

What are
hierarchical
files?

UK Data Service





Author: UK Data Service
Updated: August 2014
Version: 1

We are happy for our materials to be used and copied but request that users should:

- link to our original materials instead of re-mounting our materials on your website
- cite this as an original source as follows:

Sarah King-Hele and Jo Wathan. *What are hierarchical files?*. UK Data Service, University of Essex and University of Manchester.



Contents

| | | |
|------|---|----|
| 1. | What is a hierarchical dataset? | 3 |
| 2. | Example 1: hierarchical data in data files | 3 |
| 3. | Example 2: a more complex hierarchical dataset | 6 |
| 4. | Example 3: choosing a unit of analysis | 7 |
| 5. | Data manipulation techniques for hierarchical datasets | 9 |
| 5.1. | Creating a household dataset from an individual dataset by selecting one individual per household | 9 |
| 5.2. | Creating household level summary variables from individual level data | 10 |
| 5.3. | Attaching household data to an individual level file | 12 |

Figures

| | | |
|-----------|--|---|
| Figure 1: | an example of two households | 3 |
| Figure 2: | an example with a more complex hierarchy | 6 |

Tables

| | | |
|----------|--|---|
| Table 1: | Figure 1 data in a single file | 4 |
| Table 2: | household level file containing household variables in Figure 1 data | 5 |
| Table 3: | person/individual level file containing person level variables and the household ID for linking with the household dataset | 5 |
| Table 4: | Example 3 dataset | 7 |
| Table 5: | household and individual level questions and their answers | 7 |



1. What is a hierarchical dataset?

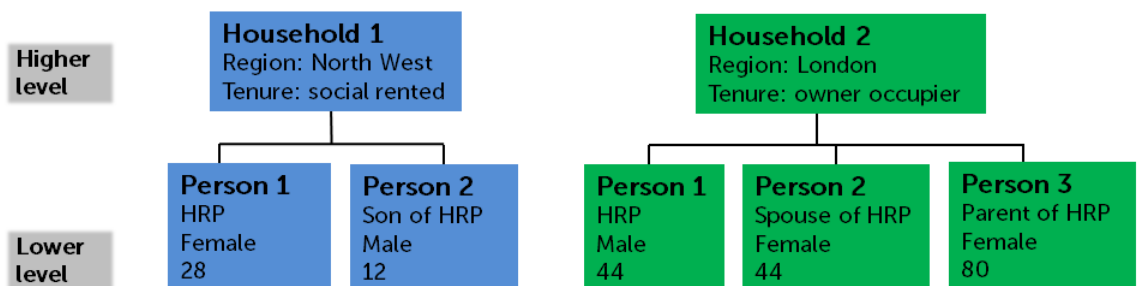
Survey data may be collected from individuals, households, families, companies or other *units*. A *hierarchical dataset* contains information about more than one unit in which one unit is nested inside another - for example, data about:

- individuals within households
- employees within businesses
- individuals within families within households

2. Example 1: hierarchical data in data files

Figure 1 shows two example households. Household 1 is in the North West, rents social housing and contains a 28 year old woman and her 12 year old son. Household 2 is in London, is owner occupied and consists of a couple in their 40s and the 80 year old mother of the husband. One person in each household is the HRP (Household Reference Person), a term commonly used in survey data to indicate a member of the household who is equivalent to a head of household¹.

Figure 1: an example of two households



In this example there are two levels of hierarchy: households are at the *higher level* and individuals at the *lower level*.






The information shown in Figure 1 can be stored as a single dataset (see Table 1) or as two datasets (see Tables 2 and 3). Both formats are very common and each format holds the same information.

Table 1 shows the information in a single dataset. Each row contains information about the individuals in the two households. This means that, by default, analyses conducted with the data in this format are at the individual/person level. For discussion about how to choose a unit of analysis with hierarchical data, see Section 3.

Each column contains a variable. *Individual level variables* contain characteristics that vary from person to person e.g. age, gender, and *household level variables* contain characteristics that vary from household to household but which are the same for all individuals within each household, e.g. region, housing tenure. This format is called 'rectangular' hierarchical data or a flat file.



