Methodological Advisory Committee
November 2001

Using the Multinomial Logistic Regression Model to Construct Design-Based Estimates for the Input-Output Survey via an Estimating Equation Approach

Authors: Stephen Carlton and Paul Schubert
c/ABS, PO Box 10, Belconnen, ACT 2616
fax: (02) 6252 8015
email: stephen.carlton@abs.gov.au, ph: (02) 6252 6426
email: paul.schubert@abs.gov.au, ph: (02) 6252 5140

PART 1: EXECUTIVE SUMMARY

1. One strategy for managing small business provider load is to ask small businesses for a reduced set of data items and then use synthetic estimation techniques to estimate totals for those data items not collected. This strategy was adopted by the Australian Bureau of Statistics (ABS) in its design of the Input-Output (I-O) Survey. Small businesses are only required to provide detailed expense items if they fall in specifically targeted industries; otherwise they provide *other operating expenses* — the sum of the detailed expense items — and a number of auxiliary variables.

2. For small businesses in industries not targeted for the I-O survey, synthetic estimation techniques are used to estimate totals of the detailed expense data items. The multinomial logistic regression model has been adopted to motivate their construction for the following reasons:

- auxiliary information can be used;
- synthetic estimates of detailed expense items produced by this method are
  - non-negative,
  - less than or equal to *other operating expenses*, and
  - sum to *other operating expenses*.

3. The multinomial logistic regression model can be used to motivate an alternative to the generalised linear regression estimator (GREG). This is the logistic generalised regression estimator (LGREG). It can be used to estimate totals of detailed expense items from the sample selected for the I-O Survey.

4. The design-based paradigm has been adopted throughout. The multinomial regression model has been used as a device to motivate estimating equations for finite population parameters which are used in the construction of estimates.

Structure of this paper

5. This paper has been divided into three sections:

- Part 1 (this part) which is an executive summary;
- Part 2 which gives an outline of the estimation problem and our proposed method of estimation, illustrates an application using a case study, and discusses our planned future development of the methodology; and
- Part 3 which gives a theoretical treatment of logistic generalised regression estimation and its application to the Input-Output Survey.

Discussion points for MAC

6. Questions for MAC to consider are listed at appropriate places within the paper and are repeated here for easy reference.

Q1: Are our proposed crude Taylor Series variance estimators good enough? Are there any better synthetic variance estimators?
Q2: We can only measure the variance part of Mean Square Error (MSE). Can we get bias estimates?
Q3: How can we take advantage of updated auxiliary information to get estimates "across years"?
Q4: Should the I-O sampling strategy change in any way?
Q5: Does the concept of variable selection for a best GREG make sense? What is "best GREG" (or LGREG)?
Any other comments that MAC has on the methodology or its application are welcome.

PART 2: THE METHODOLOGICAL PROBLEM

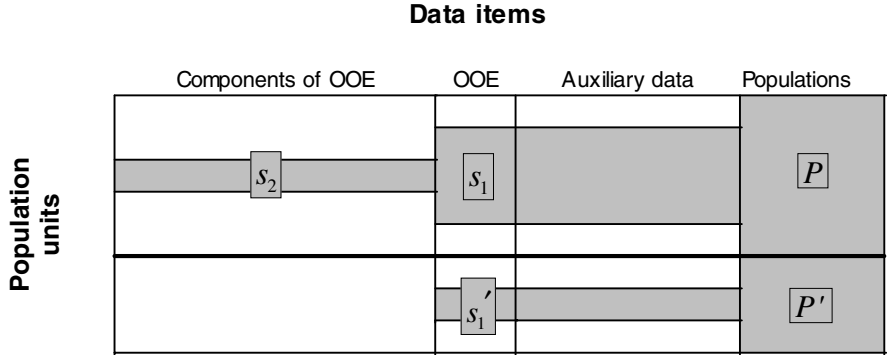
Background to the estimation problem

7. Estimates at a detailed level are needed in the compilation of the ABS' annual Input-Output (I-O) tables. Typically, data are required at the industry level for a considerable number of data items, which may be income data items, expense data items, or commodities. For the particular application this paper focuses on, estimates of about 25-30 detailed expense data items are needed for about 40 service industries. Resource and provider load constraints prohibit the collection of these detailed data from every industry every year from a large enough probability sample to get purely design-based estimates of useful precision.

8. The response to this problem has been to obtain the detailed expense data items only from large businesses every year. For the small businesses, a "rolling industry" sampling strategy has been put in place, whereby only about a quarter of all the service industries are sampled with respect to the detailed expense data items each year. Thus each industry is sampled once every four years. This makes the I-O Survey, as the survey which collects these detailed expense items is known, a partial coverage survey with a sampled component. There is still the need to estimate totals for the detailed expense items for both small and large businesses in *all* industries.

9. The sample of the I-O Survey is actually a subset of the sample of the Economic Activity Survey (EAS). The first-phase sample, EAS, collects some broad expense data items across all service industries each year, one of which is called *other operating expenses* (OOE). OOE are simply the expenses which do not fall into any of the other EAS categories. The 25-30 detailed expense data items collected in the second-phase sample (the I-O Survey) conceptually sum to OOE.

10. Diagrammatically, data availability looks like this.



11. The middle two columns represent the data that are collected by EAS. OOE is one of the EAS data items, and the others can be considered as auxiliary data for the purposes of producing estimates of the detailed expense items. The right hand column shows that the population of businesses in the service industries covered by the frame can be partitioned into two parts, P and P'. P can be thought of as the EAS population from which units can be selected for I-O. The remainder of the

population, *P'*, can be thought of as the EAS population from which units will not be selected for I-O. Note that s_1 and s_1' make up the full EAS sample (the subscipt '1' denoting the first-phase sample).

12. The detailed expense items (i.e. components of OOE, the first column in the diagram) are available only for businesses in s_2, the Input-Output sample. They are collected from a sample of the "big" businesses each year in each industry by the I-O Survey. This sample typically has a completely enumerated component. However, since these data are collected from "small" businesses on a rolling industry basis, two cases for detailed expense item availability arise.

- The first is that sample data from the small businesses are collected in that year. In this case the I-O sample, denoted by s_2 in the above diagram, is taken from the full population for that industry. In other words, the full population for that industry falls into *P* under this scenario, and none of it falls in *P'*.
- The second is that sample data from the small businesses are not collected in that year, and so the only data we have in that year for that industry are from the big businesses. In this scenario we have a partial coverage survey, where the large businesses fall into population *P* and the small businesses fall into *P'*.

13. With respect to the above diagram, note that the following relationships exist :

- $s_2 \subseteq s_1 \subseteq P_1$,
- $s_1' \subseteq P'$,
- $P \cap P' = \varnothing$,
- $P \cup P'$ is the full EAS population, and
- $s_1 \cap s_1' = \varnothing$.

