The Household, Income and Labour Dynamics in Australia (HILDA) Survey

Mark Wooden
Project Director, HILDA Survey

MELBOURNE INSTITUTE
of Applied Economic and Social Research

Income, Wealth and Expenditure Measures in HILDA

Roger Wilkins

THE UNIVERSITY OF MELBOURNE
FACULTY OF BUSINESS & ECONOMICS

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Introduction

- HILDA is the best micro data available on the finances / material circumstances of Australian households
  - Every year:
    - Annual income of individuals and households, disaggregated by component
    - Current usual labour market earnings and current government income support payments
    - Household expenditure on a wide range of consumption goods and services (since Wave 5)
    - Primary home: value, debt, mortgage repayments
  - Every four years:
    - Most assets and debts of households, disaggregated by component
  - Plus, variety of subjective data collected (e.g., how well off, satisfaction with finances, experience of financial stress)

Income

- HILDA only attempts to produce a comprehensive measure of **annual** income
- Essential elements of the income 'model' (in sequence):
  1. Income components: As reported by individuals (with imputations where necessary), plus estimation of some government benefits
  2. Gross personal income = Sum of gross (before-tax) income components
  3. Personal income taxes: Estimated based on tax rules and other information
  4. Disposable personal income = Gross personal income - Estimated personal income taxes
  5. Household income variables: Simply summations of personal income variables across all household members

Components of personal gross income

- Private income
  - Wages and salary (3 components)
  - Business income (1 component)
  - Investment income (5 components)
  - Private pensions (2 components)
  - Private transfers (2 components)
- Public transfers
  - Income support payments (18 components)
  - Non-income support payments (5 components)
  - Other (scholarships, foreign pensions, etc.) (2 components)

Also: Windfall income (8 components) [not part of gross income]

Gross income imputation and estimation

- Impute individual components for non-respondents within partially-responding households and where missing for respondents
- 12 distinct components are imputed
- Imputation method: Combination of 'Population Carryover', 'Nearest Neighbour Regression' and 'Little and Su' methods. See HILDA Technical Paper 2/09.
- 2 distinct components estimated/calculated: Family Tax Benefit and Baby Bonus. Calculations are based on eligibility criteria and payment rates.
- All 'private' income components are top-coded (and hence, so are gross and disposable income).

Disposable income

Income taxes are estimated as follows (see Technical Paper 1/09):

1. Components of income treated as taxable by the ATO are summed together
2. Obtain taxable income by subtracting estimated tax deductions, based on Taxation Statistics.
3. Apply the formulas for income tax and the Medicare levy to produce initial tax liability.
4. Estimate applicable tax offsets (Low Income, Senior Australians, Pensioner, etc.) and subtract from initial tax liability to obtain total income tax.
5. Retired persons over 65: apply actual average tax rates as indicated by Taxation Statistics.

Household income

- All household income variables are simply the sum of the personal income variables across all household members
- At most disaggregated level,14 annual components produced (with imputations)
- Definition of household (important to all financial concepts in HILDA): “a group of people who usually reside and eat together”

Quality of HILDA income data

- Population inferences: adjust for non-response / attrition using weights supplied with the data **and** imputations
- Compared with ABS income survey data, average incomes tend to be higher in HILDA (but not clear that worse).

Equivalised household disposable income (07-08 prices)					
	Mean			Median	
	HILDA	ABS		HILDA	ABS
2000-01	33,164	29,876		28,855	26,331
2002-03	33,051	30,606		29,188	26,904
2003-04	34,169	33,265		30,246	29,303
2005-06	37,540	36,446		33,073	31,545
2007-08	41,244	42,286		36,157	36,081

Wealth

- Collected in 2002, 2006, 2010 (so far)
- Obtain values for 16 asset components and 14 debt components (each of which can be examined separately)
- These are aggregated to 22 components, for each of which missing values are imputed
- Most components are measured only at the household level
- Assets measured: primary home, other property, businesses, bank accounts, investments, superannuation, trust funds, life insurance, vehicles, collectibles
- Debts measured: home, other property, business, credit cards, HECS, car loans, hire purchase loans, investment loans, other personal loans, overdue bills (since Wave 6).

Quality of wealth data

- HILDA doesn't measure everything (excludes most consumer durables)
- Highly skewed wealth distribution => any sample is likely to underestimate total wealth of Australian households, even if we get it right for all households in the sample.
- Nonetheless, comparisons with aggregate external benchmarks from the RBA and National Accounts show HILDA performs well (Bloxham and Betts, *Australian Economic Review* 2009):
  - Under-estimate currency and deposits and 'largely exclude' unfunded super
  - But very similar for real estate, superannuation (excluding unfunded super), equities, other financial assets, and household debt

Consumption expenditure

- Most information collected in the self-completion questionnaire
- Only attempt to measure at **household** level (since Wave 6, only those with responsibility for household bills asked to respond)
- 29 expenditure components
- Attempt to measure approximately 75-80% of total household expenditure
- As of Release 8, impute missing values of each expenditure component (at household level)

Expenditure components since Wave 6

- Household Questionnaire: mortgage repayments, rent, child care expenditure
- Self-Completion Questionnaire:
  - **Usual weekly** expenditure on: groceries, alcohol, tobacco, public transport and taxis, meals eaten out
  - **Usual monthly** expenditure on: petrol, men's clothing, women's clothing, children's clothing, telephone and internet
  - **Actual annual** expenditure on: holidays, private health insurance, other insurance, medical practitioner fees, medicines, utilities, home repairs/renovations, motor vehicle repairs/maintenance, education fees, new motor vehicles, second-hand motor vehicles, computers, home AV equipment, household appliances, furniture

Quality of expenditure data

Mean expenditure: % difference of HILDA from ABS Household Expenditure Survey (SCQ items only)			
Groceries	-8.7 to 18.8	Health practitioner fees	-18.1 to -7.7
Alcohol	-10.2 to 2.2	Medicines	-36.4 to -23.6
Tobacco	-8.0 to 0.3	Electricity, gas	-20.0 to -12.3
Public transport, taxis	-1.7 to 6.6	Repairs, etc. to home	-51.0 to -46.3
Meals eaten out	-18.6 to -8.6	Vehicle repairs, etc.	-0.7 to 7.8
Petrol	15.3 to 25.1	Education fees	-29.7 to -19.6
Clothing and footwear	-16.9 to -12.8	Buying vehicles	9.0 to 19.2
Telephone, internet	2.2 to 2.4	Electronic devices	-7.8 to 1.6
Holidays	16.6 to 19.4	Household appliances	-32.4 to -31.0
Health insurance	-32.1 to -23.9	Furniture	-40.1 to -38.8
Other insurance	-45.9 to -21.0		

See HILDA Discussion Paper 1/10 for more information

Analysis using HILDA financial data

- Making data comparable across time and households: Price indices and equivalence scales
- Many advantages over cross-sectional data
  - Dynamics (of income, poverty, welfare reliance, earnings, consumption, wealth, etc.)
  - Longer-term pictures of economic wellbeing (eg, 'permanent' income)
  - Causal inferences
- Dealing with measurement error in a dynamic setting: reversion to mean creates particular problems for inferences on **distributional** dynamics.

More information

- **User Manual**
- **Also refer to various technical and discussion papers on HILDA web site which provide details on methods and issues relating to constructing and imputing financial information in HILDA.**
- **Annual Statistical Report provides some basic analysis of the data (indicative of nature/properties)**

Introductory Remarks

- Funded and owned by Australian Government
- Modelled on other household panels – BHPS, SOEP
- Survey manager = Melbourne Institute of Applied Economic & Social Research (University of Melbourne)
- Fieldwork subcontractor = Roy Morgan Research
- Unit record data available (under license)
- Want to know more?
  - Articles in *The Economic Record*, June 2007 and *Australian Economic Review*, September 2010
  - www.melbourneinstitute.com/hilda/

Longitudinal Surveys in Australia

- Cohort studies
  - Longitudinal Survey of Immigrants to Australia
  - Longitudinal Surveys of Australian Youth
  - Australian Longitudinal Study of Women's Health
  - Longitudinal Study of Australian Children
  - 45 and Up Study
- Rotating panel survey
  - ABS Labour Force Survey
- Indefinite life panel survey
  - Household, Income and Labour Dynamics in Australia Survey

Key Design Features (I)

- Indefinite life panel
- Conducted annually (since 2001)
- Area-based clustered / stratified sample of dwellings
  - Selected sample = 12,252 dwellings from 488 Census CDs
- Reference population = residents of **private** dwellings
  - Remote areas excluded
  - Sample members followed into institutions (where possible)
- F2F interviews, with self-completion supplement
  - Telephone is an option
  - Switched to CAPI in W9
- All household members aged 15+ interviewed

Following Rules

- W1 household members
  = Continuing Sample Members (CSMs)
- New sample members
  - Babies / born or adopted to CSM = CSM
  - Immigrants arriving since 2002 = CSM (since w9)
  - Other hh residents = Temporary Sample Members (TSMs)
- TSMs become CSMs if has child with CSM
- Follow CSMs and their HHs (and interview all adults)
- *We do NOT follow people O/S**

Selected Areas by Region

	Total	*Met*	*Ex-met*
New South Wales	164	99	65
Victoria	121	85	36
Queensland	90	40	50
South Australia	42	30	12
Western Australia	46	34	12
Tasmania	14		
Northern Territory	4		
ACT	7		
Total	**488**		

Household Definition

- **A group of people who usually reside together and eat together**
- One-person household
  - makes provisions for food or other essentials for living without combining with others
- Multi-person household
  - group of two or more persons, living within the same dwelling, who make common provision for food or other essentials

Key Design Features (II)

- Interview involves household + individual component
  - Mean interview time for a 2-adult hh ≤ 83 minutes
- Fieldwork period
  - Main = late August to mid December
  - Tail = February / March
- Cash incentives paid
  - W1-W4 = $20 or $50 per household
  - W5-W8 = $25 per person + $25 bonus for complete hh
  - W9-W12 = $30 per person + $30 bonus

Survey Instruments

- Household Form
  - Key identifiers / Changing HH membership / HH relationships / Reasons for non-response
- Household Questionnaire
  - Collects hh level data from relevant HH member
- Continuing Person Questionnaire
  - All persons 15+ who have previously been interviewed
- New Person Questionnaire
  - All persons 15+ who have never previously been interviewed
- Self-completion Questionnaire
  - All interview respondents; 16 pp, expanded to 20 from W5

What's HILDA About?

- Multi-purpose survey
- Three main core themes
  - **Family life and household composition**
  - **Income**
  - **Labour market activity and employment**
- But wait … there's more!
  - Wealth, Health and well-being, Social participation, etc.
- Policy relevance important

What's In It? HQ / CPQ

- Core:
  - Child care
  - Housing
  - Education
  - Employment status
  - Current employment
  - Persons not in employment
  - Calendar
  - Income
  - Family formation
  - Partnering & relationships
  - Living in Australia*
  - Tracking
  - Interview situation

- Special modules:
  - W1 = Personal history**
  - W2 = Wealth
  - W3 = Retirement
  - W4 = Youth issues; Private health insurance
  - W5 = Family formation
  - W6 = Wealth
  - W7 = Retirement; Lifestyle
  - W8 = Family formation / relationships
  - W9 = Health
  - W10 = Wealth
  - W11 = Family formation; Retirement

History (collected once only)

- Country of birth / Year of arrival
- Visa category / Residence status / Citizenship
- English language
- Aboriginality
- Schooling / Qualifications
- Employment history
- Marital history
- Family background
  - parents education
  - parents occupation
  - parental separation
  - siblings
  - age left home
- Year first developed disability

Module Rotation Plan

Special Module	*w6*	*w7*	*w8*	*w9*	*w10*	*w11*	*w12*	*w13*	*w14*	*w15*
Wealth	X				X				X	
Fertility / Family			X			X				X?
Health				X				X		
Retirement		X				X				X
Education / Human capital		(a)					X			

(a) Literacy / numeracy questions included

What's In It? SCQ

- Health and well-being (SF36, Kessler 10, serious health conditions)
- Health behaviours (smoking, drinking, exercise, weight, diet)
- Social capital / relationships (satisfaction with family, social support, community participation, religion)
- Neighbourhood characteristics
- Life events
- Time use
- Finances (stressful financial events, savings habits, risk preference, h'hold expend)
- Job attributes
- Parenting (parenting stress / work family gains and strains)
- Attitudes to work / gender roles / marriage
- Personality

Wave 1 Response

- HH response
  - In-scope sample = 11,693
  - 7682 cooperating households = 66% RR
- Individual response
  - W1 individual sample = 15,127 persons
  - 13,969 respondents = 92% RR
- Sample reasonably representative, but …
  - Sydney residents under-represented
  - People from a NESB under-represented
  - Males less likely to complete a PQ
    (but no less likely to be an CSM)

Retention

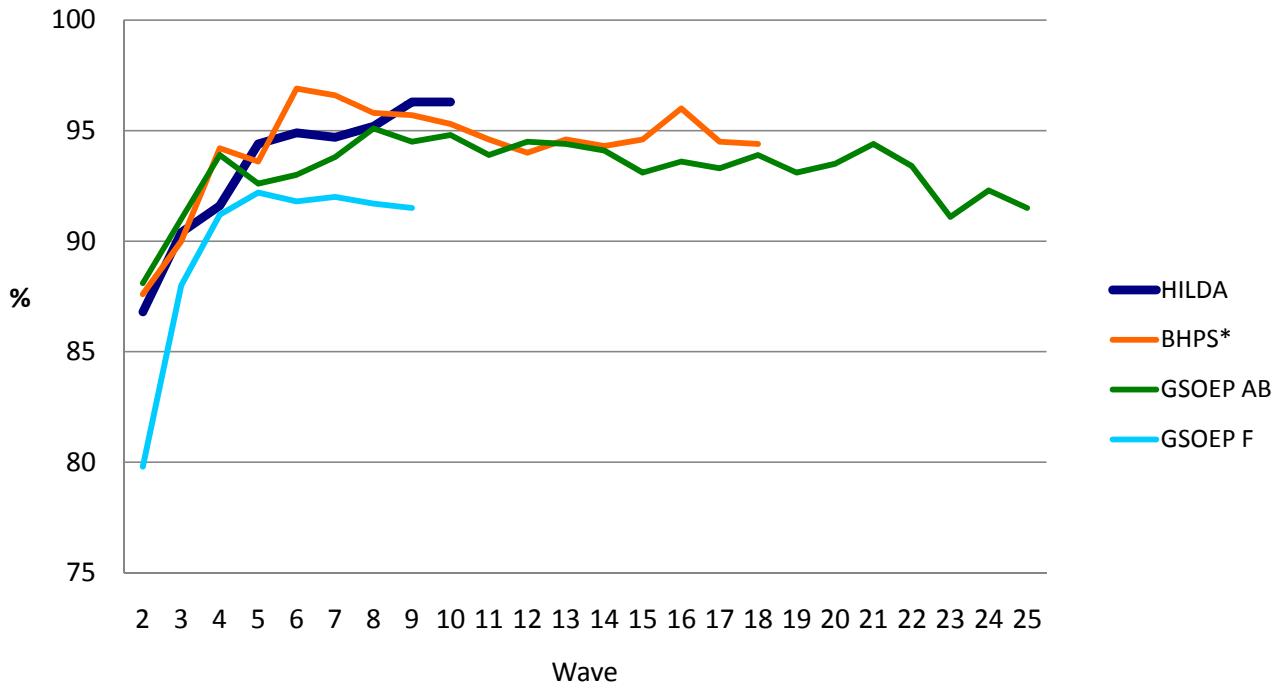
Wave	Wave-on-wave (%)	W1 respondents		Balanced panel (n)	Total sample (n)
		n	% of in-scope R		
W2	86.8	11993	86.8	11993	13041
W3	90.4	11190	82.0	10777	12728
W4	91.6	10565	78.5	9855	12408
W5	94.4	10392	77.9	9311	12759
W6	94.9	10085	76.3	8864	12905
W7	94.7	9628	73.6	8409	12789
W8	95.2	9354	72.0	8034	12785
W9	96.3	9245	71.5	7721	13301

