Feasibility assessment for using small domain estimation methods to enhance expenditure statistics for Maori

Stephen Haslett and Rachael Viles

Report prepared for Statistics New Zealand

30 June 2005

Acknowledgement

This study would not have been possible without the able assistance of a number of Statistics New Zealand staff, most notably Mike Doherty and John Lopdell, and of course the staff in the Maori Statistics Unit.

Summary

1. The intention of this feasibility study was to assess whether small area estimates of Maori expenditure were possible using existing Statistics New Zealand data sources. The success of this phase (Phase 1) of the research and the Statistics New Zealand decision on whether to proceed to Phase 2 are to be determined by two criteria :

   - Assessment of the feasibility of producing datasets with a sufficient number of matching variables to warrant further analysis.
   - Assessment of the utility of preliminary regression models based on sample survey data alone, fitted using appropriate sample survey based statistics software

   This report documents the results of the matching exercise and the preliminary regression testing.

2. There were a number of Statistics New Zealand surveys considered as possibilities for modelling, in conjunction with the 2001 Census of Population and Dwellings. These were Household Economic Survey, Household Savings Survey, Household Labour Force Survey (including Income and Child Care Supplements), and SoFIE. Of these only the Household Economic Survey (HES) and SoFIE contained any of the required expenditure information. The focus of this preliminary study is HES, because SoFIE has a strongly computer-based genesis and detailed documentation including questionnaires proved difficult to obtain.

3. Preliminary analysis of HES and the census indicates that matching on key variables will be possible, although further detailed research will be required to complete matching, as part of Phase 2 should that phase proceed. The preliminary regression models fitted to the HES survey data indicate which potential matching variables deserve greatest focus within Phase 2.

4. Phase 1 did not attempt to fit a full small area estimation model, but instead focused on the HES data and used this survey to assess whether total expenditure per capita could be adequately modelled, using software specifically designed to fit regression models to survey data. Full assessment would require further analysis using both HES and census data, but preliminary results using HES and relatively few explanatory variables give an R^2 of 0.55. The current Statistics New Zealand equal allocation for expenditure and income to people within households and, given the regression models indicate this is a first approximation only, this tends to lower R^2. In these circumstances the regression results are encouraging. To improve models, such allocation issues, at both household and individual level, warrant further consideration in any later research phase.

5. Although there are important aspects that require further research, the preliminary regression analyses based on HES data alone indicate that prediction of expenditure patterns by ethnicity is possible, and provide reasonably strong support for extension of the project to Phase 2.

1. Introduction

Small area estimation methods have been developing significantly over the last 20 years. A recent technical reference is Rao (2003), and there are earlier articles for example by Rao (1999) and Ghosh and Rao (1994). Small area methods fall within a wider context of techniques for the analysis of complex survey data – see for example Chambers and Skinner (2003) and Skinner et al (1989).

Small area methods generally utilise two data sources, most commonly a survey and a census, although sometimes two surveys usually with rather different sample sizes. The census information is used to provide improved estimates based on a statistical model.

It can be useful to distinguish two general situations:

1. The census and survey are contemporaneous and although they share matching variables useful for prediction, the variable of particular interest is measured only in the survey and not the census. A model is fitted to the survey data, and applied in the form of a predictor (which often includes random effects) to the census data. The model is fitted at respondent level, but aggregated up into small areas since at this finest level estimates are too inaccurate. Random effects may be fitted using the bootstrap (Efron and Tibshirani, 1993). Such methods are illustrated in Elbers et al (2001a), (2001(b), Jones and Haslett(2004) and Haslett and Jones (2005).
2. The census and survey are not contemporaneous but the variable interest, together with a range of matching variables is available for *both* census and survey. The statistical model is now fitted to the census data. The survey is used to update certain parameters in the model (usually the lower order ones). For further discussion see Noble, Haslett and Arnold (2002) and Noble (2004).

Small area methods have been most often used to provide finer breakdowns of statistics by area than are possible from the survey alone, but the fine subdivisions within which such statistics are formed may be subpopulations rather than areas. In this sense the term “small area estimation” may be misleading, and the alternative “small domain estimation” may be more appropriate in some circumstances, such as for expenditure patterns for Maori.

The motivation for the current project is that these methods may provide a way to significantly improve the range and quality of statistics on Maori without the additional cost and respondent burden of increasing sample size in Statistics New Zealand surveys. Other methods are available, such as use of multiple frames, and supplementary sample designs (including targeting and oversampling) but these methods have major disadvantages: even if considerable additional funding is available, accuracy of estimates for groups of particular interest may show little improvement, and those for other groups can be significantly worse. For a more complete discussion of the comparative advantages and disadvantages of each of these methods for ethnicity-based estimates in the New Zealand context, see for example Gray (2005), Haslett(2005), Wells (2005).

The detailed focus of this study, which has been determined by the Maori Statistics Unit at Statistics New Zealand, is expenditure patterns, i.e. the Maori Statistics unit has specified the priority areas for the application of small domain estimation. Sources of information have been identified by the unit and other Statistics New Zealand staff. It is important to be clear that, in the present study, ethnic groups form the small domains, and Maori is the most relevant of these, but that there is no geographically based estimates within ethnic groups being considered.

Whether it is possible to have both domain (ethnic) and area (geographical) estimates at a fine level simultaneously is not the subject of this report; the focus is on ethnicity alone.

As described by Statistics New Zealand, the particular business objectives of the current study have been:

- To work on variables the Maori Statistics Unit sees as priority for small domain estimation
- To explore the matching of variables between the 2001 Census and survey data, and to fit preliminary small area estimation models to survey data.

The original intention of this study was to fully apply small area/domain estimation techniques to estimation of Maori statistics. The project has however been divided into two phases, the first of which assesses viability and the second, which would involved the application of selected methods to conduct small area estimation should the first phase prove sufficiently successful.. The benefit of this restructuring is that Phase 2 proceeds only if Phase 1 indicates viability and that Phase 2 can be rather better specified as part of a multi-phase study.

Statistics New Zealand has specified the success of the Phase 1 study being determined by its:

- *Assessment of the feasibility of producing datasets with sufficient matching variables for further analysis.*
- *Running preliminary regression models on sample survey data using appropriate software.*
- *Documentation of results of regression testing.*

Because the model for Maori expenditure is of the first type above (i.e. there is a variable to be predicted, namely expenditure, that is not collected in the census), the initial regression results are based on modelling survey rather than census data. At the preliminary stage of the study encompassed by this report, the application of that model to census data with the addition of the required random effects at primary sampling unit (psu), household and individual level, and aggregation, is not included.

Regressions using survey data alone can often give an indication of how successful the final small area estimation model will be. Certainly high percentage of variance explained, i.e. high R^2, is encouraging and a low R^2 less so. However, it is possible, because of the random effects, to have a model with a high R^2 with strong psu (and to a lesser extent household) effects that does not give good small domain estimates, and vice versa. In this sense the current study is preliminary.

Note that the regression model fitted to survey data requires specialised algorithms, especially for assessment of standard errors and statistical significance (Skinner et al, 1989; Chambers and Skinner, 2003). The software used in this preliminary study is the statistical package Sudaan, which has such algorithms.

What is contained in this report is the feasibility assessment for small area estimation of a key expenditure variable specified by Statistics NZ, including a discussion of the matching of explanatory variables between survey(s) and census. Also included in the report is a preliminary assessment of whether the number of such matching variables and the preliminary regression models are sufficiently promising for small area estimates for Maori to be feasible.

In summary, the benefit of this research is that it has the potential to impact positively on the Official Statistics System by improving accuracy of Maori expenditure statistics, while still reducing respondent burden and hence providing a cost-savings approach to studying Maori. Given the required Statistics New Zealand support, the methods that could be developed later in Phase 2 will nevertheless be necessary to provide the technical means to improve these expenditure pattern estimates for Maori. The methods developed here in Phase 1, while useful, are not sufficient taken on their own.

2. Choice of datasets

For small area estimation to be most effective, all data are initially required at unit record (individual) level for both census and surveys. Survey information must contain design information (stratum, psu, household membership, plus sample survey weight for each respondent) so that survey based regression can be fitted in such survey analysis packages as Sudaan. For Phase 2, the 2001 census data are also required, for all ethnicities, not only Maori.

A range of survey data sets were considered. These included Household Economic Survey, Household Savings Survey, Household Labour Force Survey (including Income and Child Care Supplements), and SoFIE. Of these however only the Household Economic Survey (HES) and SoFIE contain any of the required expenditure information. HES data is available in the form of a relational database in SAS and the questionnaire is available in question by question format reproducible as hardcopy. SoFIE however has a strongly computer-based genesis and only very detailed flowcharts used for programming the computer aided interview interface were available, rather than the questionnaire itself; the structure of and variables in the computerised SoFIE database are consequently difficult to determine using statistical software alone. For this reason the sole focus of Phase 1 is the HES, within which there is detailed expenditure related information, hardcopy questionnaires, a range of variables that are potential match with those in the census, and of course ethnicity.

3. Questionnaires and related information: Sources

Statistics New Zealand questionnaires can be found at:
http://www2.stats.govt.nz/domino/external/quest/sddquest.nsf/byName?openview

Further relevant detail can be found at:

Census
Questionnaires:
http://www.stats.govt.nz/census/2001-definitions-questionnaires/chapter-10.htm
Technical information
(Database: "Survey Methods & Analytical Sup"; Subject: "Data Dictionary for Census 2001 output"; Author: Pat Coope; Date Created: 07/02/2002)
(Database: "Survey Methods & Analytical Sup"; Subject: "List of 2001 Census variables extracted from SYBASE"; Author: Andy Smith; Date Created: 06/06/2002)

Household Labour Force Survey Questionnaire:
http://www2.stats.govt.nz/domino/external/omni/omni.nsf/bf3c32a862ae07cd4c25656e000031c6/56c4386f9f792657cc256d97007e3eda?OpenDocument#Variables
HLFS Income supplement
(Database: "Survey Methods & Analytical Sup"; Subject: "New Zealand Income Survey (June 1997 Qtr -) - Technical Description"; Author: Matthew Cronin; Date Created: 12/06/2002)

Household Savings Survey - see p145 for household questionnaire:
http://www2.stats.govt.nz/domino/external/web/prod_serv.nsf/874ea91c142289384c2567a80081308e/cbcbcaf5cee17dc1cc256c3d0078dadb/$FILE/Net%20worth.pdf
HSS (contains links to Ministerial Approval Doc):
(Database: "Survey Methods & Analytical Sup"; Subject: "Household Savings Survey (HSS) 2001 - Technical Description ** DRAFT **"; Author: Travis Yates; Date Created: 15/08/2001)

Household Economic Survey (HES) - information on variables collected:
http://www2.stats.govt.nz/domino/external/omni/omni.nsf/outputs/household+economic+survey
(Database: "Survey Methods & Analytical Sup"; Subject: "HES 2000/2001 - Technical Description"; Author: Matthew Cronin; Date Created: 19/12/2001)

4. Research methodology

The initial choice of which survey datasets were appropriate lay with Statistics New Zealand, and the decision to focus on expenditure and spending pattern related information for Maori was the decision of the Maori Statistics Unit within SNZ. A statement of the Maori Statistics Unit's perspective can be found in Appendix 1.

The intention of Phase 1 was limited to assessment of whether suitable data for small area estimation exists at Statistics New Zealand, plus fitting of preliminary models. Because the full small area estimation models cannot be fitted in full in Phase 1, the required decision is asymmetric: Either SNZ decides that there is not sufficient information on ethnic spending patterns to warrant further research at this stage, or that there is sufficient information to proceed to Phase 2 and as an initial part of that second phase to undertake further feasibility testing.

The central question of Phase 1 then is whether there is sufficient match between survey and census for predictor variables to be used in regression models. The matching is required because in the type of small area model used for the expenditure data, the fitted survey model is applied to the census data using the *same* predictor variables as in the survey. The most important of the predictor variables is ethnicity itself. The expenditure information (i.e. the variables to be predicted, which need to be in the survey but are not in the 2001 census) also needed to be checked for relevance.

The research methodology used for checking matching and relevance is reasonably straight forward if somewhat time consuming to implement. First the survey questionnaire and the census questionnaire are scanned in detail to determine candidate matching variables (which are to be used as predictors). For details on candidate matches between the Household Economic survey and the 2001 census, see Appendix 2. Because the census itself does not collect expenditure data, none of these variables involve expenditure information. The matching variables are then cross-checked in the following way:

1. Categories within each variable are considered in detail, the matching categories are noted and, where matching is not clear, categories are left at their most disaggregated level.
2. Percentages or means are calculated within each category for both survey and census. Where variables are not categorised or consist of a single 'category', overall percentage or means are calculated.
3. For the survey means (which are calculated using survey weights), provisional standard errors are also calculated. These must be calculated using a sample survey package (such as Sudaan) to be correct, but provisional estimates can be calculated for example in SAS and some design effect assessed from another source used to calculate provisional confidence interval for the mean. If no design effect estimates are available (either via Sudaan or another suitable package, or from an external

source) a reasonably conservative strategy (especially for a cluster sample) is to use a design effect of one (which is the assumption behind the SAS weighted estimates of standard error).

4. Candidate variables and categories from survey and census are checked against one another. This can, for key variables, also be done using cross-tabulations of variables two or more at a time. Where discrepancies are found these may be solvable by amalgamating categories. Where categories have been left disaggregated, this information provides a basis for amalgamation to maximise match.
5. Where a suitable match can be found, the variable in question is categorised as a candidate predictor variables for the regression. Otherwise, like other non-matching variables, it cannot be used in the survey based regression model or in the model for the final small area estimates.

Detailed information on data sources for HES can be found in Appendix 5. Note that four forms are used to record the information collected in HES:

Household Questionnaire (HQ) - (1 per household)
Income Questionnaire (IQ) - (1 per adult)
Expenditure Questionnaire (EQ) - (per household)
Expenditure Diary - (1 per adult)
(plus the internet questionnaire in 2001)

Given the available matching variables, and a particular choice of expenditure variable (e.g. total expenditure per capita) available from the survey, a regression model can be fitted to the survey data. Phase 2 will depend on Statistics New Zealand's assessment of whether these survey regression models show sufficient promise as small area estimation models.

The Maori Statistics Unit's specifications focus on individual expenditure rather than family expenditure. Each choice has advantages and disadvantages in terms of small area estimation modelling. At household level the obvious choice for initial model testing was total household expenditure (TotHHExp), since if models at this level of variable aggregation were infeasible, finer subdivisions were unlikely to be particularly useful either.

Family level data requires a definition of Maori household – the definition used in Phase 1 is that the Maori household is defined as any household with at least one Maori adult (aged15+ and over). SNZ has used this definition in producing tables of Maori expenditure information (see for example Appendix 4) and the definition also conforms to the Maori Statistics Unit's specification (see Appendix 1). Other definitions are of course possible (e.g. proportion of individuals in each household who have specified Maori ethnicity as one of their ethnic categories). The influence of definition of Maori household in small area models would need to be a central consideration at Phase 2 of the project, as current Statistics New Zealand definitions have stronger effects on small area model based statistics than on design-based means.

There is a different but similar issue when using individual information. In this case ethnicity is well specified (subject to matching being possible between survey and census, of course) but some expenditure is household based (e.g. electricity, rent, rates etc) and some method of allocation of such expenditure to individuals is required. For example, one possible allocation is that such expenditure be apportioned equally to all people in the household aged 15+, and this is what has been used in Phase 1 of this study.

Small area models applied to individual rather than household data are more difficult to fit. This is not apparent at Phase 1 where the random psu, household (and individual) level random components are not used with census data (as required in any full model) to produce small area estimates per se, but only initial (sample based) regression models are considered.

At individual level (in comparison with household based models) it may also be necessary to consider additional individual level variables such as age when fitting the random component.

As a consequence (as for the question of being able to have both domain, i.e. ethnic, and area, i.e. geographical, estimates at a fine level simultaneously), whether it will be possible to fit sae models to both individual level data and household level can only be fully answered at Phase 2.

4. Matching variables in HES and the census: 2001

Appendix 2 contains candidate matching variables for HES and census 2001. Although a good match for all variables that would be used in Phase 2 for the survey regression model (and later applied to the census data) is essential, matching on ethnicity is particularly important. Within the ethnicity categories, matching for Maori is the most important (given the focus of this small area estimation project).

Using the sample survey weights for HES (designated by the variable 'weight' in the relevant HES data sets) and the census data gives the following two tables:

HES Ethnic

Variable	N	NMiss	Mean	Std Error
ethn1	7374	11	0.7364744	0.0051306
ethn2	7374	11	0.0664307	0.0029003
ethn3	7374	11	0.1414437	0.0040584
ethn4	7374	11	0.0277847	0.0019141
ethn5	7374	11	0.0129047	0.0013144
ethn6	7374	11	0.0088258	0.0010893
ethn7	7374	11	0.0030907	0.000646444
ethn8	7374	11	0.0020652	0.000528706
ethn9	7374	11	0.0044698	0.000776871
ethn10	7374	11	0.0010568	0.000378388
ethn11	7374	11	0.0054009	0.000853559
ethn12	7374	11	0.0253551	0.0018308
ethn13	7374	11	0.0208125	0.0016625
ethn14	7374	11	0.0086518	0.0010786
ethn15	7374	11	0.0192955	0.0016020

Census Ethnic

Variable	N	N iss	Mean
ethn1	3661928	158821	0.7393373
ethn2	3661928	158821	0.0758909
ethn3	3661928	158821	0.1439321
ethn4	3661928	158821	0.0315421
ethn5	3661928	158821	0.0144804
ethn6	3661928	158821	0.0111930
ethn7	3661928	158821	0.0055206
ethn8	3661928	158821	0.0016980
ethn9	3661928	158821	0.0019665
ethn10	3661928	158821	0.0021977
ethn11	3661928	158821	0.0087601
ethn12	3661928	158821	0.0298335
ethn13	3661928	158821	0.0167191
ethn14	3661928	158821	0.0142586
ethn15	3661928	158821	0.0076523

Although the standard errors here derive from SAS (and consequently assume simple random sampling) and are too small by a factor (the inverse of the design effect for HES), the Maori category (ethn3) is not significantly different at 14.1% for the HES and 14.4% for the census, even with the underestimated standard error for HES.

Other ethnicities may require some adjustment at Phase 2 (e.g. Ethnicity 2). Some differences have already been recognised by SNZ (see for example Appendix 6) and a calibration scheme has been introduced to adjust various sample percentages from HES to their census analogues. The scheme was not however introduced at the time of the 2001 HES data being used in this study. (The year of the HES used here, as noted earlier, is determined by the need to match the HES data to the date of the census in order to apply relevant sae methods.) Further, when calibration was introduced, it used the 1996 census figures for calibration, since the 2001 census figures were not yet available.

For Phase 2, the best solution to the non-match for some ethnicity groups then is to use the calibration scheme of Appendix 6, but apply the calibration to the 2001 HES so as to match the required 2001 (rather than 1996) census sub-totals. The ethnicities percentages for HES and census would then match precisely. One caution is nevertheless required. The census includes a non-residential population component and this data must be removed (because it cannot be matched with HES which excludes such people). This needs to be done *before* calibration.

For sae models, calibration changes the weighting system for the survey data slightly. While in any published small area estimates from HES it is recommended the calibration be done at the same level of detail given in Appendix 6 (so as to meet SNZ current protocols and meet consistency standards) the small area estimates themselves are likely not to be strongly sensitive to calibration. This is because we are not focussing on estimating ethnic percentages in subcategories (for which such calibration would be essential) but instead are looking at global expenditure patterns as they apply to Maori. In the latter case the sae model uses a set of regression coefficients to produce individual and/or household level estimates which are then summed. Small changes in relative weights (even of the order of 10-15% needed in this case) are unlikely to alter regression coefficients substantially.

