operational strategies. A cost-effective alternative could be to end data collection at the end of each month. However, ending data collection within one month excludes interviews completed after 31 days. Such exclusion could influence the variability of the respondents, surveillance estimates and the size of completed interviews, which could affect other operational decisions. Currently, the size of late responders is small and may not influence surveillance estimates. However, the current trend in survey responses heralds a continuous decline in survey responders, which could prolong the duration to reach respondent and the eventual increase in the proportion of late responders. Such circumstances require thorough examination of the influence of late responders on surveillance estimates in the future. This study examines whether respondents who completed the interviews within 31 days and those who completed after 31 days are different in demographics, risk behaviours, and chronic disease conditions.

2. Methods

We used the 2007 BRFSS data, which is an ongoing state-based random digit dialing (RDD) telephone survey among the non-institutionalized civilian population in the US. We divided the duration of the interview into two periods, 0-31 days and >31 days. Respondents who completed the interviews within 31 days (referred as early responders) and those completed after 31 days (referred as late responders).

1. Mohamed G. Qayad, Pranesh Chowdhury, Shaohua Hu and Lina Balluz, Division of Adult and Community Health, Behavioral Surveillance Branch, Centers for Disease Control and Prevention, Atlanta, GA 30341, U.S.A. E-mail: maq3@cdc.gov.

Demographic factors included were - gender, race, income and age. Race had four groups - white non-Hispanic, Black non-Hispanic, Hispanic and other race. Education had three levels: not a high school graduate, high school graduate, and more than high school education. Income categories were <$15,000, $15,000 - $34,999, $35,000 - $49,999 and $50,000 or more. Age had the following categories: 18 - 24 years, 25 - 44 years, 45 - 64 years, and 65 or more years. Respondents <65 years old who did not have any health plan (including health insurance, prepaid plans such as HMOs, or government plans such as Medicare) were considered not to have health plan. General health was dichotomized into good health (excellent, very good, or good health) and fair or poor health.

Health risk behaviors included were - binge drinking, current smoking, (lack of) physical activity, and (insuffisant) fruit and vegetable consumption. Binge drinking was defined as having five or more drinks for men and four or more drinks for women on at least one occasion during the preceding month. Respondents who smoked ≥100 cigarettes in their lifetime and smoked every day or some days were classified as current smokers. Physical activity had following categories - meet recommendations for physical activity, insufficient physical activity, and do not participate in physical activity. Respondents who consumed 5 or more servings of fruits and vegetables everyday were classified as meet recommendation for fruit and vegetable consumption.

Chronic conditions or illness included were Cerebro-cardio-vascular disease, hypertension, had high cholesterol, diabetes, asthma, and overweight or obesity. Respondents were considered to have myocardial infarction, or angina, or stroke or high blood pressure if they had ever been told by a doctor, nurse, or other health professional to have myocardial infraction or stroke or high blood pressure respectively. Respondents were classified as having high blood cholesterol if they had checked their blood cholesterol and was told by a health professional that their blood cholesterol was high. Respondents were classified as having diabetes if they had ever been told by a doctor that they had diabetes. Asthma was self reported and physician or health care professional diagnosed; it had three categories - current asthma, former asthma, and never asthma. Self-reported weight and height were used to calculate Body Mass Index (BMI) (BMI = weight[kg]/(height[m])*). Participants were classified as overweight if their BMI was ≥25 kg/m^2 and were classified as obese if their BMI was ≥30 kg/m^2.

We estimated the percent differences between early and late responders by demographics, health behaviors and chronic health conditions or illness. We used SUDAAN and SAS for the analysis (SAS Institute Inc., Cary, NC, USA 2004).

3. Results

In the 2007 BRFSS survey, there were 430,912 interviews completed in the U.S. We excluded 14,189 records from two states (Michigan and Louisiana) and 49 cases with missing information. We analyzed the remaining 416,674 respondents of which 394,427 (95%) were early responders, and 22,247 (5%) were late responders. We estimated weighted and unweighted percent differences between early and late responders. The absolute differences between the weighted and unweighted percentages in the variables examined ranged between 0.06% and 2.6%, except white non-Hispanics where the absolute difference was 7%. We presented the unweighted analysis for the purpose of this study.

