Methodology Advisory Committee Paper

MODELLING DETAILED BUSINESS OPERATING EXPENSES FROM ABS ECONOMIC COLLECTIONS

Peter Rossiter

November 1998

CONTENTS

1. INTRODUCTION

1.1 Background

Current strategies for improving the quality of economic statistics in Australia are contingent upon the ability of the ABS to construct a comprehensive Input-Output model of the Australian economy. Such a model must encompass the application of factors (such as labour, capital and commodity inputs) to the production of commodities and services, the flows of commodities and services between industries, and the final use of commodities and services by households, businesses and government.

For some time, the ABS has recognised that this new approach to compiling the national economic accounts has far-reaching implications for the nature, quantity and quality of the economic data which must be collected and interpreted.

One area of particular concern to the ABS is the on-going collection of detailed operating expenses from businesses. Historically, such information obtained from businesses through the key economic surveys and censuses has been rather minimal. Subsequent attempts to expand the amount of detail collected have placed an onerous reporting load upon respondents, and stringent economic conditions have made it difficult for the ABS to follow-up and ensure the quality of the responses received. In this environment, it is crucial that the ABS develop strategies

(i) to ensure that the "correct" data items are sought (taking into account the importance of the data to the ABS and the ability of the respondent to provide acceptable data);

(ii) to allocate ABS collection and follow-up resources so as to maximise the quantity of acceptable data obtained; and

(iii) to extract as much information as possible from the data provided, and produce the best economy-wide estimates possible.

A number of internal ABS investigations have already been conducted into these issues — notably (i) Aspden *et al* (1994) and Sullivan (1997); (ii) Rogers (1998) and (iii) Welsh and Szoldra (1997) and Fraser (1998). This study extends some of the theoretical issues put forward by Welsh and Szoldra with respect to the application of statistical modelling techniques to survey data, and reports empirical evidence to supplement the practical experiments carried out by Fraser.

The main objectives of this study are —

- to better understand the nature of the responses provided by individual businesses, and to explore possible sources of latent information in the data;
- to investigate ways of producing superior economy-wide estimates from the sample responses, possibly by correcting for identified deficiencies; and

- to provide feedback to ABS management and the relevant ABS collection areas on the suitability of current strategies for collecting and processing data of this type.

A case study based upon data from the 1995-96 Economic Activity Survey provides the empirical underpinning for this research.

1.2 ABS economic data collections

Section 2 reviews the ABS strategy for collecting data on detailed operating expenses. Generally, a small number of core expense items (including total operating expenses) are collected routinely by the major industry-based statistical collections. Information on the breakdown of operating expenses into detailed expense categories is then sought from a small subset of businesses selected from the core sample. The overall shares of expenditure recorded within this sub-sample are then postulated to be typical of the wider economy.

Specific information pertaining to the 1995-96 Economic Activity Survey is given in Section 5.

1.3 Model-based estimation

The two-level collection methodology described above raises an important methodological question concerning optimal estimation of the population totals for detailed operating expenses. Is the pattern of detailed expenditures reported by the small sub-sample of businesses indeed representative of the entire population? Or is it possible to make better use of the available data in forming economy-wide estimates?

If the diversity of observed expenditure patterns can be explained by employing auxiliary information or by exploiting identifiable sources of heterogeneity in the dataset, then it may be more efficient to use the following estimation approach —

1. Construct a statistical model which accurately predicts the breakdown of detailed operating expenses for businesses in the sub-sample, using whatever relevant auxiliary information may be available;

2. Use this statistical model to predict the breakdown of detailed operating expenses for all businesses in the core sample, based upon knowledge of the individual characteristics of those businesses;

3. Form total population estimates of detailed operating expenses by using the fitted values and sample weights for the core sample.

If no such explanatory information can be located, then it may be assumed that either the businesses in the sub-sample are homogeneous in their expenditure

patterns or that any variations are essentially random — implying that step 2 above is redundant. Estimation would then proceed by applying the observed patterns of expenditure (from the sub-sample) to the population estimates of total operating expenses derived from the core sample.

It is acknowledged that many businesses (especially small businesses) may have difficulty in providing accurate data on all categories of operating expenses. A cursory examination of the survey responses also reveals that few businesses provide non-zero responses in all categories, and a significant number report zero expenditure in a majority of categories. This propensity of respondents to report zero expenditures — whether valid or not — ought to be recognised explicitly in any viable statistical model of these responses.

Section 3 presents a very basic model of businesses' expenditure behaviour — the multinomial model — and shows how it may be extended to incorporate auxiliary variables and to recognise explicitly the likelihood of reporting zeros. Statistical tests are proposed for establishing whether these modifications add significantly to the explanatory power of the basic model.

1.4 Prediction of population totals

Section 4 explores the application of the models described in Section 3 to form economy-wide estimates of detailed operating expenses. While it may prove possible to find variants of the basic model which provide a substantially better explanation of *within sample* variation in expenditure patterns, this does not necessarily imply that the more complex model will provide significantly different estimates of aggregate expenditures — either within the sample or economy-wide. Conditions under which the estimates may diverge are outlined.