14. The key problem that this paper addresses is to produce the required estimates of total for each of the detailed expense data items for *P* and *P'*. This estimation problem can be dealt with by splitting the EAS population into these two components, those businesses which can be selected for the Input-Output Survey and those which will be deliberately excluded. For the first population, detailed expense item estimates can be produced directly (possibly using auxiliary information) since the relevant data items are collected. For the second population, we have proposed a synthetic estimation approach since sample information on the detailed expense data items is unavailable. This approach must satisfy the requirement that the synthetically generated detailed breakdown of OOE adds to total OOE for individual businesses.

15. This paper shows how the multinomial logistic regression model can be used to provide solutions to the above problem. It extends the results of Lehtonen and Veijanen (1998) to two-phase sample designs and synthetic estimation. All estimators have been treated as Horvitz-Thompson estimators or generalised two-phase estimates, π estimators and π^* estimators in the terminology of Särndal et al (1992).

16. As was mentioned earlier, the sampling approach for the I-O Survey is two-phase with EAS being the first phase and I-O the second phase. Stratification between the two phases is based on the same (frame) variables, but the second-phase (I-O) stratification is broader. The second-phase (I-O) sample is selected with probabilities of selection proportional to the inverse probabilities of selection of the first-phase (EAS) sample. This is done so that units within the same second-phase (I-O) strata have equal (unconditional) probabilities of selection.

17. The following quantities are used in constructing estimates:

- π_{1i}, the (first-phase) probability of the i^{th} business in P being selected in s_1, or the probability of the i^{th} business in P' being selected in s_1';
- π_{2i}, the (second-phase) probability of the i^{th} business in s_1 being selected in s_2 given s_1; and
- $\pi_i^* = \pi_{1i}\pi_{2i}$, which is used in constructing two-phase estimates.

18. To give the expressions for π_{1i} and π_{2i}, let

n_h be the first-phase sample for first-phase stratum *h*,

N_h be the first-phase population for first-phase stratum *h*,

N_g be the population for second-phase stratum *g* (*g* is some broader grouping of the first-phase strata in this application, although it doesn't have to be), and

m_g be the required second-phase sample for second-phase stratum *g*.

Then, the first-phase probability of selection is simply

$$\pi_{1i} = \left(\frac{n_h}{N_h} \right)$$

and the probability of selection in the second-phase conditional on the first-phase sample s_1 is given by

$$\pi_{2i} = \left(\frac{m_g}{N_g} \right) / \left(\frac{n_h}{N_h} \right).$$

19. "Big" businesses (generally those with employment of 200 or more) are completely enumerated in EAS, while "small" businesses are sampled. There are around 2000 businesses in the big part, while around 8000-9000 businesses (of over 0.6 million businesses in the population) are sampled in the small part in EAS. The size of the second phase sample in industries which are surveyed by I-O in any given year is typically 50-75% of the size of the EAS sample in those industries.

Proposed method of estimation

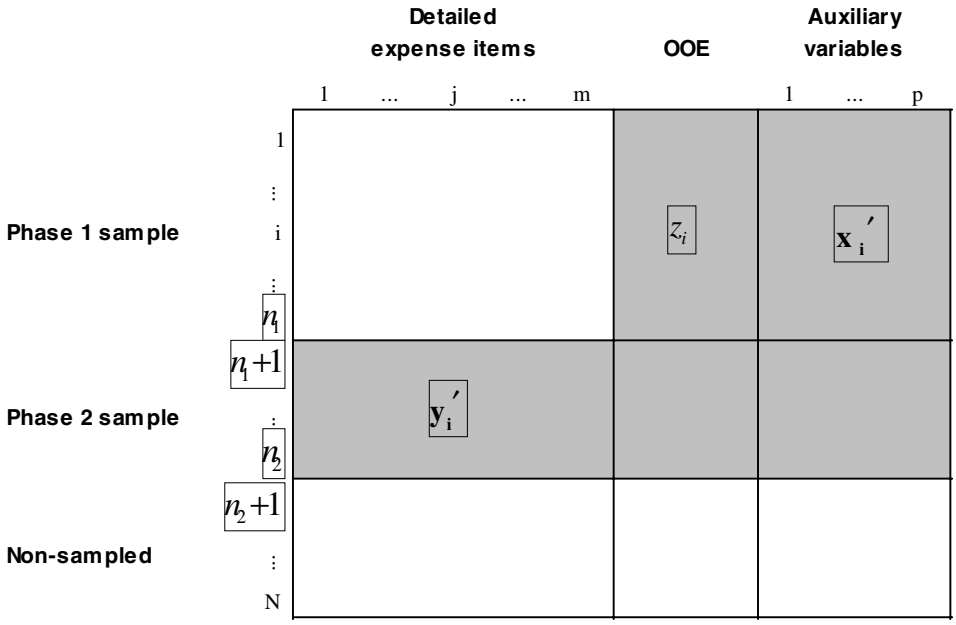
20. The two cases — estimating the total of the detailed expense data items for *P* and estimating the total of the detailed expense data items for *P'* (the main focus of this paper) — will be considered separately. In keeping with general ABS practice,

we have sought a methodology from which we can make design-based inferences. This is the fundamental difference from earlier approaches to this problem by Welsh and Szoldra (1997) and Rossiter (1998).

Estimating for P (second phase sample data available)

21. In this scenario, *P′* and s_1' in the above diagram are non-existent. The methodology up until now has been to produce a purely design-based estimate in the normal way. This method is inefficient as it does not utilise the auxiliary information available from the first-phase sample. It is possible to improve efficiency by using the auxiliary data to produce a logistic generalised regression estimator (LGREG). Lehtonen and Veijanen (1998) describe the LGREG approach under simple random sampling without replacement, and we have applied it to the two-phase case. The idea of LGREG is very similar to generalised regression estimators (GREG).

22. The following diagram introduces some of the notation used in the remainder of the paper. The EAS population could be thought of as a huge matrix, with businesses in the population forming the rows and data items (first or second phase) forming the columns. The shading indicates data availability.



23. There are *m* detailed expense data items and *p* auxiliary variables. Let $\mathbf{y}_i$ be a vector containing the values of the detailed expense items for business *i*, and $\mathbf{x}_i$ be a vector of the auxiliary data from EAS. In the above diagram, these vectors (or, more accurately, their transposes) correspond to row *i* of the "matrix". z_i, OOE for business *i*, is equal to the sum of the components of vector $\mathbf{y}_i$.

