Data Structure of HILDA

Henrik Schubert

2023-06-27

Contents

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.

Missings

In the HILDA survey, missing values are marked with negative integers, which range from -1 to -10. The different numbers indicate the reasons for the missingness.

value	label
-10	[-10] Non-responding person
-9	[-9] Non-responding household
-8	[-8] No SCQ
-7	[-7] Not able to be determined
-6	[-6] Implausible value
-5	[-5] Multiple response SCQ
-4	[-4] Refused/Not stated
-3	[-3] Don't know
-2	[-2] Not applicable
-1	[-1] Not asked

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:

Original Name	New Name	Description
`_hhrhid`	`id_hh`	Id for the household
`xwaveid`	`id`	Id for the individual
`_hhpxid`	`id_part`	Id for the partner
`hhbfxid`	`id_fath`	Id for the father
`hhbmxid`	`id_moth`	Id for the mother
`hgni`	`int`	person was interviewed (Y/N)
`hgint`	`int`	participated in an interview
`wsce`	`inc`	income from all sources
`wscei`	`inc_imp`	income all sources (imputed)
`hgyob`	`yob`	Year of birth

- `hgint` : participated in an interview
  1. = participated in the interview
  2. = did not participate in the interview
  3. = did not participate, because under age 15

Fertility major panel:

In addition to the fertility information based on the household grid, in waves 5, 8, 11, 15, 19, there is a major module on fertility information. The module contains information on past fertility (deceased children and outside the household) and intentions.

The table below summarizes the variables extractet from the fertility survey.

Original Name	New Name	Description
`tcr`	`nchild_res`	N. own resident children
`tcnr`	`nchild_non-res`	N. own non-resident children
`rcyng`	`age_res_id_child`	Age youngest own resid. child
`ncyng`	`age_non-res_chi`	Age "-" own non-resid. child
`tcn04`	`nchi_non-res_0_4`	N. own resid. child (0-4)
`tcn514`	`nchi_non-res_5_14`	N. own resid. child (0-4)
`tcn1524`	`nchi_non-res_15_24`	N. own resid. child (0-4)
`tcr04`	`nchi_res_0_4`	N. own resid. child (0-4)
`tcr514`	`nchi_res_5_14`	N. own resid. child (0-4)
`tcr1524`	`nchi_res_15_24`	N. own resid. child (0-4)
`dcany`	`dc_child_any`	Any deceased children
`dcyr`(1-10)	`yob_dc_child`	Year of birth-deceased child
`dcperm`	`dc_permission`	Permission - ask deceased ch.

2. Household data

The files are stored in the format:

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

Variables:

Original Name	New Name	Description
`_hhrhid`	`id_hh`	Id for the household
`hgxid`	`id`	Id for the individual
`hgsex`	`sex`	sex of the # household member
`hgage`	`age`	age of the # household member
`hhsize`	`hhsize`	size of the household
`hhstate`	`reg`	state of residence
`hhtype`	`hhtype`	household type
`hifditp`	`inc_disp`	disposable household income
`hifditn`	`inc_disp`	disposable household income
`hhrih`	`rel`	household relationships

- `hhrih`: household relationships
  1. [1] Couple with child < 15
  2. [2] Couple with depst (no child < 15)
  3. [3] Couple with ndepchld (no child < 15 or depst)
  4. [4] Couple without child
  5. [5] Lone parent with child < 15
  6. [6] Lone parent with depst (no child < 15)
  7. [7] Lone parent with ndechld (no child < 15 or depst)
  8. [8] Child < 15
  9. [9] Dependent student
  10. [10] Non-dependent child
  11. [11] Other family member
  12. [12] Lone person
  13. [13] Unrelated to all HH members
  14. [99] Not yet classified

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:

Original Name	New Name	Description
`_hhrhid`	`id_hh`	Id for the household
`xwaveid`	`id`	Id for the individual
`hgsex`	`sex`	sex of the # household member
`hgage`	`age`	age of the # household member
`tcnr`	`nch_nonres`	Number of non-resident child
`icniz`	`fert_int`	Fertility intentions