The modification of the basic model to recognise explicitly the incidence of zero reporting generates informative statistics which may be used to assess the extent of possible non-response within expense categories. Furthermore, if it can be established independently that non-response is an issue, the model can be adapted to impute the likely proportion of total recorded operating expenditure which would have been spent on the omitted category.

1.5 Case study

Responses from the 1995-96 Economic Activity Survey are used to illustrate the typical characteristics of data on detailed business expenses.

In addition to the data items pertaining to detailed operating expenses, three other data items — consistently recorded by all ABS economic surveys, and therefore obvious candidates for auxiliary variables — are examined:

- wages and salaries,

- employment at 30 June 1996, and
- turnover.

Typical data problems — such as missing or obviously incorrect responses and logical inconsistencies between data items — are highlighted, and responses unsuitable for statistical modelling are removed from the dataset.

As current and proposed ABS collection strategies frequently discriminate between businesses on the basis of industry classification and some concept of "size", a methodology is devised for splitting survey respondents into "small" and "large" businesses within each industry — for the purpose of establishing whether such categories capture significant heterogeneity in the dataset.

The distributions of responses (including zeros) to the various data items are examined by tabulations and histograms. Cross-tabulations, correlations and X-Y plots are employed to seek out possible relationships between data items.

See Section 5.

The models defined in Section 3 are applied to the EAS dataset in Section 6, where it is shown that modifications to the basic multinomial model to model explicitly the zero responses and to incorporate an auxiliary variable — log(Turnover) — substantially improve the explanatory power of the model. Statistically significant gains are also achieved by separately modelling subsets disaggregated by both industry and size.

Section 7 documents the extension of the fitted models from Section 6 to predict the expenditure patterns of the businesses included in the core EAS dataset.

1.6 Improvements and extensions

The case study reported in this paper does not go far enough to establish the viability of these modelling techniques in a production environment. Some of the remaining problems are outlined in Section 8.

Possible improvements to the proposed model-based estimation approach are also presented.

1.7 Conclusions

Regardless of whether the model-based approach to estimation outlined in this paper is viewed as a practical option for the future, the case study has produced a number of strong implications for future collection strategies and estimation methodologies. These are summarised in Section 9.

2. OVERVIEW OF ABS ECONOMIC COLLECTIONS

2.1 (incomplete)

Topics to be covered in this chapter include —

stratification

outline of methodology for collecting detailed operating expenses

long forms & short forms

probability of selection weights

standard estimation procedures - introduce naïve estimator

Table 2.1 Major ABS Economic Data Collections

Agricultural Finance Survey (AFS)
Construction Industry Survey (CIS)
Economic Activity Survey (EAS)
Manufacturing surveys and censuses
Mining Census
Services Industries Surveys (SIS)
Wholesale and Retail collections

Table 2.2 Core expense items collected in the Economic Activity Survey

Wages and salaries
Employer contributions to superannuation funds
Workers' compensation costs
Insurance premiums
Interest expenses
Depreciation and amortisation
Bad and doubtful debts
Purchases

and

Other operating expenses.

Table 2.3 Breakdown of "Other operating expenses"

Fringe benefits tax
Payroll tax
Land tax and land rates
Bank charges other than interest
Royalties expenses
Motor vehicle running expenses
Freight and cartage expenses
Postal, mailing and courier expenses
Telecommunication services
Repair and Maintenance expenses
Rent, leasing and hiring expenses —
    land, buildings and other structures
    motor vehicles
    other rent, leasing and hiring expenses
Audit and other accounting expenses
Legal expenses
Advertising expenses
Paper, printing and stationery expenses
Payments for data processing services
Payments for staff training services
Travelling, accommodation and entertainment expenses
Payments for other management and administrative services
Payments for cleaning services
Sales commission expenses
Commission expenses for work done on own materials
Computer software expensed
Other contract and sub-contract expenses
Other expenses

3. MODEL-BASED ESTIMATION

3.1 Introduction

Under conventional statistical practice, total operating expenses would be estimated from the main industry collections, while the proportions of total operating expenses allocated to detailed expense categories would be determined from the responses of businesses included in selective sub-samples. This approach may be rationalised on the basis of either of the following assumptions —

(a) all businesses (at least within defined post-strata) have a common pattern of expenditure; or

(b) (perhaps more plausibly) the diversity of expenditure patterns recorded in the sub-sample is typical of the diversity present (but not recorded) in the main sample and in the wider population of businesses.

However, if further insights into the diversity of observed expenditure patterns can be obtained by employing auxiliary information or by exploiting identifiable sources of heterogeneity in the dataset, then it may be more efficient to use the following alternative three-step estimation approach —

1. Construct a statistical model which accurately predicts the breakdown of detailed operating expenses for businesses in the sub-sample, using whatever relevant auxiliary information may be available;

2. Use this statistical model to predict the breakdown of detailed operating expenses for all businesses in the core sample, based upon knowledge of the individual characteristics of those businesses;

3. Form total population estimates of detailed operating expenses by using the fitted values and sample weights for the core sample.

This model-based (or model-assisted) estimation approach offers the prospect of better quality estimates from the more effective use of available data. Alternatively, failure to find any sources of explanatory information would vindicate the conventional approach to estimation.