Significant differences were observed between early and late responders in demographics, access to health-care coverage, and general health status variables (Table 1). Compared to early responders, late responders were significantly more likely to be male, to report race/ethnicity as Hispanic, to have annual income of ≥$50,000, to be younger than 45 years of age, to have less than high school education, to have access to health-care coverage, and to report good health. The absolute value of these significant differences in the variables above ranged from 1.3% to 7.6%. The percentage of Unknowns in the health-care coverage variable was 21% for late responders and 30% for early responders. The difference between early and late responders remained significant, even when we assumed the Unknowns to have a similar percentage of access to health-care coverage to those with known status in each respondent group.

A significant difference between early and late responders was also observed in health risk behaviors (Table 2). Compared to early responders, late responders were significantly less likely to meet the recommended guidelines for physical activity and daily consumption of fruits and vegetables. The absolute value of these significant differences ranged from 1.7 % to 3.1%. The differences between early and late responders remained significant even when the Unknowns were assumed to have a similar percentage to those of known status for both variables.

Table 3 shows the differences between early and late responders in chronic disease conditions or illnesses. Compared to early responders, late responders were significantly more likely to report high cholesterol, significantly less likely to report hypertension and diabetes, and were significantly less likely to be obese. The absolute value of these significant differences ranged from 1.8% to 5.8%.

Table 1
Percent differences between early responders and late responders by demographics, health-care coverage and general health, BRFSS 2007

	Length of data collection			
	Early reponders* (N = 394,427)	Late responders** (N = 22,247)	Difference (Early-late)	
Demographics	**%**	**%**	**%**	**P-Value**
Gender				
Female	62.8	60.2	2.5	0.000
Male	37.3	39.8	-2.5	
Race				
White non-Hispanic	79.1	71.5	7.6	0.000
Black non-Hispanic	7.3	8.2	-0.9	0.168
Hispanic	7.1	13.5	-6.4	0.000
Others	5.5	5.8	-0.3	0.635
Unknown	1.0	1.0	0.0	0.977
Income				
<15,000	9.7	8.7	1.0	0.146
15-34,999	26.1	24.3	1.8	0.004
35-49,999	14.1	13.4	0.8	0.252
50,000+	36.6	39.7	-3.1	0.000
Unknown	13.5	14.0	-0.4	0.496
Age				
18-24	3.6	4.9	-1.3	0.025
25-44	25.7	33.3	-7.6	0.000
45-64	40.9	40.6	0.3	0.612
65+	29.0	20.2	8.8	0.000
Unknown	0.8	1.0	-0.1	0.827
Education Level				
<High School	10.3	12.3	-2.0	0.001
High School Graduate	30.6	28.7	1.9	0.001
> High School	58.8	58.2	0.6	0.177
Unknown	0.3	0.8	-0.5	0.264
Health care coverage (<65 years)				
Yes	59.3	65.4	-6.2	0.000
No	10.8	13.2	-2.5	
Unknown	30.0	21.4	8.6	
Health Status				
Good health	80.1	81.8	-1.7	0.000
Fair or poor health	19.4	17.6	1.8	
Unknown	0.5	0.6	-0.1	

*Completed the survey within 31 days.
**Completed the survey after 31 days.

Table 2
Percent differences between early responders and late responders by health risk behaviors, BRFSS 2007

	Length of data collection			
	Early reponders* (N = 394,427)	Late responders** (N = 22,247)	Difference (Early-late)	
Risk factors	**%**	**%**	**%**	**P-Value**
Binge drinking				
Yes	11.1	11.8	-0.7	0.261
No	86.9	82.8	4.1	
Unknown	1.9	5.4	-3.4	
Smoking cigarettes				
Current smokers	18.3	17.5	0.9	0.182
Not a smoker	81.3	82.1	-0.8	
Unknown	0.4	0.5	0.0	
Physical activity recommendations				
Met recommended moderate/vigorous activity	43.4	41.8	1.7	0.000
Insufficinet physical activity	35.4	31.8	3.6	
No physical activity	14.3	11.3	3.0	
Unknown	6.9	15.2	-8.3	
Fruit & vegetable consumption				
Consumed ≥ 5 times/day	25.0	21.9	3.1	0.000
Consumed < 5 times/day	73.0	69.7	3.3	
Unknown	2.0	8.5	-6.4	

*Completed the survey within 31 days.
**Completed the survey after 31 days.

