Negotiating the Life Course

Negotiating the Life Course, Waves 1 and 2 Sampling Weights for Persons and Income Units

Trevor Breusch *

Negotiating the Life Course Discussion Paper Series

Discussion Paper DP- 016

July 2003

* Research School of Social Sciences, Australian National University

Introduction

This paper describes weighting variables for the NLC data, with the intention that the weighted sample will be representative of the Australian population, both at the person level and at the level of income units.

The sampling scheme for the NLC was random across households that have fixed-line telephone numbers in Australia. The data items collected in the NLC questionnaire relate to the households themselves, to the individuals within the households, and to the dynamics of family formation. Analyses carried out on the NLC data have been variously concerned with households, individuals or families.

Design Weights for Persons (Wave 1)

The sampling design was to accept into the sample the household that was telephoned if there was a person in the age range 18-54 available to be the respondent. If there was more than one eligible person, a randomization scheme was adopted to select the respondent. Persons in larger households therefore have a proportionally lower chance of inclusion in the sample.

The Wave 1 data file includes a variable called ‘hhwt’, which is directly proportional to the variable ‘numhh’, the number of people in the household aged 18-54. The difference is that ‘hhwt’ is scaled so that it averages to 1.0 across the sample of 2231 respondents. The bias in the design of the sampling scheme, as it operates against persons in larger households, can be corrected by weighting the responses using ‘hhwt’.

Table 1: Codebook ‘hhwt’, Wave 1

```
hhwt ---------------------------- household weighting factor
                  type:  numeric (float)

                 range:  [.5403245,3.7822716]         units:  1.000e-07
          unique values:  7                    coded missing:  0 / 2231

             tabulation:  Freq.  Value
                            726  .54032451
                           1212  1.080649
                            213  1.6209736
                             65  2.161298
                             11  2.7016227
                              3  3.2419472
                              1  3.7822716
```

Benchmarked Weights for Persons (Wave 1)

The weight variable 'hhwt' only allows for a design feature in the sampling method. It makes no allowance for other aspects of the sampling process that might lead to a biased sample, such as who owns a telephone, who is home to answer the phone, and who has the time and inclination to answer a lengthy questionnaire that is asking personal questions. A different approach is needed to correct for these potential biases.

The method used here is post-stratification and benchmarking. The idea is to cross-classify the NLC sample in several dimensions and weight each 'cell' so the weighted proportion of the sample that falls into the cell matches external estimates for the Australian population at the time. This method is similar in principle to the methods used by the ABS to calculate the weights that are provided with their surveys, such as the *Survey of Income and Housing Costs (SIHC)*. ABS methods are more complex, both because they provide inter-related weights to use at various levels of analysis (household, income unit, person) and because they stratify on extra dimensions (numbers of children and adults in the household, state, metropolitan or other region, etc.).

Benchmarking is done after the sample has been weighted to correct for the sampling design. To prevent cell sizes from getting too small, a modified form of 'hhwt' is used, where the largest number of eligible adults in a household is truncated at 'four or more'. The sample is then benchmarked on the age and sex of the respondent to *1996 Census of Population and Housing*, table B03 (Age by Sex: All Persons). Age of respondent is classified into three groups: 18-29, 30-44 and 45-54.

The variable 'ager' is used to determine the imputed age of the respondent. The eight cases of a 55-year-old respondent by 'ager' are put into the 45-54 group (id=133, 297, 467, 974, 1330, 1362,1721, 2149). Another respondent (id=867), for whom the value of 'ager' is missing because they gave the year but not month of birth, is also be placed into the appropriate age group. Table 2 displays the codebook entry for a new variable 'agegpr' representing the grouped ages of the respondents.

Table 2: Codebook 'agegpr', Wave 1

```
agegpr ------------------------------------------- age group of respondent
                  type:  numeric (float)
                 label:  agegpr

                 range:  [1,3]                        units:  1
          unique values:  3                   coded missing:  0 / 2231

             tabulation:  Freq.   Numeric  Label
                            589         1  18-29
                           1127         2  30-44
                            515         3  45-54
```

The weights derived by benchmarking the weighted sample are then combined with the truncated form of 'hhwt' to get an overall set of weights for persons, called 'pswt'. This variable is intended to be used in place of 'hhwt', since it includes the adjustment for sampling design as well as the benchmarking adjustments. The variable 'pswt' is tabulated as follows:

Table 3: Values of 'pswt' tabulated by 'agegpr', q22 (sex) and 'numhh' (truncated), Wave 1

		number eligible in houshold			
agegpr	q22	1	2	3	4 or more
18-29	female	0.5771	1.1541	1.7312	2.3083
	male	0.6740	1.3480	2.0220	2.6960
30-44	female	0.4394	0.8789	1.3183	1.7577
	male	0.5550	1.1100	1.6651	2.2201
45-54	female	0.4782	0.9565	1.4347	1.9130
	male	0.6555	1.3109	1.9664	2.6218

It can be seen that the most underrepresented persons in the NLC sample are males, particularly those in the lowest and highest age ranges, and the youngest age group of both sexes.

