N.i.D.S.

NATIONAL INCOME DYNAMICS STUDY

Wave 1

Update: September 2013

Introduction to NIDS data

SALDRU
southern africa labour and development research unit
UNIVERSITY OF CAPE TOWN

University of Cape Town • iYunivesithi yaseKapa • Universiteit van Kaapstad

The National Income Dynamics Study (NIDS) used a combination of household and individual level questionnaires. The data from the different questionnaires is recorded in separate files. The files are STATA files with one row per record. The data can be exported into most standard statistical packages.

1. General notes

Household membership, residency and panel membership: All resident household members at selected dwelling units were included in the NIDS panel, providing that at least one person in the household agreed to participate in the study. The household roster in the household questionnaire was used to identify potential participants in the study. Firstly, respondents were asked to list all individuals that have lived under this "roof" or within the same compound/homestead at least 15 days during the last 12 months OR who arrived in the last 15 days and this was now their usual residence. In addition the persons listed should share food from a common 'pot' and share resources from a common resource pool. All those listed on the household roster are considered household members.

All *resident* household members became NIDS panel members. Household members are resident if they spend at least 4 nights a week in the household. In addition, non-resident members that were "out of scope" at the time of the survey also became NIDS panel members. Out-of-scope household members were those living in institutions (such as boarding school hostels, halls of residence, prisons or hospitals) which were not part of the sampling frame.

Unique identifiers: The household identifier (*hhid*) is a six digit number that uniquely identifies all Wave 1 NIDS households. The individual identifier (*pid*) is a six digit number that uniquely identifies NIDS panel members. In the Household Roster file there are 2918 non-resident household members that were not given an individual interview, but from Version 5.0 of the data they do have *pids* allocated. Their information is only recorded in the household roster in Wave 1.

Data Structure: The questionnaires and their corresponding files are as follows:

- ***HouseholdQ*:** One record per household with data from the household questionnaire, excluding the household roster. Unique identifier: *hhid* (n= 7296).
- ***HouseholdRoster*:** One record per household member (n=31144) with data from the household roster (Section B of household questionnaire). Unique identifier for household: *hhid* (n = 7296), unique identifier for individual *pid* (n= 31141).
- ***Adult*:** One record per entry from the adult[1] questionnaire. Unique identifier for household: *hhid* (n=7289[2]), unique identifier for individual: *pid* (n=16871); 1241observations have no data beyond Section A as these individuals refused to participate in the survey. These records have a value in *w1_a_refexpl*, while participating adults are "system missing" for this variable (for reason for refusal).
- ***Proxy*:** One record per entry from the proxy[3] questionnaire. Unique identifier for household: *hhid* (n=1375), unique identifier for individual: *pid* (n=1750).
- ***Child*:** One record per entry from the child questionnaire. Unique identifier for household: *hhid* (n=4328), unique identifier for individual: *pid* (n=9605); 209 observations have no data beyond Section A as these individuals refused to participate in the survey. These records have a value in *w1_c_refexpl*, while participating children are "system missing" for this variable (for reason for refusal).

Derived variables are variables that were not asked directly of the respondent, but which were calculated or imputed from other information. For example, aggregate income and expenditure variables were constructed. Very few derived variables are included in the questionnaire files. Most of the derived variables are in the individual derived or household derived files. The STATA do-files that were used to create these files are available on **http://www.nids.uct.ac.za/home/index.php?option=com_docman&Itemid=19.** These derived data files are part of the NIDS Wave 1 dataset:

***hhderived*:** One record per household. Unique identifier for household: *hhid* (n=7296). Geographic information of the current location of households and the weights variables are included in this file.

***indderived*:** One record per NIDS panel member. There is a unique identifier for each household: *hhid* (n=7296) and a unique identifier for individual: *pid* (n=28226). Included are a number of "best" variables. These variables were created for ease of use where information is collected in more than one place, i.e. Date of Birth or education, and discrepancies arise between the different sources of information. The "best" variable represents the best-estimate based on a number of rules. For example, information from the individual questionnaire is considered more correct than the roster, but complete dates of birth are more correct than partial dates of birth.

More information about the detailed content of each of the files and variables can be found in the NIDS Wave 1 Variable Codebook.

Merging: Different pieces of information are contained in the different NIDS datasets and therefore it might be necessary to merge the datasets in order to use different variables together. When merging multiple datasets you should start by opening the file that contains all the records, for example indderived or Householdroster. Next merge in the individual level files. Individual level information can also be merged with household level information using the appropriate unique identifiers (*hhid* and *pid).*

When merging individual level data-sets together, these should be merged on the key variable *pid*. When merging in household level data, the appropriate household identifier is *hhid*.

Variable Codebook
In the variable codebook each questionnaire is treated as a section and contains a description of each variable in that data file. The format for the description is:

[Variable label]
Variable: *[variable name]*
[(Derived variable)]
[(Text variable)]
[Notes to users]

[Code list:]
[value1] [value label1]
[value2] [value label2]
...