Table 1: Figure 1 data in a single file

| Default unit of analysis (person) | HHID | Region | Tenure | PersonID | Relation to HRP | Sex | Age |
|---|------|--------|---------|----------|-----------------|-----|-----|
|  | 1 | NW | SocRent | 1 | HRP | F | 28 |
|  | 1 | NW | SocRent | 2 | Son | M | 12 |
|  | 2 | London | OwnOcc | 1 | HRP | M | 44 |
|  | 2 | London | OwnOcc | 2 | Spouse | F | 44 |
|  | 2 | London | OwnOcc | 3 | Parent | F | 80 |

Household level variables
Individual level variables

There are two ID (identity) variables in this dataset: *HHID* uniquely identifies households and *PersonID* uniquely identifies individuals within households. You can create a unique individual identifier by combining *HHID* and *PersonID*.

ID variables are important in hierarchical data because they allow you to identify which individuals are in each household. If you can link each person to their household, you can link members of the same household to each other. It is therefore possible using the information about both individuals in Household 1 to discover that Person 1 in Household 1 has her 12 year old son living with her.

A note about terminology

Note the difference between the following terms:

Individual level variable: is a variable containing characteristics that can vary from one individual to another e.g. ethnicity, age, person ID

Individual level analysis: an analysis conducted on a dataset with individuals in each row. The unit of analysis is the individual. The research question is about individual differences (as opposed to household differences) e.g. what is the relationship between poor health and age in the UK? You can include household level variables in individual level analyses with hierarchical data. The research question is still about individual differences but examines both individual and household characteristics e.g. what is the relationship between poor health, housing tenure and age in the UK?

Similarly:

Household level variable: is a variable containing characteristics that can vary from one household to another but which are the same for all individuals within each household. e.g. housing tenure, region, household ID

Household level analysis: an analysis conducted on a dataset with households in each row. The unit of analysis is the household. The research question is about household differences (as opposed to individual differences) e.g. how does household income vary by region in the UK?





The same information can be presented as two separate datasets, as shown in Tables 2 and 3. This format is more common when the data are more complex or old and is slightly more efficient as there is less replication of data than in the single file format. Each row in Table 2 contains information about households and all the variables are household level variables. Table 3 contains information about individuals in each row and all the variables are individual level variables, apart from the household ID, HHID.

Analyses conducted with the data in Table 2 are at the household level by default, and analyses conducted with the data in Table 3 are at the individual/person level. See Section 3 for discussion about choosing a unit of analysis when using hierarchical data.






If you wish to conduct analyses at the individual level using variables about individuals *and* households, you must first link together the two files using the household ID *HHID* to create a dataset that looks like that in Table 1 (see Section 5.3 for how to do this).

Table 2: household level file containing household variables in Figure 1 data

| Default unit of analysis (household) | HHID | Region | Tenure |
|---|------|--------|---------|
|  | 1 | N W | SocRent |
|  | 2 | London | OwnOcc |

Household level variables

Table 3: person/individual level file containing person level variables and the household ID for linking with the household dataset

| Default unit of analysis (person) | HHID | PersonID | Relation to HRP | Sex | Age |
|---|------|----------|-----------------|-----|-----|
|  | 1 | 1 | HRP | F | 28 |
|  | 1 | 2 | Son | M | 12 |
|  | 2 | 1 | HRP | M | 44 |
|  | 2 | 2 | Spouse | F | 44 |
|  | 2 | 3 | Parent | F | 80 |