24. For a two-phase GREG, a model

$$\mathbf{E}_\xi(\mathbf{y}_i) = \mathbf{x}_i'\mathfrak{B}$$

(where $\mathbf{E}_\xi$ denotes expectation under the model) is used to motivate a design-based estimator of the following form:

$$\hat{\mathbf{Y}}_{\mathbf{G}} = \sum_{i \in s'} \frac{\hat{\mathbf{y}}_{\mathbf{i}}}{\pi_{1i}} + \sum_{i \in s_2} \frac{(\mathbf{y}_{\mathbf{i}} - \hat{\mathbf{y}}_{\mathbf{i}})}{\pi_i^*} .$$

Here, $\hat{\mathbf{y}}_{\mathbf{i}}$ are the fitted values obtained from the second-phase sample,

$$\hat{\mathbf{y}}_{\mathbf{i}} = \mathbf{x}_{\mathbf{i}}' \hat{\mathbf{B}} ,$$

π_{1i} are the probabilities of selection of the first-phase (EAS) sample and

$$\pi_i^* = \pi_{1i} \pi_{2i} .$$

The $\hat{\mathbf{B}}$ are coefficients estimated by a weighted least squares regression of the components of $\mathbf{y}_{\mathbf{i}}$ on $\mathbf{x}_{\mathbf{i}}$ using the second-phase data, where the regression is weighted by $\frac{1}{\pi_i^*}$ .The resulting model is also applied to generate predicted values $\hat{\mathbf{y}}_{\mathbf{i}}$ for all businesses in the first-phase sample.

25. The LGREG estimator has exactly the same form as the GREG estimator, except that instead of using linear regression, a multinomial logistic model is used. Instead of using the $\mathbf{y}_{\mathbf{i}}$ directly, the proportions $p_j(\mathbf{x}_{\mathbf{i}}; \mathfrak{B})$ — the expected proportion of OOE that falls in detailed expense item *j* for business *i* — are related to the detailed expense items by

$$\mathbf{E}_\xi(\mathbf{y}_{\mathbf{i}}) = z_i \mathbf{p}(\mathbf{x}_{\mathbf{i}}; \mathfrak{B})$$

where z_i is OOE and $\mathbf{p}(\mathbf{x}_{\mathbf{i}}; \mathfrak{B})$ is a vector with j^{th} element $p_j(\mathbf{x}_{\mathbf{i}}; \mathfrak{B})$. The multinomial logistic model is then

$$\log\left(\frac{p_j(\mathbf{x}_{\mathbf{i}}; \mathbf{B})}{p_m(\mathbf{x}_{\mathbf{i}}; \mathbf{B})} \right) = \mathbf{x}_{\mathbf{i}}' \mathbf{B} .$$

Fitted values using this model are simply

$$\hat{\mathbf{y}}_{\mathbf{i}} = r_i \mathbf{p}(\mathbf{x}_{\mathbf{i}}; \hat{\mathbf{B}}) .$$

The details of exactly how $\hat{\mathbf{B}}$ is obtained are left to Part 3 of this paper.

26. Note that $\hat{\mathbf{Y}}_{\mathbf{G}}$ is a design-consistent estimator of $\sum_{i \in P} \mathbf{y}_{\mathbf{i}}$. Variances would be calculated in the same way as for GREG i.e. using the weighted residuals method (see Särndal et al (1992)).

27. To date, we have outlined the theory for this method but have not implemented it in practice, primarily because we can use purely design-based estimators (e.g. the normal two-phase Horvitz-Thompson estimator).

Estimating for P' (second phase sample data NOT available)

28. In this case, we need to estimate

$$\sum_{i \in P} \mathbf{y_i} + \sum_{i \in P'} \mathbf{y_i}$$

where the first term can be estimated as above, but the $\mathbf{y_i}$ are not directly observed from *P'*. We have used a synthetic estimation approach to generate the required estimates, viz.

$$\hat{\mathbf{Y}}_{syn} = \sum_{i \in s_1'} \frac{\hat{\mathbf{y}}_\mathbf{i}}{\pi_{1i}}.$$

29. This time, the $\hat{\mathbf{y}}_\mathbf{i}$ are synthetic values obtained via the multinomial logistic model

$$\hat{\mathbf{y}}_\mathbf{i} = z_i \mathbf{p}(\mathbf{x_i}; \hat{\mathbf{B}})$$

where $\hat{\mathbf{B}}$ is estimated from s_2. Since we have no sample data from population *P'*, we have used sample data from population *P* to generate these synthetic values.

30. For the I-O rolling industry strategy, *P* in practice generally corresponds to the big businesses and *P'* to the small businesses. Therefore, an obvious assumption that this methodology makes the way we have implemented it is that the relationship between the detailed expense items and the auxiliary variables used is the same for both large and small businesses. A scarcity of reliable historical data for the small businesses has made this assumption difficult to evaluate thoroughly.

31. An estimate obtained in this way can then be added to an estimate for population *P* (e.g. as outlined in the previous section) to give an estimate for the total population.

32. We have calculated synthetic estimates in this way. We have also developed a methodology for calculating variances (see the following section, and Part 3 of this paper for more details), although this has not yet been implemented at the time of writing.

Case study: estimating for P'

33. In this section, we give an example of how the proposed method of estimation was applied in practice. We use synthetic estimation for the food retailing industry in respect to the 1997-98 financial year as our case study. We have sampled data from the adjacent years 1996-97 and 1998-99 with which to compare our synthetic estimates.

34. For 1997-98, there were a total of $m = 26$ detailed expense data items to be estimated. Only 23 large businesses were included in the second-phase (I-O) sample, and these were a subsample of the 64 large businesses from the first-phase (EAS) completely enumerated sector in the food retailing industry. No small businesses were selected in the second-phase (I-O) sample. However, the EAS completely enumerated sector accounted for about 56% of the total estimate of OOE.

35. For this example then, *P* is the EAS completely enumerated sector and *P'* is the remainder of the population. The population size for this industry, both large and small businesses, was about 39 400. Eight first-phase (EAS) variables were considered as possible auxiliary variables to be used in the logistic regression, and the three included were *total income*, *total capital expenditure*, and *operating profit before tax*. These were chosen by

- first considering all 28 possible combinations of two of the eight first-phase variables,
- choosing the three combinations of two that resulted in the lowest values of minus twice the pseudo-loglikelihood (i.e. greatest pseudo-loglikelihood), and
- evaluating all combinations of three variables which included the three best combinations of two variables from the previous step. Again, evaluations were in terms of the value of the pseudo-loglikelihood.

