Data Structure of HILDA

Henrik Schubert

2023-06-27

Data Structure

The data is a panel from Australia running over 21 waves. The data is stored separately for each wave, which is indicated by a letter between a and u. Moreover, the data consists of three different data types.

- `raw/Eperson...` = containing all persons in all responding households and contains limited information from the HF (includes respondents, non- respondents and children)
- `raw/Household...` = household data, which consists of information for the entire household unit
- `raw/Rperson...` = containing all persons who provided an interview and contains CPQ/NPQ and SCQ information
- `raw/Combined...` = this is a combined file of the three files above. The household information and responding person information is matched to each enumerated person. **IMPORTANT: The household ID changes over waves!!!**

Identifiers

Individuals have a unique identifier for all waves which is `xwaveid`. Individuals in the same household in a particular wave have a household identifier for that wave (`_hhrhid`)5. Note that you will need to replace the underscore '_' in the variable name with the appropriate letter for the wave, 'a' for wave 1, 'b' for wave 2, etc. The household identifier will change from wave to wave as these household identifiers are randomly assigned anew each wave.

Original Name	New Name	Description	Data Sets
`_hhrhid`	`id_hh`	Id for the household	Epers, Rpers, Household
`xwaveid`	`id`	Id for the individual	Epers, Rpers
`_hhpxid`	`id_part`	Id for the partner	Epers, Rpers, household
`hhbfxid`	`id_fath`	Id for the father	Epers, Rpers
`hhbmxid`	`id_moth`	Id for the mother	Epers, Rpers

Household and Person files

Household and person-level files within a wave can be merged using `_hhrhid` (i.e. `ahhrhid` for wave 1, `bhhrhid` for wave 2, etc).7 Enumerated and responding person files within a wave can be merged by using the cross-wave identifier `xwaveid` (or alternatively the wave specific person identifier `_hhrpid`).

Partner's ID's

Partners within the household are identified by their cross-wave identifier (`_hhpxid`) or by their two-digit person number for the household (`_hhprtid`). These variables are provided on both the enumerated and responding person files and are derived using the HF relationship grid. Partners are either married or de facto and include same sex couples. If you are using `_hhprtid`, it is the person number for the household (for example, if person 02's partner is person 05, the partner identifier for person 02 will contain '05' and for person 05 it will contain '02').

Parent's ID's

Parents within the household are similarly identified in `_hhfxid` and `_hhmxid` (father's and mother's cross-wave identifiers) or `_hhfid` (father's person number) and '_hhmid" (mother's person number). A parent may be natural, adopted, step or foster (a parent's de facto partner also counts as a parent).

Data transformation

Given the threefold data structure, I decided to load the data seperately per type. Thus, the folder `code/` consists of files to load the *household*, the *enumerate person* and the *respondent person* data.

1. Enumerate person data

From this data, we get information on respondents who are residing in the houshold, but were not interviewd. The files are stored in the format:

`~/raw/Eperson_` + *wave* + `210c.dta`

Variables:

- `hbfxid` : biological Father's cross wave id
- `hbmxid` : biological Father's cross wave id
- `hhpxid` : partner's cross wave id
- `xwaveid` : person's cross wave id
- `hhrpid` : household's person id
- `hgni` : whether a person was not interviewed
- `hgint` : participated in an interview
    1. = participated in the interveiw
    2. = did not participate in the interview
    3. = did not participate, because under age 15
- `wsce` : income from all sources (unimputed)
- `wscei` : income from all sources (imputed)

2. Household data

The files are stored in the format:

`~/raw/Household_` + *wave* + `210c.dta`

Variables:

- `hgsex` (1 - 15): sex of the # household member
- `hgage` (1 - 15): age of the # household member
- `id` (1 - 15): wave id
- `hid`: household id, random household id (*changes across waves*)
- `mid`: Mother id
- `fid`: father's id
- `hh_size`: size of the household
- `_hgage`: Age last birthday at June 30 20221
- `_hgyob`: Year of birth
- `hhstate`: state residence
- `hhtype`: household type
- `mcurr`: marriage situation
- `hhrih`: household relationships
- `hifditp`: household disposible income (positive)
- `hifditn`: household disposible income (negative)

3. Respondents data

The respondents data consists of information from the interview with the household head. The PQs are administered to every member of the household aged 15 years and over. The CPQ is for people who have ever been interviewed before and the NPQ is for those who have never been interviewed before. Parental consent is sought before interviewing persons aged under 18 years who are still living with their parents. _hhpq states which type of interview was applicable and _hgwsli indicates how many weeks have elapsed since the respondent's last interview (if they are completing a CPQ). The date the PQ is completed is provided in _hhidate.

The files are stored in the format:

`~/raw/Rperson_` + *wave* + `210c.dta`

Variables:

- `_dxyr` + (1 - 10): year of birth of the deceased child
- `_dcany` + (1 - 15): any children listed at G1f
- `_psyobf` + (1 - 15): father's year of birth
- `_psyobm` + (1 - 15): mother's year of birth
- `_bsad` + (1 - 15): age difference with non-resident sibling
- `bsoy` + (1 - 15): Is non-resident children older or younger than you

Creating fertility biographies

A challenge faced in the data is that the HILDA does not have a chilbearing biography, not speaking of the birth dates for the biological children. Nonetheless, one can obtain the childbearing information by combining information on the biological children residing in the household (household data), the deceased children using the ... data, and the biological children living outside the household, which is captured in the ... data.

Variable name	Data set	Description	Transformation
`_hhrhid`	`id_hh`	# of children in HH	bio_child = rel + chil
`xwaveid`	`id`	Age of children in HH	yob = int_year - age
`_hhpxid`	`id_part`	Id for the partner	Epers, Rpers, household
`hhbfxid`	`id_fath`	Id for the father	Epers, Rpers
`hhbmxid`	`id_moth`	Id for the mother	Epers, Rpers

Data quality

A challenge that research faces when studying male fertility using surveys is related to data quality. For instance (REFERENCE) and (REFERENCE) show that the childbearing information are not correctly captured with the survey, which is particularly strong for certain groups. Thus, the problem may be emphasized for specific societal groups.

Approach to assess the data quality

In order to tackle the challenge, we use the male fertility database created by Christian Dudel and Sebastian Klüsener as gold standard to eveluate the survey-based estimates of male fertility rates. Thus, the procedure relies on the assumption that the reference data provides a good account of fertility information.

We estimate the percantage deviation from the gold-standard. This is estimated in the following way:

$$deviation = \frac{\sum_{x=15}^{55} |f(x) - g(x)|}{\sum_{x=15}^{55} f(x)},$$

where $f(x)$ is age-specific fertility rate from the gold standard data from the male fertility and $g(x)$ is the age-specific fertility rate based on the survey.

Before the investigation, we set the threshhold for sufficient quality at maximum 10 %.