Past and current ABS collection strategies have, naturally, been predicated upon the assumptions of the conventional estimation approach. These assumptions, which implicitly ignore the possibility of further systematic diversity within post-strata, offer the best prospects for minimising the size of sample that must be collected to ensure the estimates are of acceptable precision. By contrast, the identification of useful auxiliary variables may suggest the need to collect more data, or employ different sample selection techniques, to ensure that the sample is properly representative.

Preliminary investigations by Welsh and Szoldra (1997) and Fraser (1998) have established that regression-based techniques can be used to explain some of the observed variation in expenditure patterns in actual survey data. However, to date, these studies have not shown conclusively that these techniques provide statistically significant improvements over the conventional approach. In addition, a range of deficiencies in the methodologies employed make it difficult to select a suitable alternative to the conventional method.

The model-based estimation approach propounded in this paper does not claim to be definitive. It is intended only to capture the most salient features of the data — within a framework that permits statistical hypothesis testing. The primary objective is to establish conclusively whether model-based estimation is a viable technique in the context of measuring business expenses and, if so, to ascertain the key elements which must be present in such a model.

The *multinomial* model (described in Section 3.2) is perhaps the simplest statistical model for allocating of a given number of units (dollars, in this instance) to a fixed number of categories. Section 3.3 extends the basic multinomial model to incorporate auxiliary variables. Section 3.4 introduces a further modification to the model to account for the disproportionate number of reported zero expenditures which are typically found in survey data of this type. Section 3.5 provides details of the estimation of model parameters by iterative maximum likelihood techniques. Section 3.5 discusses issues of model selection and hypothesis testing.

3.2 The multinomial model

Suppose that the ABS conducts a survey to establish details of business operating expenditure, and that n businesses report non-zero aggregate expenditure on "other operating expenses" —

$$r_i > 0 \quad ; \; i = 1, \ldots, n$$

where r_i denotes expenditure (in thousands of dollars) on "other operating expenses" by the i^{th} respondent.

Each respondent also provides a detailed itemisation of other operating expenses into m expense categories —

$$c_{ij} \geq 0 \quad ; \; i = 1, \ldots, n \text{ and } j = 1, \ldots, m$$

$$\sum_{j=1}^{m} c_{ij} = r_i \quad ; \; i = 1, \ldots, n$$

where c_{ij} denotes expenditure on the j^{th} expense category by the i^{th} respondent.

The random variable of interest in this analysis is the ratio

$$S_j = \frac{C_j}{R} \quad ; j = 1, \ldots, m$$

or the share of other operating expenditure allocated to each expense category.

Specifically, it is required to find the **expected value** of S_j — which may be either a constant or a function of one or more auxiliary variables.

A common way of proceeding with this task would be to fit a regression model to the observed values of the random variable

$$s_{ij} = \frac{c_{ij}}{r_i}$$

(or some suitable transformation of s_{ij}), subject to the constraint that the fitted shares should always sum to unity. The approach taken in this study, however, is a little different. While the basic assumptions may appear at first glance to be rather restrictive, it will be shown that this alternative approach has considerable advantages from a practical perspective.

3.2.1 The basic multinomial model

Suppose the expenditure behaviour of the typical (i^{th}) business can be characterised by the multinomial (r_i ; π_1 , π_2 , ... , π_m) distribution function. That is, each thousand dollars of "other operating expenditure" is likely to be allocated to expense category 1 with probability π_1 , expense category 2 with probability π_2 , ... and expense category m with probability π_m. For this statistical model, the expected level of expenditure in category j is $r_i \pi_j$, and thus the expected share of other operating expenditure is given by the parameter π_1 . That is,

$$E\{S_j\} = \pi_j \ .$$

Under the assumption that a common probability model underlies the expenditure decisions of all businesses in the sample, the probability of observing the reported outcome for business i may be written

$$p_i = \frac{r_i!}{c_{i1}!\, c_{i2}! \cdots c_{im}!} \pi_1^{c_{i1}} \pi_2^{c_{i2}} \cdots \pi_m^{c_{im}}$$

and the *likelihood* of observing the entire dataset is then

$$lik = \prod_{i=1}^{n} p_i$$

Acknowledging that the respondents were selected from the wider population of businesses with probability of selection $1/w_i$, it is perhaps more correct to define the weighted likelihood function

$$lik = \prod_{i=1}^{n} p_i^{w_i}$$

and the corresponding *log-likelihood function*

$$L = \sum_{i=1}^{n} w_i \ln(p_i) \ .$$

The estimation procedure then consists of finding the unique values of the parameters (π_1 , π_2 , ... , π_m) which maximise the log-likelihood function, and thus provide the "best" description of the observed expenditure patterns. In this case, the solution to the maximisation problem has a trivial closed form solution, and returns the empirical weighted means, *viz.*

$$\hat{\pi}_j = \bar{s}_j = \frac{\sum_{i=1}^{n} w_i s_{ij}}{\sum_{i=1}^{n} w_i}$$

as the maximum likelihood estimators.

It is particularly convenient that the estimators arising from this basic multinomial model coincide with the estimators which would arise in conventional (non model-based) statistical estimation. As additional complexity is added to the multinomial model in following sections, it will prove possible to establish whether such modifications are statistically significant by carrying out nested hypothesis testing relative to this benchmark.