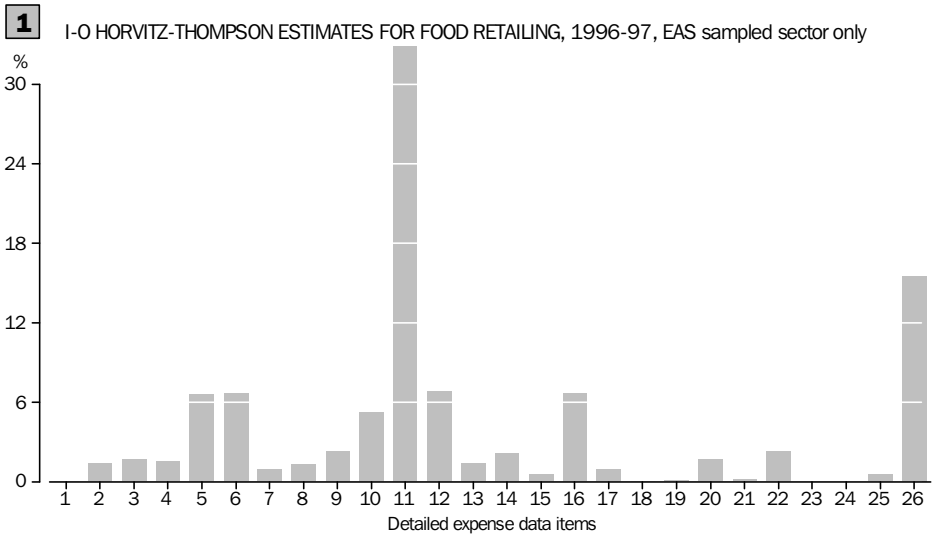
36. Note that there is no direct physical relationship between $\mathbf{x}$ and $\mathbf{y}$. The detailed expense data items are expenses such as *fringe benefits tax*, *motor vehicle running expenses*, *advertising expenses* and *stationery expenses*. It is questionable whether the available auxiliary variables from EAS are very useful in an efficient decomposition of OOE into the detailed expense data items.

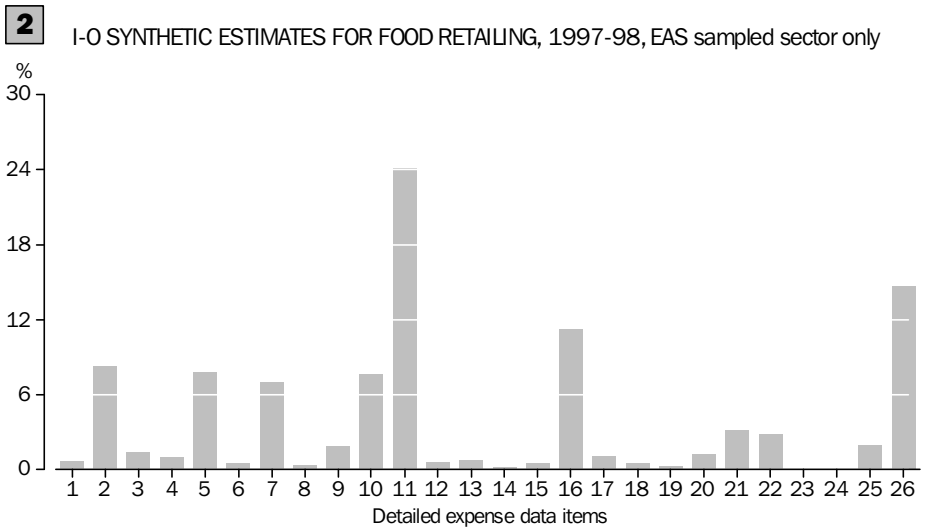
37. The SAS procedure CATMOD was used to carry out the multinomial logistic regressions. Although designed primarily for use with categorical explanatory variables, it can be used with continuous auxiliary variables as we had by specifying this fact in a DIRECT statement. Pseudo-maximum likelihood was used to estimate the regression coefficients. The SAS interactive matrix language (PROC IML) was used to manipulate the output of PROC CATMOD to produce the final synthetic estimates. PROC IML will also be needed to produce variances.

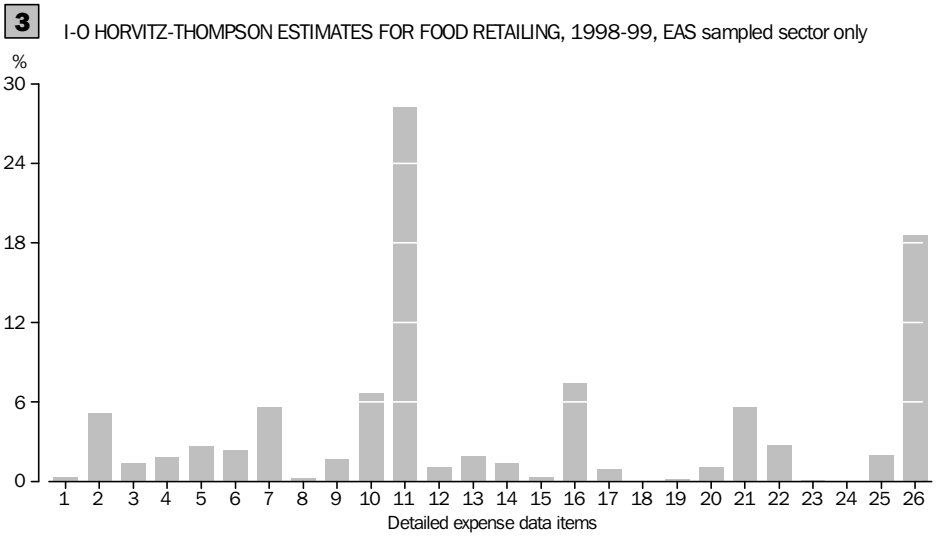
38. To incorporate the sampling weights in the PROC CATMOD calculations of the pseudo-loglikelihood and regression coefficients as specified in our proposed method of estimation, we had to do some manipulation of the input data. One way input data can be supplied is as "cell count" (frequency) data, where each observation is a cell in a contingency table. Under this set-up, each responding business has to have a separate record for each of the detailed expense data items. The variable used in the WEIGHT statement contains the value for the relevant detailed expense data item. We needed to multiply this variable by the sampling weight.

39. While this procedure gives the desired pseudo-loglikelihood and regression coefficients, other output statistics from PROC CATMOD such as standard errors of the parameter estimates and chi-square (Wald statistic) values are not correct for our application. These are a function of the variance of our estimates, and are measures we have yet to calculate.

40. The three graphs below show the synthetic estimates for *P'* for the food retailing industry for 1997-98 compared with design-based Horvitz-Thompson estimates for *P'* from neighbouring years, 1996-97 and 1998-99. The graphs display the percentage of OOE falling in each of the detailed expense data items.







41. The graphs show that the synthetic estimates from 1997-98 give a similar pattern overall to the two adjacent years, although there are some differences. For example, the percentage of OOE in data item 11 is less from the synthetic estimation, and the percentages in data items 2, 7 and 16 are higher from the synthetic estimation. It is difficult to be sure whether these differences are real

differences in business expenses across the three years, or whether they are due to deficiencies in the estimation methodology. The Horvitz-Thompson estimates from 1996-97 in particular are subject to large standard errors, as the sample size was not large (around 100).