Table 3
Percent differences between early responders and late responders by chronic conditions and illnesses, BRFSS 2007

	Length of data collection			
Diseases/chronic conditions	Early reponders* (N = 394,427) %	Late responders** (N = 22,247) %	Difference (Early-late) %	P-Value
Cerebral and CVD:				
Myocardial Infarction				
Yes	5.9	4.9	1.0	0.177
No	93.6	94.7	-1.1	
Unknown	0.5	0.4	0.1	
Angina				
Yes	6.0	4.5	1.5	0.053
No	93.1	94.7	-1.6	
Unknown	0.9	0.8	0.1	
Stroke				
Yes	3.8	2.8	1.0	0.183
No	95.9	97.0	-1.1	
Unknown	0.3	0.2	0.1	
Other illnesses/conditions:				
High cholesterol				
Yes	57.0	60.8	-3.8	0.000
No	42.3	38.4	3.8	
Unknown	0.8	0.8	0.0	
Hypertension				
Yes	35.8	30.1	5.8	0.000
No	64.0	69.8	-5.8	
Unknown	0.2	0.2	0.0	
Diabetes				
Yes	11.2	9.4	1.8	0.010
Yes-Pregnancy	0.9	1.2	-0.2	
No	86.4	88.2	-1.9	
Borderline	1.4	1.2	0.2	
Unknown	0.1	0.1	0.0	
Asthma				
Current	8.7	7.7	1.0	0.158
Former	3.8	4.0	-0.2	
Never	86.9	87.8	-0.8	
Unknown	0.6	0.6	0.1	
Overweight or Obese				
Normal weight	34.5	35.5	-1.1	
Over weight	35.0	34.7	0.4	
Obese	*26.0*	*23.6*	*2.4*	0.000
Unknown	4.5	6.2	-1.7	

*Completed the survey within 31 days.
**Completed the survey after 31 days.

4. Discussion

Our study found significant differences between early and late responders in demographic factors, and in some of the health risk behaviors and chronic disease conditions or illnesses. This shows that the composition of the two groups of responders is different with respect to these attributes. The differences observed could be due to difficulty in reaching persons working long hours and being away from their residences.

The greater likelihood of earning high income, being Hispanic, being young (18-44 years), having health-care coverage, having less than high school education, and reporting good general health among late responders fits the described characteristics of working people and healthy workers (Li and Sung 1999), (O'Neil 1979). This description is supported by their significantly lower likelihood of reporting hypertension, diabetes and obesity. But certain risk behaviors show a different profile among late responders. Late responders are less likely to meet recommended guidelines for moderate or vigorous physical activity and for daily consumption of fruits and vegetables, which may be related to late responders having long working hours and poor access to healthy foods.

The high income earners, who are mostly white non-Hispanics, and low income earners, who are mostly Hispanics and black non-Hispanics, may spend long hours in their working environments and less likely to be in their homes to receive survey calls (Voigt, Koepsell and Daling 2003). In addition, BRFSS data indicate that interviewers

make more calls on late responders, on average almost 3 times more than on early responders, which bears out the difficulty of reaching them during the 31-day survey period. The reasons for working long hours could be different in the two income groups. Hispanics, black non-Hispanics, and young age groups may have low-paying jobs and need to work long hours to make a living, while the high-income individuals may have jobs requiring them to remain at work after regular working hours.

Surveillance and epidemiological estimates based only on early or late responders should be scrutinized for possible biases prior to making any generalizations. The percentage of interviews completed after 31 days is currently small (5%) and excluding them from the analysis may have no influence on national and state level estimates. However, as the proportions of late responders are expected to increase in the future, the influence of late responders on these estimates could not be ignored (Diehr, Cain, Connell and Volinn 1990). In addition, states should examine the consequences of ending data collection at 31 days on their operations, performance indicators, data quality measures, cost-savings and other contractual agreements with their data collection contractors.