3.2.2 Grouping expense categories

The expenditure allocation process implied by the basic multinomial model is indeed very simplistic. It is assumed that each thousand dollars of expenditure will be allocated to one of the expenditure categories by a mechanism which takes no account of previous expenditures, and does not recognise the possibility of complementary purchases (ie. positive correlation between expenditures on paired expense items).

As a result of this simplicity, however, the basic multinomial model possesses a useful accounting property — that the expected share of any "group" of expense items is identically equal to the sum of the expected shares of the individual items. For example, let S_T be the share of "all taxes and charges" in "other operating expenses". Then,

$$E\{S_T\} = \sum_{l \in T} \pi_l \ .$$

A similar property relates to the estimation of the model parameters. Let π_T be the estimated share of "all taxes and charges" in "other operating expenses". Let π_1^T be the estimated share of "payroll tax" in expenditure on "all taxes and charges", obtained by fitting a separate multinomial model to the relevant subset of expense categories. Then, the share of "payroll tax" in "other operating expenses" may be computed as

$$E\{S_{PT}\} = \pi_1^T \times \pi_T$$

where this two-stage estimate is identical to the estimate which could have been computed in a single step from the complete dataset (denoted π_{PT}).

This property is perhaps unremarkable in the present context since it offers no prospect of improved computational efficiency. However, when considering more complex modifications of the basic model, it will later prove necessary to introduce groups into the estimation process so as to minimise the size of the largest model which must be fitted.

3.2.3 Post-stratification

There is good reason to expect that businesses engaged in completely unrelated activities — for example, Agriculture, Retail Trade and Education — will have vastly different patterns of operating expenditure, as the different expense categories assume differing degrees of importance. These patterns may be expected to be closer for businesses whose activities are more closely related (in some sense).

ABS economic collections are designed to ensure that information is collected from a comprehensive range of industries and economic activities. The *Australian and New Zealand Standard Industrial Classification* (ANZSIC) is a fundamental tool for grouping related activities in an hierarchical framework, and is almost without exception used as a stratification variable for sample selection. ANZSIC thus represents an obvious choice for post-stratifying the data for estimation purposes — although it remains an empirical problem to determine whether this post-stratification is useful, and, if so, whether stratification at the industry division, group, class or finer level of detail is necessary or desirable.

A simple approach to establishing whether "Industry" is a significant determinant of heterogeneity in the data is to fit the multinomial model successively at the economy wide level, the industry division level, the industry group level, and so on. Then the overall log-likelihood statistics for the entire dataset may be computed and tested for statistically significant improvements in fit, based upon the number of additional parameters required to be estimated:

$$L_{EW} \quad \text{vs} \quad \sum_{Divisions} L_d \quad \text{vs} \quad \sum_{Groups} L_g \quad \text{vs} \quad \sum_{Classes} L_c \quad \text{etc..}$$

Partitioning the dataset and the estimation process is an obvious way to proceed when trying to account for *categorical* variables such as "Industry" — although finite sample size will ultimately lead to a conflict between the objectives of homogeneity and precise parameter estimates. The complexity of competing objectives escalates rapidly as additional cross-classifying variables are added.

Not all sources of heterogeneity take the form of categorical variables. In the following section it is shown that the basic multinomial model may be modified to

simultaneously account for the effects of both *categorical* and *continuous* variables. Furthermore, this model permits the elimination of uninformative cross-classifications, thus minimising problems with degrees of freedom in parameter estimation.

3.3 Multinomial model — extended for auxiliary variables

Economic theory and/or empirical observation would suggest several factors, in addition to "industry", which may influence the pattern of operating expenditures. Examples include —

- the size of the business — whether measured in terms of the number of employees or the volume of output — and potential economies of scale;
- the type of the business — for example, whether the firm is an individual proprietorship, partnership or part of a larger corporation;
- the labour or capital intensiveness of the economic activity undertaken;
- the degree of market competition faced by the business — whether perfect competition, monopoly or something else — and the current level of demand;
- the degree of export orientation of the business; and
- the geographical location of the business — which may also proxy for a range of influences such as the regulatory environment or proximity to resources and markets.

Notice that some of these factors are defined by a range of discrete alternatives (eg. type of business, geographical location) while others may be measured by either a continuous scale or by ordered categories. "Size of business", for example, may be measured by the actual number of employees, or by "small", "medium" and "large" categories.

3.3.1 Multinomial logit model

To incorporate explanatory variables into the basic multinomial model, it is proposed to investigate the *multinomial logit* model (Theil, 1969). This model involves the following re-parameterisation:

$$\pi_{ij} = \frac{\exp(x_i' \boldsymbol{\beta}_j)}{1 + \sum_{l \neq m} \exp\left(x_i' \boldsymbol{\beta}_l\right)} \quad ; \; j = 1, \ldots m-1$$

$$\pi_{im} = \frac{1}{1 + \sum_{l \neq m} \exp\left(x_i' \boldsymbol{\beta}_l\right)}$$

where the vectors $\boldsymbol{\beta}_j$ $(j = 1, \ldots, m-1)$ contain the new parameters to be estimated. The probability of allocating expenditure to each category is now different for each respondent (π_{ij}), and is determined by the vector of business characteristics x_i —

which contains the values of the continuous explanatory variables and dummy (0,1) variables for the categorical factors.