Original Name	New Name	Description
`tchad`	`nch`	N. children ever had
`hgage`	`age`	Age at interview
`hgage`	`age`	Age at interview
`hhiage`	`age_june`	Age before wave (june)
`edagels`	`edu`	Education
`hhwtrps`	`wht_rep`	Weight $\sum_i^n w = N$
`hhwtrp`	`wht_rep2`	Weight $\sum_i^n w = 15^{10*6}$
`wschave`	`wht_rep2`	UNSPECIFIED
`lsrel*`	`rel`	relat. and marriage biography

Weights

- `hhwtrps`: This is the cross-section responding person population weight rescaled to sum to the number of responding persons in the relevant wave (i.e. 13,969 in wave 1). Use this weight when the statistical package requires the sum of the weights to be the sample size.

- `hhwtrp`: The responding person weight is the cross-section population weight for all people who responded in the relevant wave (i.e. they provided an individual interview). The sum of these responding person weights for wave 1 is 15.0 million

Creating fertility biographies

A challenge faced in the data is that the HILDA does not have a childbearing biography, not speaking of the birth dates for the biological children. Nonetheless, one can obtain the childbearing information by combining information on 1) the biological children residing in the household (**household data**), 2) the deceased children using the **fertility major module**, and the biological children living outside the household, which is captured in the **fertility major module** as well.

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

1. **Births between waves:** Births between waves (w_t, w_{t-1}) can be estimated using the *person* data and the variable `nchild`. If the value of `nchild` changes over consecutive waves, than a childbirth has taken place. The year of birth is then randomly assigned by choosing between t and $t-1$. The age of the father when giving birth is than estimated by using the `age` variable.

- Quality checks: Using `nchihld` and `nch_nonres` together
- **Issue:** some people state in later waves lower number of children than in previous waves, which is not possible
- **Issue:** some people have more than a year in between waves so that that age at birth is uncertain

2. **Births before entering interview**: Some respondents may enter the survey after having received child already. In that case, the estimation of age at parenthood has to proceed differently. Instead, we obtain the information from the household grid and the grid on children living outside the household

2.1. *Children in the household* : Using the *household data*, we obtain information on the household members. We use the `age` variable to identify children. Then we use the `relationship` variable to specify parent-child relationship.

2.2. *Children in the household* : Using the *enumerate person data*, we obtain information on all household members. We use the `yob` variable to obtain the year of birth. Then we use the variables `id_fath` and `id_moth` to identify the fathers and the mothers of the children, which are than used to merge with the *person data.* The non-resident child grid contains information on the total number of non-resident children, the age of the youngest non-resident child, and the number of non-resident in wide age classess (classes = 0-4, 5-14, 15-24). The age at parenthood is then estimated by taking the difference between the the year of birth of the parent and the year of birth of the child:

$$age - parenthood = (year_{wave} - age_{resp,wave}) - yob_{child}$$

2.3. *Deceased child grid*: The fertility major module contains information on deceased children. The modules were asked in the waves 5, 8, 11, 15 and 19, and was cleaned and saved as `data/fert_deceased.Rda`. The deceased child grid contains information on whether the respondent liked to be asked about deceased children (`dc_permission`, which is stored in the fert data), a respondet has any deceased children (`dc_child_any`) the number of deceased children (`dc_child_any`), the respective year of birth (`yob_dc_child`) and also whether questions regarding the deceased children are permitted.

```
* Issue: The fertility modules is only asked every 4 years, which may in combination with panel attriti
```

2.4. *Non-resident child grid*: The fertility major module contains information on non-resident children. The modules were asked in the waves 5, 8, 11, 15 and 19. The non-resident child grid contains information on the total number of non-resident children (`nchild_non_res`), the age of the youngest non-resident child (`age_non_res_chi`), and the number of non-resident children in certain age-classes (pattern = `nchi_non_res_0_4`,classes = 0-4, 5-14, 15-24).

```
* Issue: The fertility modules is only asked every 4 years, which may in combination with panel attriti
```

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 %.