Benchmarked Weights for Income Units (Wave 1)

The previously described weights are not suitable to use in analysis of income units in the NLC file. Both 'hhwt' and 'pswt' are at best weighting schemes for *persons*, not *income units*.

Most of the issues concerning income measurement in the NLC data have already been canvassed by Robert Ackland (NLC Workshop paper, April 2002) and Trevor Breusch and Deborah Mitchell (NLC Workshop paper, April 2002), so only a brief summary will be given here. The main conceptual problem is 'whose income?', since income sharing takes place within a household to varying degrees. Official statistical agencies such as the Australian Bureau of Statistics generally distinguish a 'household' (basically a group of people in a single dwelling who share living arrangements such as meals) from a 'family' (two or more people in a single household who are related as marriage partners or by the parent-child relationship) from an 'income unit' (a family group consisting of a one or two adults and their dependent children, if any, or a single individual not in a family relationship).

Many researchers of family incomes focus on the income unit, usually on the rationale that income sharing is high within this unit and it typically drops off sharply outside the unit. The ABS regular publication *Income Distribution*, for instance, takes this approach. This relatively narrow unit of observation has the advantage of reducing the information that is needed about other members of the wider family group or the household. Researchers using the NLC data have little choice in the matter, since that survey does not record the economic circumstances of any household members other than the respondent and his or her partner.

The treatment of income units requires clarity on several issues: the dependence status of children, the partnered status of the respondent, and identification of the 'reference person' by whom the income unit is classified.

Dependent Children

The NLC data do not include a measure of the actual economic dependence of children. This concept is difficult to measure anyway, and official statistical agencies usually adopt some simple criteria. For instance, ABS considers children in the age range 16-24 to be independent unless they are living with one or both parents and in full time education, and children 25 and over to be independent regardless of other circumstances. There is a belief abroad (fostered in part by ABS public consultations) that this definition does not reflect the contemporary mix of

part-time work and study arrangements of young adults, nor what is reported about the economic dependency of children beyond the nominal ages of adulthood. In any case this definition is not available to users of the NLC data, since there is no information about the employment, education or income status of adult children in the household unless they are the respondent.

The procedure adopted here is to assume that children who are still living with their parent(s) are economically independent once they reach the age of 18 years. While this definition is certainly very simple, it is not necessarily inferior to the ABS definition. It would be a better reflection of the realities than the ABS definition in the examples of the 16- or 17-year-old who has left school and has no job, or of the 20-year-old student who is enrolled in full-time study and is largely self-supporting. Our simple definition has the particular advantage that it can be implemented in the NLC data.

As a practical matter the determination of the number of children under the age of 18 years associated with each respondent was carried out using a modified version of Robert Ackland's code in his Stata command file 'hhdemog.do'. Ackland earlier documented this procedure to determine the numbers of children under 15 years; all that is changed here is the cut-off age. These new variables have names 'chlt18' (for 'children less than 18 years of age') and 'chge18' (for 'children aged 18 and over'). Ackland's two variables based on age 15 have also been renamed so they are more mnemonic: 'ch_n' becomes 'chlt15', while 'adch_n' becomes 'chge15'.

Table 4: Codebook 'chlt18' and 'chge18', Wave 1

```
chlt18 ------------------------------------------------- children under 18
                  type:  numeric (byte)

                 range:  [0,7]                        units:  1
         unique values:  8                    coded missing:  0 / 2231

            tabulation:  Freq.  Value
                         1091  0
                          372  1
                          483  2
                          210  3
                           57  4
                           11  5
                            4  6
                            3  7
```

```
chge18 ---------------------------------------- adult children 18 and over
                  type:  numeric (byte)

                 range:  [0,4]                        units:  1
         unique values:  5                    coded missing:  0 / 2231

            tabulation:  Freq.  Value
                         2041  0
                          139  1
                           44  2
                            5  3
                            2  4
```