In the simplest possible case, where the parameter vectors contain a single categorical variable (α_j, say), and $x_i = 1$, the *multinomial logit* model is equivalent to the basic multinomial model. It is thus possible to test whether the introduction of explanatory variables produces a statistically significant improvement in fit.

The multinomial logit model implies the following relationship between the π and β parameters:

$$\ln \frac{\pi_{ij}}{\pi_{im}} = x_i' \beta_j \quad ; j \neq m \quad .$$

This is essentially a mathematical device to ensure that the fitted proportions allocated to each category sum to unity. The use of the logarithmic transformation in this context suggests that care will be required in specifying the exact form (transformation) of the explanatory variables and in interpreting their impact.

The logit transformation is frequently used in the analysis of compositional data. The interesting aspect of this application is that the transformation is applied to the parameters of the model — rather than the underlying observed data. As the logarithmic transformation is not defined for values of zero, serious difficulties would have arisen in attempting to model the observed shares (s_{ij}) in this way, since zero responses are often encountered in practice. By contrast, it is quite acceptable to define $0 < \pi_{ij} < 1$.

3.3.2 Grouping expense categories

Once the β parameters have been estimated, the expected expenditure shares may be derived for any respondent by computing the corresponding $\pi_{ij}\,(j = 1, \ldots, m)$ parameters (which are thus conditional upon the individual characteristics of the business).

As in the case of the basic multinomial model, the expected share of any group of expense items may be computed by summing the expected shares of the individual items. However, if the group share were to be estimated directly, the resulting estimate ($\pi_{i,j\in G}(x_i)$, say) would differ from $\Sigma_{j\in G}\,\pi_{ij}(x_i)$. When averaged over the entire sample, both methods would yield the same result, but the disaggregated approach can be expected to produce a better idea of the relationships between the explanatory variables and group expenditure patterns.

In the context of the basic multinomial model, it has already been stated that the estimation process may be devolved into two or more stages — by first modelling the shares of grouped expense items, and then subsequently modelling detailed expenditure patterns within those groups — with the outcome being invariant to the number of steps employed.

This approach to estimation is particularly appealing when the number of expense categories and/or the number of explanatory variables is large. Unfortunately, the invariance property described above does not extend to the multinomial logit model — although the differences in the fitted shares are likely to be very small. As both approaches involve estimation of the same number of parameters, it is not clear that either will produce consistently superior results. Consequently, they should be viewed as closely related, but separate, models.

A third option might involve simultaneous estimation of the first and second stage parameters of the grouped data model — although this represents a more complex version of the original model, without the computational efficiency of the second.

3.4 Multinomial model — modified for "missing" data

3.4.1 "Missing" data

As noted above, one of the significant advantages of the multinomial logit model (and the nested basic multinomial model) is the ability to include zero responses directly in the model. Many competing models do not have this feature.

However, there are numerous reasons why survey respondents might report zero expenditure for a given expense category — and these reasons are generally far too numerous and complex to be adequately explained by the multinomial model.

The multinomial model may provide a satisfactory explanation for zero responses provided by businesses that have only small total outlays on "other operating expenses". It can also account for expenditure items which represent a very small proportion of total outlays. In both instances, the multinomial model is essentially recognising the likelihood that expenditure will lie between zero and $500 in certain categories (since reported expenditures are rounded to the nearest thousand dollars).

The multinomial model does *not* provide a satisfactory explanation for those businesses which will never incur certain types of expenditure — for example, businesses which may not incur payroll tax or royalty payments. Nor can it adequately account for large but infrequent expenditures (although admittedly few operating expenses fall into this category).

It is also likely that a significant number of recorded zeroes can be attributed to intentional or unintentional non-response. The following points are relevant —

- The categories most likely to be affected by non-response are those that cause most difficulty for the respondent — for example, expenditures which are either not usually present or not separately itemised in the accounts of the business.
- Where large amounts of expenditure are reported, apparently insignificant items may be overlooked or ignored.

- Confusion over the terminology used to define expenditure items may result in several possible outcomes:
  (i) total non-response for those expenditure items;
  (ii) allocation of recorded expenditure to the incorrect category;
  (iii) allocation of recorded expenditure to the "other expenses" category;
  (iv) non-response to any of the expenditure categories, but inclusion of the amount in the "total".
  In all cases, a zero will be recorded against the expenditure item in question.

To summarise, there are three broad explanations of zero expenditures —

A. Small expenditures (less than $500);
B. Expenditure items which are rarely, if ever, incurred by a business; and
C. Non-response or mis-allocation.

It is, of course, impossible to establish which of these cases applies to any given zero response. However, zero expenditures of type A are predicted by the assumptions of the multinomial model — which can therefore be used to gauge whether explanation A is *likely*, given the total operating expenditure of the business. The model can also provide some indication of the likely incidence of zero responses for a given category.

The following modification to the basic multinomial model introduces additional parameters to capture the probability that businesses may provide responses of type B or C — which will, for convenience, hereafter be referred to as "***missing***" data. (The probability of a type A response is assumed to be subsumed within the parameters of the basic model.)