Our study has a few limitations. BRFSS uses RDD methodology to select telephone numbers, which is subject to coverage bias (Rao, Link, Battaglia, Frankel, Giambo, and Mokdad 2005; Frankel, Srinath, Hoaglin, Battaglia, Smith, Wright and Khare 2003). Information collected is self-reported and may be subject to recall bias in some risk behaviors and disease estimations (Troiano, Berrigan, Dodd, Masse, Tilert and McDowell 2008; CDC 2004). In addition, we excluded two states from our analysis (Michigan and Louisiana), and extrapolation of the findings to these states should be done cautiously.

Despite these limitations, this study shows that late responders are significantly different in many respects from early responders. As the proportion of late responders may increase in the future, the influence of late responders on surveillance estimates should be examined carefully.

Acknowledgements

We would like to acknowledge State BRFSS coordinators and their contractors.

References

Centers for Disease Control and Prevention (2009a). BRFSS Turning Information into Public Health, URL http://www.cdc.gov/brfss/about.htm.

Centers for Disease Control and Prevention (2004). Access to health-care and preventive services among Hispanics and non-Hispanics – United States, 2001-2002. *Morbidity and Mortality Weekly Report*, 53, 937-941.

Diehr, P., Cain, K., Connell, F. and Volinn, E. (1990). What is too much variation? The null hypothesis in small-area analysis. *Health Services Research*, 24, 741-771.

Frankel, M.R., Srinath, K.P., Hoaglin, D.C., Battaglia, M.P., Smith, P.J., Wright, R.A. and Khare, M. (2003). Adjustments for non-telephone bias in random-digit-dialing surveys. *Statistics in Medicine*, 22, 1611-1626.

Li, C.Y., and Sung, F.C. (1999). A review of the healthy worker effect in occupational epidemiology. *Occupational Medicine*, 49, 225-229.

O'Neil, M.J. (1979). Estimating the nonresponse bias due to refusals in the telephone surveys. *Public Opinion Quarterly*, 43, 218-232.

Qayad, M.G., Balluz, L. and Garvin, W. (2009). Does continuing data collection beyond one month improve the completion and response rates in behavioral risk factor surveillance system survey? *Survey Practice Feb 2009*. URL http://surveypractice.org/2009/02/.

Rao, R.S., Link, M.W., Battaglia, M.P., Frankel, M.R., Giambo, P. and Mokdad, A.H. (2005). Assessing representativeness in RDD surveys: coverage and non-response in the Behavioral Risk Factor Surveillance System. Minneapolis: Joint Statistical Meetings. URL http://www.amstat.org/sections/SRMS/Proceedings/y2005/Files/JSM2005-000190.pdf.

SAS Institute Inc. (2004). SAS version 9.1, Cary, NC: SAS Institute, USA.

Troiano, R.P., Berrigan, D., Dodd, K.W., Masse, L.C., Tilert, T. and McDowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine and Science in Sports and Exercise*, 40, 181-8.

Voigt, L.F., Koepsell, T.D. and Daling, J.R. (2003). Characteristics of Telephone Survey Respondents According to Willingness to Participate. *American Journal of Epidemiology*, 157, 66-73.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

Survey Methodology, December 2010
Vol. 36, No. 2, pp. 229-231
Statistics Canada, Catalogue No. 12-001-X

An interesting property of the entropy of some sampling designs

Yves Tillé and David Haziza [1]

Abstract

In this short note, we show that simple random sampling without replacement and Bernoulli sampling have approximately the same entropy when the population size is large. An empirical example is given as an illustration.

Key Words: Conditional Poisson sampling; Entropy; Simple random sampling; Poisson sampling.

1. Introduction

Consider a finite population of size N and let $U = \{1, ..., k, ..., N\}$ be the set of labels of this population. A sample s is a subset of U and a sampling design is a probability law $p(.)$ on the subsets of U such that $p(s) \geq 0$ for all $s \subset U$, and

$$\sum_{s \subset U} p(s) = 1.$$

Let $\pi_k = P(k \in s)$ be the first-order inclusion probability of unit k in the sample:

$$\pi_k = \sum_{\substack{s \subset U \\ s \ni k}} p(s).$$

Similarly, let $\pi_{k\ell} = P(k \in s \text{ and } \ell \in s)$ be the second-order inclusion probability of unit k and ℓ in the sample:

