Otto-Friedrich-Universität Bamberg

Prof. Dr. Timo Schmid
timo.schmid@uni-bamberg.de

Exercises - Surveys and datasets (Chapter 3)

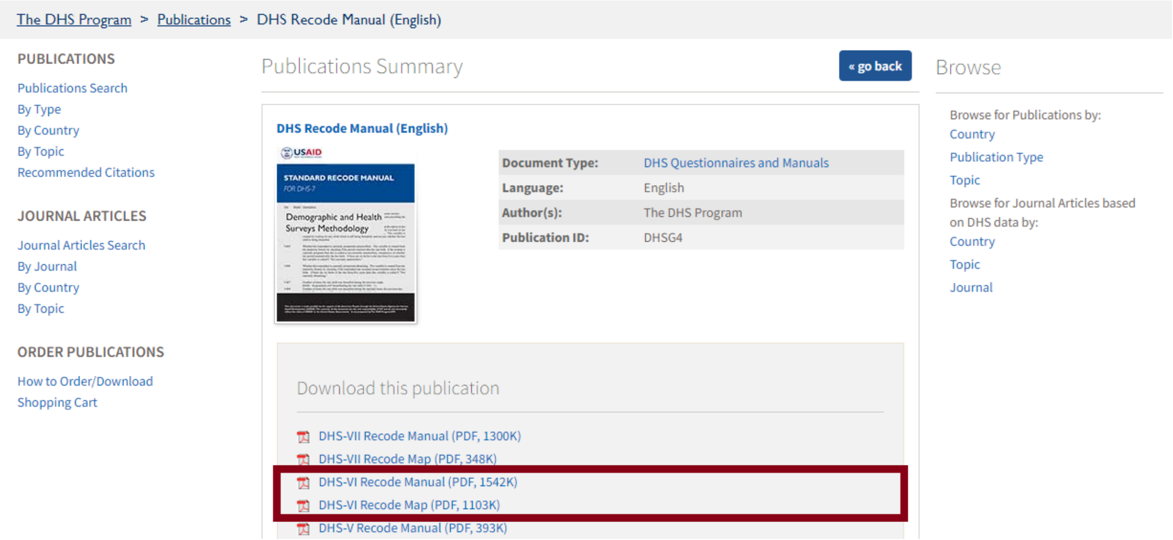
Small Area Estimation I

In the following exercise, we explore the model dataset provided by the Demographic and Health Surveys (DHS) Program. The DHS Program collects, analyzes, and disseminates representative data on population, health, HIV, and nutrition across more than 90 countries. To help users practice and become familiar with the datasets, the DHS Program provides model datasets that are freely available on its website; these datasets do not represent actual data from any country. For more information, see https://dhsprogram.com/data/Model-Datasets.cfm.

Exercise 1: The DHS model datasets contains records related to various topics. Download the following files from https://dhsprogram.com/data/Download-Model-Datasets.cfm.

- zzfulltables.zip (Model Datasets Full Report Tables and Sampling Errors),
- zzhr62dt.zip (Household Recode),
- subregional.zip (Boundaries).

Information for the variables in the datasets is available on https://dhsprogram.com/publications/publication-dhsg4-dhs-questionnaires-and-manuals.cfm. Download the following PDF files under DHS-VI Recode Manual and DHS-VI Recode Map, as illustrated below.



The DHS Program > Publications > DHS Recode Manual (English)

PUBLICATIONS
Publications Search
By Type
By Country
By Topic
Recommended Citations

JOURNAL ARTICLES
Journal Articles Search
By Journal
By Country
By Topic

ORDER PUBLICATIONS
How to Order/Download
Shopping Cart

Publications Summary

« go back

DHS Recode Manual (English)

Document Type:	DHS Questionnaires and Manuals
Language:	English
Author(s):	The DHS Program
Publication ID:	DHSG4

Download this publication

- DHS-VII Recode Manual (PDF, 1300K)
- DHS-VII Recode Map (PDF, 348K)
- DHS-VI Recode Manual (PDF, 1542K)
- DHS-VI Recode Map (PDF, 1103K)
- DHS-V Recode Manual (PDF, 393K)