Household ID variable Individual level variables

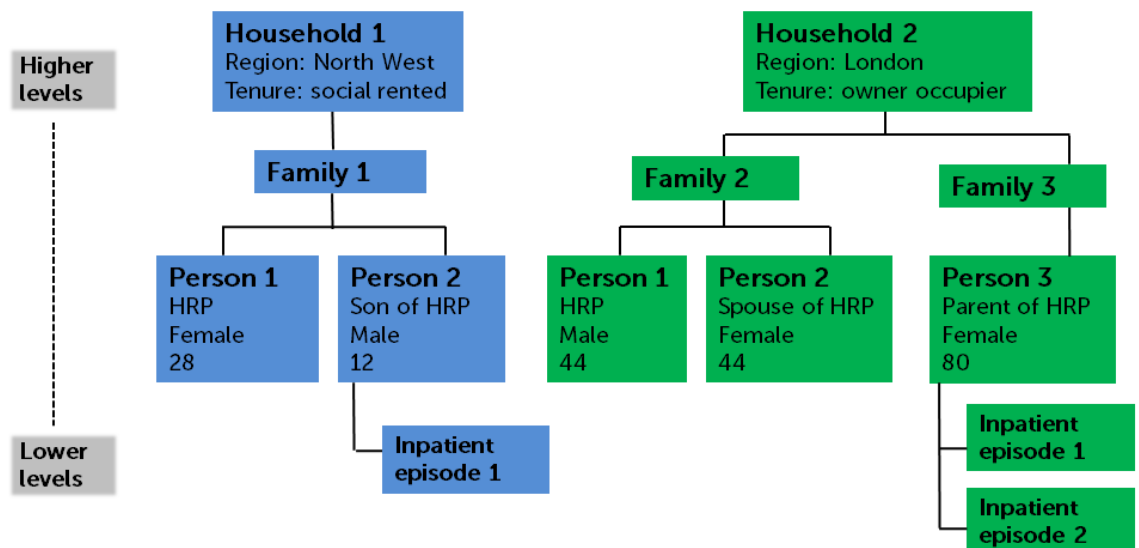


3. Example 2: a more complex hierarchical dataset

The structure of hierarchical data can be more complex. There are often intermediate units within a household such as families or 'family units' for example. Some surveys may also be used to explore units below the individual level (e.g. to look at hospital inpatient stays, or crime events).

Figure 2 shows a dataset with four levels of hierarchy. The household is the highest level of hierarchy in this example and the inpatient episode is the lowest level and there are two other levels, the 'person' level and the 'family' level.

Figure 2: an example with a more complex hierarchy



Note that not all four levels apply to all individuals:

- **there are three levels for the part of the dataset that covers individuals with no inpatient episodes:** individuals within families within households, and
- **there are four levels for people with inpatient episodes:** hospital inpatient episodes nested within individuals, nested within families, nested within households.

Note about definitions of units in surveys

Always look at the survey documentation for the definition of units in hierarchical datasets. A household or family may be defined differently in different studies, for example.

Note that many of the large UK surveys use definitions of units that are the same across a number of different datasets to make comparison of results between surveys easier. The [Office for National Statistics \(ONS\)](#) website contains definitions of 'harmonised' concepts commonly used in surveys, though you should always refer to the documentation for the survey you are using for the definitions used in that survey.









4. Example 3: choosing a unit of analysis

If data contains files at different units of analysis, then you may analyse it at any of the levels available (or at more than one level) as long as your choice makes sense.

The unit of analysis most appropriate to your research will depend on the research question that you wish to answer. It is often possible to examine similar topics at different units of analysis. For example, you could use the individual level to consider the number of people who live in poverty, or the household level to consider how many households live in poverty. Often the research question will indicate which unit to use.

Identifying the unit that you wish to use will help you to ensure that you select an appropriate dataset and that you are using the dataset correctly. Results that have differing interpretations can be obtained on the same topics depending on which unit of analysis you use. Consider the example of a dataset below. It is the same as that in Table 1 with an additional household containing a 35 year old male owner-occupier living on his own in London.



Table 4: Example 3 dataset

| Default unit of analysis (person) | HHID | Region | Tenure | PersonID | Relation to HRP | Sex | Age |
|---|------|--------|---------|----------|-----------------|-----|-----|
|  | 1 | NW | SocRent | 1 | HRP | F | 28 |
|  | 1 | NW | SocRent | 2 | Son | M | 12 |
|  | 2 | London | OwnOcc | 1 | HRP | M | 44 |
|  | 2 | London | OwnOcc | 2 | Spouse | F | 44 |
|  | 2 | London | OwnOcc | 3 | Parent | F | 80 |
|  | 3 | London | OwnOcc | 1 | HRP | M | 35 |

Household ID variable
Individual level variables

The following gives the answers to similar questions at the household and the individual levels based on the example dataset in Table 1.

Table 5: household and individual level questions and their answers

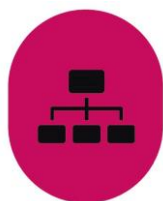
| Household level analyses  | Individual level analyses  |
|--|---|
| What proportion of households contains only 1 person? $1/3=33.3\%$ | What proportion of individuals lives alone? $1/6=16.7\%$ |
| What is the mean household size? $(2+3+1)/3=6/3=2$ | What is the average household size for individuals in the sample? $(2+2+3+3+3+1)/6=14/6=2.33$ |



The results in the two levels of analyses but with similar questions are quite different. The percentage of households with one resident is twice the percentage of individuals who live alone. Similarly individuals on average live in households of size 2.3, however because nearly a third of households are single person households the mean size *of households* is 2.