At this preliminary stage of the sae for Maori research, focus is on another aspect of the sample survey regression, again one unlikely to be substantially affected by such weighting changes, namely the percentage of variance explained by the regression model applied to the HES data.

From here on in this report (i.e. for Phase 1 only) we exclude the non-residential census population from all calculations and use the Statistics New Zealand HES variable ‘weight’ when calculating percentages from HES for *all* matching checks on all variables.

This weighting was applied and used for testing all candidate matching variables (including all categories where those variables were categorical).

Appendix 3 contains details of census and HES categories for all potential matching variables, and direct comparison of percentages within categories.

Some variables match well between HES and census and some less so. Some non-matching occurs even when allowance is made for the complex design for HES and tabulated SEs are treated as underestimates. Further study and adjustments (where possible) to improve matching would be required at Phase 2 before such variables can be used as regressor variables for ethnic based expenditure in any final small area model.

At this stage however, since it is a time consuming exercise and the time frame is rather limited, further study of matching has been postponed beyond the ‘feasibility stage’.

Instead, to assess priority for checking each of the candidate matching variables, a set of preliminary Sudaan based models have been fitted to the HES data alone. The intention is two-fold: to determine which regressor variables warrant the most immediate attention in

terms of formalising matching; and to assess whether *any* survey regression model at all gives an adequate preliminary fit, even if matching were not an issue.

The underlying logic is that if a survey regression (even assuming perfect match of census with HES) does not fit well, then there is little utility in proceeding further with the matching, since then, even in the best matching case scenario used for the preliminary survey regressions, small area modelling is unlikely to yield the required results. Alternatively, if the regression models provide good preliminary fit (as measured by R^2), further research is warranted. Of course, matching then needs to be checked further before a final survey model can be applied to census data to produce person-level or household-level expenditure predictions (which are then amalgamated into small area estimates) and this remains a time-consuming exercise.

It should be noted that, even given high R^2, it is only at the final stage of fitting a full small area estimate model that a full assessment of the viability of small area estimates for Maori expenditure would be complete, since it is theoretically possible to have high R^2 for the survey based regression and poor small area estimates via the census data. However, in general this only occurs where the regression includes subgroup effects that are directly linked to the 'areas' for small area estimates themselves, eg a Maori indicator as an explanatory variable for small area estimates of Maori, and where this alone is a major determinant of R^2. The technical explanation of why this situation causes problems is that if every Maori household has such an indicator, its standard error is not a decreasing function of increasing sample size (since it occurs for every Maori household) and so it remains a major limitation to shrinking standard errors for the corresponding small area estimates.

Nevertheless, in the interim, by way of summary, R^2 and other survey regression diagnostics, provide the best (and usually reliable) preliminary indicator of project feasibility that are available without fitting the full sae model.

5. Analysis of HES for assessing feasibility of predicting Maori expenditure

At the next stage of the project, the HES design was specified in Sudaan as a stratified cluster design and the sample based regression was fitted, making adjustment for the complex survey design.

The weighting used in this preliminary analysis was the Statistics New Zealand weighting variable 'weight', which was not rescaled to match all census ethnic percentages, and the design was specified as 'with replacement' for variance estimation purposes. Although these aspects deserve further study in Phase 2, they are not critical to the first phase general assessment of feasibility of small area estimation of expenditure patterns for Maori.

The following variables were found to be the most important in the regression for predicting total expenditure per capita, which was log transformed (and called lntotesp_pp) after data checking and before model fitting to stabilise variance, in line with standard practice):

Ethnicity:

```
ethn_NZEurop
ethn_OthEurop
ethn_NZMaori
ethn_CIMaori
ethn_PIg1
ethn_PIOthg
ethn_Asian
ethn_Oth
```

Age
    agegp
Occupation
    occup
Education
    educqual
Marital Status
    maritstat
Whether seeking work
    seekwork
Type of living accommodation
    Residencetype
Rent payments if any
    weekrentpp
Number of rooms per person in household
    roomspp
Number of cars per person
    carspp
Log income per person
    Lnincomepp
Number of people in household
    numperson

Complete definitions of each of these variables can be found in Appendix 6.

The following preliminary model was found after testing a sequence of possible models:

```
* log transfomed expenditure and income with selected interactions, but deliberately
    not including interactions with ethnicity, since these are the small domains of interest
    so increasing R**2 by using these only breaks up the sample in terms of the model so
    little additional real gain in terms of saes would be expected from such a strategy;

proc regress data=HEScomb3 DEFT1 design=wr;
 nest stratnum psunum;
 class ethn1-ethn15 sex agegp
       refpers Partner Child parent otherrel norel
             educqual
             maritstat jobavail seekwork work occup residencetype;
 model lntotexp_pp = ethn_NZEurop ethn_OthEurop ethn_NZMaori ethn_CIMaori
                     ethn_PIg1 ethn_PIOthg ethn_Asian ethn_Oth
                     agegp
                     occup educqual
                           maritstat       seekwork          residencetype
                                     weekrentpp       roomspp carspp
                           lnincomepp
                           numperson  ;
                   * variables and interactions removed from model:
                                         reduced number of ethnic groups
                                         work occup jobavail
                 sex jobhours bedroompp
                             jobhours_sq refpers Partner Child parent otherrel norel
                             agegp*occup agegp*seekwork agegp*work agegp*educqual work*educqual;
 weight weight;  * other possible weight is variable called respwgt;
                 * note that weights have not at this stage 270605 been adjusted to
                   match Maori percentage in census. this should affect R**2 only slightly.
                   see final feasibility report for details;
run;
```

Note that R^2 is approximately 0.55, even for a model that contains relatively few parameters..

These results give reasonable evidence that Phase2 is warranted. The R^2 value is perhaps not as high as in other modelling of log expenditure variables (see for example, Haslett and Jones, 2005; and Jones and Haslett, 2004). However, the allocation of household expenditure on a purely per capita basis, in line with the usual Statistics New Zealand allocation formula,

reduces the goodness of fit because every adult member in the same household is allocated the same level of expenditure (and income). The model itself indicates this allocation is not optimal, and should be improved as a preliminary to model fitting, since there are for example age group effects found in the model that the allocation formula ignores. This issue is important because the current allocation penalises the model (in terms of a lower R2 value), rather than indicating (as it should) that there is an issue with the expenditure and income allocation mechanisms. This question will have to be resolved in any further model development.

One alternative is to model expenditure at the level of households. This avoids the expenditure and income allocation problem, but raises another issue: how should ethnicity of households be designated? This question is at present addressed by Statistics New Zealand by designating households as Maori if they contain any Maori member. This however makes it more likely that large households will be designated Maori, and (in a parallel manner to expenditure allocation to individuals) blurs the ethnicity designation. Again, this matter requires further research in any later phase of this small area project.

6. Conclusions

The intention of Phase 1 study was to assess whether small area estimates of Maori expenditure were possible using existing Statistics New Zealand data sources. The success of Phase 1 and the Statistics New Zealand decision on whether to proceed to Phase 2 were to be determined by three criteria :

- *Assessment of the feasibility of producing datasets with sufficient matching variables for further analysis.*
- *Running preliminary regression models on sample survey data using appropriate software.*
- *Documentation of results of the preliminary regression testing.*

There were a number of Statistics New Zealand surveys considered as possibilities for modelling in conjunction with the 2001 Census of Population and Dwellings. These were Household Economic Survey, Household Savings Survey, Household Labour Force Survey (including Income and Child Care Supplements), and SoFIE. Of these only the Household Economic Survey (HES) and SoFIE contained any of the required expenditure information. The focus of this preliminary study was HES because SoFIE has a strongly computer-based genesis and documentation including questionnaires was difficult to obtain.

Preliminary analysis of HES and the census indicates that matching on key variables will be possible, although further detailed analysis will be required to complete matching, as part of Phase 2 (should that proceed). The preliminary regression models fitted to the HES survey data indicate which potential matching variables deserve greatest focus and effort as part of any Phase 2.

Phase 1 did not attempt to fit a ‘full’ small area model, but instead focused on the HES data and used this survey to assess whether total expenditure per capita could be adequately modelled, using software specifically designed to fit regression models to survey data. Full assessment would require further analysis using both HES and census data, but preliminary results using HES and relatively few explanatory variables give an R^2 of 0.55. The current Statistics New Zealand equal allocation for expenditure and income to people within households tends to lower R^2 and deserve further consideration., so that this regression result is encouraging.

Although important aspects require further research, the preliminary regression analyses based on HES data alone provide reasonably strong support for extension of the project to Phase 2.

References

Chambers R.L and Skinner C.J. (eds) (2003) *Analysis of Survey D*ata. Wiley.

Efron B. and Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

Elbers C., Lanjouw J.O. and Lanjouw P. (2001) Welfare in villages and towns: micro-level estimation of poverty and inequality, unpublished manuscript, The World Bank.

Elbers C., Lanjouw J.O., Lanjouw P. and Leite P.G. (2001) Poverty and inequality in Brazil: new estimates from combined PPV-PNAD data", unpublished manuscript, The World Bank.

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

Gray, A. (2005) Strategies for New Zealand household surveys overlapping Maori and Pacific peoples, Paper 16.2, IAOS Satellite Meeting, Wellington, NZ, 14-15 April 2005, http://www.stats.govt.nz/about-us/events/satellite-meeting/abstracts-and-papers.htm

Haslett, S. (2005) Multiple frames, supplementary sample designs and small domain estimation, Paper 16.4, IAOS Satellite Meeting, Wellington, NZ, 14-15 April 2005, http://www.stats.govt.nz/about-us/events/satellite-meeting/abstracts-and-papers.htm

Haslett, S. and Jones, G. (2005) *Local Estimation of Poverty in the Philippines*, Manila, June 2005, Published jointly by Philippines National Statistics Co-ordination Board and the World Bank.

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

Noble, A. D. (2004) *Small area estimation via generalized linear models*, unpublished Ph D thesis, Massey University

Noble, A., Haslett, S., and Arnold, G. (2002) Small area estimation via generalized linear models, Journal of Official Statistics, Vol 18, No 1, p45-60.

Rao J.N.K. (1999) Some recent advances in model-based small area estimation, *Survey Methodology*, 23, 175-186.

Rao J.N.K. (2003) *Small Area Estimation*. Wiley.

Skinner C.J and Holt D. and Smith T.M.F. (eds) (1989) *Analysis of Complex Survey Data*. Wiley.

Statistics New Zealand (2002) *Toward a Maori Statistics Framework*, Statistics New Zealand, Wellington

Wells, E. (2005) Targeting and screening as strategies for oversampling: cautionary examples, Paper 16.3, IAOS Satellite Meeting, Wellington, New Zealand, 14-15 April 2005, http://www.stats.govt.nz/about-us/events/satellite-meeting/abstracts-and-papers.htm

Appendix 1

PROPOSAL TO PRODUCE STATISTICS ON HOUSEHOLD SPENDING PATTERNS FOR MÄORI

Mäori Statistics Unit
Statistics New Zealand
22 March 2005

The Mäori Statistics Unit proposes that the *Small Area Estimation* project investigate the feasibility of extracting data on Mäori household spending patterns from relevant official statistical data sources. Furthermore, that the emphasis of the study be within the context of the households of which Mäori are a part of.

Background

The recent Hui Taumata focussed on three broad themes for Mäori development: people, assets and enterprises. From a statistical perspective, it became clear that there is an increasing demand for economic measures that help Mäori understand its social, economic and cultural position in society. In meeting such demands, Statistics New Zealand has a long term commitment to ensure that the statistical needs of Māori can be identified and progressively met. This includes the provision of official statistics that are relevant to Māori.

The department has developed a number of initiatives that contribute towards this. First, the development of a framework for Māori statistics that presents the department's thinking on what the statistical needs that Mäori are seeking might be. As part of the consultation process the Mäori Statistics Unit (MSU) sought comment from Mäori on the department's thinking and advice of what the priority areas may be if the framework is to be populated.

Second, as part of its statistical capability building pilot project, the department supports a small number of Mäori communities to prepare statistical profiles of their respective constituencies. In discussing specific statistical needs, it became clear that these communities were not so interested in Mäori employment, housing, child-rearing *per se*, but the effects of these issues on the families and households of which Mäori are a part from a social, economic and cultural perspective.

Due to difficulties in defining a "Mäori household" or "Mäori family" no standard output on Mäori in family or household environments is produced by the department (especially from the Census of Population and Dwellings).

As a result, the MSU proposed that the department produce a series of regional analytical reports using data from the 2001 Census. The purpose of these reports were to examine the key social, economic and cultural issues confronting Mäori children and youth by broad regional areas within the context of the families and households of which they are a part.

Furthermore, an analysis of this nature was also endorsed by Te Puni Kökiri in the most recent high level meeting between the two departments. TPK also pointed out that the demand amongst Mäori organisations was for localised rather than national level data.

Proposal

A noticeable data gap in all of this work was the household spending patterns of Mäori within the context of the families and households of which Mäori are a part. The Mäori Statistics Unit proposes that the *Small Area Estimation* project:

- investigate the feasibility of extracting data on Mäori household spending patterns from relevant data sources; and
- undertake the feasibility study within the context of the households of which Mäori are a part.

We have assumed that a data key source will be the Household Economic Survey that includes data on income and expenditure. We prefer to use the term "spending patterns" rather than expenditure. The main reason being that we are more interested in what different households spend their income on, than expenditure as such.

Framework for Analysis

This investigation differs from the norm in that whilst the individual remains the basic unit of measurement, the focus of interest is on the individual's household and family environment. By taking this approach, it could be possible to look at issues like living arrangements of children, spending patterns, employment and unemployment of parents, etc., in the context of the households and families that are affected, which in turn affects the individual.

Subject population

The subject population is Mäori ethnic group population living in private dwellings.

Household and Family

For the purposes of this project, the focus is not on whether the family or household is a "Mäori family" or "Mäori household". Rather, it is on the individual and the family or household of which he and she is a part. Thus, the problem of defining a "Mäori" family or a "Mäori" household does not arise.

Primary variables

In respect to this project four variables are especially important:

1. Age - in terms of studying the effect of spending patterns on distinct stages in the life cycle. For example the effects of spending patterns on Mäori children within the households and families that they are part of.
2. Location – to provide scope for producing information at a geographic level that meets Mäori community need.
3. Household composition and family type – to analyse the differences (if any) between various types of households and families.
4. Household expenditure group – to analyse spending patterns (in terms of $'s).

Secondary variables

We have assumed that the potential to append other variables depends largely on the success or otherwise of being able to match relevant data sources to the Census. With this in mind

- Gender of individual
- Dependency status of individual (especially children and youth)

- Role of individual in the household
- Tenure of dwelling (rented/owned by household member)
- Household crowding
- Mobility (moved/has not moved in last 5 years; in year before 2001 census)
- Ethnicity of other household members (Mäori/not Mäori)
- Highest qualification of individual
- Ability to speak te reo Mäori (speaks Mäori/does not speak Mäori)
- Median income (personal, family, household)
- Work and labour force status of individual
- Income source

Appendix 2

HES and Census Candidate Matching Variables:

The table below has HES and Census variables that might be able to be matched. Matching variables may then be used as predictors for expenditure. The table includes the variable description and dataset name, the corresponding questionnaire number (and which questionnaire e.g. HQ - Household questionnaire) and which dataset. The last column notes any possible differences between the HES and corresponding Census variable.

HES Description	HES Var name	HES Qnnaire and Qn no.	HES Dataset	Census Description	Census Var name (sybase)	Census Qnnaire and Qn no.	Census Table	Differences in Var
Age 0 1 2 ...	age	(2) - HQ	h0001p(ix)	**AGE V1.0** 000 Less than 1 Year 001 1 Year ... 120 120 Years and Over	age_code (OR birth_day_code birth_mnth_code birth_yr_code)	Q4 - IF	Individual	Derived variable from DOB and date of interview vs census night
Sex 1 Male 2 Female	sex	(3) - HQ	h0001p(ix)	**SEX V1.0** 1 Male 2 Female	sex_code	Q3 - IF	Individual	OK
Ethnic Group 11 NZ European/Pakeha 12 Other European 21 NZ Maori 31 Samoan 32 Cook Island 33 Tongan 34 Niuean 35 Tokelauan 36 Fijian 37 Other Pacific Island Groups 41 Southeast Asian 42 Chinese 43 Indian 44 Other Asian 50 Other 99 Non-response	ethnic	(4) - HQ	h0001p(ix)	**ETHNIC V3.0 Level 2** 10 European nfd 11 NZ European/Pakeha 12 Other European 21 NZ Maori 30 Pacific Peoples nfd 31 Samoan 32 Cook Island 33 Tongan 34 Niuean 35 Tokelauan 36 Fijian 37 Other Pacific Island Groups 40 Asian nfd 41 Southeast Asian 42 Chinese 43 Indian 44 Other Asian 51 Middle Eastern 52 Latin American/Hispanic 53 African (or cultural group of African origin) 54 Other 96 Repeated Value	ethnic_grp1_code ethnic_grp2_code ethnic_grp3_code ethnic_grp4_code ethnic_grp5_code ethnic_grp6_code	Q11 - IF	Individual	Can get up to a 5 digit code from the Census Up to 6 ethnic groups recorded in census, 3 in HES.