Although the term "missing" data has been applied collectively to both type B and type C responses, it is important to remember that those zeroes which arise from non-response or mis-allocation constitute a potential source of bias when forming economy-wide estimates. This topic will be addressed further in Section 4. For now, it is sufficient to assume that the extended model will provide a realistic description of *reported* expenditures.

3.4.2 Modifying the basic model

Suppose that respondent i reports an m-vector of outlays on "other operating expenses" ($\mathbf{c}_i$) whose l^{th} element is zero: $c_{il} = 0$. Without additional information, it cannot be established whether or not the response for category l is "missing" (in the sense described above). Thus, there are two possible explanations for the observed allocation of "other operating expenditure" (r_i):

(a) the respondent usually incurs expenditure on category l, but spent less than $500 during the survey period; or
(b) the response for category l is "missing", and r_i is therefore the total expenditure recorded for the remaining $m-1$ categories.

Assume it is known that $100\delta_l\%$ of all respondents provide "missing" data for category l. Then the probability of observing the reported outcome under explanation (a) may be expressed as

$$(1-\delta_l) \times \Pr\{ C_1 = c_{i1}, C_2 = c_{i2}, \ldots, C_m = c_{im} \mid R = r_i \}$$

and under explanation (b) as

$$\delta_l \times \Pr\{ C_1 = c_{i1}, C_2 = c_{i2}, \ldots, C_m = c_{im} \mid R = r_i, C_l = 0 \} \quad .$$

Under the assumptions of the multinomial model, explanation (a) states that the expenditures on the m categories arise from a multinomial (r_i ; π_1 , π_2 , ... , π_m) distribution.

Explanation (b) states that expenditure is allocated to the $m-1$ non-missing categories according to the conditional multinomial distribution with parameters

$$(r_i ; \pi_1^*, \ldots, \pi_{l-1}^*, \pi_{l+1}^*, \ldots, \pi_m^*) \qquad \text{where} \quad \pi_j^* = \frac{\pi_j}{\sum_{k \neq l} \pi_k} \quad .$$

Let M_l be an indicator variable, taking the value of unity when the response for category l is assumed to be missing and the value of zero otherwise. The total probability of observing the reported outcome for respondent i may then be written:

$$p_i = \sum_{M_l=0}^{1} [M_l \delta_l + (1-M_l)(1-\delta_l)] \times \frac{r_i!}{c_{i1}! \ldots c_{im}!} \times \left[\prod_{j=1}^{m} \pi_j^{c_{ij}} \right] \times$$

$$\left[(1-M_l)\pi_l + \sum_{j \neq l} \pi_j \right]^{-r_i}$$

where the final term is equal to one when $M_l = 0$ and

$$\left[\sum_{k \neq l} \pi_k \right]^{-r_i} = \prod_{j \neq l} \left[\frac{1}{\sum_{k \neq l} \pi_k} \right]^{c_{ij}}$$

when $M_l = 1$. (Note that $c_{il}! = 1$ and $\pi_l^{c_{il}} = 1$ when $c_{il} = 0$.)

When a respondent reports zero expenditures in multiple categories, it is required to successively partition the probability p_i to accommodate the possibility that each reported zero may be either missing or non-missing data. Accordingly, a generic definition of p_i is provided by the following:

$$p_i = \sum_{M_1} \cdots \sum_{M_m} p_{i|M}$$

where the indicator variables M_j $(j = 1, \ldots, m)$ take the initial value zero (corresponding to the case where category j is non-missing) and may increment to take the value of one when category j is assumed missing. Clearly, only indicator variables corresponding to categories with zero recorded expenditure may take the

value of one. $p_{i|M}$ is the probability of observing the recorded pattern of expenditure given the pattern of missing / non-missing data encapsulated by the m-vector M:

$$p_{i|M} = \frac{r_i!}{c_{i1}! \ldots c_{im}!} \times \prod_{j=1}^{m} [M_j \delta_j + (1-M_j)(1-\delta_j)] \pi_j^{c_{ij}} \times$$

$$\left[\sum_{j=1}^{m} (1-M_j) \pi_j \right]^{-r_i} \times \left(1 - \prod_{j=1}^{m} \delta_j \right)^{-1}$$

where the final term corrects for the fact that $r_i > 0$, and therefore it is not possible for all categories to simultaneously have missing data.

The estimation problem for the extended model is thus to find the values of the parameters $(\delta_1, \delta_2, \ldots, \delta_m, \pi_1, \pi_2, \ldots, \pi_m)$ which maximise the weighted log-likelihood function

$$L = \sum_{i=1}^{n} w_i \ln(p_i)$$

where

$$\ln(p_i) = \ln\left(\frac{r_i!}{c_{i1}! \ldots c_{im}!} \right) - \ln\left(1 - \prod_{k=1}^{m} \delta_k \right) +$$

$$\ln \sum_{M_1} \cdots \sum_{M_m} \exp\left[\sum_{j=1}^{m} \{ \ln[M_j \delta_j + (1-M_j)(1-\delta_j)] + c_{ij} \ln(\pi_j) \} - r_i \ln \sum_{k=1}^{m} (1-M_k)\pi_k \right]$$

Note that the basic multinomial model can be obtained from this extended model by applying the parameter restrictions: $\delta_1 = \delta_2 = \ldots = \delta_m = 0$. This can be seen most easily by returning to the definition of $p_{i|M}$ above. If $\delta_j = 0 \quad \forall_j$, then $p_{i|M} \equiv 0$ whenever the vector M indicates missing data in one or more categories. The only non-zero contribution to p_i thus corresponds to the case where $M_j = 0 \quad \forall_j$ — which is identical to the specification of the basic multinomial model.