42. It was noticeable when looking at the regression coefficients that the intercept term was dominating; in fact, the intercept term alone generally accounted for well over 90% of each synthetic value. However, this does not mean that the synthetic estimates using three auxiliary variables were the same as those that would have been obtained using no auxiliary variables but only an intercept term. The values of the auxiliary variables in the EAS completely enumerated sector (i.e. the data on which the logistic regression was done) were sufficiently large to have an effect, although this was not the case in the EAS sampled sector (i.e. *P',* where the synthetic values were being generated).

43. The two main problems we encountered in fitting these regressions in practice were the occurrence of zeros or very small values in the detailed expense data items, and the occurrence of zeros in the auxiliary variables. Where we had zero values for a particular detailed expense items reported for all units in our first-phase sample, we excluded the item from the logistic regression and "set" the synthetic estimate for that detailed expense item to be zero. However, because of the number of categories involved, the probabilities (the logit of which we are modelling) will be often close to zero for many of the 26 items, even where there are non-zero values observed in the first-phase sample.

44. The second problem occurred to some extent in this food retailing example where our second auxiliary variable, *total capital expenditure*, was zero for about two thirds of the responding businesses.

45. These problems could show themselves in a number of ways. For example, in the SAS output, the regression coefficient estimates were sometimes asterisked, indicating that the parameter estimates were unstable. In these cases, we also observed that one of the detailed expense items often was allocated a very high proportion of OOE (over 90%), and using other auxiliary variables resulted in a different detailed expense data item being allocated a very high proportion of OOE. In other cases, the PROC CATMOD algorithm did not converge.

46. The solution to the first problem in two separate earlier pieces of work by Welsh and Szoldra (1997) and Rossiter (1998), both of which were presented to the Methodological Advisory Committee, was to collapse the detailed expense data items into a smaller number (typically six to eight). This is not really satisfactory, as estimates for the full 26 are required separately for Input-Output tables. Since the problems tend to arise more often when we are dealing with small sample sizes, our solution generally was to combine similar industries together to effectively have a larger sample size. Occasionally, we used a lesser number of auxiliary variables in the logistic regression e.g. only one, or even just the intercept term, which is equivalent to applying to *P'* the same proportions of OOE in the detailed expenses items as observed in *P.*

47. We expect small-sample related problems to occur less often in the future, as one of the benefits of the rolling industry strategy introduced from 1998-99 is that there is a larger sample size allocated to specific industries that are sampled, rather than spreading the small sample thinly across all industries.

Future directions

Variance estimation for P'

48. Due to the way the second phase (I-O) sample is selected — so that it "looks" like a single phase stratified simple random sample (SRS) — variances that have been calculated in the past for I-O design-based estimates have simply been those appropriate for single phase stratified SRS designs. We have done the same for this project.

49. There are two components of sampling variability in our synthetic estimates. The first is the normal one associated with samples i.e. the fact that we are computing estimates and making inferences about our population of interest from a random sample. The second is that the coefficients $\hat{\mathbf{B}}$ used in generating an LGREG were themselves based on sample data, and a different sample may have resulted in different coefficients.

50. Due to complexities in calculations, our proposed variance estimator is a crude "zeroth order" Taylor Series approximation assuming single phase stratified SRS (see Part 3 for details). The zeroth order approximation captures the first component of sampling variability mentioned above but not the second. Further, we would ideally like to have an estimate of bias although it is not clear how to do this.

Q1: Are our proposed crude Taylor Series variance estimators good enough? Are there any better synthetic variance estimators?

Q2: We can only measure the variance part of Mean Square Error (MSE). Can we get bias estimates?

Making use of updated auxiliary information for P'

51. Part of the thinking behind the move to a rolling industry sampling strategy was, in years where industries are not sampled, that information from the most recent sampled year could be used in conjunction with incomplete data from the current year to provide "updated" estimates. The crude methodology that has been used to date is an example of this. What is done is simply that the proportion of OOE that falls in each of the detailed expense data items from the most recent sampled year are brought forward and applied to the OOE in the current year.

52. We would like to improve on this methodology by applying updated auxiliary information from the current year's first phase (EAS) sample to the model developed in the most recent year where second phase data are available. In other words, if we defined the population components *P* and *P'* in terms of what is covered by the I-O

sample in the current year for any given industry, then we would use $\hat{\mathbf{B}}$ coefficients from the most recent s_2 data representing P' (which may be from 1-3 years out of date) but use the current year's auxiliary data . Mathematically, if we are currently in year t and we have s_2 data from year t-1, then synthetic values would be given by

$$\hat{\mathbf{y}}_{\mathbf{i,t}} = z_{i,t}\, p(\mathbf{x}_{\mathbf{i,t}}; \hat{\mathbf{B}}_{\mathbf{t\text{-}1}}) .$$

This would allow for the decomposition of OOE into its detailed expense items to change across years, which would be an improvement over the current methodology if the auxiliary variables were useful predictors in the models. We haven't yet thought about variance estimation for such an estimator.

53. [As an aside, we note that the current crude approach can be put into the methodological framework outlined in the previous paragraph by having only one auxiliary variable which is an indicator variable denoting industry membership for each business in the first phase sample.]

Q3: How can we take advantage of updated auxiliary information to get estimates "across years"?

54. It could also be that we should change the rolling industry sampling strategy in some way to make it easier to produce these estimates. An example may stem from a problem that we have already encountered, which is to ensure there is sufficient sample to support the estimation of a large number of regression coefficients. Our estimation has been at the industry level, and we have dealt with this problem so far by collapsing across industries where necessary. Perhaps our sample design should incorporate a minimum sample size per industry to avoid this problem in the future.

Q4: Should the I-O sampling strategy change in any way?

Variable selection in a design-based framework

55. The theoretical properties of GREG (or LGREG) estimation methodology are based on the assumption that the choice of auxiliary variables does not depend on the sample chosen. In traditional model building, the choice of auxiliary information is determined via a variable selection algorithm which minimises a criterion balancing accuracy and parsimony. While a similar concept can be introduced under the design-based framework where variable selection is based on a balance between accuracy (as measured by *estimated* variance) and parsimony, it is not clear how this affects the true design mean squared error of estimates.

Q5: Does the concept of variable selection for a best GREG make sense? What is "best GREG" (or LGREG)?

Other issues

56. Any other comments that MAC has on the methodology or its application would be welcome.

PART 3: DESIGN-BASED ESTIMATORS USING ESTIMATING EQUATIONS DERIVED FROM MULTINOMIAL LOGISTIC REGRESSION

57. Estimates of totals for the detailed expense items $\sum_{i\in P}\mathbf{y_i}$ and $\sum_{i\in P'}\mathbf{y_i}$ are required to produce estimates for the full EAS population. The multinomial logistic model (discussed in the next section) is used to motivate design-based estimators which make use of the auxiliary information. An alternative prediction theory approach was adopted by Welsh and was the basis of earlier work carried out by the ABS (Welsh and Szoldra (1997)).