				97 Reponse Undentifiable 98 Response Outside Scope 99 Non-response				
Marital Status 1 Living with legal spouse 1 Living with partner 2 Not living with partner, separated 2 Not living with partner, divorced 2 Not living with partner, widowed 2 Not living with partner, never married 2 Not separated but in separate households 9 Non-response	marstat	(5) - HQ	h0001p(ix)	**LEGMARST V1.0** - 3 digit code the lowest level is shown here not further defined	legal_marital_status_code (OR social_marital_status_code)	Q21 - IF	Individual	Some differences in codes. Requires grouping. Social Marital for census incls defacto
Relation to reference person 01 Reference person (RP) 02 Spouse/partner of RP 03 Child (natural,step,adopted,foster) of RP 04 Parent (natural,step.adopted.foster) of RP 05 Other relative of RP 06 Non-relative of RP	rel2ref	(6) - HQ	h0001p(ix)	**RELAT99 V1.0** 4 digit code the highest level is shown below 01 Reference person 02 Spouse/partner of reference person 03 Child (natural, step, adopted, foster) of reference person 04 Parent (natural, step, adopted, foster) of reference person 05 Other relative of reference person 06 Non-relative of reference person 07 Guest/visitor/inmate/patient/resident 44 Don't Know 55 Refused to Answer 77 Response Unidentifiable	reln_to_occupier_code	Q19 - IF	Individual	Census may need grouping - has more categories

				88 Response Outside Scope 99 Not Stated				
Relationship to Family in Household 1 Partner or parent in 1st family 2 Son or daughter in 1st family 3 Partner or parent in 2nd family 4 Son or daughter in 2nd family 5 Partner or parent in 3rd family 6 Son or daughter in 3rd family 7 Member of 4th or other family 8 Non-family person (others in household) 9 One-person household	famrel	(7) - HQ	h0001p(ix)	**HHCOMP99 V1.0** 100 One-family household, not further defined 111 Couple only ... 431 Household of unrelated people 511 One-person household 611 Household composition unidentifiable	hhld_composn_code	Q19 - IF	Dwelling	Differences in codes. Many detailed categories in Census
Highest Educational Qualification 0 *aged under 15 or over 64* 1 *No qualifications* 2 School Certificate 3 UE, Bursary, Scholarship 4 Vocational or trade certificate 5 Bachelor degree or diploma 6 Post-graduate qualification 7 Part degree or other qualification 9 Non-response	heduqual	(9) - HQ	h0001p(ix)	**CEN.QHIGHEST V1.0** 00 No Qualification 01 Fifth Form Qualification 02 Sixth Form Qualification 03 Higher School Qualification 04 Other NZ Secondary School Qualification 05 Overseas Secondary School Qualification 06 Basic Vocational Qualification 07 Skilled Vocational Qualification 08 Intermediate Vocational Qualification 09 Advanced Vocational Qualification 10 Bachelor Degree 11 Higher Degree 97 Highest Qualification Unidentifiable 99 Not Stated	highest_qual_code	Q23 & Q24 - IF	Individual	Census requires grouping
Number of people in household	numpers	(1) - HQ	h0001h	**Number of occupants**	occupant_count_nbr	Q2 & Q19 - DF	Dwelling	
Type of dwelling	IRN8000	2.1 - EQ	h0001he	**DWELTYPE99 V1.0** - 4 digit dwelling	dwell_type_code	Q5 - DF	Dwellin	

				code highest level is shown below 1 Private Dwelling 2 Non-Private Dwelling 9 Not Elsewhere Included 99 Not Elsewhere included 9944 Don't Know 9955 Refused to Answer 9977 Response Unidentifiable 9988 Response Outside Scope 9999 Not Stated			g	
No. of bedrooms	IRN8001	2.2 - EQ	h0001he	**BEDROOMS V1.0** 01 One Bedroom 02 Two Bedrooms 03 Three Bedrooms ... 14 Fourteen or More Bedrooms 44 Don't Know 55 Refused to Answer 77 Response Unidentifiable 88 Response Outside Scope 99 Not Stated	bedroom_ count_code	Q13 - DF	Dwellin g	Census includes caravans and sleepouts as bedrooms, 1-14 or more bdrms in Census
Total no. of rooms	IRN8009	2.2 - EQ	h0001he		room_count_code	Q14 - DF	Dwellin g	Some differences e.g HES includes laundries Census excludes
Tenure of dwelling	IRN8014	2.5 - EQ	h0001he	**TENHOUSE V1.0** 10 Dwelling Owned or Partly Owned by Usual Resident(s), Mortgage Arrangements Not Further Defined 11 Dwelling Owned or Partly Owned by Usual Resident(s), Who Make Mortgage Payments 12 Dwelling Owned or Partly Owned by Usual Resident(s), Who Do Not Make Mortgage Payments	tenure_code	Q8, 9 & 11 - DF	Dwellin g	

				20 Dwelling Not Owned by Usual Resident(s), Rental Arrangements Not Further Defined 21 Dwelling Not Owned by Usual Resident(s), Who Make Rent Payments 22 Dwelling Not Owned by Usual Resident(s), Who Do Not Make Rent Payments 44 DonÆt Know 55 Refused to Answer 77 Response Unidentifiable 88 Response Outside Scope 99 Not Stated				
Person who owns this dwelling	IRN8013	3.1 - EQ	h0001he	**OWNDWEL V2.0** 1 Dwelling Owned or Partly Owned by the Usual Resident(s) 2 Dwelling is Not Owned or Partly Owned by the Usual Resident(s) 4 Don't Know 5 Refused to Answer 7 Response Unidentifiable 8 Response Outside Scope 9 Not Stated	own_dwell_code	Q10 - DF	Dwelling	OK. Small differences in wording
Amount of rent paid	IRN8103	3.4 - EQ	h0001he	**WRENTPD V1.0** 0000 No Rent Paid 0001 $1 Weekly Rent Paid 0002 $2 Weekly Rent Paid ... 9000 $9000 Weekly Rent Paid 9444 Don't Know 9555 Refused to Answer 9777 Response Unidentifiable 9888 Response Outside Scope 9999 Not Stated	weekly_rent_code	Q12 - DF	Dwelling	OK.
Dwelling has electric heating	IRN8617	12.8 - EQ	h0001he		fuel2_code	Q15 - DF	Dwellin	Needs grouping

	IRN8618 IRN8628 IRN8619 IRN8626						g	
Dwelling has mains gas	IRN8627 and ? IRN8621	12.8 - EQ	h0001he		fuel3_code	Q15 - DF	Dwellin g	Needs grouping (include fixed gas heater?)
Dwelling has gas heater	IRN8620 IRN8622 and ? IRN8621	12.8 - EQ	h0001he		fuel4_code	Q15 - DF	Dwellin g	Needs grouping (include fixed gas heater?)
Dwelling has fire	IRN8623 IRN8624 IRN8625	12.8 - EQ	h0001he		fuel5_code(?) fuel6_code(?)	Q15 - DF	Dwellin g	Needs grouping
Dwelling has water heating	IRN8629	12.8 - EQ	h0001he		fuel7_code(?) fuel8_code(?)	Q15 - DF	Dwellin g	Needs grouping. Could match with 'other fuels' in census?
Dwelling has telephone/cellpone	IRN8605 IRN8613	12.8 -EQ	h0001he	**TELEACC V1.0** 0 No Access to Telecommunication Systems 1 Access to a Telephone 2 Access to a Fax Machine 3 Access to the Internet 4 Don't Know 5 Refused to Answer 6 Repeated Value 7 Response Unidentifiable 8 Response Outside Scope 9 Not Stated	telecomm1_code	Q16 - DF	Dwellin g	Needs to be seperated out
Internet charges	IRN2890 IRN2891	12.9 - EQ	h0001he	**TELEACC V1.0** 0 No Access to Telecommunication Systems 1 Access to a Telephone 2 Access to a Fax Machine 3 Access to the Internet 4 Don't Know	telecomm1_code	Q16 - DF	Dwellin g	HES is charges, Census is availability in dwelling

				5 Refused to Answer 6 Repeated Value 7 Response Unidentifiable 8 Response Outside Scope 9 Not Stated				
Total no. of vehicles owned/used	IRN8720	13.1 - EQ	h0001he	**NUMMOTVEH V2.0** 0 No Motor Vehicle 1 One Motor Vehicle 2 Two Motor Vehicles 3 Three or More Motor Vehicles 4 Don't Know 5 Refused to Answer 7 Response Unidentifiable 8 Response Outside Scope 9 Not Stated	motor_vehicle_count	Q17 - DF	Dwelling	May need to removed borrowed vehicles from HES
Currentlyworking	IRN9000	1.1 - IQ	h0001pix	**CEN.JOBIND V1.0** **CEN.WKLFS V1.0**	job_ind_code wklfs_code (OR emp_status_code)	Q27 - IF	Individual	Some differences, HES spread across several qns. HES is currently working; Census uses last week.
Look for paid work	IRN9006	1.10 - IQ	h0001pix	**SEEKWORK99 V2.0**	seek_work_code	Q38 - IF	Individual	OK

Method of finding work	IRN9007	1.11 - IQ	h0001pix	**CEN.NUMJOBSRCH V1.0**	job_search_meth_count_code (?)	Q39 - IF	Individual	
Started a job last week	IRN9008	1.12 - IQ	h0001pix	**AVALWORK99 V2.0**	avail_start_code	Q40 - IF	Individual	OK
Occupation in 1st job	IRN9010	1.13a) - IQ	h0001pix	**NZSCO99 V1.0** 5 digit code, the highest level is shown below.	occupation99_code	Q30 - IF	Individual	
Industry of first job	IRN9013	1.13b) - IQ	h0001pix	**ANZSIC96 V4.0** (7 digit code**)**	industry_code	Q33 - IF	Individual	

1st job hours	IRN9011	1.13c) - IQ	h0001pix	**HRSWORK3NOP V1.0** 001 1 Hour Worked 002 2 Hours Worked ... 168 168 Hours Worked 169 Don't Know 170 Refused to Answer 171 Response Unidentifiable 172 Response Outside Scope 999 Not Stated	work_hrs1_code	Q35 - IF	Individual	Census - 1st job is worked most hours in HES - ?
2nd & 3rd job hours	IRN9021 IRN9031	1.13c) - IQ	h0001pix	**HRSWORK3NOP V1.0**	total_work_hrs_code (OR work_hrs1_code work_hrs2_code)	Q35 - IF	Individual	HES up to 3 job hours individually recorded, Census hours for all other jobs Census Var is total hours
Status if work less than 30 hours	IRN9009	1.15 - IQ	h0001pix		unpaid_acty1_code unpaid_acty2_code	Q41 - IF	Individual	Not sure if good match: HES qn is asked if work < 30

					| unpaid_acty9_code			hrs. Census is have you done past 4 weeks
Total regular and recurring household income	totrrinc	(derived)	h0001pix	**CEN.TOTINC V1.0**	total_income_hhld_code (OR total_income_individual_code)	Q26 - IF	Dwelling Individual	Census - gross income individual (bands) HES - gross household income (exact)
Meshblock				An alternative seven digit meshblock. These codes incorporate enumeration district (first three digits), enumeration sub-district (next two digits) and enumeration numbers (last two digits)	admin_meshblock_code	-	Dwelling	
Primary Sampling Unit Code	PSU	-	h0001he h0001p(i)		PSU	-		

For more information on the Census and HES variables and datasets see;
(Database: "Survey Methods & Analytical Sup"; Subject: "Census Metadata"; Author: Matthew Cronin; Date Created: 08/05/2002)
(Database: "Survey Methods & Analytical Sup"; Subject: "HES 2000/01 - Datasets (location & variables)"; Author: Matthew Cronin; Date Created: 10/07/2002)

Appendix 3:

Matching: Preliminary Results

The following tables, in pairs, look at the possible matching variables from HES and census. In general, direct comparisons are possible. Note however that standard errors as tabulated for HES are from SAS rather than Sudaan are a consequently for simple random sampling and are likely to be underestimates given the complex cluster based design for HES.

HES Ethnic

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
ethn1	7374	11	0.7364744	0.0051306
ethn2	7374	11	0.0664307	0.0029003
ethn3	7374	11	0.1414437	0.0040584
ethn4	7374	11	0.0277847	0.0019141
ethn5	7374	11	0.0129047	0.0013144
ethn6	7374	11	0.0088258	0.0010893
ethn7	7374	11	0.0030907	0.000646444
ethn8	7374	11	0.0020652	0.000528706
ethn9	7374	11	0.0044698	0.000776871
ethn10	7374	11	0.0010568	0.000378388
ethn11	7374	11	0.0054009	0.000853559
ethn12	7374	11	0.0253551	0.0018308
ethn13	7374	11	0.0208125	0.0016625
ethn14	7374	11	0.0086518	0.0010786
ethn15	7374	11	0.0192955	0.0016020

Census Ethnic

The MEANS Procedure

Variable	N	N Miss	Mean
ethn1	3661928	158821	0.7393373
ethn2	3661928	158821	0.0758909
ethn3	3661928	158821	0.1439321
ethn4	3661928	158821	0.0315421
ethn5	3661928	158821	0.0144804
ethn6	3661928	158821	0.0111930
ethn7	3661928	158821	0.0055206
ethn8	3661928	158821	0.0016980
ethn9	3661928	158821	0.0019665
ethn10	3661928	158821	0.0021977
ethn11	3661928	158821	0.0087601
ethn12	3661928	158821	0.0298335
ethn13	3661928	158821	0.0167191
ethn14	3661928	158821	0.0142586
ethn15	3661928	158821	0.0076523

HES Sex

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
Male	7385	0	0.4996611	0.0058187
Female	7385	0	0.5003389	0.0058187

Census Sex

The MEANS Procedure

Variable	N	NMiss	Mean
Male	3820749	0	0.4876811
Female	3820749	0	0.5123189

HES Relation to Reference Person

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
refpers	7385	0	0.3764062	0.0056381
Partner	7385	0	0.2300877	0.0048980
Child	7385	0	0.3190246	0.0054242
parent	7385	0	0.0075345	0.0010063
otherrel	7385	0	0.0280380	0.0019211
norel	7385	0	0.0389089	0.0022504

Census Relation to Reference Person

The MEANS Procedure

Variable	N	NMiss	Mean
refpers	3820749	0	0.3558648
Partner	3820749	0	0.1835316
Child	3820749	0	0.2681917
parent	3820749	0	0.0149253
otherrel	3820749	0	0.0392509
norel	3820749	0	0.0956383
other	3820749	0	0.0425974

HES 1st Job Hours Worked, by 10hr Bands

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
job1hr10	7385	0	0.0466678	0.0024546
job1hr20	7385	0	0.0489177	0.0025101
job1hr30	7385	0	0.0457617	0.0024318
job1hr40	7385	0	0.1502206	0.0041579
job1hr50	7385	0	0.1258596	0.0038600
job1hr60	7385	0	0.0402309	0.0022867
job1hr70	7385	0	0.0136934	0.0013524
job1hr80	7385	0	0.0056559	0.000872720
job1hr90	7385	0	0.0035123	0.000688475
job1hr100	7385	0	0.0015102	0.000451898
other	7385	0	0.5179699	0.0058149

Census 1st Job Hours Worked, by 10hr Bands

The MEANS Procedure

Variable	N	NMiss	Mean
job1hr10	3820749	0	0.0386168
job1hr20	3820749	0	0.0425359
job1hr30	3820749	0	0.0377952
job1hr40	3820749	0	0.1503001
job1hr50	3820749	0	0.0979965
job1hr60	3820749	0	0.0347309
job1hr70	3820749	0	0.0110787
job1hr80	3820749	0	0.0049687
job1hr90	3820749	0	0.0025134
job1hr100	3820749	0	0.0028814
other	3820749	0	0.5765825

HES Available to Start Work Last week

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
availwk	7385	0	0.0469454	0.0024616
notavailwk	7385	0	0.2407925	0.0049757
other	7385	0	0.7122621	0.0052683

Census Available to Start work last week

The MEANS Procedure

Variable	N	NMiss	Mean
availwk	3820749	0	0.0738423
notavailwk	3820749	0	0.1894506
other	3820749	0	0.7367071

HES Looked for work in the last 4 weeks

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
seekwk	7385	0	0.0291131	0.0019565
noseekwk	7385	0	0.2526743	0.0050570
other	7385	0	0.7182126	0.0052353

CENSUS Looked for work in the last 4 weeks

The MEANS Procedure

Variable	N	NMiss	Mean
seekwk	3820749	0	0.0556108
noseekwk	3820749	0	0.2380774
other	3820749	0	0.7063118

HES Number of People in Household

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
hhd1	7385	0	0.0869500	0.0032790
hhd2	7385	0	0.2472519	0.0050205
hhd3	7385	0	0.1803553	0.0044744
hhd4	7385	0	0.2540256	0.0050659
hhd5	7385	0	0.1432984	0.0040775
hhd6	7385	0	0.0576256	0.0027119
hhd7	7385	0	0.0156392	0.0014439
hhd8	7385	0	0.0104692	0.0011845
hhd9	7385	0	0	0
hhd10	7385	0	0	0
hhd11	7385	0	0.0043849	0.000768916

Census Number of People in Household

The MEANS Procedure

Variable	N	NMiss	Mean
hhd1	1368205	0	0.2323687
hhd2	1368205	0	0.3320182
hhd3	1368205	0	0.1651938
hhd4	1368205	0	0.1465036
hhd5	1368205	0	0.0727164
hhd6	1368205	0	0.0286412
hhd7	1368205	0	0.0104999
hhd8	1368205	0	0.0047456
hhd9	1368205	0	0.0022592
hhd10	1368205	0	0.0012849
hhd11	1368205	0	0.0037684

HES Marital Status

Living with partner or not living with partner

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
partner	7385	0	0.4760527	0.0058120
nopartner	7385	0	0.5231965	0.0058124
other	7385	0	0.000750815	0.000318755

CENSUS Means for Marital Status

The MEANS Procedure

Variable	N	N Miss	Mean
partner	3820749	0	0.4302690
nopartner	3820749	0	0.2774409
other	3820749	0	0.2922901

HES Age Group Means

0-14, 15-29, 30-44, 45-64, 65+

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
Child	7385	0	0.2302320	0.0048991
gp1	7385	0	0.2039630	0.0046892
gp2	7385	0	0.2331439	0.0049207
gp3	7385	0	0.2204326	0.0048241
gp4	7385	0	0.1122284	0.0036733

Census Age Group Means

0-14, 15-29, 30-44, 45-64, 65+

The MEANS Procedure

Variable	N	NMiss	Mean
child	3820749	0	0.2229490
gp1	3820749	0	0.2032062
gp2	3820749	0	0.2300313
gp3	3820749	0	0.2224260
gp4	3820749	0	0.1213875

HES Looked for work in the last 4 weeks

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
seekwk	7385	0	0.0304069	0.0019982
noseekwk	7385	0	0.2495117	0.0050358
other	7385	0	0.7200815	0.0052247

CENSUS Looked for work in the last 4 weeks

The MEANS Procedure

Variable	N	NMiss	Mean
seekwk	3820749	0	0.0556108
noseekwk	3820749	0	0.2380774
other	3820749	0	0.7063118

HES Currently Working

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
working	7385	0	0.4757677	0.0058118
notwork	7385	0	0.2940002	0.0053019
other	7385	0	0.2302320	0.0048991

CENSUS Currently Working

The MEANS Procedure

Variable	N	NMiss	Mean
working	3820749	0	0.4455179
notwork	3820749	0	0.2628165
other	3820749	0	0.2916656

HES Highest Qualification Means

Bursary Catergory includes UE, Higher School etc

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
none	7385	0	0.1623340	0.0042914
scert	7385	0	0.1224942	0.0038154
bursary	7385	0	0.0961149	0.0034301
vocat	7385	0	0.1432500	0.0040769
degree	7385	0	0.0830632	0.0032116
postgrad	7385	0	0.0235422	0.0017644
otherqual	7385	0	0.0150003	0.0014146
other	7385	0	0.3542011	0.0055658

CENSUS Highest Qualification Means

Bursary Catergory includes sixth form, Higher School etc

The MEANS Procedure

Variable	N	NMiss	Mean
none	3820749	0	0.1796048
scert	3820749	0	0.1018805
bursary	3820749	0	0.1159299
vocat	3820749	0	0.1329297
degree	3820749	0	0.0523280
postgrad	3820749	0	0.0241191
otherqual	3820749	0	0.0428440
other	3820749	0	0.1066388

HES Occupations in first job Means
only to 1 digit level

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
Managers	7385	0	0.0641292	0.0028510
Professionals	7385	0	0.0730798	0.0030288
Technicians	7385	0	0.0548087	0.0026487
Clerks	7385	0	0.0519477	0.0025826
Sales	7385	0	0.0767331	0.0030975
Agriculture	7385	0	0.0480175	0.0024881
Trades	7385	0	0.0434830	0.0023733
Operators	7385	0	0.0359889	0.0021676
Elementary	7385	0	0.0338421	0.0021043
other	7385	0	0.5179699	0.0058149

Census Occupations in main job
only to 1 digit level

The MEANS Procedure

Variable	N	N Miss	Mean
Managers	3820749	0	0.0566373
Professionals	3820749	0	0.0627374
Technicians	3820749	0	0.0499180
Clerks	3820749	0	0.0566648
Sales	3820749	0	0.0634894
Agriculture	3820749	0	0.0359927
Trades	3820749	0	0.0380478
Operators	3820749	0	0.0377091
Elementary	3820749	0	0.3550783
other	3820749	0	0.2437252

HES Means for Age 0-14 15-29 30-44 45-64 65+

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
Child	7385	0	0.2302320	0.0048991
gp1	7385	0	0.2039630	0.0046892
gp2	7385	0	0.2331439	0.0049207
gp3	7385	0	0.2204326	0.0048241
gp4	7385	0	0.1122284	0.0036733

CENSUS Means for Age 0-14 15-29 30-44 45-64 65+

The MEANS Procedure

Variable	N	NMiss	Mean
child	3820749	0	0.2229490
gp1	3820749	0	0.2032062
gp2	3820749	0	0.2300313
gp3	3820749	0	0.2224260
gp4	3820749	0	0.1213875