$$\pi_{k\ell} = \sum_{\substack{s \subset U \\ s \ni k,\, \ell}} p(s).$$

The entropy of a sampling design $p(.)$, denoted by $I(p)$, is defined as

$$I(p) = -\sum_{s \in Q} p(s) \log p(s), \qquad (1)$$

where $Q = \{s \mid p(s) > 0\}$ is the support of the sampling design $p(.)$. A sampling design has high entropy when there is a high amount of uncertainty or high amount of surprise in the sample which will be selected. In other words, when a sampling design has high entropy, it is very difficult to predict the type of sample we would obtain. Many sampling designs used in practice are high entropy designs. One notable exception is systematic sampling that has a very low entropy. The concept of entropy is useful in the context of variance estimation. When a sampling design has a high entropy, it is possible to obtain approximation of the second-order inclusion probabilities, $\pi_{k\ell}$, in terms of the first-order inclusion probabilities, which simplifies considerably the problem of variance estimation in the context of unequal probability sampling; *e.g.*, Brewer and Donadio (2003), Matei and Tillé (2005), Henderson (2006) and Haziza, Mecatti and Rao (2008).

It is well known that the sampling design with maximum entropy is Poisson sampling:

$$p_{\text{poiss}}(s) = \left(\prod_{k \in s} \pi_k \right) \left(\prod_{k \in U \setminus s} (1 - \pi_k) \right) \qquad (2)$$

for all $s \in Q$; *e.g.*, Tillé (2006). A special case of Poisson sampling is Bernoulli sampling, which is obtained from (2) by setting $\pi_k = \pi \in (0, 1)$, which leads to

$$p_{\text{bern}}(s) = \pi^{n_s} (1 - \pi)^{N - n_s}, \text{ for all } s \subset U,$$

where n_s is the random size of s. Using (1) and noting that $\sum_{s \in Q} n_s p(s) = N\pi$, the entropy of Bernoulli sampling is given by

$$I(p_{\text{bern}}) = -N(1 - \pi)\log(1 - \pi) - N\pi \log \pi, \qquad (3)$$

which is maximum when $\pi = 1/2$. In this case, we have $I(p_{\text{bern}}) = N \log 2$.

If we restrict to the class of fixed size sampling designs with first-order inclusion probabilities π_k, $k \in U$, the maximum entropy design is the so-called Conditional Poisson Sampling (CPS); (see Chen, Dempster and Liu 1994; Deville 2000; Tillé 2006). The CPS design can be implemented by repeatedly selecting samples according to Poisson sampling until the desired sample size, n (say), has been obtained. When $\pi_k = n/N$ for all $k \in U$, the CPS design reduces to simple random sampling without replacement:

$$p_{\text{srs}}(s) = \binom{N}{n}^{-1}$$

for all $s \in Q$. From (1), it follows that the entropy of simple random sampling is given by

$$I(p_{\text{srs}}) = \log N! - \log n! - \log(N - n)!. \qquad (4)$$

In other words, simple random sampling without replacement is the maximum entropy design in the class of equal probability fixed size sampling designs.

1. Yves Tillé, Institut de Statistique, University of Neuchâtel, Pierre à Mazel 7, 2000 Neuchâtel, Switzerland; David Haziza, Département de mathématiques et de statistique, Université de Montréal, Montréal, QC, Canada, H3C 3J7. E-mail: haziza@dms.umontreal.ca.

Not all sampling designs possess a high entropy. For example, the 1-in-G systematic sampling design has a very low entropy. Here, the number of samples, $G = N/n$, is assumed to be an integer value. Since $p_{\text{syst}}(s) = 1/G$ for all $s \in Q$, the entropy of systematic sampling is given by

$$I(p_{\text{syst}}) = \log N - \log n,$$

which is much smaller than (4), especially for large values of N.

2. Main result

In this section, we compare the entropy of Bernoulli sampling with that of simple random sampling without replacement. Since the support of the Bernoulli sampling designs is much larger than that of simple random sampling without replacement, we expected the entropy of Bernoulli sampling to be much larger than that of simple random sampling without replacement. Table 1 shows the entropy for simple random sampling and Bernoulli sampling for different values of N and π. Surprisingly, we found the entropy of both sampling designs for the same inclusion probabilities and the same sample size to be approximately equal. From Table 1, it is clear that both sampling designs have similar entropies, even for moderate population sizes (*e.g.*, $N = 100$), independently of the value of π. This result is somehow curious considering the strong reduction of possible samples by fixing the sample size. Indeed, recall that the size of the support is $\binom{N}{n}$ for simple random sampling without replacement, whereas it is 2^N for Bernoulli sampling. For example, for $N = 100$ and $n = 20$, the size of the support for simple random sampling without replacement is equal to $\binom{100}{20} \approx 5.36 \times 10^{20}$, whereas it is equal to $2^{100} \approx 1.26 \times 10^{30}$ for Bernoulli sampling. In other words, the size of the support of Bernoulli sampling is approximately 2.36×10^9 larger than that of simple random sampling without replacement.