Retention

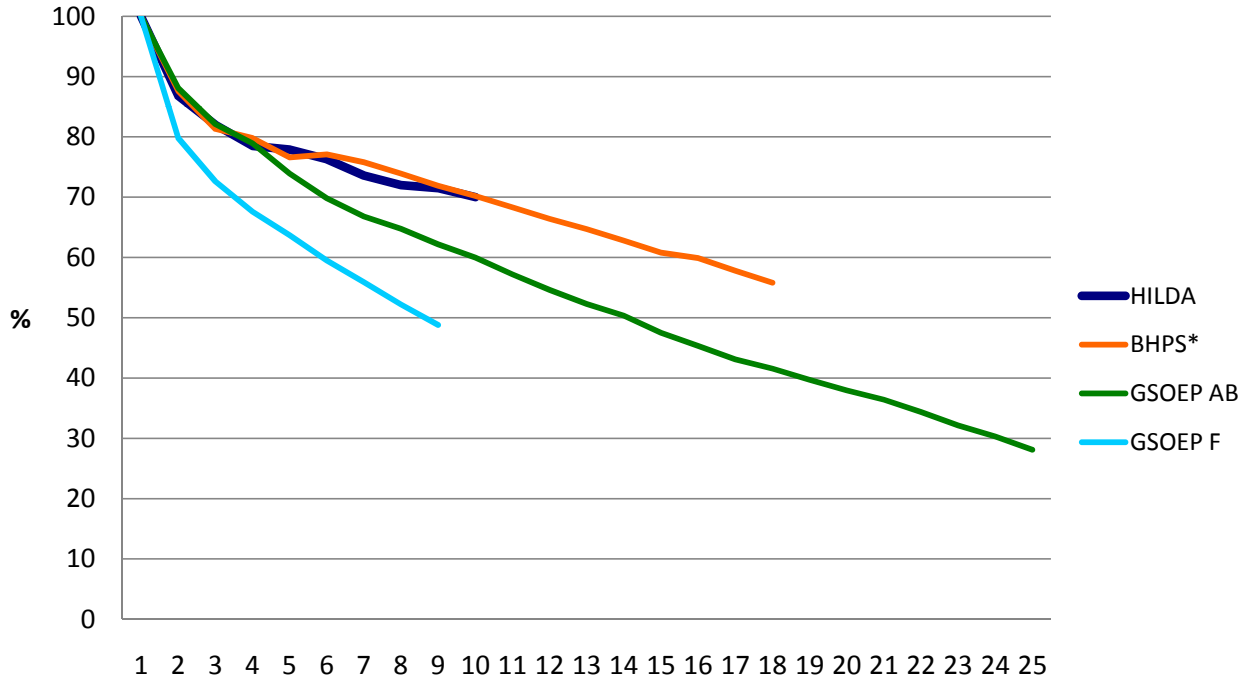
Wave	Wave-on-wave (%)	W1 respondents		Balanced panel (n)	Total sample (n)
		n	% of in-scope R		
W2	86.8	11993	86.8	11993	13041
W3	90.4	11190	82.0	10777	12728
W4	91.6	10565	78.5	9855	12408
W5	94.4	10392	77.9	9311	12759
W6	94.9	10085	76.3	8864	12905
W7	94.7	9628	73.6	8409	12789
W8	95.2	9354	72.0	8034	12785
W9	96.3	9245	71.5	7721	13301
W10	96.3	9002	70.0	7460	13526

Wave on Wave Response Rates

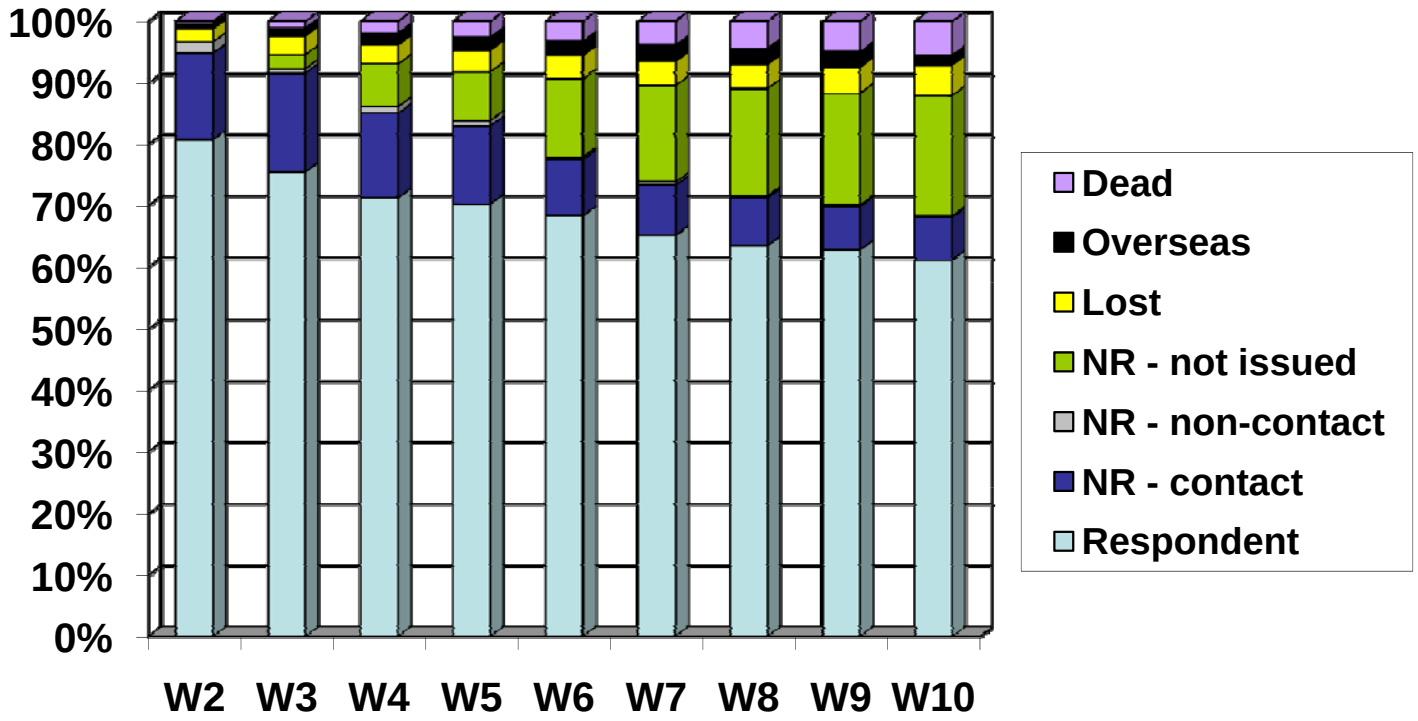
(HILDA, BHPS & GSOEP compared)



% of In-scope W1 Respondents Re-interviewed *(HILDA, BHPS & GSOEP compared)*



Fieldwork Outcomes: W1 Adults



Tracking Movers

22-23% of all hh's change address b/w each survey wave

- Pre-field office activity
  - Notifications (1800#, change of address card, email)
  - Matching to Australia Post
  - Returns to sender
  - Move indicator variable
- Other household members
- Contact information collected at previous ivw
- Neighbours
- Other community resources
- Online White Pages

Minimising Refusals

MARKETING / RESPONDENT ENGAGEMENT

- PAL and brochure, newsletter / Stat report!
- 1800 number

PERSISTENCE

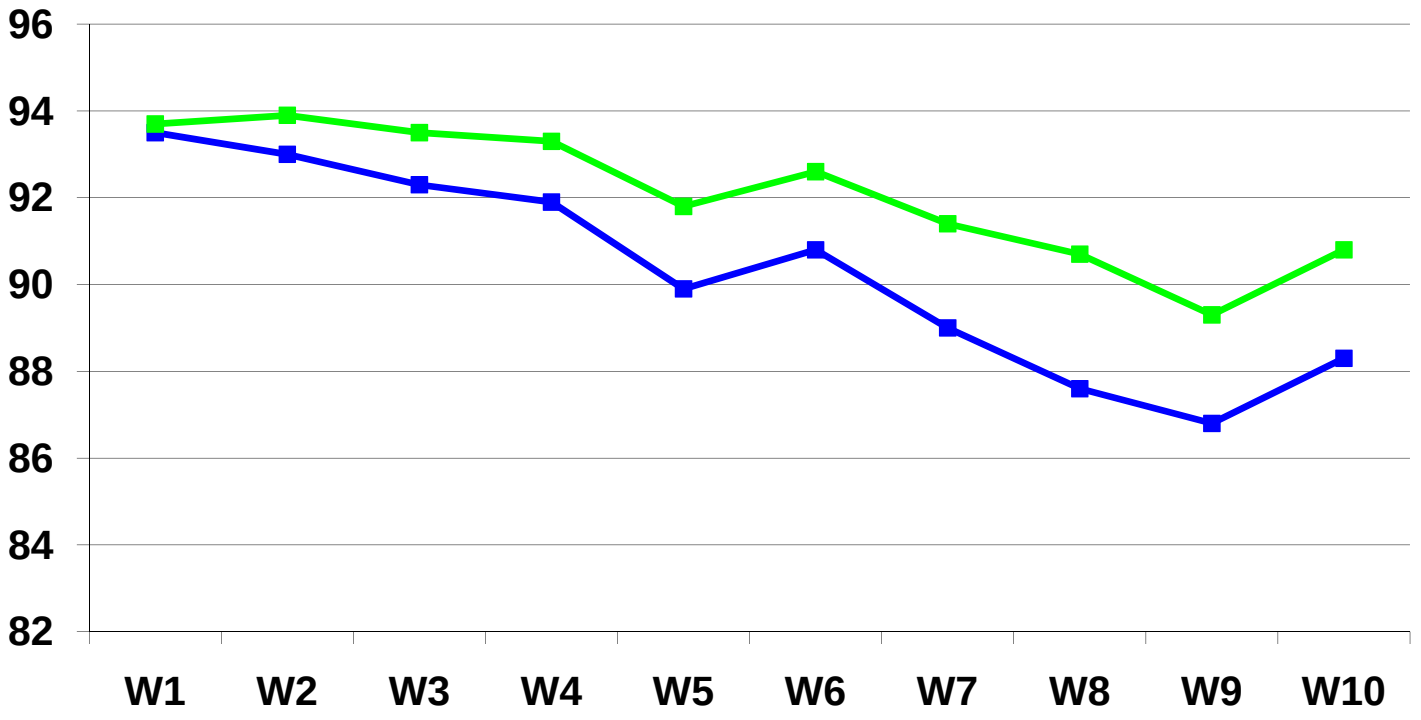
- 2-3 stage fieldwork

GOOD PEOPLE

- Selection and continuity of interviewers
- Training / interviewer engagement

INCENTIVES

SCQ response rate (% of interviewees)



Data Use

- Growing community of registered users
  - Almost 1500 licensed users
- Close to 300 journal articles using HILDA data have been published / accepted
- Many other working papers / conference papers / reports

http://www.melbourneinstitute.com/hilda/Biblio/default.html

Coming Up

- Sample expansion in W11
  - Target = 2000 new responding hh's from 125 CDs
- Matching to Death Register
- Human capital module in W12

HILDA User Training 2011:

Datafiles, conventions, documentation and derived variables

Simon Freidin
HILDA Database Manager

FACULTY OF
BUSINESS &
ECONOMICS

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Data files and documentation

Release 9 data package

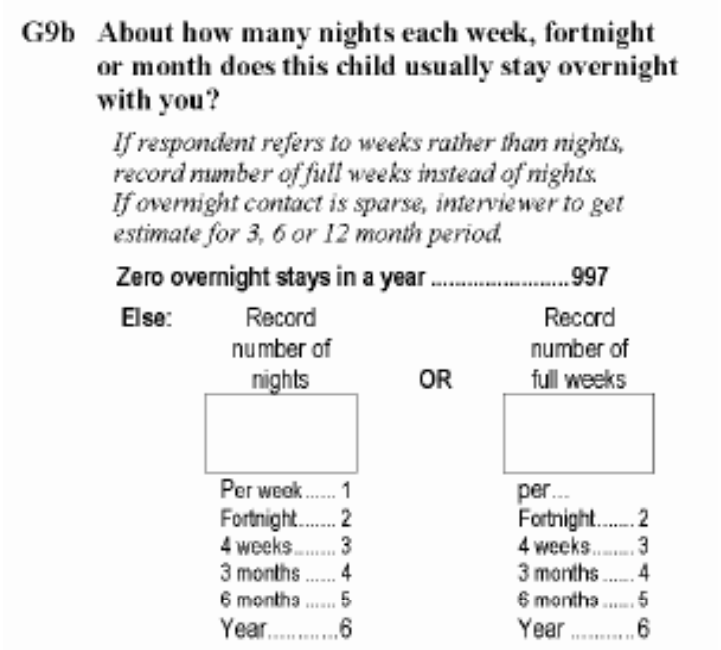
- Includes data from waves 1 to 9
- Marked up questionnaires, coding framework
- SAS, SPSS and Stata files
- One master file and one longitudinal weights file
- For each wave:
  - Household file
  - Enumerated person file
  - Responding person file
  - Combined file

Documentation

- User Manual (discussed later by Nicole)
- Marked up questionnaires
- Coding framework
- Cross-wave variable listing
- Frequencies
- Program library (discussed later by Nicole)
- PanelWhiz (discussed later by John)

Documentation – Marked up questionnaires

- Excluding derived variables



G9b About how many nights each week, fortnight or month does this child usually stay overnight with you?

If respondent refers to weeks rather than nights, record number of full weeks instead of nights. If overnight contact is sparse, interviewer to get estimate for 3, 6 or 12 month period.

Zero overnight stays in a year 997

Else:

Record number of nights		Record number of full weeks
	OR	
Per week..... 1		per...
Fortnight..... 2		Fortnight..... 2
4 weeks..... 3		4 weeks..... 3
3 months 4		3 months 4
6 months 5		6 months 5
Year.....6		Year 6

```
CNCNGTH    G9b Non-resident child overnight stays
           - answered nights or weeks

CNCNGTN    G9b nights
CNCNGTNP   G9b Non-resident child overnight stays
           - nights - period

CNCNGTW    G9b full weeks
CNCNGTWP   G9b Non-resident child overnight stays
           - weeks - period
```

Documentation – coding framework

- Subject listing

File	Variable	Data Item	Categories	Population
CHILDREN - NON RESIDENT: DERIVED VARIABLES				
RP	CNCNGT	DV: Overnight stays of non-resident child (Days per annum)	[Days]	Respondents with non-resident children
Conversion from days per week, fortnight etc to days per annum. If answered in weeks, converted to days per annum				

Modeling Retirement Behaviour Using the HILDA Survey Data

Diana Warren
13 July 2011

FACULTY OF
BUSINESS &
ECONOMICS

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Overview

- The retirement module
- Defining retirement
  - Differences between actual labour force status and self-reported retirement status
- Modeling retirement behaviour
  - A Probit model of retirement
  - Examples of variables used to explain retirement decisions
  - Estimation results
- Summary

The Retirement Module

- In Waves 3, 7, and 11, HILDA respondents aged 45 and over are asked a series of questions about retirement
- Those who are already (fully or partly) retired were asked:
  - Reasons for (fully or partly) retiring
  - How and when they made the transition to retirement
  - Whether they coordinated their retirement with their spouse
  - Changes in their standard of living since retiring
  - Changes in relationships with family and friends since retiring
- Those who did not consider themselves to be retired were asked:
  - Expectations about age of retirement
  - Expectations about how the transition to retirement will be made
  - Saving for retirement and how they expect to fund retirement
  - Factors that will be important in the decision about when to retire

The Retirement Module

- Important to consider the sequencing of questions and the population being asked each question
- Questions such as:

  "Do you expect to make a gradual transition to retirement through a bridging job?

  "At what age do you expect to begin this transition?"

  "How do you expect to make this transition?"

  are asked separately to those who are employed and those who are not employed but do not consider themselves to be retired
- To examine the intentions of those who are not retired, information from two variables must be combined

Defining Retirement

- Retirement (and partial retirement) mean different things to different people
- People might consider themselves to be retired if they:
  - Are no longer working and do not intend to work again
  - Are receiving a pension or Social Security benefits
  - Are drawing upon their superannuation
  - Have reduced their working hours and/or pay
  - Left their 'career' job, either voluntarily or involuntarily
- Studies of retirement behaviour in the 1980s and early 1990s mostly accepted individuals' own self-classifications of whether they were 'working' or 'retired'. Later studies have more commonly adopted a strictly behavioural definition of retirement, based on whether an individual is still in the labour force.