In this extended model, the π_j parameters can no longer be equated with the expected expenditure shares for the total population. Rather, they represent the expected shares of expenditure for the hypothetical respondent with no "missing" expenditure categories. As the majority of respondents report zero expenditure on one or more categories, the overall expected shares must take account of the frequency of such responses —

$$E\{S_k\} = \sum_{M_1} \cdots \sum_{M_m} \frac{\prod_{j=1}^{m} [M_j \delta_j + (1-M_j)(1-\delta_j)]}{1 - \prod_{j=1}^{m} \delta_j} \times \frac{\pi_k}{\sum_{j=1}^{m} (1-M_j) \pi_j} \quad .$$

$E\{S_k\}$ can be interpreted as a weighted average of the category k coefficients from the conditional multinomial distributions arising from all possible combinations of missing and non-missing data. The expected shares for any individual respondent may be refined by limiting the values taken by the indicator variables (M_k), according to prior knowledge of which data items have been included in "other operating expenses".

3.4.3 Grouping expense categories

Although the partitioning of the multinomial likelihood to account for the incidence of "missing" data is conceptually straightforward, the implementation can prove to be computationally difficult and time consuming. The most obvious problem lies in the number of possible combinations of missing and non-missing data which must be considered — $(2^m - 1)$ for m categories. For example, with 25 expense categories it is necessary to deal with 33,554,431 combinations in the recursive estimation algorithm. This provides a strong incentive to investigate two-stage estimation approaches, such as that described earlier for the multinomial logit model.

By combining 25 categories into five groups with five categories in each, for example, the maximum number of combinations considered by the estimation algorithm is reduced by a factor of 10^{-6}. The requirement to estimate fewer parameters simultaneously can add a few more orders of magnitude to the computational savings.

Decisions regarding the grouping of expense categories will have an impact upon the precision of the parameter estimates and the types of inference which can be applied. Different groupings will provide different characterisations of the data — which should be broadly consistent provided the original assumptions underlying the full model are satisfactory. Inconsistencies might arise in the case of expense items which are substitutes or complements, or where patterns of non-response are not consistent with the assumption of independence between categories.

To summarise, grouping of expense categories is a convenience which sacrifices some degree of precision in return for computational speed and efficiency. The extent of this sacrifice depends upon the relative importance of those combinations of missing / non-missing data which are not individually examined. (Recall that parameter estimates are invariant to grouping when there are no "missing" data — ie. in the basic multinomial model.) This suggests that there may exist algorithms for minimising the loss due to grouping, while maximising the computational savings, but this topic is beyond the scope of the current paper.

The practice of grouping related categories together may serve to minimise the impact of departures from the underlying model assumptions, and may also provide information which is significant in its own right. For example, some users may find it informative to focus upon the breakdown of group categories such as

"Taxes and charges" or "Business and contract services" — rather than have the same expense items expressed as a share of "other operating expenses".

3.4.4 Auxiliary variables

Individual respondent characteristics can be incorporated into the modified multinomial model by means of the same re-parameterisation used to define the multinomial logit model.

In the case of the multinomial logit model, the explanatory power of the auxiliary variables can be seriously dissipated by the need to account for "missing" data. In this modified version of the model, these variables are more specifically targetted towards explaining the behaviour of those respondents that actually incur each type of expense.

In addition, it is possible that the variations in the incidence of "missing" data for different classes of respondent may be explained by the use of auxiliary information — although it is likely that much of this variation may be eliminated by categorical (post-stratification) factors such as "industry" and "size".

3.5 Maximum likelihood estimation

The preceding sections have introduced four statistical models which may be used to describe the way businesses allocate expenditure to the detailed categories of "other operating expenditure". The simplest model of such behaviour is the basic multinomial (BM) model. Two significant improvements have been added. The multinomial logit (ML) model extends the basic model by introducing auxiliary variables to explain variations in expenditure patterns. The modified multinomial (MM) model acknowledges that not all businesses incur (or report) expenditures on all categories of "other operating expenses". It therefore attempts to quantify the probability that a given expense category will be present, and if so the likely proportion of "other operating expenditure" which will be allocated to it. If these quantities are found to vary systematically according to various respondent characteristics, then the modified multinomial logit (MML) model may be employed to add auxiliary variables to the analysis.

All four models may be fitted by finding the values of the model parameters which maximise the respective log-likelihood functions — although the complexity of this task varies considerably between models.

The BM model, as stated earlier, may be solved analytically. The maximum likelihood estimates of the π_j $(j = 1, \ldots, m)$ parameters are simply the empirical weighted means of the sample.