Result 1. *Let* $I(p_{\text{bern}})$ *and* $I(p_{\text{srs}})$ *be the entropy for Bernoulli sampling and simple random sampling without replacement*, *respectively given by* (3) *and* (4). *Then*,

$$\lim_{N\to\infty} \frac{I(p_{\text{srs}})}{I(p_{\text{bern}})} = 1.$$

Proof. By considering Stirling's formula (see Abramowitz and Stegun 1964, page 257)

$$\lim_{n\to\infty} \frac{n\log n - n}{\log n!} = 1,$$

we get

$$\lim_{\substack{N\to\infty \\ n\to\infty \\ N-n\to\infty}} \frac{N\log N - n\log n - (N-n)\log(N-n)}{\log\binom{N}{n}} = 1,$$

from which we obtain

$$\lim_{N\to\infty} = \frac{\log\binom{N}{N\pi}}{-N(1-\pi)\log(1-\pi) - N\pi\log\pi} = 1.$$

3. Conclusion

In this note, we showed that Bernoulli sampling and simple random sampling without replacement have very similar entropies, even for moderate population sizes. We conjecture that the same should be observed when comparing the Poisson sampling design and the CPS design for a given set on first-order inclusion probabilities. However, the proof of this result seems to be considerably more complex.

Table 1
Entropy of (Bernoulli sampling, simple random sampling) designs

N	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$	$\pi = 0.4$	$\pi = 0.5$
10	(3.3, 2.3)	(5, 3.8)	(6.1, 4.8)	(6.7, 5.3)	(6.9, 5.5)
100	(32.5, 30.5)	(50, 47.7)	(61.1, 58.6)	(67.3, 64.8)	(69.3, 66.8)
1,000	(325.1, 321.9)	(500.4, 496.9)	(610.9, 607.3)	(673, 669.4)	(693.1, 689.5)
10,000	(3,250.8, 3,246.5)	(5,004, 4,999.4)	(6,108.6, 6,103.9)	(6,730.1, 6,725.3)	(6,931.5, 6,926.6)
100,000	(32,508.3, 32,502.8)	(50,040.2, 50,034.5)	(61,086.4, 61,080.5)	(67,301.2, 67,295.2)	(69,314.7, 69,308.7)
1,000,000	(325,083, 325,076)	(500,402, 500,396)	(610,864, 610,857)	(673,012, 673,005)	(693,147, 693,140)

Acknowledgements

We thank an Associate Editor and a referee for constructive comments. Work of David Haziza was partially supported by grants from the Natural Sciences and Engineering Research Council of Canada.

References

Abramowitz, M., and Stegun, I.A. (1964). *Handbook of Mathematical Functions*. New York: Dover.

Brewer, K.R.W., and Donadio, M.E. (2003). The high entropy variance of the Horvitz-Thompson estimator. *Survey Methodology*, 29, 189-196.

Chen, S.X., Dempster, A.P. and Liu, J.S. (1994). Weighted finite population sampling to maximize entropy. *Biometrika*, 81, 457-469.

Deville, J.-C. (2000). Note sur l'algorithme de Chen, Dempster et Liu. Technical report, CREST-ENSAI, Rennes.

Haziza, D., Mecatti, F. and Rao, J.N.K. (2008). Evaluation of some approximate variance estimators under the Rao-Sampford unequal probability sampling design. *Metron*, 66, 91-108.

Henderson, T. (2006). Estimating the variance of the Horvitz-Thompson estimator. Master's thesis, School of Finance and Applied Statistics, The Australian National University.

Matei, A., and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21, 4, 543-570.