Partnered Status

The indication of whether or not the respondent has a (marriage) partner also need to be clear. Most NLC researchers have used q20 (=1 or 2 for not partnered; =3 or 4 for partnered). This also is the basis for the variable 'havepart' constructed by Ackland. However this indicator produces an anomaly where the respondent says he or she is partnered but they omit the partner from the household roster, or vice versa. If the partner does not appear in the household roster and the questions on partner's sex and birth year are both skipped the person is recorded here as not partnered; this affects six cases (id=989, 2202, 2238, 2295, 2473, 2534). Conversely the person (id=2525) who said he was not in a relationship but gave his partner's sex and included her in the household roster seems to be partnered despite his answer 'not presently in a relationship' to q22 (and his answer 'don't know' to the partner's birth year).

The actual procedure used here is to record the respondent as partnered if they included a partner in the household roster, but override this indicator if the answers given to questions q20 and q24 clearly contradict the roster. If they said they had a partner at q20 *and* gave the sex of the partner at q24 they are recorded as having a partner; if they said they had no partner at q20 *and* skipped on partner's sex at q24 they are recorded as not having a partner. This measure of partnered status is the variable 'havesp'.

Table 5: Codebook 'havesp', Wave 1

```
havesp ------------------------------------------------------ have partner
                  type:  numeric (byte)
                 label:  havesp

                 range:  [0,1]                        units:  1
         unique values:  2                    coded missing:  0 / 2231

            tabulation:  Freq.   Numeric  Label
                           826         0  not partnered
                          1405         1  partnered
                          1406
```

Table 6: Tabulation of q20 by 'havesp', Wave 1

partnered status	have partner: not partn	partnered	Total
not presently in a re	629	1	630
in a relationship but	191	0	191
living with someone b	1	178	179
married and living wi	5	1226	1231
Total	826	1405	2231

Income Units

Given the information on dependent children and partnered status, it is easy to classify the responses into four kinds of income units.

Table 7: Codebook 'iutype', Wave 1

```
iutype -------------------------------- income unit type (child dep to 18)
                  type:  numeric (byte)
                 label:  iutype

                 range:  [1,4]                        units:  1
         unique values:  4                    coded missing:  0 / 2231

            tabulation:  Freq.   Numeric  Label
                           959         1  couple with dependents
                           446         2  couple only
                           181         3  sole parent (with deps)
                           645         4  single person
```

Table 8: Tabulation of 'hhtype' by 'iutype', Wave 1

	income unit type (child dep to 18)				
household type	couple wi	couple on	sole pare	single pe	Total
lone	0	1	0	265	266
group	0	0	0	86	86
couple no kids	2	358	0	3	363
sole with kids	0	0	181	26	207
couple with kids	957	86	0	0	1043
kid of sole	0	0	0	64	64
kid of couple	0	0	0	171	171
other fam	0	1	0	30	31
Total	959	446	181	645	2231

Reference Person

In ABS income data the income unit has the personal characteristics not of the respondent as in the NLC, but rather of the 'reference person'. This concept aligns with the respondent in the case of single-adult income units, but in the case of two-adult income units the male partner (if there is one) is always regarded as the 'reference person'. In particular, income units are classified by the age of the reference person. The new variable 'agerp' contains this information for the NLC. It contains the respondent's age in the case of single-adult income units (iutype=3 and =4) and when there is a female partner in two-adult income units (iutype=1 and =2), but it contains the partner's age if the respondent is female and the partner is male. *Don't blame me for this - I am only reporting what is so.*

The change of viewpoint from respondent to reference person creates some complications because, while the NLC respondents were selected in the age range 18-54, the ages of partners go outside this range. The age of reference person is used mostly when it is grouped (and that is how we use it below for the construction of sampling weights). Two cases of 55-year-old single respondents are classified into the 45-54 group (id=297 and 467).

Table 9: Codebook 'agegp', Wave 1

```
agegp ------------------------------------------ age group of ref person
                  type:  numeric (float)
                 label:  agegp

                 range:  [1,4]                        units:  1
         unique values:  4                    coded missing:  0 / 2231

            tabulation:  Freq.   Numeric  Label
                           560         1  18-29
                          1071         2  30-44
                           522         3  45-54
                            78         4  over 55
```

Table 10: Tabulation 'agegp' by 'iutype', Wave 1

age group of ref person	income unit type (child dep to 18) couple wi	couple on	sole pare	single pe	Total
18-29	89	108	27	336	560
30-44	641	127	108	195	1071
45-54	211	151	46	114	522
over 55	18	60	0	0	78
Total	959	446	181	645	2231

Benchmarking for Income Units

Stratification is done here on three dimensions: the income unit type, the age group of the reference person, and the sex of the reference person in one-adult income units.