The multinomial logistic model

58. This is just an extension of logistic regression to the case where the response variable can take on more than two values.

59. The multinomial logistic regression model can be specified by

$$\mathbf{y_i} = \begin{bmatrix} y_{i,1} \\ y_{i,2} \\ \vdots \\ y_{i,m} \end{bmatrix} \sim multinomial\ log\ istic\left(z_i, \mathbf{p}\left(\mathbf{x_i};\mathfrak{B}\right)\right)$$

where

$$\mathbf{p}\left(\mathbf{x_i};\mathfrak{B}\right) = \begin{bmatrix} p_1\left(\mathbf{x_i};\mathfrak{B}\right) \\ p_2\left(\mathbf{x_i};\mathfrak{B}\right) \\ \vdots \\ p_m\left(\mathbf{x_i};\mathfrak{B}\right) \end{bmatrix}$$

is an $m \times 1$ vector,

$$\mathfrak{B} = [\beta_1 \mid \beta_2 \mid .. \mid \beta_{m-1}]$$

is a $p \times (m\text{-}1)$ matrix of parameters to be estimated (each β_j is a $p \times 1$ vector), and

$$log\left(\frac{p_j\left(\mathbf{x_i};\mathfrak{B}\right)}{p_m\left(\mathbf{x_i};\mathfrak{B}\right)}\right) = \mathbf{x_i}'\beta_j \quad , j=1,\ldots, m\text{-}1$$

is the generalised logit link function.

60. This gives

$$p_j\left(\mathbf{x_i};\mathfrak{B}\right) = \frac{exp\left(\mathbf{x_i}'\beta_j\right)}{1+\sum_{j=1}^{m-1} exp\left(\mathbf{x_i}'\beta_j\right)} \quad , j=1, \ldots, m\text{-}1$$

and

$$p_m\left(\mathbf{x_i};\mathfrak{B}\right)=\frac{1}{1+\sum_{j=1}^{m-1}exp\left(\mathbf{x_i}'\boldsymbol{\beta}_\mathbf{j}\right)}.$$

61. The vectors

$$\mathbf{y_i^{\#}}=\begin{bmatrix} y_{i,1} \\ y_{i,2} \\ : \\ y_{i,m-1} \end{bmatrix}$$

and

$$\mathbf{p}^{\#}\left(\mathbf{x_i};\mathfrak{B}\right)=\begin{bmatrix} p_1\left(\mathbf{x_i};\mathfrak{B}\right) \\ p_2\left(\mathbf{x_i};\mathfrak{B}\right) \\ : \\ p_{m-1}\left(\mathbf{x_i};\mathfrak{B}\right) \end{bmatrix}$$

(i.e. with the m^{th} element omitted) will also be needed when defining the score function and information matrix.

62. The observed data log-likelihood from the Input-Output sample can be defined as

$$L\left(\mathfrak{B}\right)=\sum_{i\in s_2}\sum_{j=1}^{m}y_{ij}\,log\;p_j\left(\mathbf{x_i};\mathfrak{B}\right).$$

The score function is just

$$\mathbf{S}\left(\mathfrak{B}\right)=\frac{\partial L}{\partial vec\mathfrak{B}}=\sum_{i\in s_2}\left[\mathbf{y_i^{\#}}-z_i\mathbf{p}^{\#}\left(\mathbf{x_i};\mathfrak{B}\right)\right]\otimes\mathbf{x_i},$$

a $p(m\text{-}1)$ vector. The information matrix is given by

$$\mathbf{I}\left(\mathfrak{B}\right)=-\frac{\partial^2 L}{\partial vec\mathfrak{B}\left(\partial vec\mathfrak{B}\right)'}=\sum_{i\in s_2}z_i\left\{diag\left(\mathbf{p}^{\#}\left(\mathbf{x_i};\mathfrak{B}\right)\right)-\mathbf{p}^{\#}\left(\mathbf{x_i};\mathfrak{B}\right)\mathbf{p}^{\#}\left(\mathbf{x_i};\mathfrak{B}\right)'\right\}\otimes\mathbf{x_i}\mathbf{x_i}',$$

a $p(m\text{-}1)\times p(m\text{-}1)$ matrix.

63. Since the design-based approach to estimation requires survey data to be treated as fixed values rather than observations of random variables, the above is not directly applicable. However, Binder (1983) showed how estimating equations for model parameters can be rephrased as estimating equations for finite population parameters. The next section shows how to do this.

The Binder estimating equations approach

64. Following Binder (1983), the census pseudo-loglikelihood can be defined by

$$L_0(\mathbf{B}) = \sum_{i \in P} \sum_{j=1}^{m} y_{ij} \, log \; p_j(\mathbf{x_i}; \mathbf{B})$$

with associated pseudo-score and pseudo-information matrices. $\mathbf{B}$ should not be interpreted as a matrix of model parameters (cf $\mathfrak{B}$ in previous section), but simply as the matrix argument of the function defined by L_0.

65. Following Binder's approach, the census pseudo-loglikelihood evaluated at $\mathbf{B}$ can be estimated by

$$L_\pi(\mathbf{B}) = \sum_{i \in s_1} \frac{1}{\pi_i} \sum_{j=1}^{m} y_{ij} \, log \; p_j(\mathbf{x_i}; \mathbf{B})$$

for a single-phase design (and similarly for the pseudo-score and pseudo-information matrices). For two-phase designs, the census pseudo-loglikelihood is estimated by

$$L_{\pi^*}(\mathbf{B}) = \sum_{i \in s_2} \frac{1}{\pi_i^*} \sum_{j=1}^{m} y_{ij} \, log \; p_j(\mathbf{x_i}; \mathbf{B})$$

with analogous estimators for the pseudo-score function and pseudo-information matrices.

