Tutorial #7
Additional variables

1

pairfam tutorial

7. Additional variables

Kristin Hajek & Madison Garrett, May 2022

The seventh tutorial explains all other variable types available in the anchor data in addition to the survey variables.

Preload variables

» Generated based on variables from previous wave
» Dependent interviewing
» *d0 – d663*
» Documentation in the codebooks
» Should NOT be used for analysis
» *d0*: participation in previous wave

2

What are preload variables?

Preload variables are included in the Scientific Use File from wave 2 onward. They were generated based on variables from the previous wave and contain information necessary for dependent interviewing.

The questionnaire in the current wave (more specifically: question wording, filters, etc.) depends, in part, on the articulation of these variables, which are easily recognizable by their form: a lowercase *d* followed by a number.

A detailed list of preload variables can be found in the anchor codebooks for each specific wave.

These variables should ***not*** be used for analysis;

however, the variable d0 can sometimes be useful for data preparation as it can be used to distinguish whether the anchor respondent participated in the previous wave, skipped the last wave, or took part in the survey for the first time in the current wave.

Auxiliary variables

» Generated during interview
» Used for filtering and question wording
    » *hp*: anchor has partner
    » *hpn*: name of partner mentioned
    » *erw1*: employment status
» Documentation in codebooks

3

What are auxiliary variables?

Auxiliary variables were generated during the interview and mainly used to modify filters and question wording.

For example, if a respondent reports a current partner during the survey, they will be posed further questions about this partner.

The auxiliary variable *hpn* stores the name of the current partner and is used as part of the question wording for further questions about this partner. Due to data protection laws, names were eliminated from the data and this variable only states whether a name was mentioned during the interview.

The variable *erw1* reflects whether the anchor indicated at least one occupation at the time of the current interview.

A full list of auxiliary variables is available in the codebooks for each wave.

Paradata

» Total number of contact attempts

» Interviewer gender

» Interviewer age

» Interviewer education level (from wave 5)

» Interviewer ID

» Interview duration in minutes

» Interview date (month/year)

4

What about paradata?

The following paradata is available in the anchor data sets: total number of contact attempts, interviewer gender, interviewer age, interviewer education level, interviewer ID, interview duration in minutes, and the interview date.

Generated variables

» Facilitate data usage
 → e.g., *reldur* for relationship duration
» Error-corrected
 → e.g., *sex_gen*, *doby_gen* as "best solutions"
» Syntax files (transparent, modifiable) available
» List of available variables in codebooks/question program
» Detailed documentation in Data Manual

5

Why does *pairfam* provide generated variables?

Generated variables are indicator variables generated during data processing to facilitate data usage. The generated variable *reldur*, for example, stores the relationship duration with the anchor respondent's current partner. This information is originally stored in individual Event History Calendar variables for each month of the previous year - that's a lot of variables across the panel that would need to be transformed to achieve the information included in this variable. As the EHC is very informative, but can also be difficult to use in analyses, we provide generated variables that summarize the key elements of the EHC to make analyses easier for you.

Generated variables are also error-corrected. Gender, birth year, and birth month for anchor and alteri respondents are recognizable by the suffix *_gen* and are "best solutions" over all waves. The original information of each wave is stored in the variables *original_sex* and *original_doby*. For example, if an anchor stated in three waves to be female and in one wave male, we chose the most plausible response based on the assumption of an incorrect entry – here, that the anchor is in fact female – and aligned this information over all waves in the variable *sex_gen*. Two actual changes of gender were communicated throughout the course of the panel to the interviewer. These anchor respondents received the missing code -4 for the variable *sex_gen*.

We highly recommend the usage of generated variables in relevant analyses. The

Stata syntax files for most of the provided generated variables are available as part of the Scientific Use File, making data processing more transparent. Furthermore, you can use the syntax to modify a specific generated variable and adjust it to your research question. Due to data protection measures, the syntax for some generated variables cannot be released. For example, the generation of the variable *isco*, which classifies occupations.

A list of all available variables can be found in the anchor and alteri codebooks as well as in the Question Program.

More detailed documentation concerning generated variables is available in the Data Manual.

Macrodata

» Federal state
» Municipality size classification
» Settlement structure (BIK)

Which macro variables are available?

Macrodata are retrospective context attributes concerning the anchor respondent's place of residence. Included in the Scientific Use File are federal state, municipality size at the anchor's main residence, and the settlement structure at both the anchor's and parents' main residences.

Weights

» Design weights and calibrated design weights
» For pairfam base, DemoDiff, and refreshment
» Adjustment to characteristics of general population
» Manage selective non-response (cross-sectional, longitudinal, cohort-specific)
» No weights available for step-up/alteri respondents
» Detailed documentation in the Data Manual and Technical Paper No. 17

7

Which weights are provided?