The external estimates of cell proportions are obtained from the confidentialised unit record file (CURF) of *SIHC 1996-97*. Care is taken when calculating these estimates to ensure that the appropriate weight supplied in the CURF is used. This approach ensures that the cell proportions that are calculated will match the externally referenced benchmarks that were used in constructing the CURF. The weights so calculated are given in a new variable called 'iuwt'.

Table 11: Values of 'iuwt', tabulated by 'iutype', q22 (sex) and 'agegp', Wave 1

iutype	q22	agegp 18-29	30-44	45-54	over 55
couple with depende		0.7295	0.6402	0.8330	1.1685
couple only		0.6421	0.7838	1.0110	1.2496
sole parent (with d	female	1.4701	0.8572	0.4173	
	male	2.4361	0.7779	0.7392	
single person	female	1.6607	1.3657	1.1434	
	male	1.9664	1.4218	1.0265	

The income units most underrepresented in the NLC are single people, particularly males at the younger ages, with single males in the 18-29 age group being only half as common in the NLC sample as they are in the population. (The young male sole parent group has only one member in the sample, although as indicated in the table the expected number is 2.4.) The most overrepresented groups are female single parents 45-54, who have more than twice the proportion in the sample as in the population. Couple income units are generally overrepresented in the NLC sample. The age effect is not strong for couples, although younger couples with no dependents are more heavily overrepresented than older ones, reversing the effect of age that was seen with singles. Couple only units in the oldest age range are slightly underrepresented relative to the population, but not as severely as some other groups. None of these effects is surprising when consideration is given to the twin effects of sample design and non-response due to lifestyle (particularly absence from the home). The new weight variable for income units 'iuwt' is quite different from the existing weight variable for persons 'hhwt'. The simple correlation is only –0.026.

Benchmarking Weights for Persons (Wave 2)

The principle that underlies the weights devised here for the Wave 2 sample is that the set of persons should be made representative of the population at the time of the initial sampling in Wave 1. The benchmarking method is the same as Wave 1, but with the sample restricted to the 1768 respondents who were present in Wave 2. The weighting adjustment allows for under- or overrepresentation of groups in the original sampling at Wave 1, as well as for different rates of attrition between Waves 1 and 2 for the different groups.

As in Wave 1, benchmarking for persons is done after the sample has been weighted to correct for the effect of the sample design. Again the variable 'hhwt' is used to adjust for this aspect the sample design, with the modification that the largest group in a household is 'four or more' eligible persons. *The age group of the respondent and the household size are the values from the Wave 1 data file*. The numbers of persons in the weighting categories and the resulting weights, called 'pswt2' are as follows:

Table 12: Tabulation of Wave 1 'agegpr' by q22 (sex) and 'numhh' (truncated), Wave 2 sample

		number eligible in houshold				
agegpr	q22	1	2	3	4 or more	
18-29	female	63	115	37	14	229
	male	35	96	43	12	186
30-44	female	165	332	26	6	529
	male	119	257	11	4	391
45-54	female	111	86	38	20	255
	male	66	79	21	12	178
	Total	559	965	176	68	1768

Table 13: Values of 'pswt2', tabulated by Wave 1 'agegpr', q22 (sex) and 'numhh' (truncated), Wave 2 sample

		number eligible in houshold			
agegpr	q22	1	2	3	4 or more
18-29	female	0.6382	1.2765	1.9147	2.5530
	male	0.7311	1.4622	2.1933	2.9244
30-44	female	0.4152	0.8304	1.2456	1.6607
	male	0.5533	1.1067	1.6600	2.2134
45-54	female	0.4314	0.8629	1.2943	1.7257
	male	0.6249	1.2497	1.8746	2.4995

Again we see the effects of the Wave 1 underrepresentation of males, particularly those in the lowest and highest age ranges, and underrepresentation of the youngest age group of both sexes. Comparing these weights to the Wave 1 weights, we see that proportionally more of the youngest age group have been lost in the attrition from Wave 1 to Wave 2 and proportionally more of the oldest age group have been retained. Except in the youngest age group, proportionally more males have been lost between the waves than females, although in Wave 2 it is still the youngest age group males who need the heaviest weight to be representative of the population.