66. Defining $\mathbf{B_0}$, $\mathbf{B_\pi}$ and $\mathbf{B_{\pi^*}}$ to be the maximisers of $L_0(\mathbf{B})$, $L_\pi(\mathbf{B})$ and $L_{\pi^*}(\mathbf{B})$ respectively, Binder (1983) showed that

$$Cov(\mathbf{vecB_\pi}) \approx \mathbf{I^{-1}}(\mathbf{B_0}) Cov\left(\mathbf{S}_\pi(\mathbf{B_0})\right) \mathbf{I^{-1}}(\mathbf{B_0})$$

which can be estimated by

$$\widehat{Cov}(\mathbf{vecB_\pi}) = \mathbf{I_\pi^{-1}}(\mathbf{B_\pi}) \widehat{Cov}\left(\mathbf{S}_\pi(\mathbf{B_\pi})\right) \mathbf{I_\pi^{-1}}(\mathbf{B_\pi}) .$$

67. For the two-phase situation,

$$Cov(\mathbf{vecB_{\pi^*}}) \approx \mathbf{I^{-1}}(\mathbf{B_0}) Cov\left(\mathbf{S}_{\pi^*}(\mathbf{B_0})\right) \mathbf{I^{-1}}(\mathbf{B_0})$$

which can be estimated by

$$\widehat{Cov}(\mathbf{vecB_{\pi^*}}) = \mathbf{I_{\pi^*}^{-1}}(\mathbf{B_{\pi^*}}) \widehat{Cov}\left(\mathbf{S}_{\pi^*}(\mathbf{B_{\pi^*}})\right) \mathbf{I_{\pi^*}^{-1}}(\mathbf{B_{\pi^*}}) .$$

68. Not surprisingly, $\mathbf{B_\pi}$ and $\mathbf{B_{\pi^*}}$ are known as *pseudo-maximum likelihood (PML)* estimates of $\mathbf{B_0}$. The next section shows how they are used in constructing estimators for totals of the detailed expense items.

The LGREG (Logistic Generalised Regression Estimator) for P

69. If $\mathbf{B_0}$ was known, a difference estimator for the total of the detailed expense items could be constructed i.e.

$$\widehat{\mathbf{Y}}_{\mathbf{LG}} = \sum_{i \in s_1} \frac{1}{\pi_{1i}} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_0}\right) \right] + \sum_{i \in P} z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_0}\right)$$

in the case of a single-phase design, or

$$\widehat{\mathbf{Y}}_{\mathbf{LG}} = \sum_{i \in s_2} \frac{1}{\pi_i^*} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_0}\right) \right] + \sum_{i \in s_1} \frac{1}{\pi_{1i}} z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_0}\right)$$

for a two-phase design.

70. Since this is not the case, $\mathbf{B_0}$ is estimated by $\mathbf{B_\pi}$ in the case of a single-phase design, and $\mathbf{B_{\pi^*}}$ in the case of a two-phase design.

71. The single-phase LGREG is

$$\widehat{\mathbf{Y}}_{\mathbf{LG}} = \sum_{i \in s_1} \frac{1}{\pi_{1i}} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_\pi}\right) \right] + \sum_{i \in P} z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_\pi}\right)$$

with covariance matrix

$$Cov\left(\widehat{\mathbf{Y}}_{\mathbf{LG}}\right) = \sum_{i \in P} \sum_{j \in P} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_0}\right) \right] \left[\mathbf{y_j} - z_j \mathbf{p}\left(\mathbf{x_j}; \mathbf{B_0}\right) \right]'$$

.

This can be estimated by

$$\widehat{Cov}\left(\widehat{\mathbf{Y}}_{\mathbf{LG}}\right) = \sum_{i \in s_1} \sum_{j \in s_1} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}\pi_{1ij}} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_\pi}\right) \right] \left[\mathbf{y_j} - z_j \mathbf{p}\left(\mathbf{x_j}; \mathbf{B_\pi}\right) \right]'$$

.

Lehtonen et al (1998) gave these results. It is possible using the approach given in Binder (1996) to give an improved variance estimator analogous to Särndal's g-weighted variance estimator for the generalised regression estimator (GREG). The details will not be given here.

72. The two-phase LGREG is

$$\widehat{\mathbf{Y}}_{\mathbf{LG}} = \sum_{i \in s2} \frac{1}{\pi_i^*} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_{\pi^*}}\right) \right] + \sum_{i \in s_1} \frac{1}{\pi_{1i}} z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_{\pi^*}}\right)$$

with covariance matrix

$$Cov\left(\widehat{\mathbf{Y}}_{\mathbf{LG}}\right) = \sum_{i \in P} \sum_{j \in P} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} \mathbf{y_i} \mathbf{y_j}' + E_{s_1} \left\{ \sum_{i \in s_1} \sum_{j \in s_1} \frac{\pi_{2ij} - \pi_{2i}\pi_{2j}}{\pi_i^* \pi_j^* \pi_{2ij}} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_\pi}\right) \right] \left[\mathbf{y_j} - z_j \mathbf{p}\left(\mathbf{x_j}; \mathbf{B_\pi}\right) \right]' \right\}$$

which can be estimated by

$$\widehat{Cov}\left(\widehat{\mathbf{Y}}_{\mathbf{LG}}\right) = \sum_{i \in s2} \sum_{j \in s2} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}\pi_{ij}^*} \mathbf{y_i} \mathbf{y_j}' + \sum_{i \in s2} \sum_{j \in s2} \frac{\pi_{2ij} - \pi_{2i}\pi_{2j}}{\pi_i^* \pi_j^* \pi_{2ij}} \left[\mathbf{y_i} - z_i \mathbf{p}\left(\mathbf{x_i}; \mathbf{B_{\pi^*}}\right) \right] \left[\mathbf{y_j} - z_j \mathbf{p}\left(\mathbf{x_j}; \mathbf{B_{\pi^*}}\right) \right]'$$

.

This generalises Lehtonen's results to two-phase designs. Again, Binder's approach can be used to develop an improved variance estimator but the details are omitted.

73. $\hat{\mathbf{Y}}_{\mathrm{LG}}$ should be an improvement over the simple two-phase estimate, since we have made use of auxiliary information. How much depends on the quality of the multinomial logistic approximation.

74. So far, we have only considered the part of the EAS population from which the I-O Survey will be selected. We now deal with P', the remaining part of the EAS population.

Synthetic estimation of $\sum_{i\in P'}\mathbf{y_i}$

75. For those units in EAS which are not going to be considered for the I-O sample, it is possible to construct a synthetic estimate of the total from these units by

$$\widehat{\mathbf{Y}}_{synth}=\sum_{i\in s_1'}\frac{1}{\pi_{1i}}z_i\mathbf{p}\left(\mathbf{x_i};\mathbf{B}_{\pi*}\right).$$

$\mathbf{p}\left(\mathbf{x_i};\mathbf{B}_{\pi*}\right)$ can be thought of as a predicted apportioning of OOE into the detailed expense items based on the available auxiliary information and the characteristics of those businesses from which the detailed expense items was collected.

76. The assumption has to be made that $\mathbf{B_0}$ maximises both $\sum_{i\in P}\sum_{j=1}^{m}y_{ij}\,log\;p_j\left(\mathbf{x_i};\mathbf{B}\right)$ and $\sum_{i\in P'}\sum_{j=1}^{m}y_{ij}\,log\;p_j\left(\mathbf{x_i};\mathbf{B}\right)$. This justifies the use of $\mathbf{B}_{\pi*}$, an estimate derived from sampled businesses in P, in constructing estimates for businesses in P'.

The next section shows how to produce an estimate for the full EAS population.