Self-reported Retirement Status, by Labour Force Status

- Fairly close correspondence between labour force status and self-reported retirement status

Self-reported Retirement Status, by Labour Force Status, Men and Women aged 55-70, 2003

Employment Status	Self-Reported Retirement Status				
	Completely Retired	*Partly Retired*	*Not Retired at all*	*Not applicable - never worked*	*Total*
Full time	0.1%	1.7%	97.9%	0.3%	100.0%
Part time	1.3%	31.4%	67.3%	0.1%	100.0%
Unemployed	5.2%	26.4%	66.8%	1.6%	100.0%
Not in the labour force	84.4%	4.8%	4.9%	6.0%	100.0%

Data Issues

- Differences in the questions about retirement status asked in retirement modules and main questionnaire:

D20 **Have you retired (completely) from the workforce?** (_RTCOMP)

Yes ..... 1 ➔D21a
No ..... 2 ➔D21b
Never in workforce ..... 3 ➔E1
Don't know / can't say ..... 9 ➔D21b

D21a **In which year did you retire?** (_RTYR)

Record year [ ][ ][ ][ ] ➔D22

L2a **The next set of questions are about retirement from paid employment and your plans for retirement from paid employment. Do you consider yourself to be completely retired from the paid workforce, partly retired or not retired at all?** (_RTSTAT)

Completely retired ..... 1 ➔ L2b
Partly retired ..... 2 ➔ L3
Not retired at all ..... 3 ➔ L17
Not relevant – have never been in paid work ..... 4 ➔ T1

L2b **At what age did you completely retire?** (_RTCAGE)

Record age [ ][ ] ➔L4

Don't know ..... 999 ➔ L4

- Self-reported retirement status not included in the Continuing Person Questionnaire (CPQ) in Wave 4

Example: Probit Model with Random Effects

- Random Effects Probit model, dependent variable is 0 if the individual continues working and 1 if they have retired
- Key hypothesis to be tested is that financial incentives in the retirement system have an impact on retirement decisions
- Sample:
  - Men and women aged between 55 and 70 who were paid employment when the HILDA Survey began in 2001
  - Each individual remains in the sample until they exit employment

Controlling for :

- Potential gain in lifetime retirement income from postponing retirement for one year (Accrual)
- Age, Work-limiting health condition,
- Resident dependent children, Partner's employment status,
- Home ownership, Household net worth,
- Years of education, Labour force experience,
- Job satisfaction in the previous year

Resident Dependent Children

Derived Variable ihhtype:

1	Couple family wo children or others	14	Lone parent with children < 15 w other related
2	Couple family wo children w other related	15	Lone parent with children < 15 w other not related
3	Couple family wo children w other not related	16	Lone parent with depst wo others
4	Couple family with children < 15 wo others	17	Lone parent with depst w other related
5	Couple family with children < 15 w other related	18	Lone parent with depst w other not related
6	Couple family with children < 15 w other not related	19	Lone parent with ndepchild wo others
7	Couple family with depst wo others	20	Lone parent with ndepchild w other related
8	Couple family with depst w other related	21	Lone parent with ndepchild w other not related
9	Couple family with depst w other not related	22	Other related family wo children < 15 or others
10	Couple family with ndepchild wo others	23	Other related family wo children < 15 w others
11	Couple family with ndepchild w other related	24	Lone person
12	Couple family with ndepchild w other not related	25	Group household
13	Lone parent with children < 15 wo others	26	Multi family household

```
/* Dummy Variable: resident dependent children - from household type */
use "C:\data\Wave 9\Rperson_i90c.dta", clear
recode ihhtype (-10/-1 = .)  (1/3 10/12 19/26 = 0) (4/9 13/18 = 1), gen(depchd)
label variable depchd  "Resident dependent children"
```

Years of Education

History variable iedhigh: Highest level of education achieved

1	Postgrad - masters or doctorate
2	Grad diploma, grad certificate
3	Bachelor or honours
4	Adv diploma, diploma
5	Cert III or IV
6	Cert I or II
7	Cert not defined
8	Year 12
9	Year 11 and below
10	Undetermined

```
/* continuous variable: years of education */
recode iedhigh (1 = 18) (2 = 16) (3 = 15) (4 = 13) (5 8 = 12)(6 7 9 = 11) (10 = .), gen(yrsed)
label variable yrsed "Years of education“

/* four categories: highest level of education */
recode iedhigh (1/3 = 1) (4/7 = 2) (8 = 3) (9 = 4) (10 = .), gen(educ)
label variable educ "Highest level of education"
label define educ 1 "Degree" 2 "Certificate or Diploma" 3 "Year 12" 4 "Year 11 or below"
```

Labour Force Experience

History variable iehtjb: Time in paid work – years

For new persons, this is calculated from the months or years reported working in the employment history section of their interview (i.e., _EHTJBYR + _EHTJBMT). For continuing persons, this variable is calculated as the number of years reported working in their employment history section of their first interview (_EHTJB) plus time spent in paid work since last interview (calculated from calendar).

History variable iehtse: Time since full-time education - years

_EHTSE is the sum of _EHTJB, _EHTUJ and _EHTO (time in paid work, time not working and looking for work, time not working or looking for work). If _EHTJB, _EHTUJ and _EHTO are missing, _EHTSE is calculated as time since FT education as of last interview + time (years) since wave last interview.

```
/* labour force experience : proportion of time employed since leaving full-time education*/
gen exp = (iehtjb/iehtse)*100 if (iehtjb >= 0 & iehtse >= 0)
label variable exp "Work experience - time in paid work since leaving ft education"
```

Note: In some cases where respondents did not respond in every wave, time in paid work since last interview cannot be determined (these cases aregiven a code of -7). Therefore, labour force experience since leaving full-time education cannot be determined.

Work-limiting Health Condition

K5a **Looking at SHOWCARD K5, do you have any long-term health condition, impairment or disability (such as these) that restricts you in your everyday activities, and has lasted or is likely to last, for 6 months or more?** (_HELTH)

Yes 1
No ..... 2 ➔K19a

K7 **[Does your condition / Do your conditions] limit the type of work or the amount of work you can do?** (_HELTHWK)

Yes 1
No ..... 2 ➔K9
Unable to do any work ..... 3 ➔K9

K8 **Using the scale on SHOWCARD K8, could you pick a number between 0 and 10 to indicate how much your condition[s] limit[s] the amount of work you can do?**

An answer of 0 means "not at all" and an answer of 10 means you are "unable to do any work". (_HELTHDG)

Enter number from 0 to 10 [ ]

```
/* work -limiting health condition */
recode ihelth (-10/-1 = .)(1=1)(2=0), gen(wlhc)
replace wlhc = 0 if (ihelthwk == 2 | ihelthdg == 0)
label variable wlhc "Work-limiting health condition"
```

Partner's Employment Status

```
use "C:\data\Wave 9\Rperson_i90c.dta", clear
sort xwaveid
save "C:\temp\rperson.dta", replace

/* creating a small file of partner information */
use "C:\temp\rperson.dta", clear
/*keep only the variables wanted - partner cross-wave id, age, sex, labour force status */
keep ihhpxid ihgage ihgsex iesbrd
/* drop if no valid partner cross-wave id */
drop if ihhpxid == ""
/* rename the partner variables (swap person id with partner's id )*/
rename ihhpxid xwaveid
rename ihgage partage
rename ihgsex partsex
rename iesbrd partlfs
sort xwaveid
save "C:\temp\partner.dta", replace
```

Example: Partner's Employment Status

```
/* merge partner information onto file */
use "C:\temp\rperson.dta"
sort xwaveid
merge 1:1 xwaveid using "C:\temp\partner.dta“

/* drop cases from partner file that were not matched (_merge == 2) */
/* then drop _merge variable */
drop if _merge==2
drop _merge

/* set to zero for those who are not partnered */
recode partlfs (-10/-1 = .)(1 = 1)(2/3 = 2)
replace partlfs = 0 if (imrcurr >= 3 & imrcurr <= 6)
label variable partlfs 0 “Not partnered” 1 “Partner employed” 2  “Partner not employed"
```

Probit Model with Random Effects

- Unobserved heterogeneity: observations may be conditionally different in ways that are not accounted for by the explanatory variables included in the model
- In the case of retirement, unobserved heterogeneity may reflect work ethic or preferences for work and leisure
- To account for potential unobserved heterogeneity, a random effects model is used

```
/* convert xwaveid from string to numeric */
gen xid = real(xwaveid)
/* tell stata you have panel data, with individuals denoted by xid and years by wave */
xtset xid wave
xtprobit retired accrual netw age homeown othhhinc wlhc reskids i.prtempstat yrsed exp jobsat /*
*/ if sex == 1 & age >= 55 & age <= 70, re
margins, dydx(*) predict(pu0)
```

Estimation Results

Mean Marginal Effects from Probit Regressions with Random Effects, Men and Women aged 55-70

	Men	Women
Accrual ($'0,000)	-0.070***	-0.031*
Household wealth ($'00,000)	-0.001**	-0.002
Age	0.006**	0.008**
Own home	0.067***	0.047**
Other household income	0.001	0.0002
Work-limiting health condition	0.159***	0.054*
Resident dependent children	-0.057***	-0.018
Partner's Employment Status (Control = No Partner)		
Partner employed	-0.054**	0.006
Partner not employed	0.042	0.104***
Years of education	-0.008*	-0.007
Labour force experience	-0.006***	-0.001**
Job satisfaction in previous year	0.0001	-0.001***

* significant at 10%; ** significant at 5%; *** significant at 1%.

Summary

- Retirement modules in wave 3, 7 and 11 provide information about retirement behaviour and retirement intentions
- Issues to keep in mind:
  - Population being asked each question
  - Differences in self-reported and behavioural definition of retirement
  - Potential loss of observations when using History Variables and matching partner information
- SPSS, SAS and STATA programs for matching partner information are available on the HILDA web site:

  http://www.melbourneinstitute.com/hilda/doc/programlibrary.html

Documentation – cross-wave listing

File	Variable	Label	Wave 1	Wave 2	Wave 3
RP	_NCNGTH	Non-resident child overnight stays - answered nights or weeks	PQ H9	PQ G9	PQ G9b
RP	_NCNGTN	nights	PQ H9	PQ G9	PQ G9b
RP	_NCNGTNP	Non-resident child overnight stays - nights - period	PQ H9	PQ G9	PQ G9b
RP	_NCNGTW	full weeks	PQ H9	PQ G9	PQ G9b
RP	_NCNGTWP	Non-resident child overnight stays - weeks - period	PQ H9	PQ G9	PQ G9b

Documentation - Frequencies

cncngtnp G9b Non-resident child overnight stays - nights - period

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Week	60	.5	20.5	20.5
	2 Fortnight	118	.9	40.3	60.8
	3 4 weeks	51	.4	17.4	78.2
	4 3 months	18	.1	6.1	84.3
	5 6 months	1	.0	.3	84.6
	6 Year	45	.4	15.4	100.0
	Total	293	2.3	100.0	
Missing	-1 Not asked	12435	97.7		
Total		12728	100.0		

¢

Master file

- Lists all persons **ever** in a responding household
- Contains
  - Identifiers
  - Interview and household outcomes
  - Some person information
- 27,532 records and 73 variables (Release 9)
- File placed with data for latest wave

Household file

- Lists all **responding** households for particular wave (General Release)
- Includes household level information from
  - Household Form
  - Household Questionnaire
- Commonly used file for household-level analysis
- 7682 records and 711 variables in Wave 1

Enumerated person file

- Lists all **persons in responding households** for particular wave
  - Respondents
  - Non-responding adults
  - Children
- Includes person information from Household Form and some derived variables
- 19,914 cases and 190 variables in Wave 1
- Commonly used file for income dynamics or to derive summary HH info to match onto other files

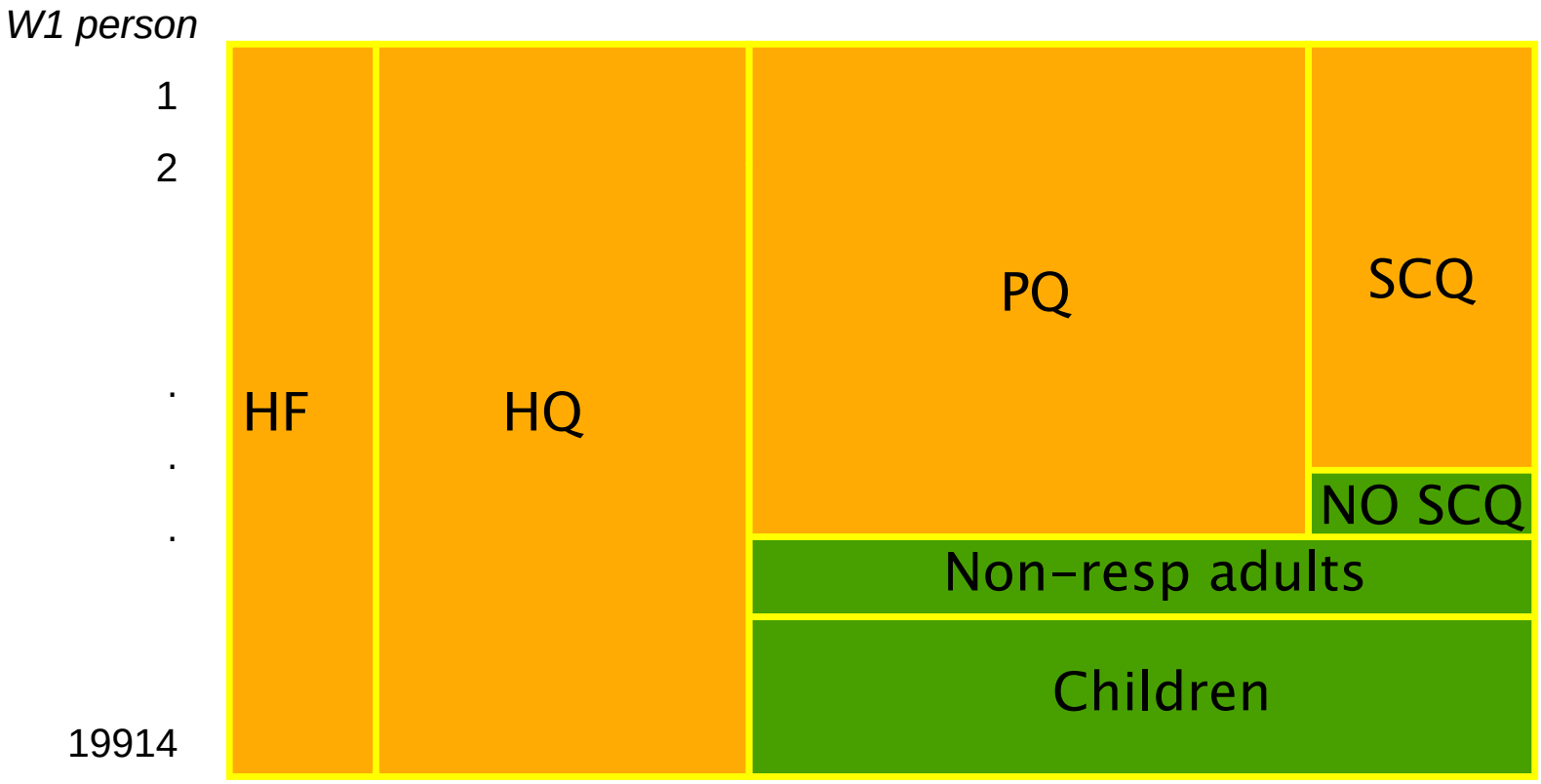
Responding person file

- Lists all **responding persons** for particular wave
- Includes person information from
  - Person Questionnaire
  - Self-Completion Questionnaire
- Commonly used file for person level analysis
- 13,969 cases and 2330 variables (about 850 from the calendar) in Wave 1

Combined file

- Full file of all enumerated persons
  - Household information attached to all individuals in household
  - Responding person information attached only to those individuals interviewed
- Commonly used file for person level analysis
  - but watch for missing data of non-responding adults!
- 19914 cases and 3100 variables in Wave 1

Combined file (c’td)



Longitudinal weights file

- Contains longitudinal weight for different types of balanced panels
- More about this in the weighting section
- 27532 cases and 145 variables in Release 9