Benchmarked Weights for Income Units (Wave 2)

Again the idea is that the set of income units should be made representative of the population at the time of the initial sampling in Wave 1. Here is a cross-tabulation of the age group of the reference person and the income unit type *in the Wave 1 data file* for the 1768 responses in the Wave 2 sample and the new weight variable, called 'iuwt2'.

Table 14: Tabulation of Wave 1 'agegp' by 'iutype', Wave 2 sample

age group of ref person	income unit type (child dep to 18) couple wi	couple on	sole pare	single pe	Total
18-29	71	78	17	233	399
30-44	525	96	89	148	858
45-54	186	127	40	91	444
over 55	16	51	0	0	67
Total	798	352	146	472	1768

Table 15: Values of 'iuwt2' tabulated by Wave 1 'iutype', q20 (sex) and 'agegp', Wave 2 sample

		agegp			
iutype	q22	18-29	30-44	45-54	over 55
couple with depende		0.7247	0.6194	0.7489	1.0417
couple only		0.7045	0.8217	0.9526	1.1650
sole parent (with d	female	1.8932	0.7996	0.3779	
	male	1.9305	0.9247	0.7030	
single person	female	1.8285	1.4006	1.0747	
	male	2.3245	1.4984	1.0677	

It can be seen that young single person income units are even more underrepresented in the Wave 2 sample than in the first wave. Older couples, particularly those with dependents, and older single parents, have been particularly well retained in the sample between the waves. These groups were mostly over-represented in the first wave, so now they require even smaller weights for the Wave 2 sample to be representative of the population.