Design weights as well as calibrated design weights are available for the *pairfam* base, *DemoDiff*, and refreshment sample anchor respondents.

Weights provide factors to adjust the observed data to characteristics of the general population and manage selective non-response by assigning observations with characteristics of higher selectivity a higher analysis weight.

Therefore, both cross-sectional survey participation bias and longitudinal panel attrition bias for the following waves can be tackled.

Third, a correction of cohort-specific non-response aiming to represent actual cohort sizes in the population of interest can be integrated.

The design weight corrects disproportionate sampling across cohorts and the combination of multiple selection frames including *DemoDiff* and the wave 11 refreshment sample.

The calibrated design weight calibrates the design weights to reference characteristics, thereby correcting both baseline and longitudinal survey non-response.

No weights are available for *step-up* respondents or alteri respondents.

A detailed documentation of the weighting factors can be found in the Data Manual and in Technical Paper No. 17.

Flag & tag variables

Flag variables:

» Inconsistent interview entries

e.g., marriage before beginning of relationship, net income larger than gross income

Tag variables:

» Inconsistencies between waves

e.g., birth data/gender differ from previous wave for anchor/partner/child(ren)
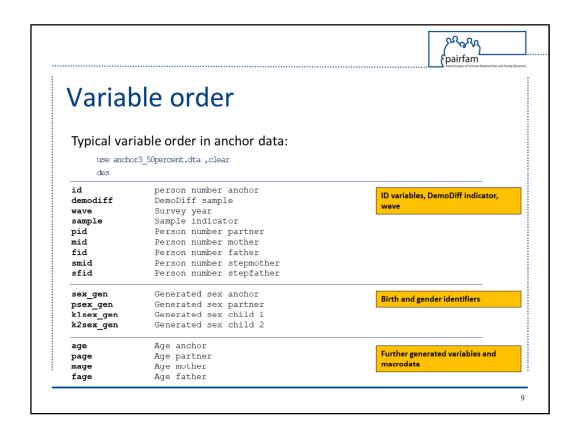
8

What are flag and tag variables?

Flag variables mark inconsistent interview entries in one wave, for example if a marriage reportedly began before the relationship with the same partner, or if a respondent's net income is higher than the gross income.

Tag variables, on the other hand, mark inconsistencies between waves, for example if the birth data or gender differs from previous waves for anchor, partners, or children.

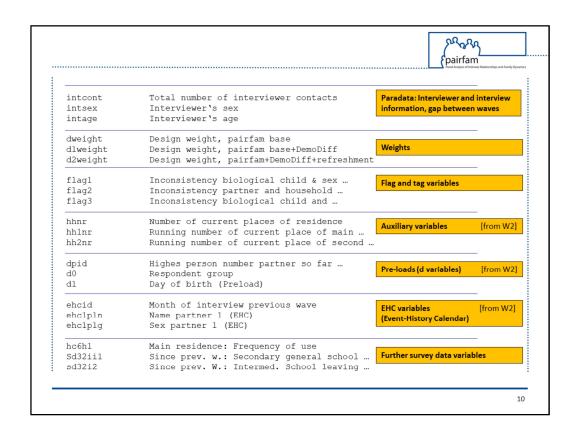A full list of flag and tag variables is available in the Data Manual.

Variable order

Typical variable order in anchor data:

```
use anchor3_50percent.dta ,clear
des
```

id	person number anchor	**ID variables, DemoDiff indicator, wave**
demodiff	DemoDiff sample	
wave	Survey year	
sample	Sample indicator	
pid	Person number partner	
mid	Person number mother	
fid	Person number father	
smid	Person number stepmother	
sfid	Person number stepfather	
sex_gen	Generated sex anchor	**Birth and gender identifiers**
psex_gen	Generated sex partner	
k1sex_gen	Generated sex child 1	
k2sex_gen	Generated sex child 2	
age	Age anchor	**Further generated variables and macrodata**
page	Age partner	
mage	Age mother	
fage	Age father	

9

Is there a specific order to the variables?

All anchor data sets have the same variable order. Note that this table does not include the full list of variables, but only an excerpt of each variable type.

Identification variables are listed first, followed by sample indicators and the survey year.

Next are the generated birth and gender identifiers, followed by further generated variables, for example age, and macro variables.

Afterwards, paradata are listed, for example the number of interviewer contacts and interviewer's gender and age.

The weighting variables then follow along with the flag and tag variables.

Next are the auxiliary and preload variables.

Last but not least – in fact the majority of variables – the Event History Calendar variables followed by the rest of survey data variables.

Next up:
Tutorial #8 – Missing
values and filter missings

This concludes the description of available variable types in the *pairfam* data.

The next tutorial will cover missing values and filter missings.