Estimating a total for the full population, $P\cup P'$

77. This is done by adding the estimated totals for P and P'. There are two possibilities:

$$\widehat{\mathbf{Y}}_1=\sum_{i\in s_2}\frac{1}{\pi_i^*}\mathbf{y_i}+\sum_{i\in s_1'}\frac{1}{\pi_{1i}}z_i\mathbf{p}\left(\mathbf{x_i};\mathbf{B}_{\pi*}\right)$$

which is the direct two-phase estimate of $\sum_{i\in P}\mathbf{y_i}$ plus the synthetic estimate for $\sum_{i\in P'}\mathbf{y_i}$, or:

$$\widehat{\mathbf{Y}}_2=\sum_{i\in s2}\frac{1}{\pi_i^*}\left[\mathbf{y_i}-z_i\mathbf{p}\left(\mathbf{x_i};\mathbf{B}_{\pi*}\right)\right]+\sum_{i\in s_1\cup s_1'}\frac{1}{\pi_{1i}}z_i\mathbf{p}\left(\mathbf{x_i};\mathbf{B}_{\pi*}\right).$$

which is the two-phase LGREG of $\sum_{i\in P'}\mathbf{y_i}$ plus the synthetic estimate for $\sum_{i\in P'}\mathbf{y_i}$.

78. The first estimator has already been used, and the second may be investigated in the future. We only deal with covariance matrix estimation for the first estimator.

Covariance matrix estimation for $\widehat{\mathbf{Y}}_1$

79. Firstly, take the Taylor series approximation

$$\mathbf{p}(\mathbf{x_i};\mathbf{B}_{\pi*}) \approx \mathbf{p}(\mathbf{x_i};\mathbf{B}_0) + \left.\frac{\partial \mathbf{p}(\mathbf{x_i};\mathbf{B})}{(\partial vec\mathbf{B})'}\right|_{\mathbf{B}=\mathbf{B}_0} vec(\mathbf{B}_{\pi*} - \mathbf{B}_0)$$

and define

$$\mathbf{p}^0(\mathbf{x_i};\mathbf{B}_{\pi*}) = \mathbf{p}(\mathbf{x_i};\mathbf{B}_0)$$

$$\mathbf{p}^1(\mathbf{x_i};\mathbf{B}_{\pi*}) = \mathbf{p}(\mathbf{x_i};\mathbf{B}_0) + \left.\frac{\partial \mathbf{p}(\mathbf{x_i};\mathbf{B})}{(\partial vec\mathbf{B})'}\right|_{\mathbf{B}=\mathbf{B}_0} vec(\mathbf{B}_{\pi*} - \mathbf{B}_0)$$

80. The simplest way to approximate the covariance matrix of $\widehat{\mathbf{Y}}_1$ (and it is probably very crude) is by considering

$$\widehat{\mathbf{Y}}_1^0 = \sum_{i\in s2} \frac{1}{\pi_i^*}\mathbf{y_i} + \sum_{i\in s_1'} \frac{1}{\pi_{1i}} z_i \mathbf{p}^0(\mathbf{x_i};\mathbf{B}_{\pi*})$$

The covariance matrix of is just the sum of $Cov(\hat{\mathbf{Y}}_{\pi^*})$ and $Cov\left(\sum_{i\in s_1'} \frac{1}{\pi_{1i}} z_i \mathbf{p}^0(\mathbf{x_i};\mathbf{B}_{\pi*})\right)$.
This gives

$$Cov\left(\widehat{\mathbf{Y}}_1^0\right) = \sum_{i\in P}\sum_{j\in P} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} \mathbf{y_i}\mathbf{y_j}' + E_{s_1}\left\{\sum_{i\in s_1}\sum_{j\in s_1} \frac{\pi_{2ij} - \pi_{2i}\pi_{2j}}{\pi_i^*\pi_j^*\pi_{2ij}} \mathbf{y_i}\mathbf{y_j}'\right\} + \sum_{i\in P'}\sum_{j\in P'} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}} z_i z_j \mathbf{p}(\mathbf{x_i};\mathbf{B}_0)\mathbf{p}(\mathbf{x_j};\mathbf{B}_0)'$$

which can be estimated by

$$\widehat{Cov}\left(\widehat{\mathbf{Y}}_1^0\right) = \sum_{i\in s_2}\sum_{j\in s_2} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}\pi_{ij}^*} \mathbf{y_i}\mathbf{y_j}' + \sum_{i\in s_2}\sum_{j\in s_2} \frac{\pi_{2ij} - \pi_{2i}\pi_{2j}}{\pi_i^*\pi_j^*\pi_{2ij}} \mathbf{y_i}\mathbf{y_j}' + \sum_{i\in s_1'}\sum_{j\in s_1'} \frac{\pi_{1ij} - \pi_{1i}\pi_{1j}}{\pi_{1i}\pi_{1j}\pi_{1ij}} z_i z_j \mathbf{p}(\mathbf{x_i};\mathbf{B}_{\pi*})\mathbf{p}(\mathbf{x_j};\mathbf{B}_{\pi*})'$$

81. This is misleading as a quality measure since it does not include an estimate of the bias sums of squares and crossproducts matrix

$$\sum_{i\in P'}\left[\mathbf{y_i} - z_i\mathbf{p}(\mathbf{x_i};\mathbf{B}_0)\right]\left[\mathbf{y_i} - z_i\mathbf{p}(\mathbf{x_i};\mathbf{B}_0)\right]'$$

caused by the use of synthetically generated values for businesses in s_1'. The estimation of the above is left as an open question.

82. It is possible to derive a covariance matrix estimate for

$$\widehat{\mathbf{Y}}_1^1 = \sum_{i \in s2} \frac{1}{\pi_i^*} \mathbf{y_i} + \sum_{i \in s_1'} \frac{1}{\pi_{1i}} z_i \mathbf{p}^1 \left(\mathbf{x_i}; \mathbf{B}_{\boldsymbol{\pi}*} \right)$$

but it is far more complicated in appearance than $\widehat{Cov}(\hat{\mathbf{Y}}_1^0)$ and suffers from the same defect in that it makes no attempt to "capture" the design bias incurred by using synthetically generated values. The details are not given here.

Bibliography

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* **51,** 279-292.
Binder, D. A. (1996). Linearization methods for single phase and two phase samples: a cookbook approach *Survey Methodology* **22,** 17-22.
Hidiroglou, M and Särndal, C. (1998) Use of auxiliary information for two-phase sampling. *Survey Methodology* **24** 11-20
Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology* **24** 51-55.
Rossiter, P. (1998). Modelling Detailed Business Operating Expenses. Unpublished paper prepared for ABS Methodological Advisory Committee meeting, November 1998.
Särndal, C., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer-Verlag.
Welsh, A. and Szoldra, E. (1997). Allocation of income/expenditure in Input-Output tables. Unpublished paper prepared for ABS Methodological Advisory Committee meeting, November 1997.