Browse

Browse for Publications by:
Country
Publication Type
Topic
Browse for Journal Articles based on DHS data by:
Country
Topic
Journal

Exercise 2: Load the household survey data (`ZZHR62FL.DTA`) into a new `R` script. The data includes more than 7000 variables. For further analyses, remove any unnecessary variables. Follow the steps below:

a. Many variables in the dataset contain only missing values. Investigate why these variables have only missing values by comparing `hh012` and `hv101_n`. Remove any variables that contain only `NA` values from the dataset.

b. We will use the household basic data, household members' basic data, and household characteristics variables for further exercises. Remove all other variables.

c. Remove any unnecessary index variables (`hvidx_1` to `hvidx_24`).

Exercise 3: To gain insights into the variables in the dataset, answer the following questions based on the dataset and the variable description file.

a. What do the variables `hv005` and `hv012` represent?

b. `hv025` denotes the type of place of residence (urban/rural) for each household. Which unique values are included in `hv025`?

c. Specify the number of regions included in the dataset.

d. How many households are living in each region?

e. Examine the distribution of rural and urban households in each region.

Exercise 4: Next, we want to examine the average number of children who are 5 years old or younger (`hv014`). Follow the steps and answer the questions.

a. Calculate the mean of `hv014`.

b. How is the weighted mean defined?

c. Calculate the weighted mean of `hv014` using the sample weights `hh005`. Do the mean and the weighted mean yield the same result?

d. Calculate the mean and weighted mean of `hv014` for rural and urban areas. Compare the results.

Exercise 5: Lack of access to clean water is one of the most critical issues, especially in poorer countries. The DHS program collects information from the households to provide relevant indicators and summary statistics. Table 2.1 in `2. Housing Characteristics and Household Population.pdf` (available in `zzfulltables.zip`) provides descriptive statistics on various household water variables in rural and urban areas. To reproduce this table using the household dataset, follow these steps.

a. Identify the variable that contains information about the source of drinking water.

b. Determine the number of unique values in this variable and compare them with those in Table 2.1.

c. Adjust the variable so that it aligns with the expressions in the table.

d. Create the weighted frequency table (using the function `fre()`) for the adjusted variable and compare the results with Table 2.1 (column `Total` under `Households`).

e. Calculate the weighted frequency of households for rural and urban areas using the same variable.

f. Calculate the weighted frequencies of individuals in the population in urban, rural, and total areas, using the household member count (`hv012`) for the calculation.

g. Repeat the steps from a. to f. with variable `hv204` and reproduce the table for the time to obtain drinking water in Table 2.1.

Exercise 6: We are now interested in the district-specific percentages of households with improved drinking water sources. The synthetic district indicator is contained in `ind_district.RData`. Follow the steps to obtain and map the results.

a. Load `ind_district.RData` and add the district indicator to the household dataset.

b. Create a binary variable `imp_water` to indicate whether the household has an improved water source. The definition of improved versus non-improved sources can be found in Table 2.1 of `2. Housing Characteristics and Household Population.pdf`.

c. Calculate the weighted percentage of households with an improved drinking water source for each district. Save the results in a data frame.

d. Load the shape file `model_subregional_boundaries.shp` and integrate the data frame from step c. into the map data using the function `merge()`.

e. Plot the percentage of households in each district obtained from c. on the map.

Exercise 7: The household dataset also includes information about household members. For example, `hv108_n` represents the education level of the `n`th household member in years. Using these variables, we can create useful variables at the household level. Create a new variable `avg_educ` which represents the average completed education level in years for each household based on `hv108_n`.

Exercise 8: `hv271` contains the wealth index as a measure of relative wealth on a continuous scale. We aim to analyze the wealth index using a linear regression model. Follow the steps and answer the questions.

a. Fit a linear regression model with the independent variables `imp_water, avg_educ, hv025`.

b. Does access to an improved drinking water source have a significant effect on the wealth index?

c. Interpret the β coefficients of the model.

d. Check whether the model assumptions are satisfied.

e. Find other potential independent variables. Also, consider creating new variables as described in the previous exercise. How can we determine an optimal set of independent variables?