Conventions

- Missing data
  - -1 Skipped (not asked)
  - -2 Not applicable
  - -3 Don’t know
  - -4 Refused
  - -5 Invalid multi response (SCQ only)
  - -6 Value implausible (after intensive checking)
  - -7 Unable to determine value
  - -8 No SCQ matched
  - -9 Non-responding HH
  - -10 Non-responding person (Combined File only)

Conventions (c'td)

- Variable naming
  - 1st char: Wave identifier (W1=A, W2=B, W3=C)
  - 2nd & 3rd char: General subject area
  - 4th to 8th char: Specific subject
- For example,
  - aesbrd = W1 labour force status broad
  - bhhtype = W2 household type
  - cedq110n = number doctoral degrees obtained

Confidentialising

- Data affected
  - Limited geography – state and remoteness area (area has been randomised)
  - Top-coded age, income & wealth
  - 2 digit ANZSCO & ANZSIC
  - Day and month of birth not provided

Creating your own dataset

Identifiers

Wave 1:

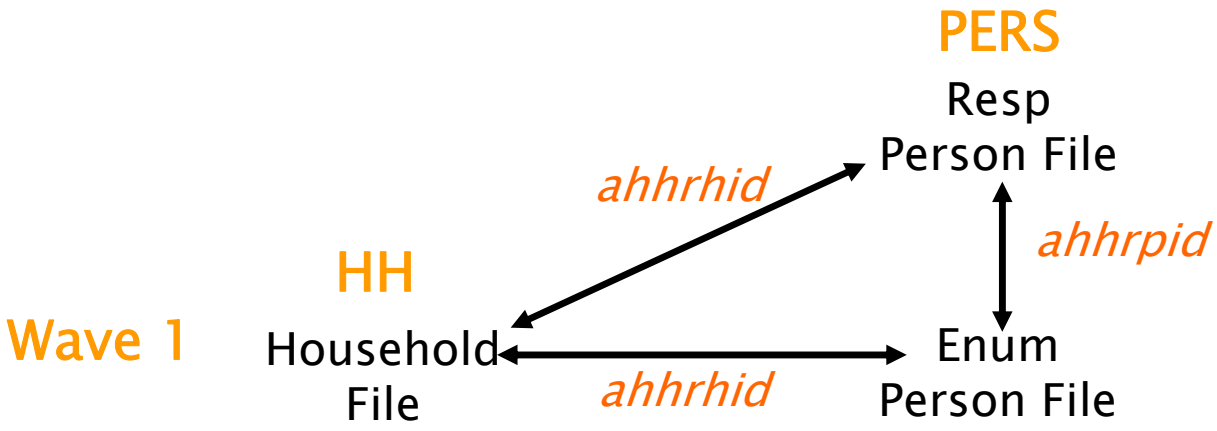
_hhrhid	001011	(6 digits)
_hhrpid	00101101	(8 digits)
xwaveid	0100001	(7 digits)

Wave 2+:

_hhrhid	00011	(5 digits)
_hhrpid	0001101	(7 digits)
xwaveid	0100001	(7digits)

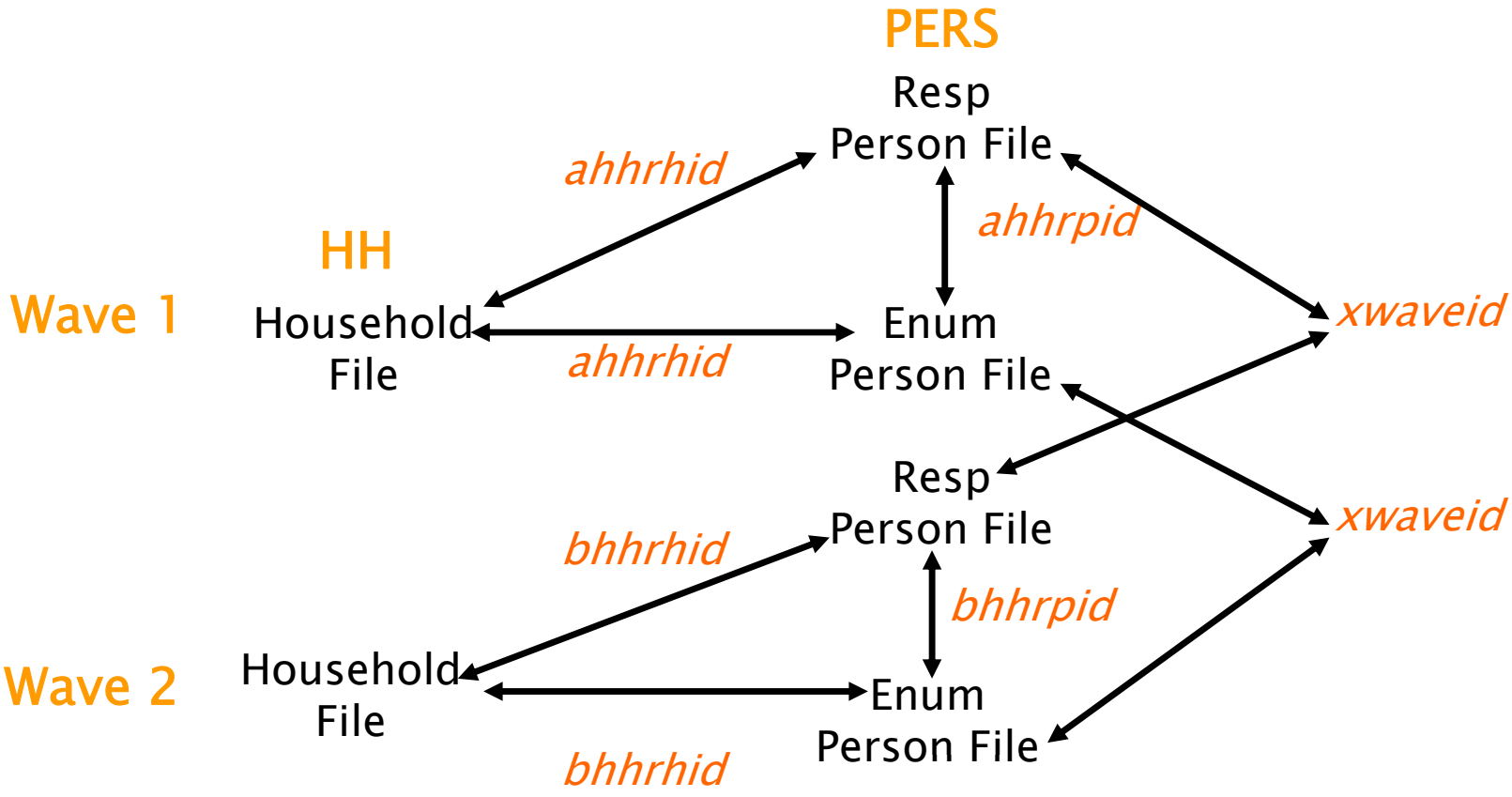
Also have cross-wave partner id (_hhpxid), mother id (_hhmxid), father id (_hhfxid) if they live in same household

Matching files – within a wave



Or use combined file

Matching files – across waves



Longitudinal files

- Two choices: Wide or Long

WIDE

LONG

Wide longitudinal files

- Use ‘wide’ longitudinal files when looking at descriptive change statistics
- ‘Wide’ example

Wave 1			Wave 2			
Xwaveid	aesbrd	aedhigh	atifei	besbrd	bedhigh	btifei
1234501	1	3	46000	1	2	53000
1234502	2	3	12000	1	3	35000

Long longitudinal files

- Use ‘long’ longitudinal file when statistical package requires one record per person per time point
- ‘Long’ example

Xwaveid	esbrd	edhigh	tifei	wave
1234501	1	3	46000	1
1234501	1	2	53000	2
1234502	2	3	12000	1
1234502	1	3	35000	2

Creating longitudinal files

- Example SPSS, SAS and Stata programs to construct ‘long’ longitudinal files can be found in Program Library on the HILDA web site
- Or use PanelWhiz
- Choice between balanced or unbalanced panel
- Choice between including deaths and overseas movers or not

Advice on statistical packages

- Depends on the experience of the user, but my recommendations are…
- Construct own dataset with variables of special interest to you (eg of matching files in SPSS, SAS and Stata in Program Library)
- Use SAS or SPSS for:
  - Data manipulation
  - Simple analysis
- Use Stata for:
  - More complex analysis

Things to watch out for

Responding vs enumerated

- Not every adult in responding household provided an interview
- Identified by: _hhresp >=63 and _hhresp <=68
- Need to be mindful of this when calculating total household information from person file (eg number of people in household receiving benefits)

Skips

- Skips are identified by –1
- Check questionnaire for skips
- For example,
  - A3 (whether English is first language) in W1 only asked if not born in Australia
  - If constructing variable for whether English is first language, then need to include those born in Australia

Positive and negative variables

- Income can be negative for
  - Business income
  - Rental income
  - Total income
- Wealth can be negative for:
  - Household net wealth
  - Home equity
  - Other property equity
  - Total property equity
  - Business equity
- Provided as two variables
  - positive and negative components
- Subtract negative from positive (after treating missing values if necessary)
- For example: abifip - abifin

Multiple variables available

- For example, labour force status
  - HF – 1 person in HH, simple questions
  - PQ – derived variable
  - PQ income section – wage and salary earner
  - SCQ – hours worked

Calendar

- Structure

	Occurs	July	Aug	Sept	Oct	Nov	Dec
Study	1	1 1 1	1 1 1	1 1 1	1 1 1	1 0 0	0 0 0
Job 1	1	0 0 0	1 1 1	1 1 1	0 0 0	0 0 0	0 0 0
Job 2	1	0 0 0	0 0 0	0 0 0	0 0 0	1 1 1	1 1 1
Job 3	0	-1-1-1	-1-1-1	-1-1-1	-1-1-1	-1-1-1	-1-1-1

- Tips for use
  - List calendar variables
  - Turn data into spells
  - Calendar finish generally in month/week prior to ivw

Caution…

Do not combine data from different releases!

- Each release contains all data in prior waves
- Manage the different releases you receive by
  - Overwriting
  - Storing in new spot
- See User Manual for indication of changes
  - Eg: cleaning, correcting derivations, new variables

Derived Variables

Overview

- When derived variables are created:
  - Collapse information back into one variable when question asked in multiple ways to help respondent
  - Complicated derivations
  - Missing data are imputed
  - Convert open-end answers to standard codeframes
  - Match external data to derive applicable measures
  - Code ‘Other, specify’ answers

Overview (c'td)

- All derived variables have prefix 'DV:' or 'History:' in the variable label.
- A description of how the variables was derived is supplied in the coding framework
- User Manual has a list of all derived variables
  - Check before constructing own variables
  - If derive own, compare against provided variables – feedback welcome!

Relationship identifiers

- Use cross-wave identifier
  - Mother identifier *(_hhmxid*)
  - Father identifiers *(_hhfxid*)
  - Partner's identifier *(_hhpxid*)

- Or, use identifiers from relationship grid
  - Given as person number in the household (eg 01, 02, etc)
  - Need to add person number to household number to get matchable identifier
  - Example

    HH id 12345, partner's person number 02

    Partners full id: 1234502

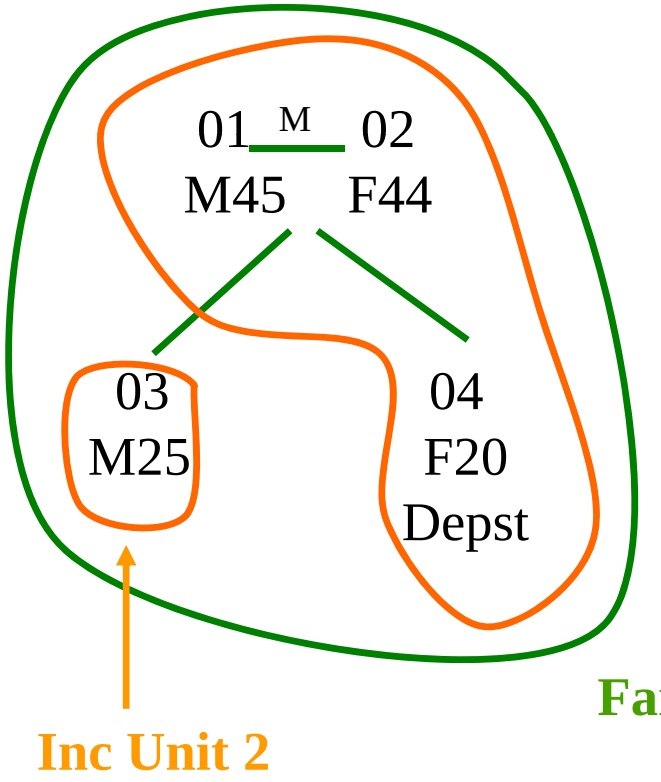
Relationships in households

- Complete relationship grid for everyone in the household (complex to use and **NO** order to how people are listed in HH)
- Derived variables from grid:
  - Relationship in household (_*hhrih*)
  - Family identifier (_*hhfam*)
  - Income unit identifier (_*hhiu*)
  - Family type (_*hhfty*)
  - Household type (_*hhtype*)

Relationships in households (c'td)

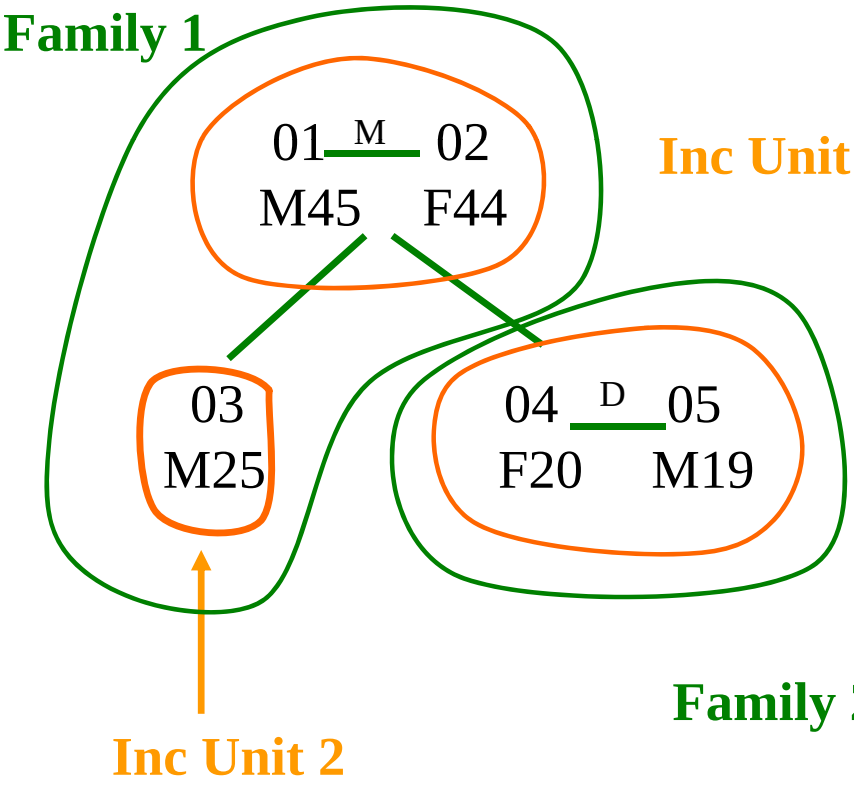
- Household->Family->Income unit->Person
- Household
  - Reside in same dwelling and make common provision for food and other essentials for living
- Key links in family units
  - Married/de facto relationship
  - Parent-child relationship
- Also for income units
  - Age of children