Suppose that k auxiliary variables are available for inclusion in the ML model, and that the parameterisation also includes a constant term. Then it is necessary to

solve for $(k+1)(m-1)$ parameters, since β_m is typically defined as a normalising vector. The solution to this problem requires that the partial derivatives of the log-likelihood function with respect to all parameters (the gradient vector) equal zero, and that the matrix of second partial derivatives (the Hessian matrix) is negative definite. Since the likelihood equations in this case are nonlinear, it is necessary to employ an iterative numerical optimisation procedure. Standard statistical software packages are available to perform this task — although some may have difficulty in handling the *weighted* log-likelihood function specified here.

The MM and MML models are, in principle, solved in exactly the same way as the ML model. However, due to the complexity of the log-likelihood specification for these models, issues of programming efficiency and numerical accuracy may arise for models involving large numbers of parameters and/or observations. The following points are relevant —

- Most optimisation software will compute numerical approximations to the gradient and Hessian functions. However, considerable gains in efficiency can be achieved by providing the program with the exact analytic formulae.
- Complex functions of matrices (eg. the likelihood function for the MM model) can be written quite economically in matrix programming languages such as GAUSS or SAS/IML. However, this economy of expression may conflict with computational efficiency when the data structures are large (say >100,000 elements). Sequential (rather than simultaneous) processing of records can sometimes prove optimal.
- When total expenditure by a business is large, the probability of observing any given allocation of detailed expenditures can be very small — often below the level of accuracy of the software or the computer. The addition of two or more such small numbers (as required for the MM and MML models) requires careful attention.

All models may be estimated by use of the Constrained Maximum Likelihood (CML) module associated with the matrix programming language GAUSS.

3.6 Model selection and hypothesis testing

The ML, MM and MML models have been proposed as extensions or improvements on the basic multinomial (BM) model. The empirical task is now to establish whether the additional complexity of these models can deliver a significant gain in explanatory power. Furthermore, it is desirable to establish whether one model specification is likely to outperform all others.

Since all models are estimated by the technique of maximising the log-likelihood function for a given sample of businesses, it is natural to seek out the method that results in the greatest likelihood. However, this simple approach tends to overlook

the fact that the more complex models may have many times the number of parameters of the simpler models.

One method of choosing between competing models is to test whether the addition of extra parameters (implied by the more complex model) does in fact result in a *statistically significant* improvement in fit, relative to the simpler model.

It is possible to perform such pairwise comparisons between most of the four models by testing simple parameter restrictions upon the more complex model as follows:

(a) ML vs BM — Test whether the parameters pertaining to the auxiliary variables are jointly equal to zero;

(b) MM vs BM — Test whether the δ_j $(j = 1, \ldots, m)$ coefficients are jointly zero;

(c) MML vs BM — Test whether the parameters pertaining to the auxiliary variables *and* the δ_j $(j = 1, \ldots, m)$ coefficients are jointly equal to zero;

(d) MML vs ML — Test whether the δ_j $(j = 1, \ldots, m)$ coefficients are jointly zero;

(e) MML vs MM — Test whether the parameters pertaining to the auxiliary variables are jointly equal to zero.

Under the assumption of asymptotic normality of the maximum likelihood estimates, it can be shown that

$$-2\,[L(\beta_R) - L(\beta_U)] \sim \chi^2_k$$

where $L(\beta_U)$ is the maximum value of the log-likelihood function in the absence of any parameter restrictions and $L(\beta_R)$ is the maximum value when the restrictions apply. k is the number of parameter restrictions imposed. This test is known as the *Likelihood Ratio Test* (LRT).

A simpler (although perhaps less theoretically sound) procedure for selecting the "best" model from a range of competing alternatives has been proposed by Akaike (1973). Akaike's Information Criterion (AIC), which is defined by

$$AIC = -\tfrac{2}{N} L(.) + \tfrac{2p}{N} \quad ,$$

includes a penalty for the loss of degrees of freedom (p is the dimension of the parameter vector). The preferred model is the one which has the smallest AIC.

The LRT and AIC can also be used to eliminate uninformative variables from the ML and MML models, and to establish whether post-stratification of the sample is worthwhile.

10. BIBLIOGRAPHY

AKAIKE, HIROTUGU "Information Theory and an Extension of the Maximum Likelihood Principle" in B.N PETROV and F. CSAKI (eds.) *2nd International Symposium on Information Theory*, pp. 267-281. (Budapest: Akademiai Kiado, 1973)

ASPDEN, CHARLES; MALTI JAIN; PAUL SULLIVAN and FRED VON REIBNITZ "Study into Source Data and the Quarterly National Accounts: Report to Management", *internal ABS report*, July 1994.

FRASER, BRUCE "Options for Using Linear Models to Improve Estimates of Input-Output Expense Items and Reduce Respondent Load", *internal ABS investigation (in progress)*, June 1998.

ROGERS, RUSSELL "Study of Methodologies to Support Input/Output Requirements for Detailed Expenses Data from the Annual Economic Collections", *internal ABS report (draft)*, June 1998.

SULLIVAN, PAUL "Review of Source Data Collection and Use for Compilation of National Accounts using Input/Output Framework", *internal ABS report*, December 1997.

THEIL, HENRI "A Multinomial Extension of the Linear Logit Model", *International Economic Review,* vol. 10, pp. 251-259, 1969.

WELSH, ALAN and EDWARD SZOLDRA "Allocation of Income/Expenditure in Input/Output Tables", *internal ABS report*, September 1997.