Tillé, Y. (2006). *Sampling Algorithms*. New York: Springer.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following people who have provided help or served as referees for one or more papers during 2010.

P. Ardilly, *INSEE*
P. Beatty, *National Center for Health Statistics*
J.-F. Beaumont, *Statistics Canada*
C. Bocci, *Statistics Canada*
G. Brackstone
J. van den Brakel, *Statistics Netherlands*
J.M. Brick, *Westat Inc*
P.D. Brick, *Westat Inc*
C. Calder, *Ohio State University*
P. Cantwell, *U.S. Bureau of the Census*
R. Chambers, *Centre for Statistical and Survey Methodology*
P. Dick, *Statistics Canada*
A.H. Dorfman, *U.S. Bureau of Labour Statistics*
G. Dubreuil, *Statistics Canada*
M. Elliott, *University of Michigan*
J.L. Eltinge, *U.S. Bureau of Labor Statistics*
G. Falk, *University of Virginia*
M. Fay, *National Institute of Allergy and Infectious Diseases*
W.A. Fuller, *Iowa State University*
J. Gambino, *Statistics Canada*
N. Ganesh, *National Opinion Research Center at the University of Chicago*
S. Ghosh, *Alberta Health Services-Cancer care*
C. Girard, *Statistics Canada*
S. Godbout, *Statistics Canada*
R. Griffin, *U.S. Census Bureau*
D. Haziza, *Université de Montréal*
Y. He, *Harvard Medical School*
S. Heeringa, *University of Michigan*
M. Hidiroglou, *Statistics Canada*
S. Holan, *University of Missouri*
A. Holmberg, *Statistics Sweden*
D. Hubble, *Westat Inc*
B. Hulliger, *University of Applied Sciences Northwestern Switzerland*
J. Jones, *Office for National Statistics, United Kingdom*
D. Judkins, *Westat Inc*
D. Kasprzyk, *Mathematica Policy Research*
P. Kelly, *Statistics Canada*
P. Kott, *RTI International*
P. Lahiri, *JPSM, University of Maryland*
M.D. Larsen, *George Washington University*
P. Lavallée, *Statistics Canada*
J. Legg, *Amgen Inc., USA*
R. Little, *University of Michigan*
S. Lohr, *Arizona State University*
P. Lynn, *Univerity of Essex*
M. Maia, *Catholic University of Portugal*
D.J. Malec, *U.S. Census Bureau*
J. Maples, *U.S. Census Bureau*
E. Martin, *U.S. Census Bureau*
K. Miller, *U.S. National Center for Health Statistics*
T. Mulcahy, *National Opinion Research Center*
G. Nathan, *Hebrew University*
S.F. Nielsen, *Copenhagen Business School, Denmark*
A. Nigam, *Institute of Applied Statistics and Development Studies*
J. Opsomer, *Colorado State University*
Z. Patak, *Statistics Canada*
D. Pfeffermann, *Hebrew University*
N.G.N. Prasad, *University of Alberta*
M. Pratesi, *Università di Pisa*
J.N.K. Rao, *Carleton University*
J. Reiter, *Duke University*
L.-P. Rivest, *Université Laval*
S. Rubin-Bleuer, *Statistics Canada*
J. Ryten
N. Salvati, *Università di Pisa*
C.-E. Särndal, *Université de Montréal*
N. Schenker, *National Center for Health Statistics*
F.J. Scheuren, *National Opinion Research Center*
P. do N. Silva, *Escola Nacional de Ciências Estatísticas*
P. Smith, *Office for National Statistics*
T.M.F. Smith, *University of Southampton, UK*
E. Stasny, *Ohio State University*
D. Steel, *University of Wollongong*
L. Stokes, *Southern Methodist University*
A. Théberge, *Statistics Canada*
M. Thompson, *University of Waterloo*
S. Thompson, *Simon Fraser University*
D. Toth, *U.S. Bureau of Labor Statistics*
C. Tucker, *U.S. Bureau of Labor Statistics*
V.J. Verma, *Università degli Studi di Siena*
W.E. Winkler, *U.S. Census Bureau*
K.M. Wolter, *Iowa State University*
C. Wu, *University of Waterloo*
W. Yung, *Statistics Canada*
P.A. Zandbergen, *University of New Mexico*
A. Zaslavsky, *Harvard Medical School*

Acknowledgements are also due to those who assisted during the production of the 2010 issues: Céline Ethier of Statistical Research and Innovation Division, Christine Cousineau and Teresa Jewell of Household Survey Methods Division, Nick Budko and Sophie Chartier of Business Survey Methods Division, Matthew Belyea, Louise Demers, Anne-Marie Fleury, Roberto Guido, Liliane Lanoie, Darquise Pellerin of Dissemination Division, and Jeff Jodoin of Client Services Division.