Family example 1



Couple family with dependent student without others

Family example 2

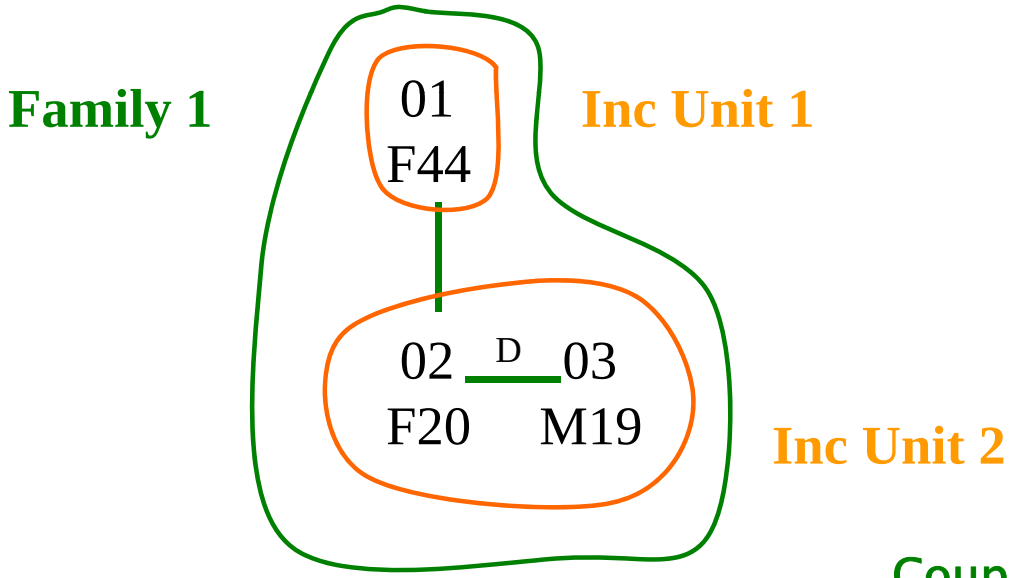


Multi-family household

1: Couple family with non-dependent child without others

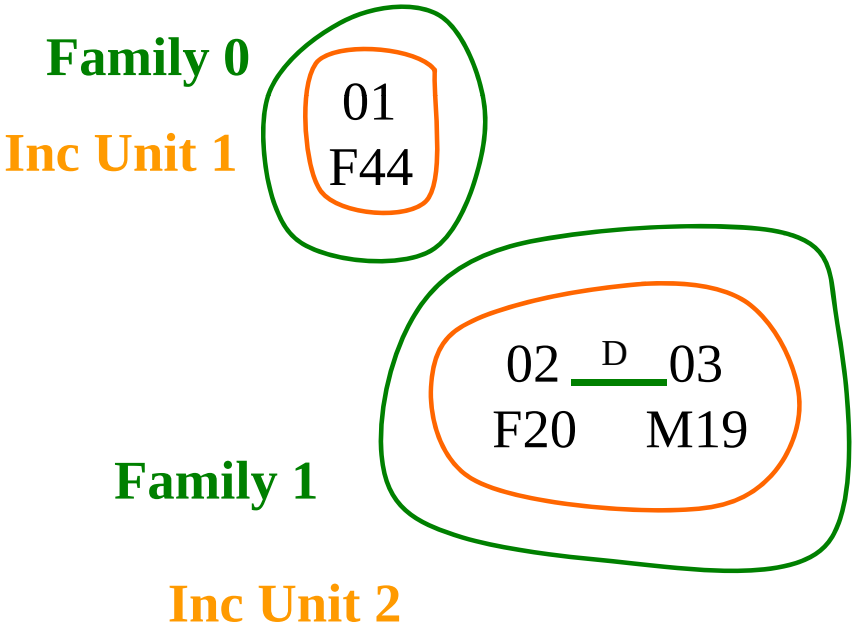
2: Couple family without children or others

Family example 3



Couple family without children with other related

Family example 4

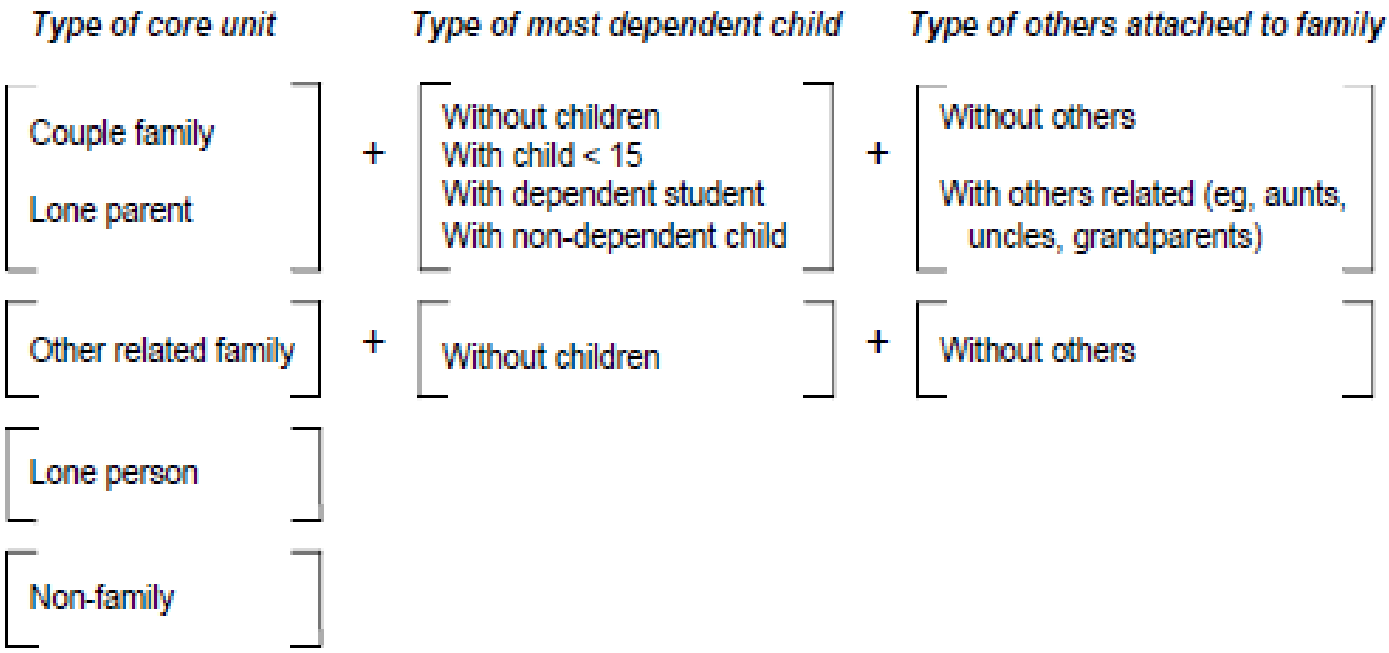


Couple family without children with other not related

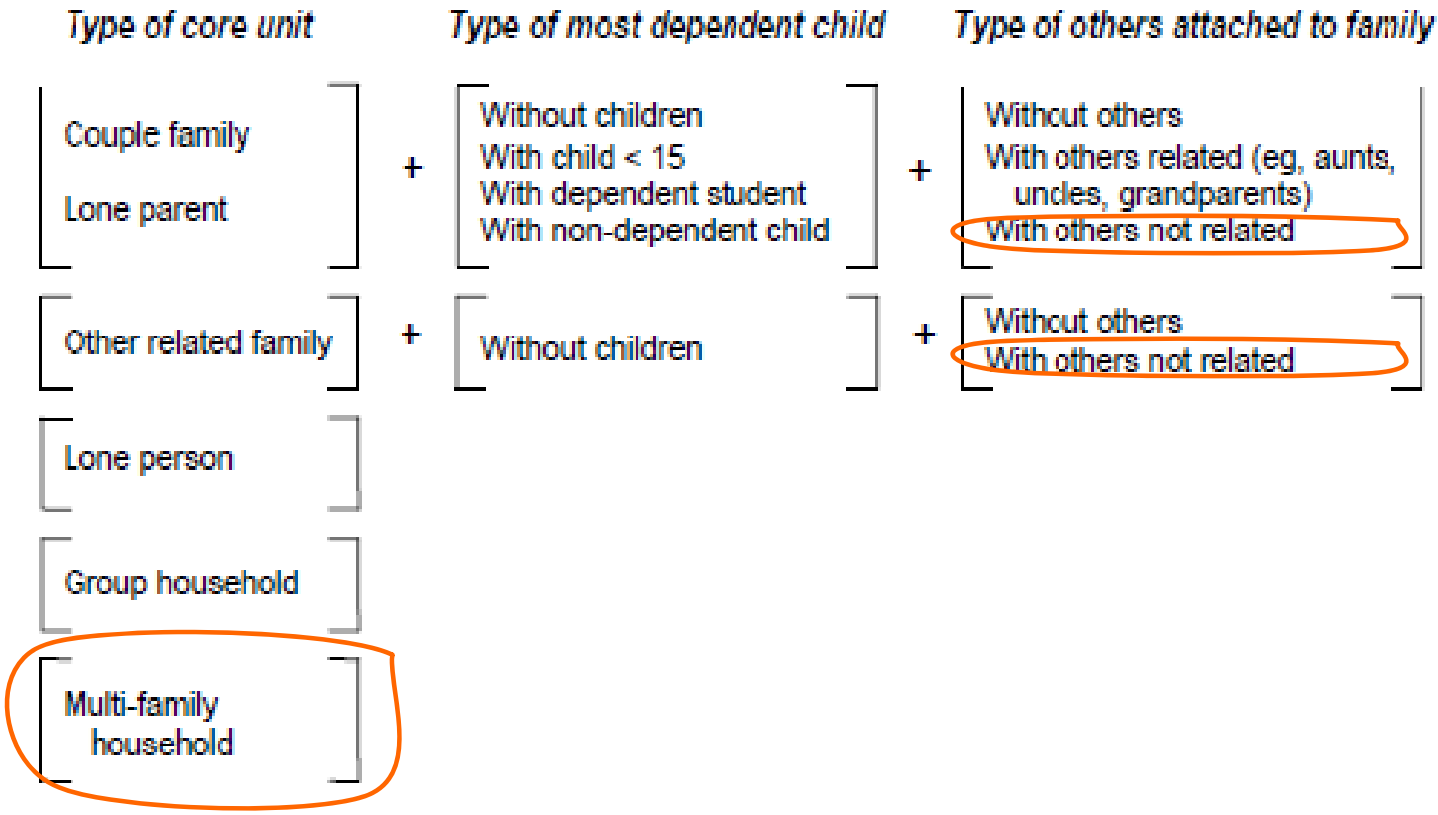
Family 0: Non-family member

Family 1: Couple family without children without other related

Construction of family type description



Construction of household type



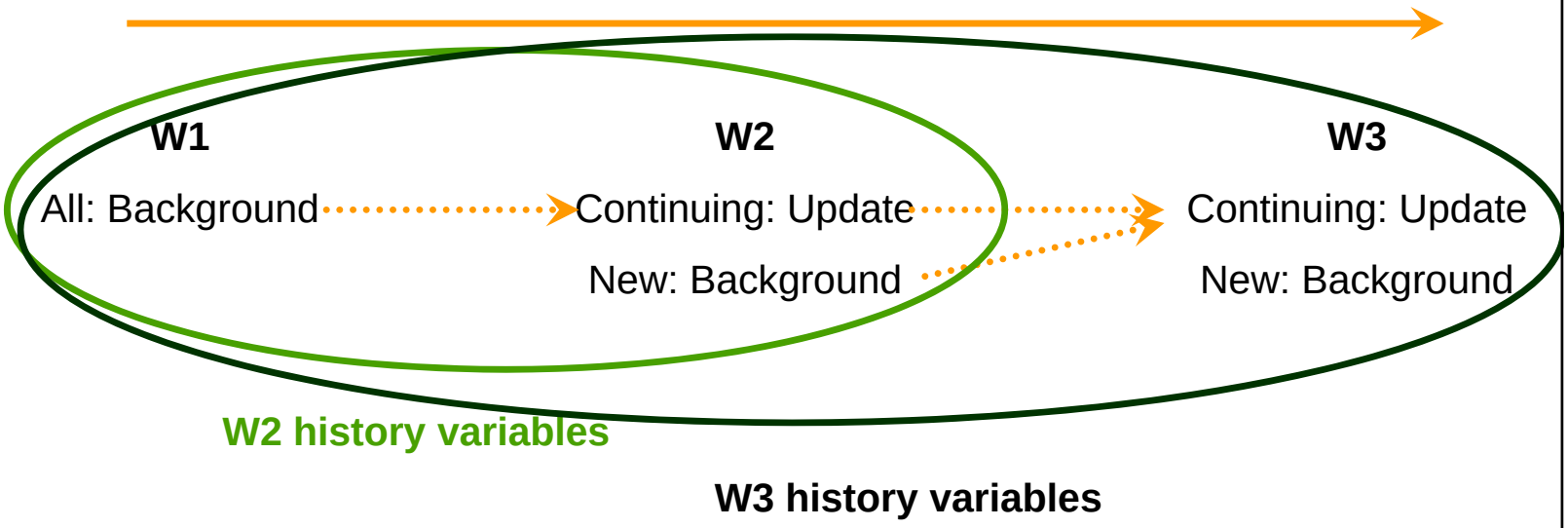
Household and family type

- Household type = family type, except for:
  - Multifamily HH
  - Group households (family number = 0, all non-family members, separate income units)
  - HH with ‘other not-related’

History variables

- W1: background from all respondents
- W2+: update from continuing respondents, plus background from new respondents
- W2 history variables
  - W1 background + update for continuing
  - W2 background for new
- W3 history variables
  - W2 history + update for continuing
  - W3 background for new

History variables (c’td)



Modeling Retirement Behaviour Using the HILDA Survey Data

Diana Warren
13 July 2011

FACULTY OF
BUSINESS &
ECONOMICS

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Overview

- The retirement module
- Defining retirement
  - Differences between actual labour force status and self-reported retirement status
- Modeling retirement behaviour
  - A Probit model of retirement
  - Examples of variables used to explain retirement decisions
  - Estimation results
- Summary

The Retirement Module

- In Waves 3, 7, and 11, HILDA respondents aged 45 and over are asked a series of questions about retirement
- Those who are already (fully or partly) retired were asked:
  - Reasons for (fully or partly) retiring
  - How and when they made the transition to retirement
  - Whether they coordinated their retirement with their spouse
  - Changes in their standard of living since retiring
  - Changes in relationships with family and friends since retiring
- Those who did not consider themselves to be retired were asked:
  - Expectations about age of retirement
  - Expectations about how the transition to retirement will be made
  - Saving for retirement and how they expect to fund retirement
  - Factors that will be important in the decision about when to retire

The Retirement Module

- Important to consider the sequencing of questions and the population being asked each question
- Questions such as:

  "Do you expect to make a gradual transition to retirement through a bridging job?

  "At what age do you expect to begin this transition?"

  "How do you expect to make this transition?"

  are asked separately to those who are employed and those who are not employed but do not consider themselves to be retired
- To examine the intentions of those who are not retired, information from two variables must be combined

Defining Retirement

- Retirement (and partial retirement) mean different things to different people
- People might consider themselves to be retired if they:
  - Are no longer working and do not intend to work again
  - Are receiving a pension or Social Security benefits
  - Are drawing upon their superannuation
  - Have reduced their working hours and/or pay
  - Left their 'career' job, either voluntarily or involuntarily
- Studies of retirement behaviour in the 1980s and early 1990s mostly accepted individuals' own self-classifications of whether they were 'working' or 'retired'. Later studies have more commonly adopted a strictly behavioural definition of retirement, based on whether an individual is still in the labour force.