Appendix 1: Stata Code for Waves 1 and 2

Note that this code is written to do the calculations for income units first and for persons second.

```
capture log close
log using "c:\statawrk\NLC_weights\weights1.log", replace

*********************************************************
* Name:     weights1.do
* Purpose:  Forms weights for Income Units and Persons in NLC, Wave 1
* Version:  v3.0 with simplified age ranges.
* Modified: June 2003
* Author:   Trevor Breusch
* Infile:   NLC main file d1015.dta
*           SIHC CURF ids96.dta
*           R Ackland's (modified) hhdemog2.dta
*
* Outfile:  weights1.dta
* Stata ver: 7
*********************************************************;

      # delimit ;
      set more off;

      version 7;
      cd "c:\statawrk\NLC_weights";

***********************************************;
** Open the CURF and calculate some proportions;
***********************************************;

      use "c:\statawrk\NLC_Income_Mar03\ids96.dta", clear;

* First reduce the file to Income Units as the observations and
* delete the variables that refer to persons.
*
* The idea is to generate a unique code number for each Income Unit
* and drop all observations for the second and later appearances of
* the same number.

* Note the use of base-6 arithmetic (although base-5 would have
* worked too). There are at most 5 Families in a Household, and at
* most 5 Income Units in a Family.
*
* Note that the label on 'iunou' is wrong: it counts Income Units in
* the Family, not in the Household as it says.;

      gen incunit =hhidu*36+famnou*6+iunou;
      drop if incunit[_n]==incunit[_n-1];
      drop recper-wtpsn;
      drop incunit;

* Recode the agegroups for better comparison with NLC.;
* The method distinguishes income units that have a female respondent
* (hence she is 18-54) and a male partner who is older than 54.;

    gen agegpru=.;
replace agegpru=0 if ageru>=1&ageru<=3  ;  * 15-17 ;
replace agegpru=1 if ageru>=4&ageru<=11 ;  * 18-29 ;
replace agegpru=2 if ageru>=12&ageru<=14;  * 30-44 ;
replace agegpru=3 if ageru>=15&ageru<=16 ; * 45-54 ;
replace agegpru=4 if ageru>=17&ageru<=29&agepu>=4&agepu<=16; * rp>=55 but
part 18-54;
```

```
replace agegpru=5 if ageru>=17&ageru<=29&agepu>0&(agepu<=3|agepu>=17);  *
rp>=55 and part out range;

* Find the proportions in each type of income unit;

tab agegpru iutype [aw=wtunit] if agegpru>=1&agegpru<=4, matcell(C1) row;

* Find the proportions for gender within the singles groups;

tab agegpru sexru [aw=wtunit] if iutype==3&(agegpru>=1&agegpru<=4),
matcell(C3) row;
tab agegpru sexru [aw=wtunit] if iutype==4&(agegpru>=1&agegpru<=4),
matcell(C4) row;

mat list C1; mat list C3; mat list C4;

************************************************;
** Get some variables from hhdemog.dta (modified);
************************************************;

use "..\nlc\wave1\hhdemog2.dta", clear;

label var chlt18 "children under 18";
label var chge18 "adult children 18 and over";
keep id chlt18 chge18;

sort id;
compress;
save temp.dta, replace;

************************************************;
** Now load the NLC Wave 1 data and code some variables;
************************************************;

use "..\nlc\wave1\d1015.dta", clear;

** Merge the data from hhdemog.do;

sort id;
merge id using temp.dta;
tab _merge;
drop _merge;

** New hhwt to replace faulty rounded one in nlc97;

egen sumn=sum(numhh);
replace hhwt=(numhh/sumn)*_N;

** New hhwtm, restricting to maximum weight of 4 persons;

    gen numhh2=numhh if numhh<=4;
replace numhh2=4 if numhh>4;
egen sumn2=sum(numhh2);
gen hhwtm=(numhh2/sumn2)*_N;

label var hhwtm "hhwt with hhold size max of 4";

codebook hhwt hhwtm;

** Partnered;

gen havesp=0;
for num 2/9: replace havesp=1 if q282aX==2;
```

```
* Need to modify this variable for some peculiar cases. Rule is partnered
* if say partnered and partner's sex stated;

replace havesp=0 if (q20==1|q20==2)&q24==-9; *partner sex skipped;
replace havesp=1 if (q20==3|q20==4)&(q24==1|q24==2); *partner sex stated;

label var havesp "have partner";
label define havesp 0 "not partnered" 1 "partnered";
label values havesp havesp;

*******************************************************;
** Income Unit type. Same coding as SIHC CURF;
*******************************************************;

gen iutype=.;

replace iutype=1 if havesp==1&chlt18>0;  *couple with dep kids;
replace iutype=2 if havesp==1&chlt18==0; *couple no dep kids;
replace iutype=3 if havesp==0&chlt18>0;  *sole parent dep kids;
replace iutype=4 if havesp==0&chlt18==0; *single;

label var iutype "income unit type (child dep to 18)";

label define iutype
1 "couple with dependents"