ELECTRONIC PUBLICATIONS AVAILABLE AT
PUBLICATIONS ÉLECTRONIQUES DISPONIBLE À
www.statcan.gc.ca

ANNOUNCEMENTS

Nominations Sought for the 2012 Waksberg Award

The journal *Survey Methodology* has established an annual invited paper series in honour of Joseph Waksberg to recognize his contributions to survey methodology. Each year a prominent survey statistician is chosen to write a paper that reviews the development and current state of an important topic in the field of survey methodology. The paper reflects the mixture of theory and practice that characterized Joseph Waksberg's work.

The recipient of the Waksberg Award will receive an honorarium from Westat. The paper will be published in a future issue of *Survey Methodology*.

The author of the 2012 Waksberg paper will be selected by a four-person committee appointed by *Survey Methodology* and the American Statistical Association. Nomination of individuals to be considered as authors or suggestions for topics should be sent before February 28, 2011 to the chair of the committee, Elizabeth Martin (betsy@folhc.org).

Previous Waksberg Award honorees and their invited papers are:

2001 Gad **Nathan**, "Telesurvey methodologies for household surveys – A review and some thoughts for the future?". *Survey Methodology*, vol. 27, 1, 7-31.

2002 Wayne A. **Fuller**, "Regression estimation for survey samples". *Survey Methodology*, vol. 28, 1, 5-23.

2003 David **Holt**, "Methodological issues in the development and use of statistical indicators for international comparisons". *Survey Methodology*, vol. 29, 1, 5-17.

2004 Norman M. **Bradburn**, "Understanding the question-answer process". *Survey Methodology*, vol. 30, 1, 5-15.

2005 J.N.K. **Rao**, "Interplay between sample survey theory and practice: An appraisal". *Survey Methodology*, vol. 31, 2, 117-138.

2006 Alastair **Scott**, "Population-based case control studies". *Survey Methodology*, vol. 32, 2, 123-132.

2007 Carl-Erik **Särndal**, "The calibration approach in survey theory and practice". *Survey Methodology*, vol. 33, 2, 99-119.

2008 Mary E. **Thompson**, "International surveys: Motives and methodologies". *Survey Methodology*, vol. 34, 2, 131-141.

2009 Graham **Kalton**, "Methods for oversampling rare subpopulations in social surveys". *Survey Methodology*, vol. 35, 2, 125-141.

2010 Ivan P. **Fellegi**, "The organisation of statistical methodology and methodological research in national statistical offices". *Survey Methodology*, vol. 36, 2, 123-130.

2011 Danny **Pfeffermann**, Manuscript topic under consideration.

Members of the Waksberg Paper Selection Committee (2010-2011)

Elizabeth A. Martin (Chair)
Mary Thompson, *University of Waterloo*
J.N.K. Rao, *Carleton University*
Steve Heeringa, *University of Michigan*

Past Chairs:

Graham Kalton (1999 - 2001)
Chris Skinner (2001 - 2002)
David A. Binder (2002 - 2003)
J. Michael Brick (2003 - 2004)
David R. Bellhouse (2004 - 2005)
Gordon Brackstone (2005 - 2006)
Sharon Lohr (2006 - 2007)
Robert Groves (2007 - 2008)
Leyla Mojadjer (2008 - 2009)
Daniel Kasprzyk (2009 - 2010)

The Canadian Journal of Statistics **La revue canadienne de statistique**

CONTENTS TABLE DES MATIÈRES

Volume 38, No. 2, June/juin 2010

Special Issue: Inferences in Generalized Linear Longitudinal Mixed Models

Volume 38, No. 3, September/septembre 2010

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 26, No. 1, 2010

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 26, No. 2, 2010

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 26, No. 3, 2010

All inquires about submissions and subscriptions should be directed to jos@scb.se