Variable names: The variable names are constructed as follows: *w1_questionnairetype_sectionQuestion*
For example:

w1_a_bhgen3

- w1 – wave 1
- a - adult questionnaire
- bh – birth history section of the questionnaire
- gen3 - gender of child 3

Text variables are non-numeric response, such as responses to "Other(Specify)".
Code list gives the list of valid responses.
Notes to users can contain information on the variable, including construction, panel formatting, interviewer instructions, and confidentiality and relevant time periods.

Non-response codes: Non response codes are usually indicated with negative numbers. The only exception is dates where four digits were used for years and two digits for months.

-3 – Missing	3333 – Missing year	33 – Missing month
-5 – Not Applicable	5555 – Not Applicable year	55 – Not Applicable month
-8 – Refused	8888 – Refused year	88 – Refused month
-9 – Don't Know	9999 – Don't know year	99 – Don't know month

Missing (-3) indicates that a question was supposed to have been answered, but was not. A system missing (.) indicates that a skip pattern was enforced and that no data had to be collected.

Date format: The full date in all date variables was split into day, month and year and is a numeric variable. In all datasets, the day in date of birth variables has been excluded in order to protect the confidentiality of the respondents.

Multiple Mention Responses: Multiple mentions resulted where respondents were given the option to give more than one response to a question. The responses are recorded in ascending order in the multiple mention variables, for example the first response is recorded in the first variable, the second response in the second variable and so on. The number of variables is determined by the respondent who selected the most options.

2. Sampling and Weights

Before analysis and report-writing on the NIDS data could begin it was necessary to calculate sampling weights. Professor Martin Wittenberg at the University of Cape Town was asked to calculate these weights for NIDS. Technical Paper Number 2 "Calculating the NIDS weights" details the methodologies and assumptions made when calculating the weights.

This is essentially a two stage procedure. In the first stage, the design weights were calculated as the inverse of the probability of inclusion. In the second, the weights were calibrated to the 2008 mid-year population estimates. Two sets of weights are thus provided, the design weights and the post-stratification weights.

The basis of the calculation of the design weights is the information that Stats SA provided to NIDS about the process of two-stage sampling from the Master sample. Two sets of calculations were necessary in deriving the design weights. First there is a calculation of the probability of sampling each PSU and, second, there is a calculation about the probability of including each specific household in each PSU in the NIDS sample. The latter corrects for household non-response.

The second set of weights are the post-stratification weights. These weights adjust the design weights such that the age-sex-race marginal totals in the NIDS data match the population estimates produced by Stats SA for the Mid Year Population Estimates for 2008. In addition, we imposed the constraint that the population distribution by provinces should correspond to that released in those population estimates and that the total weights should add up to the estimated total population of 49,561,256 (updated in 2013 for new mid-year population estimates). Finally, a further constraint imposed was that the weights should be constant within households.

The recommended weight variable for household and individual analysis is **w1_wgt**. This variable includes sampling and post stratification corrections as described above. The NIDS sample included a stratified and clustered sample design. The stratification was done by District Council and therefore the stratification variable is **w1_hhdc**. Clustering was done by PSU and therefore the cluster variable is **w1_cluster**. In Stata the recommended svyset command is **svyset [pw= w1_wgt], strata (w1_hhdc) psu(w1_hhcluster)**.

3. Updated Wave 1 Weights: September 2013

Together with Wave 3 of the National Income Dynamics Study, updates to Wave 2 and Wave 1 have been released. Since the information on the sample for these waves has changed a little (e.g. age information has been improved, some households have been removed) it has been necessary to recalculate **all** the weights previously released as well. Indeed since a few households have been removed from Wave 1 even the "design weights correcting for nonresponse" will be slightly different in the affected clusters.

Nevertheless the **methods** used, i.e. the algorithms underpinning the calculations, have not been changed. This means that the revised weights will be very similar in most cases to the ones released previously. Indeed because the algorithms have not been changed, the documentation released with previous weights should be consulted as well for further information.

The **calibrated weights**, however, have changed in that all calibration has happened to the revised mid-year population estimates as released by Statistics South Africa in 2013. This was necessary to ensure that the population totals (and totals within particular provinces and age groups) did not jump discontinuously as a result of the upward revision of South Africa's overall population size. In practice this means that the calibrated weights for 2008 and 2010 will now gross up to slightly larger totals than before.

See the Wave 3 User Manual for more details of the latest weights.

4. Data inconsistencies

Our policy has been to represent what was captured in-field as faithfully as possible. Due to the nature of fieldwork and respondent's interpretation of questions this sometimes leads to internal inconsistencies. Where we were unable to verify the information with the respondent post-field we have left the information as it was recorded in-field. We encourage our users to apply their expertise in dealing with these anomalies. Further data verification continues and updated versions of the data will be released regularly. We therefore urge users to register so that they can receive notification of the updates.