Self-reported Retirement Status, by Labour Force Status

- Fairly close correspondence between labour force status and self-reported retirement status

Self-reported Retirement Status, by Labour Force Status, Men and Women aged 55-70, 2003

Employment Status	Self-Reported Retirement Status				
	Completely Retired	*Partly Retired*	*Not Retired at all*	*Not applicable - never worked*	*Total*
Full time	0.1%	1.7%	97.9%	0.3%	100.0%
Part time	1.3%	31.4%	67.3%	0.1%	100.0%
Unemployed	5.2%	26.4%	66.8%	1.6%	100.0%
Not in the labour force	84.4%	4.8%	4.9%	6.0%	100.0%

Data Issues

- Differences in the questions about retirement status asked in retirement modules and main questionnaire:

D20 **Have you retired (completely) from the workforce?** (_RTCOMP)

Yes ..... 1 ➔D21a
No ..... 2 ➔D21b
Never in workforce ..... 3 ➔E1
Don't know / can't say ..... 9 ➔D21b

D21a **In which year did you retire?** (_RTYR)

Record year [ ][ ][ ][ ] ➔D22

L2a **The next set of questions are about retirement from paid employment and your plans for retirement from paid employment. Do you consider yourself to be completely retired from the paid workforce, partly retired or not retired at all?** (_RTSTAT)

Completely retired ..... 1 ➔ L2b
Partly retired ..... 2 ➔ L3
Not retired at all ..... 3 ➔ L17
Not relevant – have never been in paid work ..... 4 ➔ T1

L2b **At what age did you completely retire?** (_RTCAGE)

Record age [ ][ ] ➔L4

Don't know ..... 999 ➔ L4

- Self-reported retirement status not included in the Continuing Person Questionnaire (CPQ) in Wave 4

Example: Probit Model with Random Effects

- Random Effects Probit model, dependent variable is 0 if the individual continues working and 1 if they have retired
- Key hypothesis to be tested is that financial incentives in the retirement system have an impact on retirement decisions
- Sample:
  - Men and women aged between 55 and 70 who were paid employment when the HILDA Survey began in 2001
  - Each individual remains in the sample until they exit employment

Controlling for :

- Potential gain in lifetime retirement income from postponing retirement for one year (Accrual)
- Age, Work-limiting health condition,
- Resident dependent children, Partner's employment status,
- Home ownership, Household net worth,
- Years of education, Labour force experience,
- Job satisfaction in the previous year

Resident Dependent Children

Derived Variable ihhtype:

1	Couple family wo children or others	14	Lone parent with children < 15 w other related
2	Couple family wo children w other related	15	Lone parent with children < 15 w other not related
3	Couple family wo children w other not related	16	Lone parent with depst wo others
4	Couple family with children < 15 wo others	17	Lone parent with depst w other related
5	Couple family with children < 15 w other related	18	Lone parent with depst w other not related
6	Couple family with children < 15 w other not related	19	Lone parent with ndepchild wo others
7	Couple family with depst wo others	20	Lone parent with ndepchild w other related
8	Couple family with depst w other related	21	Lone parent with ndepchild w other not related
9	Couple family with depst w other not related	22	Other related family wo children < 15 or others
10	Couple family with ndepchild wo others	23	Other related family wo children < 15 w others
11	Couple family with ndepchild w other related	24	Lone person
12	Couple family with ndepchild w other not related	25	Group household
13	Lone parent with children < 15 wo others	26	Multi family household

```
/* Dummy Variable: resident dependent children - from household type */
use "C:\data\Wave 9\Rperson_i90c.dta", clear
recode ihhtype (-10/-1 = .)  (1/3 10/12 19/26 = 0) (4/9 13/18 = 1), gen(depchd)
label variable depchd  "Resident dependent children"
```

Years of Education

History variable iedhigh: Highest level of education achieved

1	Postgrad - masters or doctorate
2	Grad diploma, grad certificate
3	Bachelor or honours
4	Adv diploma, diploma
5	Cert III or IV
6	Cert I or II
7	Cert not defined
8	Year 12
9	Year 11 and below
10	Undetermined

```
/* continuous variable: years of education */
recode iedhigh (1 = 18) (2 = 16) (3 = 15) (4 = 13) (5 8 = 12)(6 7 9 = 11) (10 = .), gen(yrsed)
label variable yrsed "Years of education“

/* four categories: highest level of education */
recode iedhigh (1/3 = 1) (4/7 = 2) (8 = 3) (9 = 4) (10 = .), gen(educ)
label variable educ "Highest level of education"
label define educ 1 "Degree" 2 "Certificate or Diploma" 3 "Year 12" 4 "Year 11 or below"
```

Labour Force Experience

History variable iehtjb: Time in paid work – years

For new persons, this is calculated from the months or years reported working in the employment history section of their interview (i.e., _EHTJBYR + _EHTJBMT). For continuing persons, this variable is calculated as the number of years reported working in their employment history section of their first interview (_EHTJB) plus time spent in paid work since last interview (calculated from calendar).

History variable iehtse: Time since full-time education - years

_EHTSE is the sum of _EHTJB, _EHTUJ and _EHTO (time in paid work, time not working and looking for work, time not working or looking for work). If _EHTJB, _EHTUJ and _EHTO are missing, _EHTSE is calculated as time since FT education as of last interview + time (years) since wave last interview.

```
/* labour force experience : proportion of time employed since leaving full-time education*/
gen exp = (iehtjb/iehtse)*100 if (iehtjb >= 0 & iehtse >= 0)
label variable exp "Work experience - time in paid work since leaving ft education"
```

Note: In some cases where respondents did not respond in every wave, time in paid work since last interview cannot be determined (these cases aregiven a code of -7). Therefore, labour force experience since leaving full-time education cannot be determined.

Work-limiting Health Condition

K5a **Looking at SHOWCARD K5, do you have any long-term health condition, impairment or disability (such as these) that restricts you in your everyday activities, and has lasted or is likely to last, for 6 months or more?** (_HELTH)

Yes 1
No ..... 2 ➔K19a

K7 **[Does your condition / Do your conditions] limit the type of work or the amount of work you can do?** (_HELTHWK)

Yes 1
No ..... 2 ➔K9
Unable to do any work ..... 3 ➔K9

K8 **Using the scale on SHOWCARD K8, could you pick a number between 0 and 10 to indicate how much your condition[s] limit[s] the amount of work you can do?**

An answer of 0 means "not at all" and an answer of 10 means you are "unable to do any work". (_HELTHDG)

Enter number from 0 to 10 [ ]

```
/* work -limiting health condition */
recode ihelth (-10/-1 = .)(1=1)(2=0), gen(wlhc)
replace wlhc = 0 if (ihelthwk == 2 | ihelthdg == 0)
label variable wlhc "Work-limiting health condition"
```

Partner's Employment Status

```
use "C:\data\Wave 9\Rperson_i90c.dta", clear
sort xwaveid
save "C:\temp\rperson.dta", replace

/* creating a small file of partner information */
use "C:\temp\rperson.dta", clear
/*keep only the variables wanted - partner cross-wave id, age, sex, labour force status */
keep ihhpxid ihgage ihgsex iesbrd
/* drop if no valid partner cross-wave id */
drop if ihhpxid == ""
/* rename the partner variables (swap person id with partner's id )*/
rename ihhpxid xwaveid
rename ihgage partage
rename ihgsex partsex
rename iesbrd partlfs
sort xwaveid
save "C:\temp\partner.dta", replace
```

Example: Partner's Employment Status

```
/* merge partner information onto file */
use "C:\temp\rperson.dta"
sort xwaveid
merge 1:1 xwaveid using "C:\temp\partner.dta“

/* drop cases from partner file that were not matched (_merge == 2) */
/* then drop _merge variable */
drop if _merge==2
drop _merge

/* set to zero for those who are not partnered */
recode partlfs (-10/-1 = .)(1 = 1)(2/3 = 2)
replace partlfs = 0 if (imrcurr >= 3 & imrcurr <= 6)
label variable partlfs 0 “Not partnered” 1 “Partner employed” 2  “Partner not employed"
```

Probit Model with Random Effects

- Unobserved heterogeneity: observations may be conditionally different in ways that are not accounted for by the explanatory variables included in the model
- In the case of retirement, unobserved heterogeneity may reflect work ethic or preferences for work and leisure
- To account for potential unobserved heterogeneity, a random effects model is used

```
/* convert xwaveid from string to numeric */
gen xid = real(xwaveid)
/* tell stata you have panel data, with individuals denoted by xid and years by wave */
xtset xid wave
xtprobit retired accrual netw age homeown othhhinc wlhc reskids i.prtempstat yrsed exp jobsat /*
*/ if sex == 1 & age >= 55 & age <= 70, re
margins, dydx(*) predict(pu0)
```

Estimation Results

Mean Marginal Effects from Probit Regressions with Random Effects, Men and Women aged 55-70

	Men	Women
Accrual ($'0,000)	-0.070***	-0.031*
Household wealth ($'00,000)	-0.001**	-0.002
Age	0.006**	0.008**
Own home	0.067***	0.047**
Other household income	0.001	0.0002
Work-limiting health condition	0.159***	0.054*
Resident dependent children	-0.057***	-0.018
Partner's Employment Status (Control = No Partner)		
Partner employed	-0.054**	0.006
Partner not employed	0.042	0.104***
Years of education	-0.008*	-0.007
Labour force experience	-0.006***	-0.001**
Job satisfaction in previous year	0.0001	-0.001***

* significant at 10%; ** significant at 5%; *** significant at 1%.

Summary

- Retirement modules in wave 3, 7 and 11 provide information about retirement behaviour and retirement intentions
- Issues to keep in mind:
  - Population being asked each question
  - Differences in self-reported and behavioural definition of retirement
  - Potential loss of observations when using History Variables and matching partner information
- SPSS, SAS and STATA programs for matching partner information are available on the HILDA web site:

  http://www.melbourneinstitute.com/hilda/doc/programlibrary.html

HILDA User Training 2011:

Web-based tools, weighting and imputation

Nicole Watson
HILDA Deputy Director, Survey Methodology

FACULTY OF
BUSINESS &
ECONOMICS

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Web-based tools

Documentation – User Manual

- Data
  - Structure
  - Identifiers
  - Matching files, creating longitudinal files
  - Conventions
- Summary of derived variables
- Documentation options
- Sample and collection

Documentation – Program library

- SAS, SPSS, Stata example programs
  - Match household and responding person file for a wave
  - Add partner variables
  - Create long longitudinal files
  - Create wide longitudinal files
- SAS macro to strip the first letter from the variable name
- Code contributed by users

Documentation – Other tools

- On-line data dictionary
  - melbourneinstitute.com/hildaddictionary/onlinedd/Default.aspx
- Questionnaires
  - http://melbourneinstitute.com/hilda/questionnaires/default.html

Documentation – Data quality resources

- Discussion/Technical papers
  - Impact of transition to CAPI and new fieldwork provider
  - Expenditure data
  - Occupation and industry coding
  - Height and weight data
  - Recall bias in employment data
  - Trade union membership
  - Kessler 10 scale
  - Attrition analysis
  - Re-engaging non-respondents

- HILDA User Manual

  - Section 6: Data quality issues
  - Section 8: Data collection

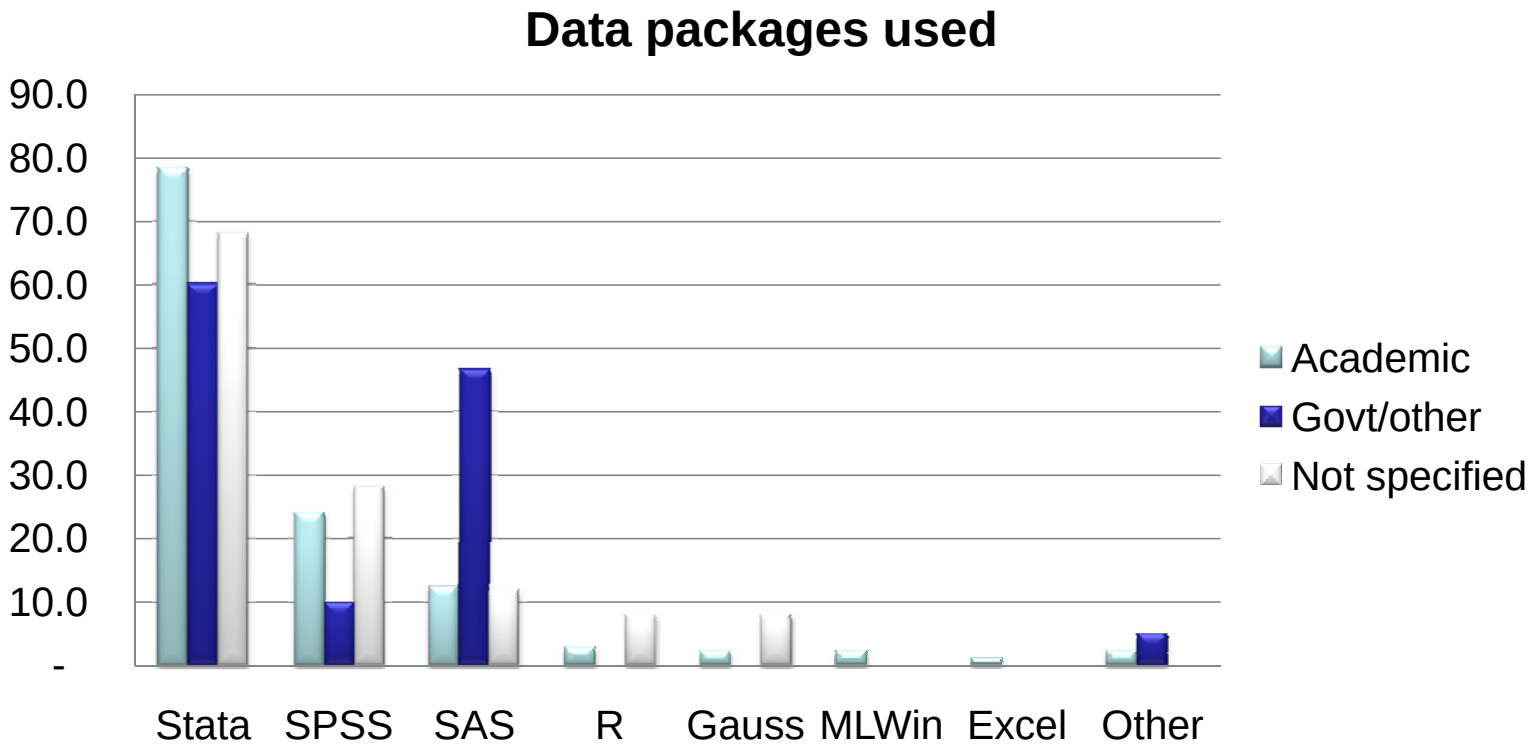
- FaHCSIA data quality statement
  - To be released soon

- Research repository

  melbourneinstitute.com/hilda/FLoSse.html

Statistical Packages

Deciding which one to use…



Weighting

Outline

- Overview
- Steps in weighting
- Benchmarking
- Replicate weights
- Advice on using the weights
- Final remarks

Overview

Purpose of weights:

- Weighted estimates sum to the population rather than the sample
- Adjusts for
  - Unequal probabilities of selection
  - Unit non-response
  - Wave non-response
- Used in
  - Tabulations (almost always)
  - Regressions (sometimes)

What weights do not do…

- Weights do not adjust for
  - Stratification
  - Clustering of the sample
- Weights are only part of the picture in analysing data from complex surveys
  - Weights help correct for biases in estimates
  - Standard errors around those estimates need to be calculated assuming a complex design

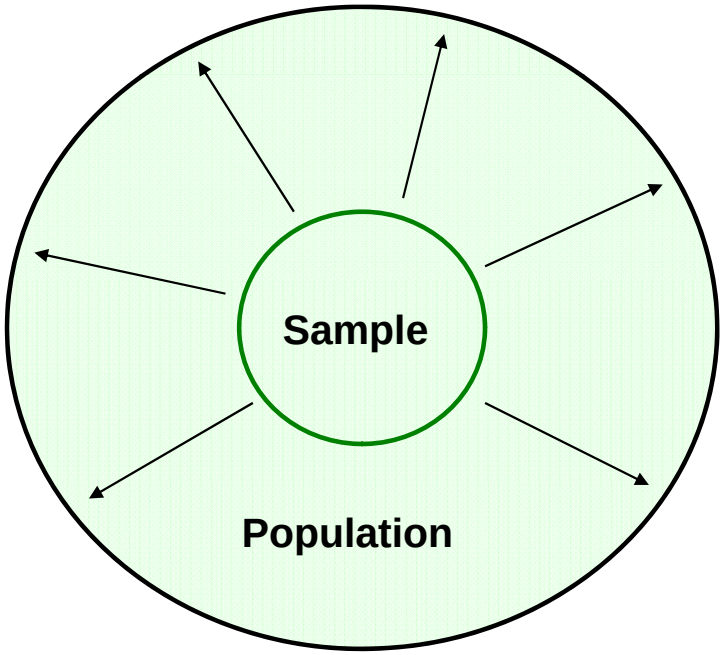
Weights provided

	Household	Enumerated Person	Responding Person
Longitudinal		✓	✓
Cross-sectional	✓	✓	✓

Three main steps

1. Design weights
2. Non-response / attrition adjustment
3. Calibration / benchmarking

Weighting process



- IDEAL WORLD: no non-response
- STEP 1: weight sample up to population by inverse of probability of selection (design weight)