Weighting

Note that if you are analysing the data at a different level from that intended by the data producer, your results may not be representative of the population if a survey weight is needed and there is no appropriate weight provided with the data. See our [*What is weighting?*](#) guide for more information about weighting survey data.



5. Data manipulation techniques for hierarchical datasets

Once you have your research question(s), if the data are not already in the format you need for your analyses, you must rearrange them to create a dataset you can use. The examples use the data from Figure 1 with household data at the higher level and individual data at the lower level. The same techniques apply to other hierarchical datasets.






5.1. Creating a household dataset from an individual dataset by selecting one individual per household

If you have an individual level dataset containing a household ID variable, you can create a household level dataset (to do household level analyses) by selecting one individual per household (e.g. the HRP) and using their data to represent the household.

This approach works well for household level variables (e.g. region or housing tenure) but for inherently individual characteristics such as ethnicity and sex, selecting one person to stand for a whole household may make less sense. However, note that the variable 'age' in an individual level dataset is an individual level variable, but when you create a household dataset using only the HRP in each household, 'age' now means 'age of HRP', a household level variable. 'Age of HRP' may be a useful concept to you or not (this will depend on your research question).

This process starts with the following file (the one dataset version of hierarchical data shown in Table 1):

Table 1: Figure 1 data in a single file



| Default unit of analysis (person) | HHID | Region | Tenure | PersonID | Relation to HRP | Sex | Age |
|---|------|--------|---------|----------|-----------------|-----|-----|
|  | 1 | NW | SocRent | 1 | HRP | F | 28 |
|  | 1 | NW | SocRent | 2 | Son | M | 12 |
|  | 2 | London | OwnOcc | 1 | HRP | M | 44 |
|  | 2 | London | OwnOcc | 2 | Spouse | F | 44 |
|  | 2 | London | OwnOcc | 3 | Parent | F | 80 |

Household ID variable Individual level variables

After selecting only PersonID=1 and dropping all the individual level variables the file created is:



Table 2: household level file containing household variables

| Default unit of analysis (household) | HHID | Region | Tenure |
|---|------|--------|---------|
|  | 1 | N W | SocRent |
|  | 2 | London | OwnOcc |

Household level variables

Finding a variable that identifies an individual in each household

There will normally be some variable that will allow you to select one individual per household even if there is no HRP or Head of household identified in the data. For example, if individuals are numbered PersonID=1, 2... x in each household (where x is the size of the household) as in Examples 1 and 2 above, then just choose PersonID=1 in each household. If HRPs are defined in the data, you may see the HRP identified in variables of different formats such as:

- a binary variable that indicates whether the respondent is the HRP or not
- a variable which indicates the person number of the HRP. You will then need to test whether the respondent's own person number is the same as that of the HRP
- a variable that indicates the respondent's relationship to the HRP

To create a household file from individual file using SPSS or Stata, see:






- Section 5.1 of our 'What is SPSS for Windows?' guide
- Section 7.1 of our 'What is Stata?' guide

5.2. Creating household level summary variables from individual level data

With the data from Figure 1 which contains an age variable for each household member, you could create a new household level variable entitled 'age of oldest person in the household' derived by using the 'age' variables for each individual and the HHID variable. The original data must be in the flat file format (as shown in Table 1).

The new variable is either created as a new household level dataset with HHID or as an additional variable in the original data.



So the original data looks like this:

| Default unit of analysis (person) | HHID | Region | Tenure | PersonID | Relation to HRP | Sex | Age |
|---|------|--------|---------|----------|-----------------|-----|-----|
|  | 1 | NW | SocRent | 1 | HRP | F | 28 |
|  | 1 | NW | SocRent | 2 | Son | M | 12 |
|  | 2 | London | OwnOcc | 1 | HRP | M | 44 |
|  | 2 | London | OwnOcc | 2 | Spouse | F | 44 |
|  | 2 | London | OwnOcc | 3 | Parent | F | 80 |

Household ID variable Individual level variables








And the new variable 'age of oldest person in household' is created in either a new household dataset:

| Default unit of analysis (household) | HHID | Age of oldest person in HH |
|---|------|----------------------------|
|  | 1 | NW |
|  | 2 | London |