HES Means for HHD reg recur Income

The MEANS Procedure

Variable	N	N Miss	Mean	Std Error
loss	2808	0	0.0044832	0.0012610
zero	2808	0	0.000813699	0.000538188
less5k	2808	0	0.0039385	0.0011822
less10k	2808	0	0.0157785	0.0023521
less15k	2808	0	0.0762763	0.0050101
less20k	2808	0	0.0826646	0.0051976
less25k	2808	0	0.0960533	0.0055617
less30k	2808	0	0.0820895	0.0051811
less40k	2808	0	0.1236388	0.0062129
less50k	2808	0	0.0999646	0.0056615
less70k	2808	0	0.1705738	0.0070994
less100k	2808	0	0.1377059	0.0065040
more100k	2808	0	0.1060192	0.0058108
other	2808	0	0	0

CENSUS Means for HhldIncome

The MEANS Procedure

Variable	N	N Miss	Mean
loss	1368205	0	0.0026334
zero	1368205	0	0.0038357
less5K	1368205	0	0.0116540
less10K	1368205	0	0.0293925
less15K	1368205	0	0.0782997
less20K	1368205	0	0.0686600
less25K	1368205	0	0.0455538
less30K	1368205	0	0.0770140
less40K	1368205	0	0.0869329
less50K	1368205	0	0.0836198
less70K	1368205	0	0.1306931
less100K	1368205	0	0.0902241
more100K	1368205	0	0.0922691
other	1368205	0	0.1817023

HES total number of rooms in dwelling Means 3-19 total rooms

Variable	N	NMiss	Mean	Std Error
room3	2808	0	0.000872830	0.000557383
room4	2808	0	0.0072538	0.0016017
room5	2808	0	0.0260051	0.0030039
room6	2808	0	0.0486352	0.0040600
room7	2808	0	0.0889471	0.0053730
room8	2808	0	0.1294570	0.0063363
room9	2808	0	0.2832828	0.0085048
room10	2808	0	0.1595375	0.0069115
room11	2808	0	0.0833919	0.0052183
room12	2808	0	0.0648756	0.0046489
room13	2808	0	0.0507298	0.0041420
room14	2808	0	0.0217767	0.0027548
room15	2808	0	0.0179925	0.0025089
room16	2808	0	0.1595375	0.0069115
room17	2808	0	0.0833919	0.0052183
room18	2808	0	0.0648756	0.0046489
room19	2808	0	0.0507298	0.0041420

CENSUS Number of Bedrooms Means 1-19 and 20+ rooms excludes non-private dwellings

Variable	N	NMiss	Mean
room1	1359843	0	0.0050741
room2	1359843	0	0.0144980
room3	1359843	0	0.0464862
room4	1359843	0	0.0998681
room5	1359843	0	0.1803841
room6	1359843	0	0.2597918
room7	1359843	0	0.1576513
room8	1359843	0	0.0956721
room9	1359843	0	0.0483548
room10	1359843	0	0.0230004
room11	1359843	0	0.0095327
room12	1359843	0	0.0050013
room13	1359843	0	0.0025591
room14	1359843	0	0.0019907
room15	1359843	0	0.0014833
room16	1359843	0	0.0010097
room17	1359843	0	0.000508882
room18	1359843	0	0.000311065
room19	1359843	0	0.000255912
room20	1359843	0	0.0465664
other	1359843	0	0.0462774

HES Number of vehicles owned/used per household
0 to 3 or more

The MEANS Procedure

Variable	N	N Miss	Mean	Std Error
car0	2808	0	0.1079058	0.0058561
car1	2808	0	0.3841921	0.0091807
car2	2808	0	0.2931857	0.0085922
car3	2808	0	0.2147164	0.0077504

CENSUS Number of Motor Vehicles per household
0 to 3 or more vehicles

The MEANS Procedure

Variable	N	N Miss	Mean
car0	1368205	0	0.0971923
car1	1368205	0	0.3898802
car2	1368205	0	0.3402940
car3	1368205	0	0.1240779
other	1368205	0	0.0424439

HES Dwelling Type Means

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
house	2808	0	0.8433717	0.0068600
flats2	2808	0	0.0889985	0.0053744
flats3	2808	0	0.0523556	0.0042042
flats3storey3	2808	0	0.0096312	0.0018434
jointshop	2808	0	0.0045196	0.0012660
bach	2808	0	0.0011235	0.000632283

Census Dwelling Type Means
Excludes non-private

The MEANS Procedure

Variable	N	NMiss	Mean
house	1359843	0	0.7574970
flats2	1359843	0	0.0971531
flats3	1359843	0	0.0465429
flats3storey3	1359843	0	0.0111947
jointshop	1359843	0	0.0123463
bach	1359843	0	0.0064787
otherprivate	1359843	0	0.0655311

HES Weekly Rent Grouping Means

less $100, $100-$199, $200-$299, $300-$399, $400-499, $500+

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
less100	763	0	0.2217332	0.0150488
less200	763	0	0.4725106	0.0180857
less300	763	0	0.2449856	0.0155801
less400	763	0	0.0389083	0.0070053
less500	763	0	0.0143884	0.0043140
over500	763	0	0.0074739	0.0031201

CENSUS Weekly Rent Grouping Means

less $100, $100-$199, $200-$299, $300-$399, $400-499, $500+
excludes non-private dwellings and other(no response etc)

The MEANS Procedure

Variable	N	NMiss	Mean
less100	361589	0	0.2293847
less200	361589	0	0.4318494
less300	361589	0	0.2455191
less400	361589	0	0.0632818
less500	361589	0	0.0172903
over500	361589	0	0.0126746

HES Number of Bedrooms

0-7 Bedrooms

The MEANS Procedure

Variable	N	NMiss	Mean	Std Error
bdroom0	2808	0	0.000181341	0.000254149
bdroom1	2808	0	0.0398609	0.0036925
bdroom2	2808	0	0.2176350	0.0077884
bdroom3	2808	0	0.5159787	0.0094325
bdroom4	2808	0	0.1858710	0.0073423
bdroom5	2808	0	0.0334293	0.0033928
bdroom6	2808	0	0.0061419	0.0014747
bdroom7	2808	0	0.000901900	0.000566581

CENSUS Number of Bedrooms

1-14 Bedrooms excludes non-private dwellings

The MEANS Procedure

Variable	N	N Miss	Mean
bdroom1	1359843	0	0.0523413
bdroom2	1359843	0	0.1958329
bdroom3	1359843	0	0.4542524
bdroom4	1359843	0	0.1946938
bdroom5	1359843	0	0.0432322
bdroom6	1359843	0	0.0083017
bdroom7	1359843	0	0.0028400
bdroom8	1359843	0	0.0485056
other	1359843	0	0.0444485

HES Tenure of Dwelling

The MEANS Procedure

Variable	N	N Miss	Mean	Std Error
rent	2808	0	0.2808481	0.0084825
rentfree	2808	0	0.0423024	0.0037991
mortgage	2808	0	0.3138376	0.0087588
nomortgage	2808	0	0.3630119	0.0090762

CENSUS Tenure
Own - no mortgage info, Dont Own - no rent info

The MEANS Procedure

Variable	N	N Miss	Mean	Std Error
own	1368205	0	0.0086478	0.000079157
mortgage	1368205	0	0.3239829	0.000400096
nomortgage	1368205	0	0.3022573	0.000392609
dont_own	1368205	0	0.0107447	0.000088141
rent	1368205	0	0.2623072	0.000376069
rentfree	1368205	0	0.0282173	0.000141568
other	1368205	0	0.0463469	0.000179734

Appendix 4:

Average household expenditure by income group of households for Maori households only.

Maori estimates for Table 1																						
	Under 14,900		14,900 to 20,699		20,700 to 25,899		25,900 to 32,399		32,400 to 40,599		40,600 to 51,099		51,100 to 62,299		62,300 to 76,699		76,700 to 101,099		101,100 or more		All income groups	
Expenditure group and subgroup	estimate	rse	estimate	rse	estimate	rse	estimate	rse	estimate	rse	estimate	rse	estimate	rse	estimate	rse	estimate	rse	estimate	rse	estimate	rse
Food	**48.31**	**57.4**	**75.90**	**49.9**	**73.61**	**58.6**	**90.24**	**45.0**	**104.84**	**50.1**	**123.09**	**37.8**	**114.65**	**51.5**	**143.62**	**47.4**	**152.75**	**54.6**	**182.84**	**67.0**	**108.30**	**13.3**
Cereals, Cereal products	6.53	60.3	8.63	52.7	10.58	63.3	12.80	51.7	13.26	49.1	13.55	43.4	15.74	58.2	14.05	55.8	18.06	56.3	28.94	68.2	13.46	14.8
Farm Products	5.81	67.1	11.34	54.2	11.20	57.5	11.14	44.7	13.95	50.4	12.48	38.1	13.86	52.3	13.49	59.6	14.39	56.7	20.66	67.5	12.43	16.0
Fish	0.63	115.6	0.82	87.4	1.70	108.4	2.12	97.0	1.45	77.3	1.67	61.5	2.86	85.8	2.24	83.9	4.39	81.2	1.92	131.0	1.94	30.1
Fruit	3.14	132.9	2.10	62.2	3.24	66.4	3.33	49.0	5.56	65.8	6.17	63.9	7.00	64.4	6.50	62.4	10.66	55.4	13.21	66.7	5.60	19.3
Meals away from home	9.31	120.4	13.55	78.2	7.62	68.0	19.88	74.0	11.90	55.5	22.82	46.0	18.09	56.3	35.33	55.1	34.70	62.8	46.84	80.2	21.15	20.9
Meat	4.65	76.0	12.00	59.9	15.20	77.0	11.37	54.3	24.09	49.1	23.24	53.3	15.83	68.0	22.03	60.0	17.18	66.2	16.21	69.3	16.65	18.5
Other food stuffs	6.66	94.1	13.71	94.0	7.04	73.2	9.15	62.9	9.66	71.5	16.72	63.0	10.99	53.1	21.59	84.7	12.84	61.3	11.20	70.6	12.40	25.4
Poultry	2.14	103.0	2.09	78.2	2.93	69.6	2.96	66.0	5.34	72.3	4.00	77.1	6.53	68.3	5.04	71.0	6.70	71.9	4.59	86.0	4.15	23.0
Sweet products,spreads,beverag	6.02	69.8	7.32	55.6	7.57	56.4	13.46	48.6	10.68	54.6	14.08	44.4	16.69	56.8	15.90	55.0	23.40	57.2	27.61	76.4	13.60	17.1
Veges	3.42	74.0	4.34	62.3	6.53	51.8	4.02	50.2	8.95	65.4	8.37	57.2	7.05	60.0	7.44	50.2	10.43	62.4	11.66	67.6	6.91	17.1
Housing	**121.96**	**56.8**	**148.51**	**51.5**	**119.86**	**53.6**	**156.93**	**50.7**	**122.86**	**51.1**	**149.13**	**35.3**	**220.05**	**49.6**	**233.22**	**58.1**	**267.13**	**61.1**	**255.38**	**85.3**	**173.25**	**15.7**
Mortgage Payments	28.87	131.6	7.49	122.1	38.13	80.9	40.70	103.1	33.28	79.1	25.78	81.4	64.19	72.8	93.93	76.3	155.94	74.1	151.51	108.8	56.24	26.8
Payments to local authorities	5.81	129.1	2.49	81.2	7.78	60.7	8.88	97.8	8.07	61.1	4.88	66.0	12.15	73.4	10.98	57.8	20.14	66.0	22.25	63.9	9.33	22.7
Property Maintenance Goods	3.08	145.7	0.86	102.4	2.24	99.9	13.79	142.9	8.36	109.1	5.02	82.0	10.00	100.8	6.52	70.3	18.39	88.7	38.37	98.1	9.22	40.3
Property Maintenance Service	1.72	153.0	8.12	182.5	10.08	154.2	5.40	111.9	5.57	112.9	4.03	129.9	9.19	149.0	37.15	189.6	23.52	103.5	23.98	130.9	12.03	77.1
Rent	82.49	58.6	129.55	56.8	61.62	74.5	88.17	58.5	67.58	75.9	109.43	41.1	124.53	69.3	84.64	74.4	49.15	103.4	19.27	199.9	86.43	20.2
Household operations	**53.80**	**56.2**	**57.73**	**54.6**	**56.21**	**48.1**	**65.03**	**70.1**	**69.65**	**45.6**	**82.24**	**37.1**	**75.04**	**56.1**	**104.26**	**55.4**	**123.95**	**53.8**	**144.45**	**69.2**	**79.72**	**14.1**
Domestic Fuel & power	15.44	62.3	22.32	52.5	21.59	50.8	25.29	64.8	21.74	45.3	20.58	37.1	23.73	50.4	26.02	46.0	29.61	55.5	31.34	63.0	23.36	14.0
Floor Coverings	.	.	.	.	.	.	.	.	0.58	202.4	.	.	.	.	.	.	2.89	172.4	0.76	199.9	0.32	131.5
Furnishings	0.55	125.7	0.21	199.6	0.14	203.4	0.72	135.3	1.81	177.8	1.11	127.9	1.01	151.0	1.21	144.8	5.64	135.6	3.92	144.0	1.40	55.6

Furniture	3.66	126.0	6.73	93.0	6.40	88.0	2.55	109.3	4.19	106.7	6.19	105.8	9.40	123.0	7.09	96.1	20.91	100.7	18.52	125.3	7.34	33.7
Home Appliances	14.24	97.1	5.39	129.5	4.02	95.8	4.46	110.4	15.31	85.2	15.65	53.9	11.11	79.7	15.56	116.8	17.95	84.0	26.20	77.9	12.20	30.2
Household equipment & utensils	0.83	188.4	1.63	123.6	0.67	131.9	0.54	90.8	1.47	125.7	2.62	81.9	1.89	109.5	3.11	95.7	2.82	112.2	3.11	94.6	1.81	35.9
Household services	15.22	64.0	15.48	75.5	16.41	52.9	26.77	107.9	16.80	49.7	28.96	58.7	21.50	62.4	38.44	52.8	33.56	55.6	43.68	82.9	25.49	20.9
Household supplies	2.93	64.7	3.70	59.8	6.08	65.2	4.61	65.1	6.86	73.1	5.76	45.9	6.06	60.7	7.85	66.4	7.70	58.3	12.66	67.8	6.07	18.5
household textiles	0.92	177.2	2.26	160.6	0.90	114.7	0.08	150.5	0.88	116.4	1.39	73.5	0.34	136.5	4.98	132.5	2.87	100.8	4.26	131.9	1.73	55.0
Apparel	**3.74**	**132.6**	**17.18**	**88.9**	**2.49**	**103.4**	**9.34**	**102.3**	**7.20**	**138.4**	**21.26**	**63.1**	**6.60**	**95.2**	**11.83**	**71.0**	**23.63**	**87.9**	**46.62**	**83.2**	**13.43**	**32.3**
Footwear not otherwise classi	.	.	.	.	.	.	0.04	200.6	0.25	209.5	1.77	179.7	.	.	0.45	198.8	.	.	2.85	182.8	0.49	113.5
Childrens Clothing	0.90	144.5	5.17	154.3	.	.	0.15	137.1	1.44	174.5	6.08	96.9	1.69	169.2	0.82	112.0	0.76	144.3	7.44	164.3	2.26	57.6
Childrens Footwear	.	.	2.58	141.6	.	.	0.28	162.4	.	.	2.24	162.4	0.16	197.8	0.08	199.7	0.37	157.3	.	.	0.64	95.3
Clothing not otherwise classif	0.77	127.7	1.90	125.6	0.44	225.1	0.07	199.6	1.04	138.2	2.55	105.0	0.31	207.2	3.10	198.4	2.27	176.8	8.01	135.9	1.74	63.6
Clothing supplies and services	.	.	0.38	200.1	.	.	0.47	147.1	.	.	0.85	199.6	0.63	175.0	.	.	0.29	142.2	7.87	157.5	0.69	95.1
Footwear supplies & services	.	.	.	.	.	.	.	.	0.05	201.7	0.15	198.3	.	.	1.36	173.7	.	.	0.05	201.1	0.20	148.3
Mens Clothing	0.22	164.7	3.32	117.1	0.74	155.9	0.52	156.7	0.51	140.1	1.74	107.6	0.44	206.9	2.23	127.1	3.72	125.1	2.27	148.8	1.46	51.7
Mens Footwear	.	.	1.49	147.1	.	.	0.38	160.0	0.36	128.4	1.54	129.9	0.34	206.9	0.11	199.1	0.94	169.7	4.30	125.3	0.77	59.3
Womens Clothing	1.84	158.2	2.34	197.9	1.31	96.3	6.27	131.4	3.55	193.9	3.37	123.8	1.66	138.3	3.62	104.6	13.74	93.7	13.65	152.5	4.60	48.8
Womens Footwear	.	.	.	.	.	.	1.17	124.4	.	.	0.97	198.2	1.38	207.2	0.06	202.8	1.53	145.3	0.18	199.3	0.57	79.2
Transportation	**10.87**	**78.8**	**43.34**	**62.5**	**54.90**	**71.1**	**59.37**	**65.8**	**61.53**	**60.3**	**116.62**	**48.7**	**75.65**	**58.4**	**135.61**	**63.8**	**227.01**	**63.7**	**216.23**	**70.1**	**93.02**	**21.7**
Overseas travel	0.86	209.9	.	.	.	.	1.10	164.2	1.02	209.5	8.24	99.5	19.02	118.7	12.77	105.3	12.43	113.8	15.87	145.6	6.52	46.7
Private transport costs N.E.C	0.42	91.4	3.99	133.8	3.17	154.0	3.30	105.0	1.16	91.2	5.36	80.4	3.20	71.6	2.51	79.0	7.15	94.6	3.97	115.7	3.38	33.0
Public transport within NZ	0.82	168.2	2.59	105.7	1.86	173.9	1.78	90.8	2.96	99.2	4.99	81.8	7.05	140.8	4.52	121.9	10.44	80.6	13.33	87.2	4.38	35.3
Road Vehicles	.	.	6.74	127.9	20.85	137.1	6.02	123.4	17.61	95.0	52.73	76.7	7.02	178.6	56.22	111.2	119.50	97.7	103.12	108.1	34.38	49.0
Vehicle ownership expenses	8.77	86.1	30.03	64.7	29.03	62.3	47.17	70.8	38.77	61.6	45.30	39.6	39.36	60.9	59.59	68.0	77.49	68.9	79.95	83.9	44.37	19.7
Other Goods Group	**35.71**	**74.7**	**37.19**	**51.9**	**41.88**	**70.1**	**55.59**	**56.8**	**55.65**	**55.4**	**78.05**	**43.1**	**56.24**	**51.3**	**120.38**	**55.9**	**120.78**	**62.8**	**118.50**	**67.4**	**70.46**	**17.9**
Alcohol	6.57	110.2	9.88	84.4	4.21	112.4	8.40	92.0	6.86	87.0	26.67	57.8	2.43	96.0	28.91	80.4	17.54	90.9	22.61	96.4	13.93	33.8
Goods N.E.C	1.30	159.0	0.45	189.2	1.94	167.0	0.27	141.7	0.60	121.5	0.49	150.2	2.41	143.3	6.23	147.5	2.76	91.3	1.79	120.6	1.77	69.3
Leisure & Recreational Groups	0.76	122.1	2.48	115.6	2.76	136.3	2.71	63.6	7.04	96.4	10.95	76.6	4.99	78.9	14.39	68.2	22.74	90.2	12.50	98.5	7.82	31.2
Medical Goods	9.35	171.8	1.79	101.8	4.39	88.3	1.42	71.0	2.10	78.0	1.23	78.3	6.22	93.2	13.39	77.9	3.14	87.8	10.85	89.8	4.93	45.8
Personal Goods	0.99	124.5	4.82	94.2	4.19	91.4	6.06	132.4	6.94	76.2	5.83	69.7	6.24	81.0	13.41	83.0	11.81	84.3	6.73	72.8	6.81	32.0
Pets Racehorses & Livestock	0.54	105.6	1.72	99.5	2.81	87.3	2.96	67.4	2.90	88.2	3.97	71.4	5.94	108.6	2.45	82.5	7.21	71.5	6.22	78.5	3.45	25.5
Publications,Stationary & offi	6.09	134.8	5.10	103.3	10.90	117.6	16.89	116.1	6.33	112.7	5.87	77.2	5.71	107.1	11.39	121.0	23.41	72.0	26.50	86.2	10.97	34.9
Recreational Vehicles	.	.	.	.	.	.	.	.	.	.	0.55	142.6	0.40	200.1	0.81	189.1	0.20	176.1	-5.82	.	-0.05	.
Tobacco products	8.21	90.6	9.19	67.4	6.71	64.9	13.58	78.7	19.59	70.5	18.44	50.9	16.91	74.4	23.66	84.1	17.64	85.1	32.46	107.6	16.30	25.9