Design weight in HILDA wave 1

P(select household) = P(select CD) *
P(select dwelling | select CD) *
P(select household | select dwelling)

$$p_{hs} = 488 \frac{\hat{N}_{di}}{\sum_j \hat{N}_{dj}} \frac{n_{di}}{N_{di}} \frac{n_{hi}}{N_{hi}}$$

where

$\hat{N}_{di}$ =estimated number of dwellings in CD

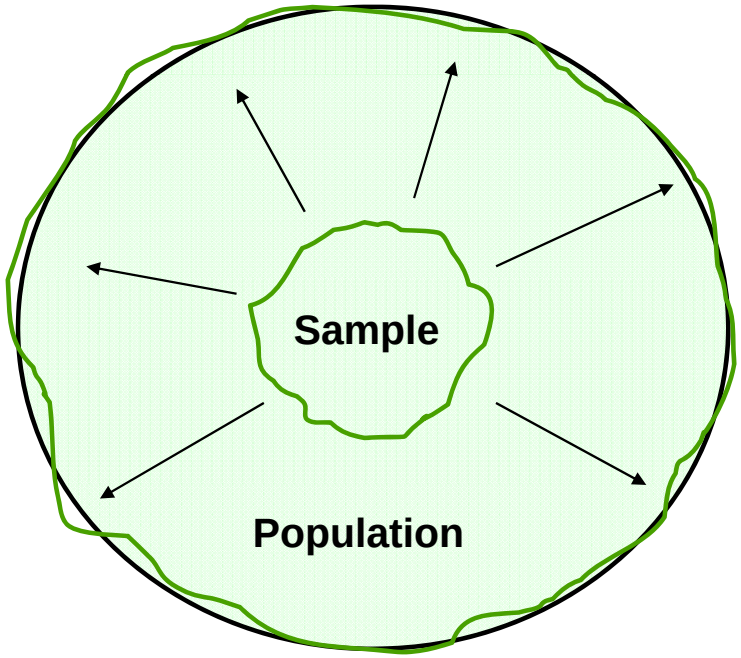
N_{di} =actual number of dwellings in CD

n_{di} =number of dwellings selected in CD

N_{hi} =number of households in dwelling

n_{hi} =number of households selected in dwelling

Weighting process (c’td)



- REAL WORLD: non-response
- STEP 2: adjust design weight for non-response

Non-response adjustment in HILDA wave 1 (households)

Internal (collected)

- Security features
  - Locked gate – no intercom access**
  - Locked gate – intercom access**
  - Security guard/doorman/on-site manager/gatekeeper
  - Security door**
  - No trespassing sign
  - Evidence of a dangerous dog*
  - No junk mail sign/no hawkers sign**
  - Bars on windows
- Dwelling type*
- External condition of dwelling**
- Proportion of highrises in area

** sig at 1%, * sig at 5%

External (known about CD)

- Geographic location**
- Proportion of different family and household types
- Median age of persons in the area
- Density of area (population per km^2)**
- Proportion of people speaking a language other than English
- Proportion of different dwelling types
- Proportion of people with different employment states
- Average household size
- Median weekly household income**

$$w_{1h} = \frac{1}{p_{hs}} \frac{1}{\hat{p}_{hr}}$$

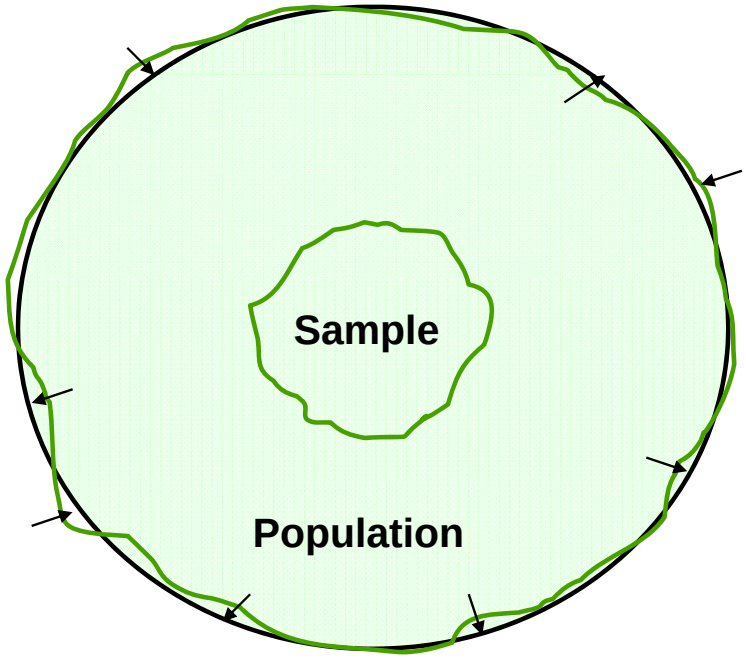
Non-response adjustment in HILDA wave 1 (responding persons)

- Geographic location**
- Labour force status**
- Sex**
- Age group**
- Number of adults in HH**
- Number of children in HH**
- Marital status**
- English ability**
- Dwelling type**
- Long term health condition
- Housing tenure

** sig at 1%, * sig at 5%

$$w_{1i} = w_{1hb} \frac{1}{\hat{p}_{ir}}$$

Weighting process (c'td)



- STEP 3: benchmark to known population totals
  - Also called post-stratification or calibration
  - For HILDA, we use the SAS macro GREGWT (written by ABS Methodology Division) to adjust to multiple benchmarks simultaneously

Benchmarking in HILDA wave 1

Household and enumerated person benchmarks (integrated)		
Number	*Variable Group*	*Level*
1	Number of adults and children in household	Household
2	State by part of State	Household
3	Sex by age group	Person
4	State by part of State	Person
5	Labour force status	Person
6	Marital status	Person

Responding person benchmarks		
Number	*Variable Group*	*Level*
1	Sex by age group	Person
2	State by part of State	Person
3	State by labor force status	Person
4	Marital status	Person
5	Household composition (number of adults and children)	Person

Longitudinal responding person weight

- Same process for longitudinal weights
  1. Design weight
  2. Adjust for non-response
  3. Benchmark
- Include deaths and overseas movers

- Longitudinal balanced panel from wave 1 to T

1. & 2.

$$w_{L_{1\_T}i} = w_{1ib} \frac{1}{\hat{p}_{ri_{2\_T}}}$$

Design weights, non-response and b'mark adjustment for W1

Non-response adjustment for attrition W2 to WT

3. Benchmark $w_{L_{1\_t}i}$ to wave 1 person characteristics to give $w_{L_{1\_t}ib}$

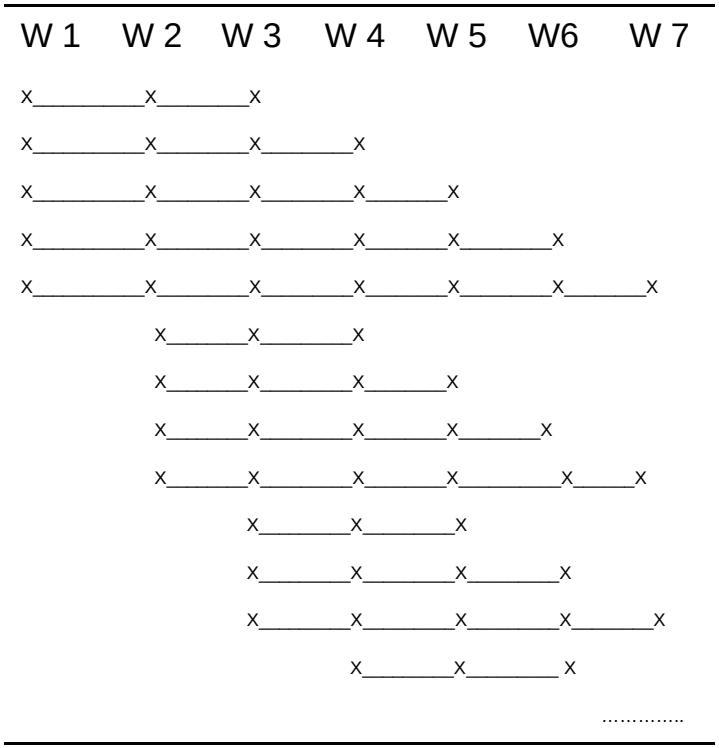
Non-response adjustment for attrition

- Person characteristics
  - Age**
  - Sex
  - Marital status**
  - Ability in speaking English
  - Employment status and hours**
  - Number of children the person has
  - Country of birth**
  - Highest level of education achieved**
  - Relationship in household**
  - Health status
  - Likelihood of moving
  - Number of times moved in last 10 years*
  - Whether reference person in household
- Interview situation
  - Cooperation**
  - Assisted interview
  - Difficulties during interview (english, eyesight, hearing, reading)
  - Suspicious**
  - Understanding
  - Influence of others
  - Length of interview*
  - SCQ returned**
- Household characteristics
  - Location (State by part of State)**
  - Remoteness area*
  - SEIFA index of disadvantage**
  - Dwelling type**
  - Dwelling condition*
  - Bedrooms**
  - Number of calls made**
  - Partial household responding**
  - Number of adults**
  - Number of children
  - Household type*
  - Housing tenure*
  - Benefit recipient in household**
  - Household income
  - Interview time in household*
- Household changes
  - Splits**
  - Moved between waves**

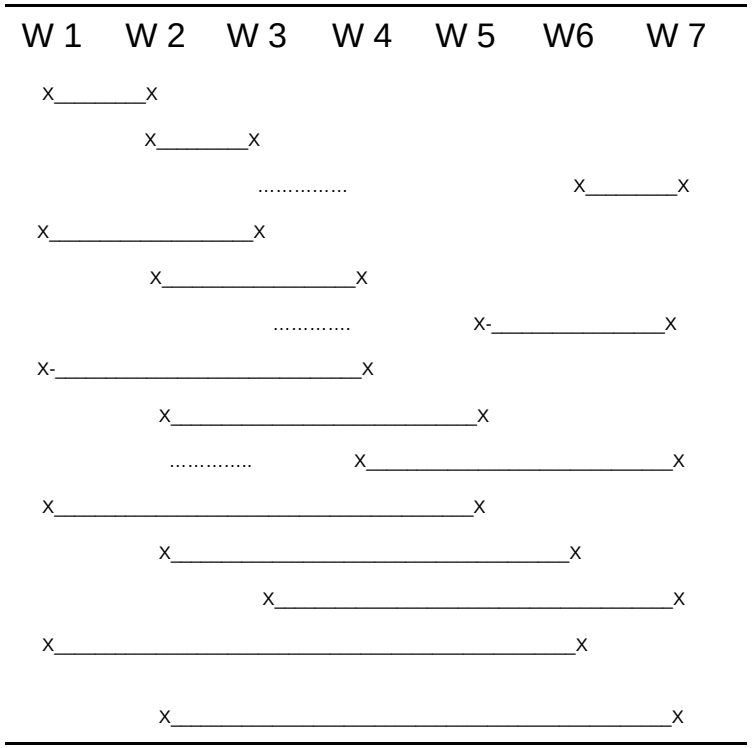
** sig at 1%, * sig at 5%

Balanced panels

Balanced continuous panels



Balanced wave combination panels



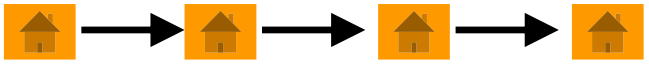
Cross-sectional weights for Wave 2 onwards

- Reminder of following rules
  - W1 household members
    - Continuing Sample Members (CSMs)
  - Follow and interview CSMs and their HHs
  - New sample members
    - Temporary Sample Members (TSMs)
  - TSMs become CSMs when:
    - New baby born/adopted to CSM
    - Parent of new CSM

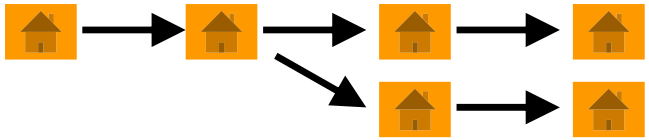
Adjustment to weights…

- …for new sample members

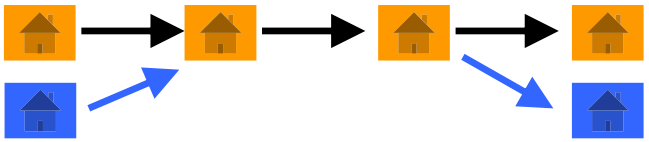
No change



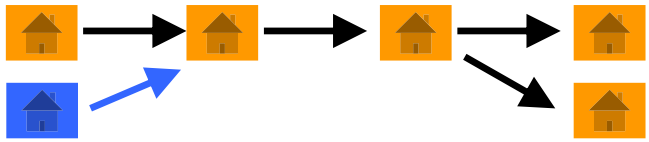
Splits



New entrants



New entrants converted to PSMs



Special inclusions/exclusions

- Moved overseas or died – still receive a longitudinal weight
- Zero weights
  - Remote areas of Australia excluded from the benchmarking process
    - ABS classification updated after wave 1 – causes some w1 zero weights
    - Affects people who moved to a remote area
  - Non-private dwellings for cross-sectional weights

Benchmarking

- Also called post-stratification or calibration
- Why do it?
  - Sample frame never perfect
  - Non-response adjustment uses modelled probabilities, and also isn't perfect

Revisions to benchmarks

- **Estimated residential population**
  - Births
  - Deaths
  - Interstate and intrastate migration
  - Overseas migration
  - Census
- **Estimated households**
  - Estimated residential population
  - Census
- **Labour force estimates**
  - Change in estimation methodology
  - Change in benchmarks

Benchmarking

Household weight benchmarks		
Number	*Variable Group*	*Level*
1	Number of Adults and Children in Household	Household
2	State by part of state	Household
3	Sex, Age Group	Person
5	State by part of state	Person
6	Labour Force Status	Person
7	Social Marital Status	Person

Responding person weight benchmarks		
Number	*Variable Group*	*Level*
1	Sex by age group	Person
2	State by part of state	Person
3	State by labor force status	Person
4	Marital statue	Person
5	Household composition (number of adults and children)	Person

Replicate weights

- 45 replicate weights have been provided
  - Essentially replicate the weighting process but on a subset of the data
- Use to calculate standard errors

Data

Weights

Population	*Weight*	*Replicate Weights*
Responding People (longitudinal)	_lnwtrp	_rwln1 to _rwln45
	_lr*t1_tn*	_lr*t1_tn1* to _lr*t1_tn45*
	_lr*t1tn*	_lr*t1tn1* to _lr*t1tn45*
Enumerated People (longitudinal)	_lnwte	_rwlne1 to _rwlne45
	_le*t1_tn*	_le*t1_tn1* to _le*t1_tn45*
	_le*t1tn*	_le*t1tn1* to _le*t1tn45*
Households (cross-sectional)	_hhwth	_rwh1 to _rwh45
Responding People (cross-sectional)	_hhwtrp	_rwrp1 to _rwrp45
Enumerated People (cross-sectional)	_hhwte	_rwe1 to _rwe45

Sample Design Characteristics

Characteristic	*Variable*
Stratum	_hhstrat
Cluster	_hhraid

Advice on using weights

- Means, ratios, totals etc for the population?
  - Must use weights
- Which weight to use depends on which question you are answering
  - How many people live in poor households in 2007?
  - What proportion of people have changed their employment status between 2005 and 2007?
- Static versus change
  - Cross-section versus longitudinal weights

Using weights in regression models?

- If use weights in models
  - Guards against
    i) mis-specification of model holding in the population, and
    ii) informative sample designs
  - BUT estimates tend to have larger standard errors
- Run with and without
  - If substantial differences, reconsider specification – what additional information is in the weights that is not in your model?
  - If no difference, use model without weights as will tend to have lower standard errors
- Further info

  Pfeffermann, D. (1993),'The Role of Sampling Weights When Modeling Survey Data', *International Statistical Review*, Vol. 61, No. 2, pp. 317-337.