2 "couple only"
3 "sole parent (with deps)"
4 "single person" ;

label values iutype iutype;

*******************************************************:
** Age and sex of the reference person (male if couple IU);
*******************************************************;

    gen agerp=ager;
replace agerp=agesp if (iutype==1|iutype==2)&q22==2&q24==1;
label var agerp "age of reference person";

    gen agegp=.;
replace agegp=0 if agerp<18;
replace agegp=1 if agerp>=18&agerp<=29;
replace agegp=2 if agerp>=30&agerp<=44;
replace agegp=3 if agerp>=45&agerp<=54;
replace agegp=4 if agerp>=55;

* Two 55 year old respondents will be reclassified as 54 yo to
* simplify the groupings;
replace agegp=3 if id==297|id==467;

label define agegp
0 "under 18"
1 "18-29"
2 "30-44"
3 "45-54"
4 "over 55";

label values agegp agegp;
label var agegp "age group of ref person";

* Proportions in each type of income unit;

tab agegp iutype, matcell(N1) row;
mat list N1;
```

```
*********************************************************;
* Weights for the Income Units;
*********************************************************;

* Sample frequencies are zero if agegp=4 and iutype=3 or 4.
* These require special treatment;

   gen iuwt=.;
   mat R1=(1,1,1,1\1,1,1,1\1,1,1,1\1,1,1,1);

sca sumC1=0; sca sumN1=0;

forval i=1/4 {; forval j=1/4 {;
   sca sumC1=sumC1+C1[`i',`j'];
   sca sumN1=sumN1+N1[`i',`j'];
}; };

forval i=1/3 {; forval j=1/4 {;
   mat R1[`i',`j']=(C1[`i',`j']/sumC1)/(N1[`i',`j']/sumN1);
   replace iuwt=R1[`i',`j'] if agegp==`i'&iutype==`j';
}; };

* Now to handle the last row which has the zeros;

mat R1[4,1]=(C1[4,1]/sumC1)/(N1[4,1]/sumN1);
mat R1[4,2]=(C1[4,2]/sumC1)/(N1[4,2]/sumN1);

forval i=1/4 {;
   replace iuwt=R1[4,`i'] if agegp==4&iutype==`i';
};

* Modify these weights for the gender ratio in each agegroup
* for iutype=3 and 4. These iutypes have only 3 age groups;

tab agegp q22 if iutype==3, matcell(N3) row;
tab agegp q22 if iutype==4, matcell(N4) row;
mat list N3;
mat list N4;

* For iutype=3, weight each cell separately (although cells
* have very small counts);

mat R3=(1,1\1,1\1,1);

forval i=1/3 {; forval j=1/2 {;
   mat R3[`i',`j']=(C3[`i',`j']/(C3[`i',1]+C3[`i',2]))/
   (N3[`i',`j']/(N3[`i',1]+N3[`i',2]));
   replace iuwt=iuwt*R3[`i',`j'] if agegp==`i'&q22==`j'&iutype==3;
}; };

* For iutype=4, weight each cell sepatately (although cells
* have very small counts);

mat R4=(1,1\1,1\1,1);

forval i=1/3 {; forval j=1/2 {;
   mat R4[`i',`j']=(C4[`i',`j']/(C4[`i',1]+C4[`i',2]))/
   (N4[`i',`j']/(N4[`i',1]+N4[`i',2]));
   replace iuwt=iuwt*R4[`i',`j'] if agegp==`i'&q22==`j'&iutype==4;
}; };

* Check there is no rounding error. IU weights should average=1.0000;

summ iuwt;
```

```
label var iuwt "weight for income units W1";

*************************************************;
** Age groups for Individuals;
*************************************************;

** Age group of respondent;

* One person has ager missing (assume birthday before interview
* date of 03-NOV-96);
replace ager=1997-1950 if id==867;

    gen agegpr=.;
replace agegpr=1 if ager>=18&ager<=29;
replace agegpr=2 if ager>=30&ager<=44;
replace agegpr=3 if ager>=45&ager<=54;
replace agegpr=4 if ager>=55&ager~=.;

label define agegpr
1 "18-29"
2 "30-44"
3 "45-54"
4 "over 55";

label values agegpr agegpr;
label var agegpr "age group of respondent";

* Eight 55 year old single respondents are reclassified as 54 yo to
* make the weight adjustments come out;

      replace agegpr=3 if ager==55;

*****************************************************;
** Weights for Persons, weighting first by hhwtm (modified form
*  maximum of 4 persons/household);
*****************************************************;

tab agegpr q22 [aw=hhwtm] if agegpr>=1&agegpr<=3, matcell(N5) row;
mat list N5;

* Population values from Census 1996
* Rows `i' are agegpr=1/3, cols `j' are Q22=1/2;

mat C96=(   1591993, 1582414\
            2033996, 2083361\
            1128244, 1109180);

sca sumC96=0; sca sumN5=0;

forval i=1/3 {; forval j=1/2 {;
   sca sumC96=sumC96+C96[`i',`j'];
   sca sumN5=sumN5+N5[`i',`j'];
}; };

mat R5=(1,1\1,1\1,1);
gen pswt=.;

forval i=1/3 {; forval j=1/2 {;
   mat R5[`i',`j']=(C96[`i',`j']/sumC96)/
      (N5[`i',`j']/sumN5);
   replace pswt=hhwtm*R5[`i',`j'] if agegpr==`i'&q22==`j';
}; };

* Check there is no rounding error. Person weights should average=1.0000;
```

```
summ pswt;

label var pswt "weight for persons W1";

*****************************************************;
** Finish up Wave 1 and save;
*****************************************************;

keep id q22 hhtype chlt18 chge18 havesp iutype agegp agegpr hhwt hhwtm iuwt