Toiletries	1.90	75.4	1.77	67.8	3.98	80.2	3.29	75.9	3.29	93.4	4.05	72.3	5.01	70.0	5.75	67.0	14.32	73.0	4.65	73.3	4.55	26.2
Other Services Group	**31.81**	**90.4**	**51.89**	**94.7**	**47.97**	**67.3**	**49.91**	**58.7**	**57.26**	**55.3**	**78.78**	**49.0**	**78.51**	**55.5**	**136.08**	**46.8**	**183.64**	**60.5**	**253.12**	**69.7**	**86.41**	**18.4**
Accomodation Services	2.01	151.0	0.14	130.5	0.08	203.4	0.48	147.1	.	.	2.16	141.0	1.56	144.9	4.44	89.3	8.35	165.2	2.73	127.6	2.03	60.7
Contributions to Savings	2.78	122.6	3.70	106.5	4.68	136.2	2.21	91.6	5.53	83.8	13.09	65.2	11.06	72.2	25.30	55.3	45.27	69.1	70.11	74.5	14.70	28.8
Educational & Tuitional servic	0.49	173.3	18.73	197.0	7.32	123.2	8.16	117.5	14.57	127.4	14.99	139.4	2.68	104.1	7.79	126.5	18.70	104.3	11.94	128.7	10.46	64.1
Financial, Insurance & Legal S	8.86	79.0	7.20	55.8	17.28	67.3	14.55	62.8	12.39	49.3	16.04	55.2	18.71	53.8	30.17	49.1	30.71	67.0	55.56	69.3	19.13	16.3
Health Services	3.17	107.8	6.91	100.1	4.45	100.9	7.43	104.1	3.66	98.6	3.91	91.7	13.35	74.2	14.64	68.3	22.05	86.5	12.78	79.5	8.60	27.8
Leisure Services	1.83	120.6	7.96	102.6	2.86	126.7	5.04	93.3	9.94	68.8	13.76	73.4	15.95	87.5	17.21	55.1	18.39	73.6	47.11	123.2	12.15	32.2
Outgoings N.E.C	10.86	142.8	5.23	131.3	8.23	92.6	5.80	59.6	7.74	67.4	10.90	110.3	12.57	75.6	23.54	69.3	22.60	81.1	46.05	105.0	13.36	33.3
Personal Services	1.20	203.0	0.30	150.5	0.38	169.1	0.71	140.5	0.55	112.2	1.61	127.1	0.38	177.5	0.87	133.5	8.75	99.4	3.61	157.2	1.54	60.2
Services N.E.C	0.62	203.0	1.60	200.1	2.57	140.3	4.31	198.9	2.00	104.3	1.28	89.2	0.36	198.4	9.04	113.7	5.49	135.2	0.97	141.7	3.08	60.9
Vocational Services	.	.	0.13	165.3	0.11	233.9	1.23	157.8	0.89	111.7	1.05	66.5	1.90	82.7	3.09	60.1	3.34	76.1	2.26	107.2	1.37	33.9

From: HES 2006\Methodology\Sample size\maoriestimates.xls

Appendix 5

HES 2000/01 Datasets (Location & Variables)

This document provides details of:
- the location of the 'new' HES datasets, the old HES datasets and the asset datasets
- the main differences between the 'old' HES datasets and the 'new' HES datasets'
- a list of the variables on the household (h), person level (pix) and household expenditure (he) datasets.

Note: Integrated weighting has been used to calculate the weights.

Location of SAS datasets

All output HES 2000/01 datasets are stored on the following directory:
J:\RO\D_lmhe\hes datasets.

In addition to the HES 2000/01 datasets there are asset datasets. The asset datasets are located on the following directory:
J:\RO\D_lmhe\asset

Note:
i) The HES 1997/98 datasets that have been reweighted using integrated weighting are also stored on this directory. These are the household datasets (eg H0001h and H9798H), the person level datasets (eg H0001pix and H9798pix) and the household expenditure datasets (eg H0001he and H9798he).

ii) The 'old' HES datasets are located 'S:\RO\D_Imhe\old_hes'. These are the household datasets (eg c9798H), the personal datasets (eg c9798pix) and the household expenditure datasets (eg c9798he). These datasets should no longer be used. (note, the variable weight on these datasets is the response factor weight).

iii) Additional HES datasets created during the different stages of the weighting estimation can be found in the directories mentioned in the following document

(Database: "Social Statistics Project"; Subject: "Where do the programs and data live?"; Author: Aroha Harris; Last Editor: Tracey Kale; Date Created: 06/09/1999; Date Modified: 08/05/2002).

Overview of the list of available SAS datasets

There are three main SAS datasets (two of these SAS datasets have sub datasets):

1. **h0001h**: contains ***household characteristics*** information for each of the ***2,808*** households surveyed. There is one record per household.

2. **h0001he**: contains ***household expenditure*** information, including expenditure on each IRN (or item eg irn0001 = apples). There is one record per IRN, which results in many records per household.

   Dataset **h0001hea** is a subset of this dataset (it has the same number of records but only has a select few of the variables). There is one record per IRN, which results in many records per household.

3. **h0001pix**: contains ***personal*** characteristics information for each of the ***7,385*** persons in the survey as well as personal income information for each person aged 15 or over. The weight variable attached to each personal record is the weight assigned to their household, ie each person in a household will have the same weight. There is one record per person.

   Dataset **h0001p** is a subset of this dataset, it has the same number of records but doesn't have the IRN income amount variables (in other words it contains the personal characteristics information). However it does have an additional variable, the PSU variable. There is one record per person.

   Dataset **h0001pi**: is similar to the **h0001he** dataset, but is person level income information rather than household level expenditure information. For each person there is one record per IRN, which results in many records per person.

   Each of these datasets can be merged using the common variable 'address'.

Details of the available SAS datasets

h0001h:
Contains household characteristics information for each of the ***2,808*** households surveyed. There is one record per household. This dataset has the following variables:

SAS VARIABLE	TYPE	DESCRIPTION	Codes
address	char	household address reference number	
adnotwk	char	number of adults not working	
ftwork	char	number of adults working full-time	
hhtype	**char**	**Household composition (household composition by number of dependent children)**	11 Couple only 21 Couple with one dependent child 22 Couple with two dependent children 23 Couple with three or more dependent children 24 All other 'couples with child(ren) only' households* 31 One parent with dependent child(ren) only 32 All other 'One parent with child(ren) only' households** 41 Other one-family households 51 One-person household 61 All other households 99 *Household Unidentifiable*
~~**oldhhtype**~~	~~**char**~~	~~**the old household type variable**~~	Note: This is actually not on the dataset, contrary to information stored on the Labour Market & HED Workgroup database.
intrvwr	char	interviewer code	
itemct	num	item count (not needed). This is the number of IRN items in a household	
month	char	survey month (reference to when the interview took place): 01=Jul, 02=Aug, ,....,12=Jun	
numpers	char	number of people in household	
persno	char	person number	
~~prticptn~~	~~char~~	~~household participation code (not needed)~~	Note: This is actually not on the dataset, contrary to information stored on the Labour Market & HED Workgroup database.
psu	char	primary sampling unit	
ptwork	char	number of adults working part-time	
~~rectyp~~	~~char~~	~~record type (not needed)~~	Note: This is actually not on the dataset, contrary to information stored on the Labour Market & HED Workgroup database.
tothhexp	**num**	**total household expenditure (excluding net capital outlay)**	
totexpnco	**num**	**total household expenditure**	

		(including net capital outlay)
totirinc	num	total irregular household income
totrrinc	num	total regular and recurring household income
weight	**num**	**household integrated weight**
respwgt	**num**	**household response factor weight (sample selection weight that has been adjusted by applying the response factors to adjust for non-response)**
year	char	survey year (reference to when the interview took place): 00 (2000) or 01 (2001)

h0001he:

Contains household expenditure information, including expenditure on each item reference number (IRN) (or item eg irn0001 = apples). For each household there is one record per IRN, which results in many records per household. This dataset has the following variables:

SAS VARIABLE	TYPE	DESCRIPTION
address	char	household address reference number
amt	num	IRN amount ($)
irn	num	item reference number (IRN)
irnsourc	num	IRN source: 01(Expenditure Questionnaire), 03(Expenditure Diary)
month	char	survey month (reference to when the interview took place): 01=Jul, 02=Aug, ,....,12=Jun
psu	char	primary sampling unit
purchmth	num	month of purchase
~~rectyp~~	~~char~~	~~record type (not needed)~~ Note: This is actually not on the dataset, contrary to information stored on the Labour Market & HED Workgroup database.
storetyp	num	storetype
year	char	survey year (reference to when the interview took place): 00 (2000) or 01 (2001)

Note: **h0001hea**: is a cut down version of the **h0001he** dataset, it has the same number of records but only a select few of the variables (note the different SAS name for the variable that represents the IRN amount). For each household there is one record per IRN, which results in many records per household. This dataset has the following variables:

SAS VARIABLE	TYPE	DESCRIPTION
address	char	household address reference number
amount	num	IRN amount ($)
irn	num	item reference number (IRN)
storetyp	num	storetype

h0001pix:
Contains personal characteristics information for each person in the survey as well as personal income information for each person aged 15 or over. The weight variable attached to each personal record is the weight assigned to their household ie each person in a household will have the same weight. There is one record per person. This dataset has the following variables:

SAS VARIABLE	TYPE	DESCRIPTION	Codes
address	char	address	household address reference number
age	num	age	
ethnic	char	ethnic group	11 NZ European/Pakeha 12 Other European 21 NZ Maori 31 Samoan 32 Cook Island 33 Tongan 34 Niuean 35 Tokelauan 36 Fijian 37 Other Pacific Island Groups 41 Southeast Asian 42 Chinese 43 Indian 44 Other Asian 50 Other 99 Non-response
famrel	num	relationship to family in household	1 Partner or parent in 1st family 2 Son or daughter in 1st family 3 Partner or parent in 2nd family 4 Son or daughter in 2nd family 5 Partner or parent in 3rd family 6 Son or daughter in 3rd family 7 Member of 4th or other family 8 Non-family person (others in household) 9 One-person household
ftpteduc	num	enrolment in full or part-time education	0 none 1 full-time 2 Part-time 9 Non-response
heduqual	num	highest educational qualification	0 *aged under 15 or over 64* 1 *No qualifications*

			2 School Certificate 3 UE, Bursary, Scholarship 4 Vocational or trade certificate 5 Bachelor degree or diploma 6 Post-graduate qualification 7 Part degree or other qualification 9 Non-response
intvrwr	char	interviewer code	
irn9000 to	num	income amount ($) for irn9000	
irn9721	num	income amount ($) for irn9721	
itemct	num	item count (not needed). This is the number of IRN items in a household	
marstat	num	marital status	1 Living with legal spouse 1 Living with partner 2 Not living with partner, separated 2 Not living with partner, divorced 2 Not living with partner, widowed 2 Not living with partner, never married 2 Not separated but in separate households 9 Non-response
month	char	survey month (reference to when the interview took place): 01=Jul, 02=Aug, ,....,12=Jun	
persno	char	person number	
~~prticptn~~	~~char~~	~~household participation code~~	Note: This is actually not on the dataset, contrary to information stored on the Labour Market & HED Workgroup database.
~~psu~~	~~char~~	~~primary sampling unit~~	Note: This is actually not on the dataset, contrary to information stored on the Labour Market & HED Workgroup database. However, psu is on the **h0001p** dataset. It can also be sourced from the **h0001h** and **h0001he** datasets.
~~rectyp~~	~~char~~	~~record type (not needed)~~	Note: This is actually not on the dataset, contrary to information stored on the Labour Market & HED Workgroup database.
rel2ref	num	relationship to reference person	01 Reference person (RP) 02 Spouse/partner of RP

			03 Child (natural,step,adopted,foster) of RP 04 Parent (natural,step.adopted.foster)of RP 05 Other relative of RP 06 Non-relative of RP
school	num	enrolment in any sort of education	00 *none* 01 Playcentre 02 Kindergarden or Kohanga re 03 State or integrated primary or intermediate 04 State or integrated secondary 05 Private primary or intermediate 06 Private secondary 07 Community college 08 Correspondence school 09 Technical Correspondence Institute (Open Polytechnic) 10 Polytechnic 11 University 12 Occupational training college 13 Private tertiary institution 98 Other (please describe) 99 *non-response*
sex	num	sex	1 Male 2 Female
weight	**num**	**integrated weight**	
respwgt	**num**	**response factor weight (sample selection weight that has been adjusted by applying the response factors to adjust for non-response)**	
year	char	survey year (reference to when the interview took place)	00 (2000); or 01 (2001)

Note: **h0001p**: is a cut down version of the **h0001pix** dataset, it has the same number of records but doesn't have the IRN income amount variables. However it does have an additional variable, the PSU variable. There is one record per person. This dataset has the following variables:

SAS VARIABLE	TYPE	DESCRIPTION	Codes
address	char	address	household address reference number
age	num	age	
ethnic	char	ethnic group	11 NZ European/Pakeha 12 Other European 21 NZ Maori 31 Samoan 32 Cook Island 33 Tongan 34 Niuean 35 Tokelauan 36 Fijian 37 Other Pacific Island Groups 41 Southeast Asian 42 Chinese 43 Indian 44 Other Asian 50 Other 99 Non-response
famrel	num	relationship to family in household	1 Partner or parent in 1st family 2 Son or daughter in 1st family 3 Partner or parent in 2nd family 4 Son or daughter in 2nd family 5 Partner or parent in 3rd family 6 Son or daughter in 3rd family 7 Member of 4th or other family 8 Non-family person (others in household) 9 One-person household
ftpteduc	num	enrolment in full or part-time education	0 none 1 full-time 2 Part-time 9 Non-response
heduqual	num	highest educational qualification	0 *aged under 15 or over 64* 1 *No qualifications*

			2 School Certificate 3 UE, Bursary, Scholarship 4 Vocational or trade certificate 5 Bachelor degree or diploma 6 Post-graduate qualification 7 Part degree or other qualification 9 Non-response
intvrwr	char	interviewer code	
itemct	num	item count (not needed). This is the number of IRN items in a household	
marstat	num	marital status	1 Living with legal spouse 1 Living with partner 2 Not living with partner, separated 2 Not living with partner, divorced 2 Not living with partner, widowed 2 Not living with partner, never married 2 Not separated but in separate households 9 Non-response
month	char	survey month (reference to when the interview took place): 01=Jul, 02=Aug, ,....,12=Jun	
persno	char	person number	
psu	char	primary sampling unit	
rel2ref	num	relationship to reference person	01 Reference person (RP) 02 Spouse/partner of RP 03 Child (natural,step,adopted,foster) of RP 04 Parent (natural,step.adopted.foster)of RP 05 Other relative of RP 06 Non-relative of RP
school	num	enrolment in any sort of education	00 *none* 01 Playcentre 02 Kindergarden or Kohanga re 03 State or integrated primary or intermediate

			04 State or integrated secondary 05 Private primary or intermediate 06 Private secondary 07 Community college 08 Correspondence school 09 Technical Correspondence Institute (Open Polytechnic) 10 Polytechnic 11 University 12 Occupational training college 13 Private tertiary institution 98 Other (please describe) 99 *non-response*
sex	num	sex	1 Male 2 Female
weight	**num**	**integrated weight**	
respwgt	**num**	**response factor weight (sample selection weight that has been adjusted by applying the response factors to adjust for non-response)**	
year	char	survey year (reference to when the interview took place)	00 (2000); or 01 (2001)

Note: **h0001pi**: is similar to the **h0001he** dataset, but is person level income information rather than household level expenditure information. For each person there is one record per IRN, which results in many records per person. The dataset has the following variables:

SAS VARIABLE	TYPE	DESCRIPTION
address	char	household address reference number
amt	num	IRN amount ($)
irn	num	item reference number (IRN)
irnsourc	num	IRN source: 02(Income Questionnaire)
month	char	survey month (reference to when the interview took place): 01=Jul, 02=Aug, ,....,12=Jun
persno	char	person number
psu	char	primary sampling unit
purchmth	num	month of purchase
storetyp	num	storetype
year	char	survey year (reference to when the interview took place): 00 (2000) or 01 (2001)

Asset datasets

For documentation relating to the structure of the HES, ASSET and ADDKEY datasets
(Database: "Lab Mkt and HED Workgroup"; Subject: "Structure of the HES, ASSET and ADDKEY data sets"; Author: Caroline Brooking; Date Created: 27/06/2000; Date Modified: 07/06/2002).

Item Reference Number (IRN) details

In the 2000/01 HES there are three types of item reference numbers (IRNs):

HES expenditure IRN - this classification is used in the 2000/01 HES. It is used for coding expenditure from both the Expenditure Questionnaire and the Expenditure Diary.

It lists all the IRNs that can be used for the coding expenditure, sales and trade-ins. The following list shows the highest level categories of the classification:

0 Food
1 Housing
2 Household operation
3 Apparel
4 Transportation
5 Other goods
6 Other services
7 Refunds, sales and trade-ins

HES non-expenditure IRN - this classification is used in the 2000/01 HES in the Expenditure and Income questionnaires. The non-expenditure IRNs (8000 series) are used to code any data that does not relate directly to spending (eg lender, number of loans, landlord, etc, rather than amount borrowed, amount paid, etc.). In addition these IRNs are non-income IRNs. Therefore they do not relate directly to income, but as for expenditure, are used to code related information (eg benefit vs amount received from benefit).

HES income IRN - this classification is used for coding any income data and some income related information from the 2000/01 HES Income questionnaire. This classification includes all IRNs in the 9000 series.

Differences between the 'old' and 'new' HES datasets

The main differences (or changes) from the 'old' HES datasets are:

i. The weight variable
what was the variable "weight" is now called "respwgt" and the new integrated weight is called "weight"
this is on both the household and personal incomes datasets

ii. Household composition
An additional variable has been added for household composition (to match the classification used in 2000/01). This variable is called hhtype. The old household type variable is now called (oldhhtye)
this variable is on the household dataset

iii. Total household expenditure
an additional variable has been added to the household dataset for total household expenditure which **excludes net capital outlay (this is called TOTHHEXP)**
total household expenditure including net capital outlay is called **(TOTEXPNCO)**

Key HES 2000/01 documents - coding manual, questionnaires, interviewers manual etc

(Database: "Lab Mkt and HED Workgroup"; Subject: "key 2001 documents - coding manual, questionnaires, interviewers manual etc"; Author: Jacqui Gebbie; Date Created: 22/08/2001; Date Modified: 15/05/2002).

Standard tables and detailed expenditure analysis publications

(Database: "Lab Mkt and HED Workgroup"; Subject: "Standard Tables and Detailed Expenditure Analysis Publications"; Author: Jacqui Gebbie; Date Created: 16/10/2001; Date Modified: 08/02/2002).