Only part of the story…

- Complex sample design, eg HILDA has
  - Stratification
  - Systematic selection
  - Households clustered in areas
  - People clustered in households
  - Unequal probabilities of selection
- Weights used to calculate estimates
- When calculating standard errors, need
  - Strata, cluster, weights
  - Replicate weights (replicates sample design on subset of data – HILDA has 45 replicates)

Methods of complex SE calculation

- Method categories:
  - Linearisation techniques
  - Resampling
- User manual recommends two:
  - Taylor Series Linearisation
  - Delete a group Jackknife

Delete-a-group Jackknife

- Method of producing replicate weights for the sample
- 45 replicates provided for all weights on the dataset.
- Takes into account the complex design of the survey and the way it has evolved from wave 1
- Simple formula for the variance using the replicate weights

$$Var(\hat{x}) = \left(\frac{R-1}{R} \right) \sum_{r=1}^{R} \left(\hat{x}^{(r)} - \hat{x} \right)^2$$

Delete-a-group Jackknife – replicates

- Aim to mirror the overall sample design within each replicate group
- Ordered CD's systematically allocated to a replicate group
  - CD 1 to 45 allocated to replicate group 1 to 45 then repeated
- Allocated group of CD's dropped from each replicate
  - Replicate group size 44/45 of main sample
- Replicate group undergoes weighting process
  - units in dropped CD's receive zero weight

Taylor Series linearisation

- Use Taylor's Theorem to create a linear approximation to a non-linear statistic
- Calculate the variance of the linear approximation
- Benefits
  - Well developed theory
  - Can be applied in general sample designs
  - Software readily available

Sample design issues

- Taylor Series
  - Software packages assume a stratified cluster design
  - Doesn’t take into account benchmarking effects or systematic selection
- Replicate groups don’t correctly reflect the sample design
  - Cluster size needs to be taken into account
  - Systematic selection skips

More information

- Technical papers
  - Wave 1 weighting
  - Wave 2 weighting
  - HILDA Standard Errors User Guide
  - Guide to Standard Errors for Cross-sectional estimates (Wave 1)

melbourneinstitute.com/hilda/data/technical_papers.html

Imputation

Overview

- Three types of non-response:
  - item non-response
  - wave non-response
  - unit non-response
- Level: person and household
- Imputation: A method of ‘completing’ data
- Allows for improved data use

Overview (c'td)

- What do we impute in Release 9:
  - Income
  - Wealth
    - Wave 2 and wave 6 special module
    - Home value (every wave)
  - Expenditure
  - Ad-hoc

    Date of birth, wave 2 employment status

Size of missing data problem- Income

Variable	*W1*	*W2*	*W3*	*W4*	*W5*	*W6*	*W7*	*W8*	*W9*
Financial year income (non-zero cases only)									
Wages and salaries	7.9	6.9	5.5	3.8	4.5	4.6	5.1	4.6	5.9
Aust Gov't pensions	1.4	1.7	1.0	1.5	1.1	0.9	1.0	1.9	3.6
Aust Gov't allowances	3.0	2.1	2.1	2.6	2.0	0.9	1.2	1.9	4.6
Foreign govt pensions	0.5	2.7	0.0	0.5	2.4	0.5	1.0	2.2	2.3
Business income	29.1	28.6	27.4	19.4	21.6	18.6	19.8	19.0	20.3
Interest income	19.5	18.6	13.9	11.0	11.3	12.8	11.6	11.2	14.4
Dividends and royalties	14.6	14.5	11.9	9.2	10.2	11.3	11.3	11.3	13.5
Rent income	20.3	14.7	14.9	11.3	10.5	10.3	10.2	10.5	11.1
Private pensions	6.3	4.6	3.2	4.1	4.8	3.9	4.1	4.0	4.9
Private transfers	8.0	22.9	15.5	14.4	20.7	13.4	18.5	20.1	18.6
Total FY income	**15.7**	**14.9**	**12.1**	**9.6**	**10.7**	**10.3**	**10.4**	**10.6**	**12.3**

Size of missing data problem- Wealth(w2,6)

Person-level wealth

	W 2	*W6*
Joint bank accounts	9.8	6.0
Own bank accounts	4.6	3.3
Superannuation, retirees	20.1	19.7
Superannuation, not retired	17.3	27.5
HECS debt	10.6	7.6
Joint credit card debt	10.1	7.6
Own credit card debt	3.6	3.1
Other Debt	2.4	1.8

Household-level wealth

	W2	*W6*	*W6 band*
Children's bank accounts	6.2	4.6	
Business value	20.1	17.5	7.8
Cash investments	11.6	12.3	7.1
Equity investments	15.3	13.3	4.4
Collectibles	14.0	15.1	8.1
Home value	7.6	4.2	
Other property value	4.6	0.5	
Life insurance	24.9	28.5	16.9
Trust funds	35.7	35.8	26.7
Vehicles: Value	2.3	1.5	
Business debt	22.9	11.6	8.1
Home: All debt	7.6	4.2	
Other property: Debt	7.1	5.9	

Size of the missing data problem (Home value)

- Home Value

	W1	*W2*	*W3*	*W4*	*W5*	*W6*	*W7*	*W8*	*W9*
Home value (households) (non-zero cases only)									
Home value	5.9	7.6	5.6	4.0	3.3	4.2	2.6	3.0	3.6

Size of missing data problem - Expenditure

	Wave								
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Responding persons									
Usual payments/repayments per month (collected in the HQ)									
Rent	0.3	0.4	0.4	0.3	0.3	0.4	0.5	0.5	0.6
First mortgage	1.9	1.5	1.4	1.3	1.2	0.9	1.2	1.4	1.7
Second mortgage	0.7	0.5	0.5	0.5	0.5	0.5	0.6	0.6	1.0
Weekly household expenditure (collected in the HQ)									
Work-related child care, term-time (school aged)	0.3	0.3	0.1	0.2	0.1	0.1	0.1	0.1	0.2
Work-related child care, holidays (school aged)	0.3	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.2
Work-related child care (not yet at school)	0.1	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.2
Non-work-related child care (school aged)	-	0.4	0.0	0.2	0.1	0.0	0.1	0.1	0.8
Non-work-related child care (not yet at school)	-	0.1	0.0	0.1	0.1	0.0	0.2	0.2	0.9
All groceries	1.2	-	1.0	0.9	0.9	-	-	-	-
Groceries for food and drink	2.0	-	1.7	1.2	1.2	-	-	-	-
Meals eaten outside	0.9	-	1.0	0.9	0.8	-	-	-	-

Size of missing data problem - Expenditure

	Wave								
Variable	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Responding persons									
Annualised household expenditure (collected in the SCQ)									
Groceries	-	-	-	-	15.1	14.5	16.5	18.0	18.7
Alcohol	-	-	-	-	15.9	15.4	17.1	18.8	19.3
Cigarettes and tobacco	-	-	-	-	16.4	16.2	17.8	19.0	19.7
Public transport and taxis	-	-	-	-	16.5	16.9	18.3	19.5	20.0
Meals eaten out	-	-	-	-	15.1	15.1	16.8	18.4	19.2
Leisure activities	-	-	-	-	15.9	-	-	-	-
Motor vehicle fuel	-	-	-	-	15.6	14.6	16.6	18.4	19.1
Men's clothing and footwear	-	-	-	-	-	15.7	17.5	19.1	19.7
Women's clothing and footwear	-	-	-	-	-	16.4	18.1	19.3	20.1
Children's clothing and footwear	-	-	-	-	-	17.2	18.3	20.1	20.8
Clothing and footwear	-	-	-	-	16.6	-	-	-	-
Telephone rent and calls	-	-	-	-	16.0	-	-	-	-
Telephone rent and calls, internet charges	-	-	-	-	-	14.7	16.7	18.2	19.0
Holiday and holiday travel costs	-	-	-	-	15.8	15.1	17.3	18.9	19.4

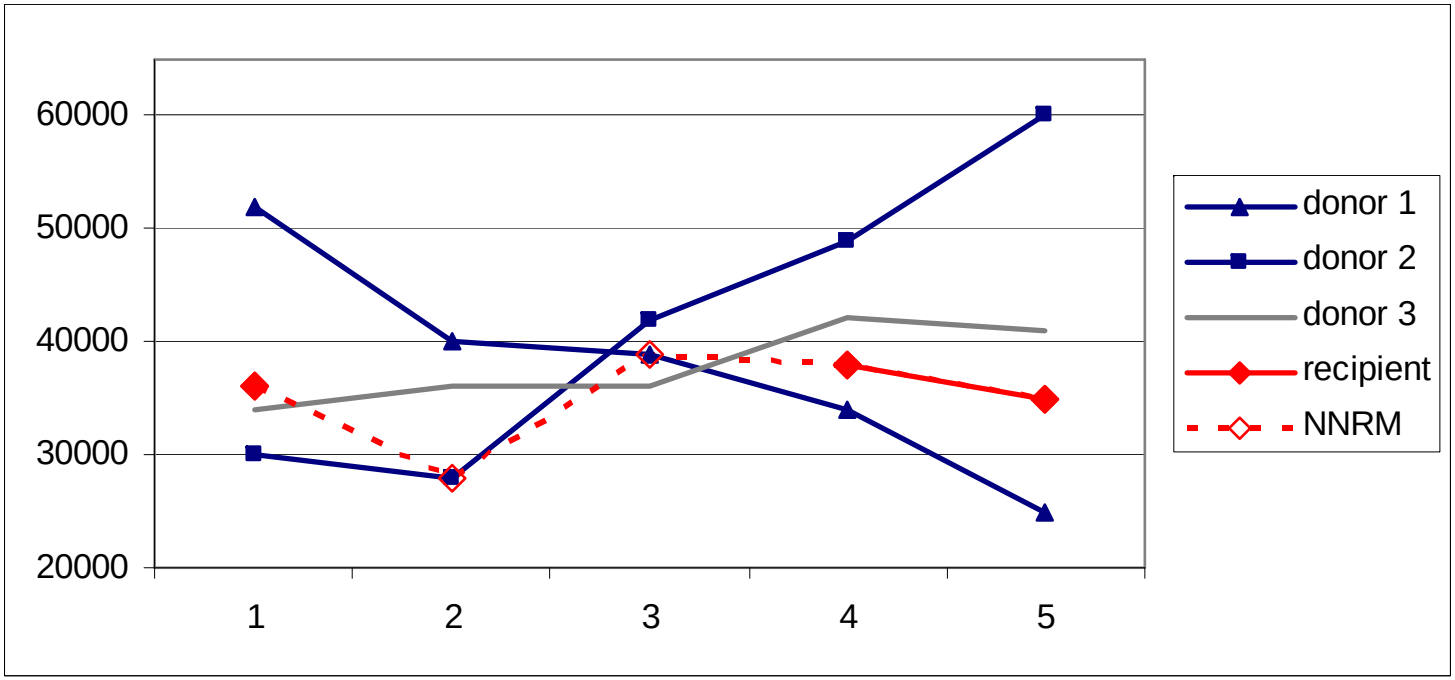
Methods

- Income and expenditure
  - Nearest neighbour regression method
  - Population carry-over
  - Little & Su method
- Wealth
  - Nearest neighbour regression method
  - Little & Su method
- Age, wave 2 employment status
  - Hot deck method

Nearest neighbour regression method

- Cross-sectional method
- Process:
  - Use data from respondents to create a multiple regression model
  - Calculate the predicted values for both donors and non-respondents.
  - Order the predicted values and choose the NN to a non-respondents
  - Donor's actual value used to impute missing value
- Restrict donors to imputation classes
- Donor used maximum of 2 times

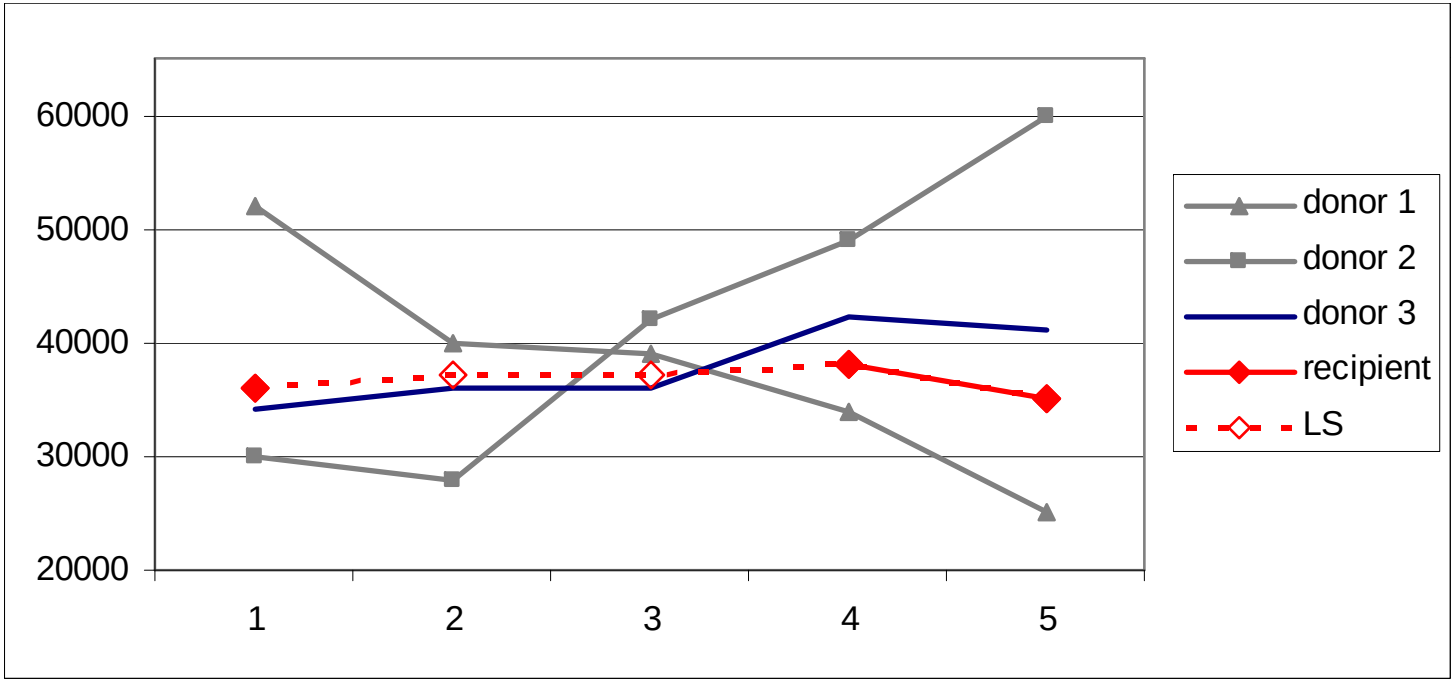
Nearest neighbour regression method



Little and Su method

- Longitudinal method
- The method adjusts for how the person differs from the sample average in a wave, and also how the wave differs from the average of the other waves.
- Used for repeat measures on a single variable
  - Needs at least one wave non-missing
- Use the single donor for ALL waves requiring imputation
- Donor chosen from similar age range

Little and Su method



Population carry-over

- Longitudinal method
- Before applying the Little and Su method when imputing income (for non-respondents) and expenditure
- Carry the information forward or backward
  - using complete cases to decide the probability with which
    - the last wave value is carried forward or
    - the next wave value is carried backwards
- Only carry over whether zero or non-zero

Other imputation

- Age imputation:
  - Simple hot deck method
  - Imputation class: sex, household size, relationship in households, household type, partner age and parent age.
- Wave 2 employment status
  - Employment status of non-respondents
  - Derive from wave 3 response to the calendar questions
  - Hot deck imputation

Longitudinal households

- Longitudinal households:
- Households are linked longitudinally for imputation purpose
- Indicator variables:
- _*hhlink*: household linkage indicator for home value imputation
- _*hwlink*: household linkage indicator for wealth imputation (household link between wave 2 and 6)
- _*hxylink*: household linkage indicator for expenditure imputation

Datasets

- Dataset includes
  - Original variables
  - Imputed variables
  - Imputation flag
- All imputed variables have the suffix ‘i’
- All imputation flag variables have the suffix ‘f’
- Up to the user if they want to use it

More information

- Technical papers
  - Imputation methods (income and wealth)
  - Expenditure imputation
  - Evaluation of alternative income imputation methods

melbourneinstitute.com/hilda/data/technical_papers.html

melbourneinstitute.com/hilda/data/discussion_papers.html

Questions?

Where to go for help…

- User Manual
- Technical papers (survey methodology, weighting, imputation, data quality)
- Questionnaires (waves 1-11)
- HILDA website: melbourneinstitute.com/hilda/
- HILDA team: hilda-inquiries@unimelb.edu.au