pswt;

codebook chlt18 chge18 havesp iutype agegp agegpr;
tab2 hhtype iutype;
tab2 agegp iutype;
cor hhwt hhwtm iuwt pswt;

by iutype, s: tab iuwt agegp if q22==1;
by iutype, s: tab iuwt agegp if q22==2;
by q22, s: tab pswt agegpr;

sort id;
compress;
desc;

save temp1.dta, replace;

*****************************************************;
* Attrition for Wave 2;
* The idea here is to weight so that the set of respondents
* remaining in Wave 2 is a representative sample at the time
* of Wave 1;

*****************************************************;
* Find out who in Wave 1 was present in Wave 2;
*****************************************************;

use "c:\statawrk\NLC_Events_Jan03\Panel_Mar03", clear;
gen both=(ak__2~=.);
label var both "Present in Wave 2";
keep id both;

sort id;
save temp2.dta, replace;

*****************************************************;
* Now re-work the Wave 1 weights for the respondents who
* are present in Wave 2;
*****************************************************;

use temp1.dta, clear;
merge id using temp2.dta;
tab _merge;
drop _merge;

keep if both==1;
mat drop N1 N3 N4 N5 R1 R3 R5;
sca drop sumN1 sumN5 sumC96;
tab agegp iutype, matcell(N1) row;
mat list N1;

*****************************************************;
* Weights for the Income Units. Allows attrition to Wave2;
*****************************************************;
```

```
* Sample frequencies are zero if agegp=4 and iutype=3 or 4.
* These require special treatment;

   gen iuwt2=.;
   mat R1=(1,1,1,1\1,1,1,1\1,1,1,1\1,1,1,1);

sca sumN1=0;

forval i=1/4 {; forval j=1/4 {;
   sca sumN1=sumN1+N1[`i',`j'];
}; };

forval i=1/3 {; forval j=1/4 {;
   mat R1[`i',`j']=(C1[`i',`j']/sumC1)/(N1[`i',`j']/sumN1);
   replace iuwt2=R1[`i',`j'] if agegp==`i'&iutype==`j';
}; };

* Now to handle the last row which has the zeros;

mat R1[4,1]=(C1[4,1]/sumC1)/(N1[4,1]/sumN1);
mat R1[4,2]=(C1[4,2]/sumC1)/(N1[4,2]/sumN1);

forval i=1/4 {;
   replace iuwt2=R1[4,`i'] if agegp==4&iutype==`i';
};

* Modify these weights for the gender ratio in each agegroup
* for iutype=3 and 4. These iutypes have only 3 age groups;

tab agegp q22 if iutype==3, matcell(N3) row;
tab agegp q22 if iutype==4, matcell(N4) row;
mat list N3;
mat list N4;

* For iutype=3, weight each cell separately (although cells
* have very small counts);

mat R3=(1,1\1,1\1,1);

forval i=1/3 {; forval j=1/2 {;
   mat R3[`i',`j']=(C3[`i',`j']/(C3[`i',1]+C3[`i',2]))/
   (N3[`i',`j']/(N3[`i',1]+N3[`i',2]));
   replace iuwt2=iuwt2*R3[`i',`j'] if agegp==`i'&q22==`j'&iutype==3;
}; };

* For iutype=4, weight each cell sepatately (although cells
* have very small counts);

mat R4=(1,1\1,1\1,1);

forval i=1/3 {; forval j=1/2 {;
   mat R4[`i',`j']=(C4[`i',`j']/(C4[`i',1]+C4[`i',2]))/
   (N4[`i',`j']/(N4[`i',1]+N4[`i',2]));
   replace iuwt2=iuwt2*R4[`i',`j'] if agegp==`i'&q22==`j'&iutype==4;
}; };

* Check there is no rounding error. IU weights should average=1.0000;

summ iuwt2;
label var iuwt2 "weight for income units W2";

************************************************************;
* Weights for Persons, weighting first by hhwtm2 (modified form
* maximum of 4 persons/household). Allows attrition to Wave2;
```

```
*****************************************************************;

* First rebalance hhwtm to average to 1.0 on the reduced sample;

egen sumhhwtm=sum(hhwtm);
gen hhwtm2=(hhwtm/sumhhwtm)*_N;

tab agegpr q22 [aw=hhwtm2] if agegpr>=1&agegpr<=3, matcell(N5) row;
mat list N5;

* Population values from Census 1996
* Rows `i' are agegpr=1/3, cols `j' are Q22=1/2;

mat C96=(   1591993, 1582414\
            2033996, 2083361\
            1128244, 1109180);

sca sumC96=0; sca sumN5=0;

forval i=1/3 {; forval j=1/2 {;
   sca sumC96=sumC96+C96[`i',`j'];
   sca sumN5=sumN5+N5[`i',`j'];
}; };

mat R5=(1,1\1,1\1,1);
gen pswt2=.;

forval i=1/3 {; forval j=1/2 {;
   mat R5[`i',`j']=(C96[`i',`j']/sumC96)/
      (N5[`i',`j']/sumN5);
   replace pswt2=hhwtm2*R5[`i',`j'] if agegpr==`i'&q22==`j';
}; };

* Check there is no rounding error. Person weights should average=1.0000;

summ pswt2;
label var pswt2 "weight for persons W2";

*****************************************************************;
** Finish up;
*****************************************************************;

keep id iuwt2 pswt2;
sort id;
save temp3.dta, replace;

use temp1.dta, clear;
merge id using temp3.dta;
tab _merge;
drop _merge;
sort id;

cor hhwt hhwtm iuwt pswt iuwt2 pswt2;

by iutype, s: tab iuwt2 agegp if q22==1;
by iutype, s: tab iuwt2 agegp if q22==2;
by q22, s: tab pswt2 agegpr;

sort id;
compress;
desc;
save weights2.dta, replace;
log close;
```