Appendix 6:

Specifications for Integrated Weighting in the HES

Introduction

The purpose of integrated weighting is to assign a final weight to all people and households in the sample. This weight (WEIGHT) has the following properties:

- All individuals within a household receive the same weight.
- The sum of WEIGHT for people in any *adult* (15+) sex*age group is equal to the population estimate of non-institutionalised, usually resident NZ population in that sex*age group, living in permanent private dwellings.
- The sum of WEIGHT across any *adult* (15+) sex*age group living in a particular Regional Council Authority is equal to the population estimate of non-institutionalised, usually resident NZ population in that sex*age group, living in permanent private dwellings in that Regional Council Authority.
- The sum of WEIGHT across all Maori aged 15-30 is equal to the population estimate of non-institutionalised, usually resident Maori population aged 15-30 who live in permanent private dwellings.
- The sum of WEIGHT across all Maori aged 30+ is equal to the population estimate of non-institutionalised, usually resident Maori population aged 30+ who live in permanent private dwellings.
- The sum of WEIGHT across all households containing exactly 2 adults is equal to the estimate of New Zealand resident households living in permanent private dwellings and containing exactly 2 adults.
- The sum of WEIGHT across all households that do not contain exactly 2 adults is equal to the estimate of New Zealand resident households living in permanent private dwellings and that does not contain exactly 2 adults.

I have taken the steps used to implement integrated weighting in the HLFS and modified them for the HES. Changes I have made are highlighted by **blue, bold text**. Note: Non-response adjustment should be performed before integrated weighting takes place.

Steps to be taken

1. Obtain benchmarks

Create a dataset Censpopn which contains the Pop&Demog population estimates for the benchmarks we are calibrating to.

Censpopn should have 1 variable (CENSEST) and **45** observations (2 sex***13** adult age groups **+ 15 region rows + 1 Maori 30- row + 1 Maori 15-30 row + 1 two adult household row + 1 other household row).** The rows should contain these benchmarks in this order.

The age*sex groups should be sorted by age*sex. They should only include age*sex groups where AGE1>3 (**(Note: the definition for AGE1 is the same as the definition used in the HLFS except that categories 16 (75-79) and 17 (80+) have been collapsed into one category).**

Regional council authority groups should be sorted by regional council authority number. Regional Council Authority 18 should not be included (hence there will only be 15 region rows).

Survey Methods will provide this dataset.

2. Create macro variables AGECAT, RESET20, COLS, CELLS

- AGECAT = number of AGE categories **= 13**
  Under adult-integrated weighting, there are **13** adult categories of derived variable AGE1: taking the values 4 (15-19yrs) up to **16 (75+)**

- RESET20 = **0**
  If this variable is set to zero negative weights will remain negative. If this variable is set to one negative weights will be set as zero.

- **REGCAT = the number of regional council authorities less one**
  %let REGCAT=16-1;

- COLS = number of SEX/AGE indicators **+ number of region indicators + 2 Maori indicators**

```
%let COLS  = %eval( 2 * &agecat + &regcat + 2);
```

- **T_COLS = indicates the column of the 'two adult household' indicator.**
  %let T_COLS=%eval(&COLS + 1)

- **AUG_COLS = indicates the column of the 'other household' indicator. This is also the last column.**
  %let AUG_COLS=%eval(&COLS + 2);

- **CELLS = number of cells in the matrix containing all of the benchmarks we are calibrating to.**

```
%let CELLS = %eval( &AUGCOLS * &AUGCOLS );
```

3. Create hes_tbmp

Create a dataset for the HES with the same attributes as tblimpps (but with AGE1 adjusted for the collapsed age groups), call this dataset

hes_tbmp. This dataset should have

- **the variables refnum, person, child, respwgt, eth2, age, age1, region (region council authority).**
- **one observation per responding person selected into the HES.**
- non-response adjustment already applied to the selection weight (adjusted weight is respwgt).

Below is some code which I used to create an example dataset from the datasets that I had available. I anticipate that this will be changed when the more appropriate datasets are known.

Note that some of the variable names used are not those usually used in the HES. I have renamed them to the HLFS as this will minimise the changes that need to be made in the TAS code.

```
To do this we need two datasets
   person - a row for each person.  columns identifying person, household identifier, sex, age, ehtnicity
   household - rows for each houshold. columns=household identifier and a household weight, respwgt, which has been adjusted for
   non-response.

dm 'clear log';
dm 'clear output';

libname hes 's:\ro\d_lmhe\hes';

proc format;
  value eth
  1='European'
  2='NZ Maori'
  3='Pacific Islands'
  4,9='Other';

  value agef
  0-4=1
  5-9=2
  10-14=3
  15-19=4
  20-24=5
  25-29=6
  30-34=7
  35-39=8
  40-44=9
  45-49=10
  50-54=11
  55-59=12
  60-64=13
  65-69=14
  70-74=15
  75-high=16;
run;

data pall(
  keep=address persno weight age age1 sex eth2 child
```

```
  rename=(address=refnum persno=person weight=respwgt eth2=ethnic)
  );

  merge hes.c9798p(in=ina keep=address age sex ethnic persno)
        hes.c9798h(in=inb keep=address weight);
  by address;
  if ina and inb;

  a=input(substr(ethnic,1,1),1.);
  b=input(substr(ethnic,3,1),1.);
  c=input(substr(ethnic,5,1),1.);

  select ;
    when (a=2|b=2|c=2) eth2=2;   * Maori ;
    when (a=3|b=3|c=3) eth2=3;   * Pacific Islands ;
    when (a=4|b=4|c=4) eth2=4;   * Asian ;
    when (a=5|b=5|c=5) eth2=4;   * Other ;
    when (a=1|b=1|c=1) eth2=1;   * European ;
    when (a=9)         eth2=9;   * non-response ;
  end;

  age1=put(age,agef.);
  if age1 in ('1','2','3') then child=1;
  else child=0;
run;
```

4. Create some extra indicator variables

In hes_tmp create a variable r_count which has the following values for Regional Council.

Regional Council	r_count
01	1
02	2
03	3
04	4
05	5
06	6

07	7
08	8
09	9
12	10
13	11
14	12
15	13
16	14
17	15
18	16

In hes_tbmp create the variables ymao (young Maori) and omao (Old Maori) where:

```
ymao=1 if eth2=2 and 7<AGE1>3, ymao=0 otherwise.
omao=1 if eth2=2 and AGE1>=7, omao=0 otherwise.
```

5. Construct indicators for Age/Sex, Region, Maori 15-30, Maori 30+.

This should be an array with one column for each adult SEX/AGE category, one column for each region, and one column for each of the Maori categories (i.e. *&COLS* columns in total). Each row should have a 1 in only one of the Age/Sex columns (indicating their Age and Sex), a 1 in only one of the region columns, a one in the appropriate Maori column if they are Maori, and 0s in the rest of the columns.

Only individuals in hes_tbmp with CHILD=0 and RESPWGT>0 should be included.

Using the hes_tbmp dataset, the following code indicates steps to create this array. The array is added to the hes_tbmp dataset, and stored as *xmatrix*.

```
data XMATRIX (drop = AGE1 CHILD I COLNUM);
  set hes_tbmp (keep = REFNUM PERSON AGE AGE1 SEX r_count ymao omao
                       CHILD LGR LFSTATUS ESTATUS RESPWGT);
   array COL(&COLS);
     length age2 3;
     age2 = (age1 - 3);
```

```
    do I = 1 to &COLS.;
       col(I) = 0;
    end;
       /* sex indicator */
    select(SEX);
      when( 1 ) colnum = AGE2;
      when( 2 ) colnum = ( &AGECAT + AGE2 );
      otherwise;
      end;
    COL(colnum) = 1;
    /* regional indicator */
    if r_count not in (16) then do;
      colnum2=r_count+&agecat+&agecat;
        col(colnum2)=1;
    end;
       /* maori indicator */
       col(&COLS-1)=omao;
       /* young maori indicator */
    col(&COLS)=ymao;
       output;
  run;
```

6. Count the number of people in each household in the age*sex, region, and the two Maori columns

Using the dataset *xmatrix*, count the number of individuals in each household in the age*sex columns , the region column, and the two Maori columns.

```
proc summary nway data = XMATRIX;
  class REFNUM;
  var COL1-COL&COLS;
  output out = Ujs (drop = _TYPE_
                    rename=_FREQ_=noinhhld)
         sum = U1-U&COLS;
run;
```

The output of this step is a dataset, ***Ujs***, which has one observation for every household in ***xmatrix.*** Ujs has the following structure:

REFNUM	NOINHHLD	U1	U2	U3	------	U&cols-3	U&cols-2	U&cols-1	U&cols

						(3)	(2)	(1)	(0)
R20100XXXXX	2	1	0	1	------	0	0	1	0
R20100YYYYY	3	0	1	2	------	3	1	1	0

(0) Column which will be used to calibrate to the number of Maori aged 15-30.
(1) Column which will be used to calibrate to the number of Maori aged 30+.
(2) Column which will be used to calibrate to the number of adults in r_count 15 (region 17).

7. Multiply Inclusion weights by household averages

Merge the Ujs back onto ***xmatrix*** by REFNUM. The resulting dataset, ***pzmatrix***, should have one observation per person in ***xmatrix***.

Create two variables (two and other) which indicate whether or not the persons household contains exactly two adults.

Create &**AUGCOLS** columns, COL1 to COL&**AUGCOLS**, where for each individual,

$$COL_i = \frac{U_i}{\text{NoInHhld}_i}$$

Create &**AUGCOLS** columns, PCOL1 to PCOL&**AUGCOLS**, where for each individual,

$$PCOL_i = COL_i * \text{RESPWGT}$$

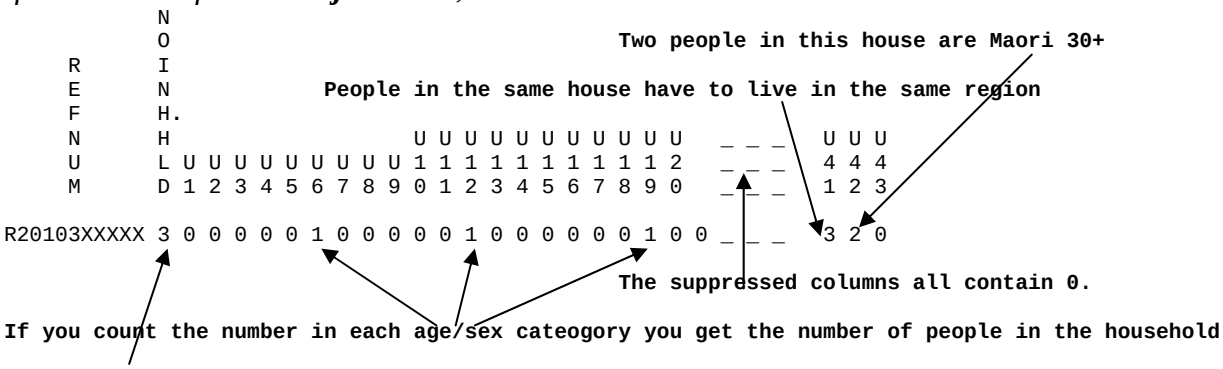
```
data PZMATRIX (drop = U1-U&cols I);
  merge Ujs XMATRIX;
  by REFNUM;
  array U(&cols.);
  array COL(&AUGCOLS.);
  array PCOL(&AUGCOLS.);
  length NOINHHLD 3;

  /* create household type indicators */
```

```
  if noinhhld=2 then do;
       two=1;
      other=0;
  end;
     else do;
          other=1;
          two=0;
  end;
  do I = 1 to &COLS.;
       COL(I)  = U(I) / NOINHHLD;
       PCOL(I) = COL(I) * &wgt.;
  end;

       COL(&T_COLS.) = two/NOINHHLD;
       PCOL(&TCOLS.)= COL(&T_COLS.) * &wgt.;
       COL(&AUGCOLS.) = other/NOINHHLD;
       PCOL(&AUGCOLS.)= COL(&AUGCOLS.) * &wgt.;
 run;
```

Example:
If this is a line from the ***Ujs*** *dataset,*



For the three people who live in the household, COL1 to COL45 would show (Note that COL21-COL40 all contain zero):

```
          N
          O    R
```

```
          R     I  P  E
          E     N  E  S                         C C  C  C  C C C C C  C   C  _ _ _  C  C   C C  C
          F     H  R  P  C C C C C  C  C C C  O O  O  O  O O O O O  O   O  _ _ _  O  O   O O  O
          N     H  S  W  O O O O O  O  O O O  L L  L  L  L L L L L  L   L  _ _ _  L  L   L L  L
          U     L  O  G  L L L L L  L  L L L  1 1  1  1  1 1 1 1 1  1   2  _ _ _  4  4   4 4  4
          M     D  N  T  1 2 3 4 5  6  7 8 9  0 1  2  3  4 5 6 7 8  9   0  _ _ _  1  2   3 4  5

R20103xxxxx 3 11 66  0 0 0 0 0 0.33 0 0 0 0 0 0.33 0  0 0 0 0 0 0.33 0  _ _ _  1 0.66 0 0  1
R20103xxxxx 3 12 66  0 0 0 0 0 0.33 0 0 0 0 0 0.33 0  0 0 0 0 0 0.33 0  _ _ _  1 0.66 0 0  1
R20103xxxxx 3 13 66  0 0 0 0 0 0.33 0 0 0 0 0 0.33 0  0 0 0 0 0 0.33 0  _ _ _  1 0.66 0 0  1
```

This household does not contain exactly two adults so every one is in the 'other household' category.

PCOL1 to PCOL45 would show (note PCOL21-PCOL40 are equal to 0):

```
                N
                O     R
          R     I  P  E                       P P P  P P P P P P P  P        P  P  P P P
          E     N  E  S  P P P P P P  P P P C C C  C C C C C C C  C  _ _ _  C  C  C C C
          F     H  R  P  C C C C C C  C C C O O O  O O O O O O O  O  _ _ _  O  O  O O O
          N     H  S  W  O O O O O O  O O O L L L  L L L L L L L  L  _ _ _  L  L  L L L
          U     L  O  G  L L L L L L  L L L 1 1 1  1 1 1 1 1 1 1  2  _ _ _  4  4  4 4 4
          M     D  N  T  1 2 3 4 5 6  7 8 9 0 1 2  3 4 5 6 7 8 9  0  _ _ _  1  2  3 4 5

R20103xxxxx 3 11 66  0 0 0 0 0 22 0 0 0 0 0 22 0 0 0 0 0 0 22 0  _ _ _  66 44 0 0 66
R20103xxxxx 3 12 66  0 0 0 0 0 22 0 0 0 0 0 22 0 0 0 0 0 0 22 0  _ _ _  66 44 0 0 66
R20103xxxxx 3 13 66  0 0 0 0 0 22 0 0 0 0 0 22 0 0 0 0 0 0 22 0  _ _ _  66 44 0 0 66
```

8. Carry out matrix multiplication

Set dataset ***Z*** to be COL1 - COL&**AUGCOLS**
Set dataset ***PZ*** to be PCOL1 - PCOL&**AUGCOLS**

```
data Z(keep=col1-col&AUGCOLS)
     PZ(keep=pcol1-pcol&AUGCOLS);
     set pzmatrix;
run;
```

Create a matrix ***bigZ*** from dataset ***Z*** (dimension n*p, where p=&COLS and n is the number of people in ***pzmatrix***)

Create a matrix ***PbigZ*** from dataset ***PZ*** (dimension n*p)

Create a matrix ***x*** from dataset ***censpopn*** (dimension p*1)

Carry out the following matrix multiplication, creating matrix ***Wiw*** (dimension n*1)**:**

$$Wiw = \text{PbigZ}\{ (\text{bigZ})^{T} (\text{PbigZ}) \}^{-1} \text{ x}$$

Store this as a SAS dataset, ***Wiw***.

```
PROC IML;

  use Z var _ALL_;
  read ALL into bigZ;

  use PZ var _ALL_;
  read ALL into PbigZ;

  use censpopn var _ALL_;
  read all into x;

  Wiw=PbigZ*inv(t(bigZ)*PbigZ)*x;

  create Wiw from Wiw[colname='Wiw'];
  append from Wiw;
  close Wiw;
```

Wiw has one variable, WIW, which contains an integrated weight value for every individual in the ***hes_tbmp*** **dataset with RESPWGT>0 and CHILD=0.**

Rename WIW as WEIGHT: the name for the integrated weight.

9. Merge integrated weights onto final datasets

We now have the final set of weights. As all individuals within a household have the same weight, these weights can be merged onto output datasets by refnum (or address).

Potential output datasets may be similar to *C9798h.sd2* and *C9798p.sd2*. These are stored in the folder *J:\RO\D_lmhe\hes*.

File origin:

Topic by Gareth Minshall

Database:Survey Methods & Analytical Sup

Subject: ➲ Integrated weights Specs
Categories: ➲ Household Collections\HES\HES 2000/01
Attention: Gareth Minshall, Michael Eglinton, Ramesh Bhula
Other Editors: Gareth Minshall, Matthew Cronin, Michael Eglinton

Note from Gareth 8/10/01: It was decided to calibrate to various child benchmarks after these specs were created (incl children in region, add age benchmarks 0-4, 5-9, 10-14). The specifications have not been updated. There are no plans to update them.

HES_int_wgting.doc

The specifications for Integrated Weighting in the HLFS have been modified for use in the HES.

.