Household ID variable + new summary variable

Or as an additional variable in the original dataset:

| Default unit of analysis (person) | HHID | Region | Tenure | PersonID | Relation to HRP | Sex | Age | Age of oldest person in HH |
|---|------|--------|---------|----------|-----------------|-----|-----|----------------------------|
|  | 1 | NW | SocRent | 1 | HRP | F | 28 | 28 |
|  | 1 | NW | SocRent | 2 | Son | M | 12 | 28 |
|  | 2 | London | OwnOcc | 1 | HRP | M | 44 | 80 |
|  | 2 | London | OwnOcc | 2 | Spouse | F | 44 | 80 |
|  | 2 | London | OwnOcc | 3 | Parent | F | 80 | 80 |

Household ID variable Individual level variables New summary variable

It is possible to extend this kind of analysis using other functions. Other examples of the kinds of summary variables you could create with this set of variables include:

- the age of the youngest person in the household – using the minimum (min) over household ID
- the average age of males in the household – using the average (mean) over the household ID
- the number of people aged 65+ in the household (or aged 10-16 years) – by creating a new variable to equal 1 if people are aged 65+ (or aged 10-16 years) and 0 otherwise, and then counting (count) over household ID

Summary variables can summarise any aspect of a household as long as the data are available to create the variable. For example:

- the number of people in part-time employment in the household
- the average income of households with an HRP with a disability (and the average income of households with an HRP without a disability)



To create a summary variable using SPSS or Stata, see:

- Section 5.2 of our 'What is SPSS for Windows?' guide
- Sections 7.2 and 7.3 of our 'What is Stata?' guide



5.3. Attaching household data to an individual level file

Suppose you have data at a higher level (e.g. household) that you wish to attach to a lower level (e.g. individual) in a hierarchy. Such a situation routinely arises when dealing with hierarchical data if:






- Your data are held in multiple data tables which need to be matched in order to use them together.
- You have created a household level dataset (as demonstrated in Section 5.2 above) from an individual level dataset, and now want to re-attach this data to the individual file.

The household level and person level tables each contain a household ID variable. This ID enables you to relate the household level records in the household file to the cases in the person level file. In this way, you can make a single data file that contains data from both tables. In effect, this is moving from the 2-file format of hierarchical data shown in Tables 2 and 3 to the 1-file format in Table 1.

This process starts with the following files (the two dataset version of hierarchical data shown in Tables 2 and 3):

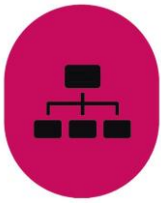
| Default unit of analysis (household) | HHID | Region | Tenure |
|---|------|--------|---------|
|  | 1 | NW | SocRent |
|  | 2 | London | OwnOcc |

Household level variables






| Default unit of analysis (person) | HHID | PersonID | Relation to HRP | Sex | Age |
|---|------|----------|-----------------|-----|-----|
|  | 1 | 1 | HRP | F | 28 |
|  | 1 | 2 | Son | M | 12 |
|  | 2 | 1 | HRP | M | 44 |
|  | 2 | 2 | Spouse | F | 44 |
|  | 2 | 3 | Parent | F | 80 |

Household ID variable Individual level variables

After linking the data by adding in household variables to the individual file, the file created is



that in Table 1:

| Default unit of analysis (person) | HHID | Region | Tenure | PersonID | Relation to HRP | Sex | Age |
|---|------|--------|---------|----------|-----------------|-----|-----|
|  | 1 | NW | SocRent | 1 | HRP | F | 28 |
|  | 1 | NW | SocRent | 2 | Son | M | 12 |
|  | 2 | London | OwnOcc | 1 | HRP | M | 44 |
|  | 2 | London | OwnOcc | 2 | Spouse | F | 44 |
|  | 2 | London | OwnOcc | 3 | Parent | F | 80 |

Household ID variable
Individual level variables

To do this you would take the maximum value of age by the household. Both SPSS and Stata allow you to calculate values over an ID variable like household ID.

To attach household data to an individual file using SPSS or Stata, see:

- Section 5.3 of our 'What is SPSS for Windows?' guide
- Section 7.3 of our 'What is Stata?' guide

ⁱ Where there are joint householders, the HRP is defined by the Office for National Statistics (who produce many of the UK large-scale surveys) as the individual with the highest income. If the joint householders have the same income, the eldest is selected. The HRP can be used in analysis to 'represent' the household in terms of income, say, or to select a single member of each household.

11 August 2014

T +44 (0) 1206 872143
E help@ukdataservice.ac.uk
W ukdataservice.ac.uk

The UK Data Service delivers
quality social and economic
data resources for researchers,
teachers and policymakers.

© Copyright 2014
University of Essex and
University of Manchester

UK Data Service