Retention: Indefinite
Composed: 23/09/2000 12:14 pm
Modified: 25/06/2002 04:04 pm
Keywords:

Importance: Routine

By: Matthew Cronin

Appendix 6:

SAS program for creating HES datasets and running preliminary survey-based regression models for expenditure

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;
*                                                                           *;
*  Program Name:                                                            *;
*                                                                           *;
*  Author:         * RViles & SHaslett *                                    *;
*                                                                           *;
*  Date Written:   9 May 2005                                               *;
*                                                                           *;
*  Purpose:        This program was written to create                       *;
*                  datasets of HES and Census candidate                     *;
*                  matching variables for Small Area                        *;
*                  Estimation project for Maori                             *;
*                                                                           *;
*                                                                           *;
*                                                                           *;
*                                                                           *;
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *;

* modified 10 June 2005 so as to create databases from which all relevant matches
  HES and census, can be generated - SJH;
* modified 24 June 2005 from match_var_meansSJH1.sas to get a list of available exact variables names
  for sudaan analaysis from HES;
* modified 27 June 2005 to get potentail matching variables into Sudaan compatible form;

* as for HES4sudaan3.sas but with occupation variable fixed and with
  educqual included in the regression. note that occup can be added to regression too only if
  missing values are recoded as zeros;
* definition of amalgamated ethnic categories also fixed;


/*HES datasets are here but at SNZ only*/
libname HES 'J:\RO\D_lmhe\hes datasets';  run;
/*Store our work here*/
libname SAE 'J:/s_sm/sm2/HES_SAE'; run;



* HES personal datasets;


* extract psu code and merging var address from personal HES dataset h0001p;


options obs=1; run;
 proc print data=hes.h0001p; run;
options obs=max; run;


data h0001p ;
   set hes.h0001p (keep= address persno psu weight);
run;


* Extract matching variables from personal HES dataset h0001pix;


options obs=1; run;
 proc print data=hes.h0001pix; run;
options obs=max; run;


data h0001pix  (rename=(famrel=rel2hhd ));
```

```
  set hes.h0001pix (keep=address persno age sex ethnic marstat rel2ref famrel heduqual
                                                IRN9000  IRN9006  IRN9007  IRN9008  IRN9010  IRN9013 IRN9011
                        IRN9021  IRN9031   IRN9009 Persno );
    label IRN9000='Currently working' IRN9006='Look for paid work' IRN9007='Method look work'
                IRN9008='Start Job Last week' IRN9009='Status work < 30hrs'
          IRN9010='Occupation 1st job' IRN9011='1st job hrs' IRN9013='Industry 1st job'
          IRN9021='2nd job hrs' IRN9031='3rd job hrs';
run;


*** combine all relevant personal data from HES;

proc sort data=h0001p;   by address persno; run;
proc sort data=h0001pix; by address persno; run;

data h0001pcomb;
 merge h0001p h0001pix; by address persno; run;


************ HES household datasets;

**** Extract matching variables from household HES dataset h0001h;


options obs=100; run;
 proc print data=hes.h0001h; run;
options obs=max; run;


data h0001h; * (drop=ethnicity1 ethnicity2 ethnicity3 i);;
  set hes.h0001h; * (keep=address numpers);
  * ADDRESS   AMT   IRN  TOTHHEXP  TOTEXPNCO  MONTH  PERSNO  ITEMCT  TOTRRINC  TOTIRINC
    WEIGHT   RESPWGT   ADNOTWK   FTWORK   HHTYPE   NUMPERS   PTWORK   YEAR  INTVRWR  PSU;

   label tothhexp='Total household exp (excl outlay)'
         totexpnco='Total household exp (incl outlay)';

        totexp_pp = tothhexp / numpers;
        totexpnco_pp = TOTEXPNCO / numpers;
        totrrinc_pp = TOTRRINC / numpers;
        totirinc_pp= TOTIRINC / numpers;
run;


proc sort data=h0001h; by address irn; run;
options obs=100; run;
 proc print data=hes.h0001h; run;
options obs=max; run;



**** Extract matching variables from household HES dataset h0001he;

options obs=10;
 proc print data=hes.h0001he; run;
options obs=max;

options obs=1; run;
 proc print data=h0001hh; run;
option obs=max; run;


   * note that this household level data can now be merged with personal level data
     keeping in mind that variables not ending in pp are still household aggregates
     and should not as such be used in any regression modelling;

proc sort data=h0001pcomb; by address; run;
proc sort data=h0001h    ; by address; run;

libname HEScomb 'C:\shaslett\data';  run;

data HEScomb1;
  merge h0001h h0001pcomb; by address;
   length ethnic_num $ 15;
   length ethn1-ethn15 8;
   length numperson 8;
   length maritstat 8;
   length jobhours 8;
   length educqual 8;

   * ethnicity;
               eth1H=substr(ethnic,1,2);
               eth2H=substr(ethnic,3,2);
```

```
              eth3H=substr(ethnic,5,2);
            if eth1H ='11' then ethnicity1=100000000000000;
            if eth1H ='12' then ethnicity1=010000000000000;
            if eth1H ='21' then ethnicity1=001000000000000;
            if eth1H ='31' then ethnicity1=000100000000000;
            if eth1H ='32' then ethnicity1=000010000000000;
            if eth1H ='33' then ethnicity1=000001000000000;
            if eth1H ='34' then ethnicity1=000000100000000;
            if eth1H ='35' then ethnicity1=000000010000000;
            if eth1H ='36' then ethnicity1=000000001000000;
            if eth1H ='37' then ethnicity1=000000000100000;
            if eth1H ='41' then ethnicity1=000000000010000;
            if eth1H ='42' then ethnicity1=000000000001000;
            if eth1H ='43' then ethnicity1=000000000000100;
            if eth1H ='44' then ethnicity1=000000000000010;
            if eth1H ='50' then ethnicity1=000000000000001;
            if eth1H ='99' then ethnicity1=000000000000000;
             if  eth2H = '11' then ethnicity2=100000000000000;
             if  eth2H = '12' then ethnicity2=010000000000000;
             if  eth2H = '21' then ethnicity2=001000000000000;
             if  eth2H = '31' then ethnicity2=000100000000000;
             if  eth2H = '32' then ethnicity2=000010000000000;
             if  eth2H = '33' then ethnicity2=000001000000000;
             if  eth2H = '34' then ethnicity2=000000100000000;
             if  eth2H = '35' then ethnicity2=000000010000000;
             if  eth2H = '36' then ethnicity2=000000001000000;
             if  eth2H = '37' then ethnicity2=000000000100000;
             if  eth2H = '41' then ethnicity2=000000000010000;
             if  eth2H = '42' then ethnicity2=000000000001000;
             if  eth2H = '43' then ethnicity2=000000000000100;
             if  eth2H = '44' then ethnicity2=000000000000010;
             if  eth2H = '50' then ethnicity2=000000000000001;
             if  eth2H = '99' then ethnicity2=000000000000000;
            if eth3H = '11' then ethnicity3=100000000000000;
            if eth3H = '12' then ethnicity3=010000000000000;
            if eth3H = '21' then ethnicity3=001000000000000;
            if eth3H = '31' then ethnicity3=000100000000000;
            if eth3H = '32' then ethnicity3=000010000000000;
            if eth3H = '33' then ethnicity3=000001000000000;
            if eth3H = '34' then ethnicity3=000000100000000;
            if eth3H = '35' then ethnicity3=000000010000000;
            if eth3H = '36' then ethnicity3=000000001000000;
            if eth3H = '37' then ethnicity3=000000000100000;
            if eth3H = '41' then ethnicity3=000000000010000;
            if eth3H = '42' then ethnicity3=000000000001000;
            if eth3H = '43' then ethnicity3=000000000000100;
            if eth3H = '44' then ethnicity3=000000000000010;
            if eth3H = '50' then ethnicity3=000000000000001;
            if eth3H = '99' then ethnicity3=000000000000000;

          ethnic_num=sum(ethnicity1,ethnicity2,ethnicity3);
          do i=1to 15;
           if substr(ethnic_num,i,1)=' ' then substr(ethnic_num,i,1)='0';
          end;

* ethnic;    *use ethn1-ethn15 for Sudaan;

     array ethn   ethn1-ethn15;
          do _i_=1 to 15;
            ethn=substr(ethnic_num,_i_,1);
          end;
          if eth1h='99' and eth2h='  ' and eth3h='  ' then do over ethn; ethn=.; end;

       * ethn1  = NZ European / Pakeha
       * ethn2  = Other European
       * ethn3  = NZ Maori
       * ethn4  = Samoan
       * ethn5  = Cook Island Maori
       * ethn6  = Tongan
       * ethn7  = Niuean
       * ethn8  = Tokolauan
       * ethn9  = Fijian
       * ethn10  = Other Pacific Isalnd Groups
       * ethn11  = Southeast Asian
       * ethn12  = Chinese
       * ethn13  = Indian
       * ethn14  = Other Asian
       * ethn15  = Other;

* sex;       * use sex for Sudaan;

      if sex=1 then Male=1;
          if sex=2 then Male=0;
```

```
      if sex=1 then Female=0;
              if sex=2 then Female=1;

* age;
              if       0<=age< 5  then agegpc=' 0- 4';
              else if  5<=age<10  then agegpc=' 5- 9';
          else if 10<=age<15  then agegpc='10-14';
      else if 15<=age<20  then agegpc='15-19';
          else if 10<=age<15  then agegpc='10-14';
      else if 15<=age<20  then agegpc='15-19';
              else if 20<=age<25  then agegpc='20-24';
      else if 25<=age<30  then agegpc='25-29';
              else if 30<=age<35  then agegpc='30-34';
      else if 35<=age<40  then agegpc='35-39';
              else if 40<=age<45  then agegpc='40-44';
      else if 45<=age<50  then agegpc='44-49';
              else if 50<=age<55  then agegpc='50-54';
      else if 55<=age<60  then agegpc='55-59';
              else if 60<=age<65  then agegpc='60-64';
      else if 65<=age<70  then agegpc='65-69';
              else if 70<=age<75  then agegpc='70-74';
      else if 75<=age<80  then agegpc='75-79';
              else if 80<=age<85  then agegpc='80-84';
      else if 85<=age<90  then agegpc='85-89';
              else if 90<=age<95  then agegpc='90-94';
      else if 95<=age<100 then agegpc='95-99';
              else if age>=100    then agegpc=' 100+';
*    age;
         if age le 14 then agegpchild=1;
          else agegpchild=0;
         if age ge 15 and age le 29 then agegp15_29=1;
          else agegp15_29=0;
         if age ge 30 and age le 44 then agegp30_44=1;
          else agegp30_44=0;
         if age ge 45 and age le 64 then agegp45_64=1;
          else agegp45_64=0;
         if age ge 65 then agegp_over64=1;
          else agegp_over64=0;

     * age;       * use for Sudaan;

             if age le 14 then agegp=1;
         else if age ge 15 and age le 29 then agegp=2;
         else if age ge 30 and age le 44 then agegp=3;
         else if age ge 45 and age le 64 then agegp=4;
         else if age ge 65 then agegp=5;

* 'relation to reference person' var in HES ;
      if rel2ref = '01' then refpers=1;
             else if rel2ref ne '01' then refpers = 0;
          if rel2ref = '02' then Partner=1;
             else if rel2ref ne '02' then partner=0;
          if rel2ref = '03' then Child=1;
             else if rel2ref ne '03' then child=0;
          if rel2ref = '04' then parent=1;
             else if rel2ref ne '04' then parent =0;
          if rel2ref = '05' then otherrel =1;
             else if rel2ref ne '05' then otherrel=0;
          if rel2ref = '06' then norel=1;
             else if rel2ref ne '06' then norel=0;

* number of people in household;
     if numpers='01' then hhd1=1;
     else if numpers ne '01' then hhd1=0;
     if numpers='02' then hhd2=1;
     else if numpers ne '02' then hhd2=0;
     if numpers='03' then hhd3=1;
     else if numpers ne '03' then hhd3=0;
     if numpers='04' then hhd4=1;
     else if numpers ne '04' then hhd4=0;
     if numpers='05' then hhd5=1;
     else if numpers ne '05' then hhd5=0;
     if numpers='06' then hhd6=1;
     else if numpers ne '06' then hhd6=0;
     if numpers='07' then hhd7=1;
     else if numpers ne '07' then hhd7=0;
     if numpers='08' then hhd8=1;
     else if numpers ne '08' then hhd8=0;
     if numpers='09' then hhd9=1;
     else if numpers ne '09' then hhd9=0;
     if numpers='10' then hhd10=1;
     else if numpers ne '10' then hhd10=0;
     if numpers='11' then hhd11=1;
```

```
        else if numpers ne '11' then hhd11=0;

* numeric code for number of people ;  *use in Sudaan - implicit and possibly explicit;

                numperson=numpers;

*       Marital Status;
        if marstat= '1' then partnerms=1;
        else partner = 0;
        if marstat= '2' then nopartnerms=1;
        else nopartner = 0;
        if marstat= '9' then other=1;
        else other = 0;

*       Marital Status;  * use for Sudaan;

       if marstat= '1' then maritstat=1;
       else if marstat= '2' then maritstat=2;
       else if marstat= '9' then maritstat=3;

*   Hours worked in 1st Job;
        if irn9011 ge 1 and irn9011 le 010 then job1hr10=1;
         else job1hr10=0;
        if irn9011 gt 010 and irn9011 le 020 then job1hr20=1;
         else job1hr20=0;
        if irn9011 gt 020 and irn9011 le 030 then job1hr30=1;
         else job1hr30=0;
        if irn9011 gt 030 and irn9011 le 040 then job1hr40=1;
         else job1hr40=0;
        if irn9011 gt 040 and irn9011 le 050 then job1hr50=1;
         else job1hr50=0;
        if irn9011 gt 050 and irn9011 le 060 then job1hr60=1;
         else job1hr60=0;
        if irn9011 gt 060 and irn9011 le 070 then job1hr70=1;
         else job1hr70=0;
        if irn9011 gt 070 and irn9011 le 080 then job1hr80=1;
         else job1hr80=0;
        if irn9011 gt 080 and irn9011 le 090 then job1hr90=1;
         else job1hr90=0;
        if irn9011 gt 090 and irn9011 le 98 then job1hr100=1;
         else job1hr100=0;
        if irn9011 gt 98 or irn9011 =. then other=1;
         else other=0;
* jobhours;  * for use in Sudaan;
         jobhours=irn9011;
         if jobhours=. then jobhours=0;
         jobhours_sq=jobhours**2;

        *available to start a job last week;
        if irn9008 =1 then availwk=1;
         else availwk=0;
        if irn9008 =2 then notavailwk=1;
         else notavailwk=0;
        if irn9008 not in (1,2) then jobavother=1;
         else jobavother=0;
* job availability;  * for use in Sudaan;
        if availwk=1 then jobavail=1;
    if notavailwk=1 then jobavail=0;
        if jobavother=1 then jobavail=2;

*       seek work;
        if irn9006 =1 then seekwk=1;
         else seekwk=0;
        if irn9006 =2 then noseekwk=1;
         else noseekwk=0;
        if irn9006 not in (1,2) then other=1;
         else other=0;
* seeking work; * for use in Sudaan;
    if irn9006 =1 then seekwork=1;
        if irn9006 =2 then seekwork=0;
        if irn9006 not in (1,2) then seekwork=2;

*       working;
        if irn9000=1 then working =1;
         else working=0;
        if irn9000=2 then notwork=1;
         else notwork=0;
        if irn9000=. then other=1;
         else other=0;
* working; * for use in Sudaan;
    if irn9000 =1 then work=1;
        if irn9000 =2 then work=0;
        if irn9000 not in (1,2) then work=2;
```

```
*       highest qual;
        if heduqual='1' then none=1;
         else none=0;
        if heduqual ='2' then scert=1;
         else scert=0;
        if heduqual ='3' then bursary=1;
         else bursary=0;
        if heduqual ='4' then vocat=1;
         else vocat=0;
        if heduqual ='5' then degree=1;
         else degree=0;
        if heduqual ='6' then postgrad=1;
         else postgrad=0;
        if heduqual ='7' then otherqual=1;
         else otherqual=0;
        if heduqual ='9' then other=1;
         else other=0;
        if heduqual ='0' then notscope=1;
         else notscope=0;
 * highest educational qualification: * for use in Sudaan
               1=none 2=scert 3=bursary 4=vocat 5=degree 6=postgrad
            7=otherqual 9=other 0=notscope;

        educqual=heduqual;

*       Occupation;
        irn901O=put(irn901O,$2.); *make into character var?;
        occupation=substr(irn901O,1,1);
        if occupation=1 then Managers=1;
         else Managers=0;
        if occupation=2 then Professionals=1;
         else Professionals=0;
        if occupation=3 then Technicians=1;
         else Technicians=0;
        if occupation=4 then Clerks=1;
         else Clerks=0;
        if occupation=5 then Sales=1;
         else Sales=0;
        if occupation=6 then Agriculture=1;
         else Agriculture=0;
        if occupation=7 then Trades=1;
         else Trades=0;
        if occupation=8 then Operators=1;
         else Operators=0;
        if occupation=9 then Elementary=1;
         else Elementary=0;
        if occupation=. then other=1;
         else other=0;
* occupation;  * for use in Sudaan;
   if occupation=. then occup=0; else occup=occupation;

* income;
        if totrrinc lt 0 then loss=1;
        else loss=0;
        if totrrinc=0 then zero=1;
        else zero=0;
        if totrrinc gt 0 and totrrinc le 5000 then less5k=1;
        else less5k=0;
        if totrrinc gt 5000 and totrrinc le 10000 then less10k=1;
        else less10k=0;
        if totrrinc gt 10000 and totrrinc le 15000 then less15k=1;
        else less15k=0;
        if totrrinc gt 15000 and totrrinc le 20000 then less20k=1;
        else less20k=0;
        if totrrinc gt 20000 and totrrinc le 25000 then less25k=1;
        else less25k=0;
        if totrrinc gt 25000 and totrrinc le 30000 then less30k=1;
        else less30k=0;
        if totrrinc gt 30000 and totrrinc le 40000 then less40k=1;
        else less40k=0;
        if totrrinc gt 40000 and totrrinc le 50000 then less50k=1;
        else less50k=0;
        if totrrinc gt 50000 and totrrinc le 70000 then less70k=1;
        else less70k=0;
        if totrrinc gt 70000 and totrrinc le 100000 then less100k=1;
        else less100k=0;
        if totrrinc gt 100000  then more100k=1;
        else more100k=0;
        if totrrinc =. then other=1;
        else other=0;
* income per person; * for use in Sudaan;
        incomepp=totrrinc/numperson;
```

```
* define Sudaan design information;

   strat=substr(address,1,3);
   * psu is called psu;

run;

* household level varaibles to be added to dataset;

data HEShhld1 (keep=address
                   residencetype
                   weekrent
                   numbedroom
                   numroom
                   numcars
                   houseownership);
  set sae.HES_var_test;
* residence type;
 if irn=8000 then do;
        if amt=1 then house=1;
         else house=0;
        if amt=2 then flats2=1;
         else flats2=0;
        if amt=3 then flats3=1;
         else flats3=0;
        if amt=4 then flats3storey3=1;
         else flats3storey3=0;
        if amt=5 then jointshop=1;
         else jointshop=0;
        if amt=6 then bach=1;
         else bach=0;
 end;
 * residence type; * for use in Sudaan;
    if irn=8000 then do;
          residencetype=amt;
          if residencetype<=0 or residencetype>6 then residencetype=0;
        end;

 *weekly rent;
 if irn=8103 then do;
        weekrent=amt/52;
        if weekrent=. then weekrent=0;
        if weekrent=0 then rentzero=1;
        else rentzero=0;
        if weekrent lt 100 then rentless100=1;
        else rentless100=0;
        if weekrent ge 100 and weekrent lt 200 then rentless200=1;
        else rentless200=0;
        if weekrent ge 200 and weekrent lt 300 then rentless300=1;
        else rentless300=0;
        if weekrent ge 300 and weekrent lt 400 then rentless400=1;
        else rentless400=0;
        if weekrent ge 400 and weekrent lt 500 then rentless500=1;
        else rentless500=0;
        if weekrent ge 500 then rentover500=1;
        else rentover500=0;
  end;
   if weekrent=. then weekrent=0;

* bedrooms;
  if irn=8001 then do;
        if amt =0 then bdroom0=1;
        else bdroom0=0;
        if amt =1 then bdroom1=1;
        else bdroom1=0;
        if amt =2 then bdroom2=1;
        else bdroom2=0;
        if amt =3 then bdroom3=1;
        else bdroom3=0;
        if amt =4 then bdroom4=1;
        else bdroom4=0;
        if amt =5 then bdroom5=1;
        else bdroom5=0;
        if amt =6 then bdroom6=1;
        else bdroom6=0;
        if amt =7 then bdroom7=1;
        else bdroom7=0;
  end;
  if irn=8001 then numbedroom=amt;

* rooms;
  if irn=8009 then do;
```

```
    if amt =1 then room1=1;
        else room1=0;
        if amt =2 then room2=1;
        else room2=0;
        if amt =3 then room3=1;
        else room3=0;
        if amt =4 then room4=1;
        else room4=0;
        if amt =5 then room5=1;
        else room5=0;
        if amt =6 then room6=1;
        else room6=0;
        if amt =7 then room7=1;
        else room7=0;
        if amt =8 then room8=1;
        else room8=0;
        if amt =9 then room9=1;
        else room9=0;
        if amt =10 then room10=1;
        else room10=0;
        if amt =11 then room11=1;
        else room11=0;
        if amt =12 then room12=1;
        else room12=0;
        if amt =13 then room13=1;
        else room13=0;
        if amt =14 then room14=1;
        else room14=0;
        if amt =15 then room15=1;
        else room15=0;
        if amt =16 then room16=1;
        else room16=0;
        if amt =17 then room17=1;
        else room17=0;
        if amt =18 then room18=1;
        else room18=0;
        if amt =19 then room19=1;
        else room19=0;
 end;
  if irn=8009 then numroom=amt;

* cars;
 if irn=8720 then do;;
        if amt=0 then car0=1;
        else car0=0;
        if amt =1 then car1=1;
        else car1=0;
        if amt =2 then car2=1;
        else car2=0;
        if amt ge 3 then car3=1;
        else car3=0;
  end;
  if irn=8720 then  numcars=amt;

* house ownership;
  if irn=8014 then do;
        if amt =1 then rent=1;
        else rent=0;
        if amt =2 then rentfree=1;
        else rentfree=0;
        if amt =3 then mortgage=1;
        else mortgage=0;
        if amt=4 then nomortgage=1;
        else nomortgage=0;
  end;
*house ownership; * for use in Sudaan;
 if irn=8014 then do;
    houseownership=amt;
    if houseownership<=0 or houseownership>4 then houseownership=0;
 end;

run;

proc sort data=HEShhld1; by address; run;

proc means data=HEShhld1 noprint;
 var  residencetype weekrent numbedroom numroom numcars houseownership;
 output out=HEShhld2 sum=residencetype weekrent numbedroom numroom numcars houseownership;
 by address;
run;

* proc print data=HEShhld2; run;
```

```
proc sort data=HEScomb1; by address; run;
proc sort data=HEShhld2; by address; run;

data HEScomb2a;
 merge HEScomb1 HEShhld2; by address;

*rent per person; * for use in Sudaan;
   weekrentpp=weekrent/numperson;

* bedrooms per person;  * for use in Sudaan;
    bedroomspp=numbedroom/numperson;

* rooms per person;  * for use in Sudaan;
    roomspp=numroom/numperson;

* cars per person; * for use in Sudaan;
    carspp=numcars/numperson;
run;

 * make permanent SAS dataset;

data HEScomb.HEScomb;
  set HEScomb2a;
  run;



/* * check how many strata under definition above;

   data tst1; set HEScomb.HEScomb; strat=substr(address,1,3); run;
   proc sort data=tst1; by strat; run;
   data tst2 (keep=strat);
    set tst1; by strat;
     if first.strat then output;
   run;
  * check number of observations in tst2 is 122, which is the number of strata
  * actually equals 120;

*/


* expenditure per person;
*       totexp_pp
        totexpnco_pp
* income per person;
*       totrrinc_pp
        totirinc_pp;
* notes on regressor variables:
   ethn1-ethn15 (since multiple response possible) 0/1
   sex (male = 1, female=2)
   agegp (1=age le 14, 2=age ge 15 and age le 29, 3=age ge 30 and age le 44,
          4=age ge 45 and age le 64, 5=age ge 65)
   refpers Partner Child parent otherrel norel (relation to head of household: 0/1)
   maritstat (1=partnered,2=no partner,3=other)
   jobhours   jobhours_sq
   jobavail (0=notavail, 1=availwk, 2=other)
   seekwork (0=no, 1=yes, 2=other)
   work (ie working: 0=no, 1=yes, 2=other)
   educqual
                  (1=none 2=scert 3=bursary 4=vocat 5=degree 6=postgrad
             7=otherqual 9=other 0=notscope)
   occup
     (1=Managers=1 2=Professionals 3=Technicians 4=Clerks 5=Sales 6=Agriculture
      7=Trades 8=Operators 9=Elementary 0=other)
   residencetype (1=house 2=flats2 3=flats3 4=flats3storey3 5=jointshop 6=bach)
   weekrentpp
   bedroomspp
   roomspp
   carspp
   houseownership (1=rent 2=rentfree 3=mortgage 4=nomortgage)
   incomepp
   numperson

* possibly important interactions:
   agegp*occup
   agegp*seekwork
   agegp*work
   agegp*educqual
  Note too link between the work variable and income;


data HEScomb2;
 set HEScomb.Hescomb;
```

```
  length stratnum 8;
  length psunum 8;
  length irn9010c $ 2 ;
   stratnum=strat; psunum=psu;
   if totexp_pp>0 then lntotexp_pp=log(totexp_pp);
   if incomepp>0  then lnincomepp=log(incomepp);
*    highest qual;
       if heduqual='1' then none=1;
        else none=0;
       if heduqual ='2' then scert=1;
        else scert=0;
       if heduqual ='3' then bursary=1;
        else bursary=0;
       if heduqual ='4' then vocat=1;
        else vocat=0;
       if heduqual ='5' then degree=1;
        else degree=0;
       if heduqual ='6' then postgrad=1;
        else postgrad=0;
       if heduqual ='7' then otherqual=1;
        else otherqual=0;
       if heduqual ='9' then other=1;
        else other=0;
       if heduqual ='0' then notscope=1;
        else notscope=0;
         * highest educational qualification: * for use in Sudaan
               1=none 2=scert 3=bursary 4=vocat 5=degree 6=postgrad
               7=otherqual 9=other 0=notscope;
       * educqual=heduqual;
 *     Occupation;
       * irn9010=put(irn9010,$2.); *make into character var?;
       irn9010c=irn9010;
       occupation=substr(irn9010c,1,1);
       if occupation=1 then Managers=1;
        else Managers=0;
       if occupation=2 then Professionals=1;
        else Professionals=0;
       if occupation=3 then Technicians=1;
        else Technicians=0;
       if occupation=4 then Clerks=1;
        else Clerks=0;
       if occupation=5 then Sales=1;
        else Sales=0;
       if occupation=6 then Agriculture=1;
        else Agriculture=0;
       if occupation=7 then Trades=1;
        else Trades=0;
       if occupation=8 then Operators=1;
        else Operators=0;
       if occupation=9 then Elementary=1;
        else Elementary=0;
       if occupation=. then other=1;
        else other=0;
     * occupation;  * for use in Sudaan;
      if occupation=. then occup=0; else occup=occupation;
*    age;
           if age le 14 then child=1;
            else child=0;
           if age ge 15 and age le 29 then gp1=1;
            else gp1=0;
           if age ge 30 and age le 44 then gp2=1;
            else gp2=0;
           if age ge 45 and age le 64 then gp3=1;
            else gp3=0;
           if age ge 65 then gp4=1;
            else gp4=0;
*       working;
       if irn9000=1 then working =1;
        else working=0;
       if irn9000=2 then notwork=1;
        else notwork=0;
       if irn9000=. then otherw=1;
        else otherw=0;
   * working; * for use in Sudaan;
       if irn9000 =1 then work=1;
       if irn9000 =2 then work=0;
       if irn9000 not in (1,2) then work=2;

run;

proc sort data=HEScomb2; by stratnum psunum; run;

/*
```

```
proc print data=HEScomb2;
 var totexp_pp numbedroom numroom numcars bedroomspp roomspp carspp incomepp numperson;
run;
*/


* untransfomed;

proc regress data=HEScomb2 DEFT1 design=wr;
 nest stratnum psunum;
 class ethn1-ethn15 sex agegp
       refpers Partner Child parent otherrel norel educqual
            maritstat jobavail seekwork work occup residencetype;
 model totexp_pp = ethn1-ethn15 sex agegp
                   refpers Partner Child parent otherrel norel
                                educqual
                        maritstat jobavail seekwork work occup residencetype
                     jobhours jobhours_sq weekrentpp bedroomspp roomspp carspp
                           incomepp
                           numperson;
 weight weight;  * other possible weight is variable called respwgt;
                  * note that weights have not at this stage 270605 been adjusted to
                    match Maori percentage in census. this should affect R**2 only slightly.
                    see final feasibility report for details;
run;
/*
* note some negative values for income;
* data on this variable and others needs further editing;

proc plot data=HEScomb2;
 plot totexp_pp*incomepp;
run;
*/


* log transfomed expenditure and income;

proc regress data=HEScomb2 DEFT1 design=wr;
 nest stratnum psunum;
 class ethn1-ethn15 sex agegp
       refpers Partner Child parent otherrel norel
            educqual
            maritstat jobavail seekwork work occup residencetype;
 model lntotexp_pp = ethn1-ethn15 sex agegp
                     refpers Partner Child parent otherrel norel
                                      educqual
                          maritstat jobavail seekwork work occup residencetype
                     jobhours jobhours_sq weekrentpp bedroomspp roomspp carspp
                           lnincomepp
                           numperson;
 weight weight;  * other possible weight is variable called respwgt;
                  * note that weights have not at this stage 270605 been adjusted to
                    match Maori percentage in census. this should affect R**2 only slightly.
                    see final feasibility report for details;
run;


* log transfomed expenditure and income with selected interactions, but deliberately
    not including interactions with ethnicity, since these are the small domains of interest
    so increasing R**2 by using these only breaks up the samplke in terms of teh model so
    little additional real gain in terms of saes would be expected from such a strategy;

proc regress data=HEScomb2 DEFT1 design=wr;
 nest stratnum psunum;
 class ethn1-ethn15 sex agegp
       refpers Partner Child parent otherrel norel
            maritstat jobavail seekwork work occup residencetype;
 model lntotexp_pp = ethn1-ethn15    agegp
                           maritstat       seekwork          residencetype
                                                occup  educqual
                                    weekrentpp       roomspp carspp
                           lnincomepp
                           numperson  ;
                    * variables and interactions removed from model:
                                       work occup jobavail
                   sex jobhours bedroompp
                               jobhours_sq refpers Partner Child parent otherrel norel
                               agegp*occup agegp*seekwork agegp*work agegp*educqual work*educqual;
 weight weight;  * other possible weight is variable called respwgt;
                  * note that weights have not at this stage 270605 been adjusted to
                    match Maori percentage in census. this should affect R**2 only slightly.
                    see final feasibility report for details;
run;
```

```
         * ethn1  = NZ European / Pakeha
         * ethn2  = Other European
         * ethn3  = NZ Maori
         * ethn4  = Samoan
         * ethn5  = Cook Island Maori
         * ethn6  = Tongan
         * ethn7  = Niuean
         * ethn8  = Tokolauan
         * ethn9  = Fijian
         * ethn10  = Other Pacific Isalnd Groups
         * ethn11  = Southeast Asian
         * ethn12  = Chinese
         * ethn13  = Indian
         * ethn14  = Other Asian
         * ethn15  = Other;

data HEScomb3;
 set HEScomb2;
  if ethn1=1 then ethn_NZEurop=1;
  if ethn2=1 then ethn_OthEurop=1;
  if ethn3=1 then ethn_NZMaori=1;
  if ethn5=1 then ethn_CIMaori=1;
  if ethn4=1 or ethn6=1 or ethn7=1 or ethn8=1 or ethn9=1 then ethn_PIg1=1;
  if ethn10=1 then ethn_PIOthg=1;
  if ethn11=1 or ethn12=1 or ethn13=1 or ethn14=1 then ethn_Asian=1;
  if ethn15=1 then ethn_Oth=1;
  array ethnarr   ethn_NZEurop ethn_OthEurop ethn_NZMaori ethn_CIMaori
              ethn_PIg1 ethn_PIOthg ethn_Asian ethn_Oth;
   do over ethnarr; if ethnarr=. then ethnarr=0; end;
run;

/*
proc print data=HEScomb3;
  var ethn_NZEurop ethn_OthEurop ethn_NZMaori ethn_CIMaori
              ethn_PIg1 ethn_PIOthg ethn_Asian ethn_Oth;
run;
*/

* log transfomed expenditure and income with selected interactions, but deliberately
    not including interactions with ethnicity, since these are the small domains of interest
    so increasing R**2 by using these only breaks up the sample in terms of the model so
    little additional real gain in terms of saes would be expected from such a strategy;

proc regress data=HEScomb3 DEFT1 design=wr;
 nest stratnum psunum;
 class ethn1-ethn15 sex agegp
       refpers Partner Child parent otherrel norel
            educqual
            maritstat jobavail seekwork work occup residencetype;
 model lntotexp_pp = ethn_NZEurop ethn_OthEurop ethn_NZMaori ethn_CIMaori
                     ethn_PIg1 ethn_PIOthg ethn_Asian ethn_Oth
                     agegp
                     occup educqual
                           maritstat       seekwork          residencetype
                                    weekrentpp       roomspp carspp
                           lnincomepp
                           numperson  ;
                   * variables and interactions removed from model:
                                         reduced number of ethnic groups
                                         work occup jobavail
                 sex jobhours bedroompp
                             jobhours_sq refpers Partner Child parent otherrel norel
                             agegp*occup agegp*seekwork agegp*work agegp*educqual work*educqual;
 weight weight;  * other possible weight is variable called respwgt;
                 * note that weights have not at this stage 270605 been adjusted to
                   match Maori percentage in census. this should affect R**2 only slightly.
                   see final feasibility report for details;
run;
```

Appendix 7:

SAS'/Sudaan: Preliminary model for sample survey regression from HES for expenditure

```
                              S U D A A N
            Software for the Statistical Analysis of Correlated Data
           Copyright        Research Triangle Institute          July 2004
                                  Release 9.0.0


Number of observations read       :   7385    Weighted count:  3804401
Observations used in the analysis :   7343    Weighted count:  3779753
Denominator degrees of freedom    :    617


Maximum number of estimable parameters for the model is 44

File HESCOMB3 contains  737 Clusters
 735 clusters were used to fit the model
Maximum cluster size is  28 records
Minimum cluster size is   1 records

Weighted mean response is 9.425728


Multiple R-Square for the dependent variable LNTOTEXP_PP: 0.555253
```

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable LNTOTEXP_PP: LNTOTEXP_PP
by: Independent Variables and Effects.

Independent Variables and Effects	Beta Coeff.	DEFF Beta #1	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0
Intercept	5.61	1.88	0.31	5.00	6.22	18.05
ETHN_NZEUROP	0.13	3.82	0.03	0.06	0.20	3.71
ETHN_OTHEUROP	0.17	3.35	0.04	0.08	0.25	3.73
ETHN_NZMAORI	-0.07	4.83	0.04	-0.14	-0.01	-2.11
ETHN_CIMAORI	-0.16	3.62	0.07	-0.30	-0.02	-2.26
ETHN_PIG1	-0.05	4.30	0.06	-0.16	0.07	-0.84
ETHN_PIOTHG	0.45	2.13	0.19	0.08	0.82	2.40
ETHN_ASIAN	-0.00	7.15	0.08	-0.16	0.16	-0.02
ETHN_OTH	0.11	3.47	0.06	-0.01	0.23	1.76
AGEGP						
1	0.16	1.48	0.07	0.02	0.30	2.27
2	-0.06	1.02	0.10	-0.26	0.15	-0.55
3	-0.15	1.01	0.10	-0.35	0.05	-1.45
4	-0.16	0.98	0.10	-0.36	0.04	-1.55
5	0.00	.	0.00	0.00	0.00	.
OCCUP						
0	0.10	1.17	0.06	-0.03	0.22	1.56
1	0.05	1.47	0.04	-0.03	0.13	1.29
2	0.01	1.43	0.04	-0.06	0.08	0.25
3	0.07	1.48	0.04	-0.01	0.15	1.83
4	0.07	1.32	0.04	0.00	0.14	1.98
5	0.03	1.41	0.03	-0.04	0.10	0.91
6	-0.07	1.48	0.04	-0.15	0.01	-1.72
7	0.02	2.20	0.05	-0.09	0.12	0.29
8	-0.10	1.33	0.04	-0.18	-0.01	-2.31
9	0.00	.	0.00	0.00	0.00	.
EDUCQUAL						
0	-0.29	1.03	0.11	-0.50	-0.08	-2.71
1	0.00	2.16	0.05	-0.10	0.11	0.09
2	0.05	2.08	0.05	-0.05	0.16	0.97
3	0.18	2.17	0.06	0.07	0.29	3.18
4	0.13	2.10	0.05	0.02	0.24	2.41
5	0.18	2.15	0.06	0.07	0.29	3.09
6	0.32	1.80	0.06	0.19	0.45	5.02
7	0.22	1.49	0.06	0.10	0.34	3.53
9	0.00	.	0.00	0.00	0.00	.
MARITSTAT						
1	-0.38	1.20	0.17	-0.72	-0.04	-2.20
2	-0.45	1.22	0.17	-0.79	-0.11	-2.57

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable LNTOTEXP_PP: LNTOTEXP_PP
by: Independent Variables and Effects.

Independent Variables and Effects	P-value T-Test B=0
Intercept	0.0000
ETHN_NZEUROP	0.0002
ETHN_OTHEUROP	0.0002
ETHN_NZMAORI	0.0355
ETHN_CIMAORI	0.0244
ETHN_PIG1	0.4021
ETHN_PIOTHG	0.0168
ETHN_ASIAN	0.9845
ETHN_OTH	0.0793
AGEGP	
1	0.0236
2	0.5795
3	0.1489
4	0.1223
5	.
OCCUP	
0	0.1197
1	0.1977
2	0.8008
3	0.0673
4	0.0479
5	0.3657
6	0.0856
7	0.7693
8	0.0211
9	.
EDUCQUAL	
0	0.0070
1	0.9264
2	0.3314
3	0.0015
4	0.0162
5	0.0021
6	0.0000
7	0.0004
9	.
MARITSTAT	
1	0.0281
2	0.0104

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable LNTOTEXP_PP: LNTOTEXP_PP
by: Independent Variables and Effects.

Independent Variables and Effects	Beta Coeff.	DEFF Beta #1	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0
MARITSTAT						
3	0.00	.	0.00	0.00	0.00	.
SEEKWORK						
0	-0.15	1.14	0.06	-0.26	-0.04	-2.67
1	-0.17	1.06	0.06	-0.29	-0.05	-2.77
2	0.00	.	0.00	0.00	0.00	.
RESIDENCETYPE						
1	0.67	2.22	0.03	0.61	0.73	22.39
2	0.63	2.12	0.04	0.56	0.71	17.32
3	0.60	2.05	0.05	0.51	0.69	13.12
4	0.67	1.92	0.08	0.51	0.83	8.27
5	0.70	4.74	0.21	0.30	1.10	3.42
6	0.00	.	0.00	0.00	0.00	.
WEEKRENTPP	0.00	2.87	0.00	0.00	0.00	3.23
ROOMSPP	0.04	6.12	0.01	0.02	0.07	3.35
CARSPP	0.16	2.57	0.02	0.12	0.20	8.23
LNINCOMEPP	0.36	3.11	0.02	0.32	0.40	16.30
NUMPERSON	-0.06	11.76	0.02	-0.11	-0.01	-2.56

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable LNTOTEXP_PP: LNTOTEXP_PP
by: Independent Variables and Effects.

Independent Variables and Effects	P-value T-Test B=0
MARITSTAT	
3	.
SEEKWORK	
0	0.0079
1	0.0058
2	.
RESIDENCETYPE	
1	0.0000
2	0.0000
3	0.0000
4	0.0000
5	0.0007
6	.
WEEKRENTPP	0.0013
ROOMSPP	0.0009
CARSPP	0.0000
LNINCOMEPP	0.0000
NUMPERSON	0.0106

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983) Working Correlations: Independent Link Function: Identity
Response variable LNTOTEXP_PP: LNTOTEXP_PP

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	43	132006.36	0.0000
MODEL MINUS INTERCEPT	43	350.55	0.0000
INTERCEPT	.	.	.
ETHN_NZEUROP	1	13.75	0.0002
ETHN_OTHEUROP	1	13.92	0.0002
ETHN_NZMAORI	1	4.44	0.0355
ETHN_CIMAORI	1	5.09	0.0244
ETHN_PIG1	1	0.70	0.4021
ETHN_PIOTHG	1	5.75	0.0168
ETHN_ASIAN	1	0.00	0.9845
ETHN_OTH	1	3.09	0.0793
AGEGP	4	6.10	0.0001
OCCUP	9	3.54	0.0003
EDUCQUAL	8	14.25	0.0000
MARITSTAT	2	7.03	0.0010
SEEKWORK	2	4.17	0.0159
RESIDENCETYPE	5	109.18	0.0000
WEEKRENTPP	1	10.43	0.0013
ROOMSPP	1	11.21	0.0009
CARSPP	1	67.79	0.0000
LNINCOMEPP	1	265.82	0.0000
NUMPERSON	1	6.57	0.0106