Computer editing and imputation of missing values
(Indian Experience)

A set of collected data often contains some errors. It is necessary to make the data-set valid, consistent and realistic before using them. Various measures may be taken in this regard. For example,

(i) Manual scrutiny of data, if data are collected using paper-schedule /questionnaire,

(ii) In-built checks in data transcription software,

(iii) Computer scrutiny which comprises scrutiny of data by computer, study of scrutiny findings manually, suggesting suitable corrections and incorporating those corrections in the data-file, and

(iv) Computer editing (auto-correction).

Three broad categories of errors must be distinguished here: format errors, identification errors and content errors. Format errors involve various types of misspecification in which the information is recorded. For example, it is recoded in whole number whereas it was supposed to be recorded in three decimal places. Or, the information is recorded in wrong unit. An identification error occurs when a particular unit is given the identification code of another sample unit resulting in incorrect identification or duplication of identification. Format errors or identification errors can be taken care of during manual editing but they cause serious problems in computer editing. It is, therefore, particularly advisable that these errors are removed from the data-file by manual editing before computer edit program is applied.

Why Computer editing (auto-correction)?

Generally, some errors still remain in data even after the Computer scrutiny stage is over. This happens due to following reasons.

(1) By principle, only obvious corrections are incorporated during validation; no arbitrary subjective correction is incorporated in data during validation.
(2) Sometimes, a missing or not acceptable entry needs to be replaced by an imputed figure. This involves a lot of searching and/or computations which will take huge amount of time if someone tries to do them manually.

This Courseware has been received with thanks from the National Academy of Statistical Administration (NASA), CSO, India for use in the SAARCSTAT Web portal. Necessary editing has been done, wherever needed, without distortion of facts and figures.

For example, if value (or quantity) of consumption of a commodity needs to be imputed then it may be done by using the quantity-figure (or value) and average price of that commodity; average should be based on sufficient number of observations from similar households. For this, one has first to separate out the schedules of similar households, then to note down the value and quantity figures observed for that commodity in those households and then to calculate the average price. This is a very time-consuming task. It is not feasible to take such action during validation. Also, there is a chance of doing mistake.

(3) Some corrective actions may be missed during validation; either during scrutiny of error lists or during updating of data.

(4) Person who updates the data may make a mistake in understanding the corrective action suggested by the scrutiniser.

(5) Some fresh errors may occur during validation due to human-error.

These are taken care of at this last stage called computer editing.

Computer editing – scope & implementation

In one sense, computer editing is totally different from other types of data validation. In all the previous stages of data validation, the inconsistencies in the data are indicated to the data processing personnel and it is for them to manually edit or not the data file by incorporating the necessary corrections in consultation with the actual entries in the schedule. In computer editing, scrutinising as well as editing works are done by the computer without any human intervention in between. For this, editing rules indicating corrective action for each validation rule are also judiciously framed and programs are developed accordingly. Validated data are thoroughly examined while developing the computer edit rules and review of data and trial runs are needed to finalize the edit rules. Any mistake in framing the edit rules may spoil the data and therefore, auto correction software is prepared with due care at every level.

Computer editing falls into following basic categories:

1. *Logical edits: Logical edits infer missing or inconsistent values from other* information available for that person/household/enterprise on the basis of a set of rules. For example, household-size is made equal to the number of person-wise records of demographic block of that household. Education level of a child may be inferred to be 'illiterate' if age is less than 4.

2. Hot-deck or dynamic allocation: In the cases where logical edit is not possible, hot deck techniques may be used. This involves searching the data file for a "donor" record which shares key characteristics with the missing, illegible or inconsistent case. For example, if 'sex' of a person is missing or illegible and could not be logically inferred from other characteristics, the sex of the most proximate person record which had same

household type and size, race, relation to head of household, age, marital status and occupation may be allocated to him/her.
3. Cold-deck or static imputation: In this technique, imputation is done using mean or median or modal value derived from the entire set of similar records. Example cited in paragrph (2) of page 1 is an example of cold-deck imputation. Replacement values for some selected parameters are computed beforehand for groups of similar records, say, records of same stratum pertaining to a particular commodity. Replacement value for that item is then same for all households of that stratum.
4. Deletion of irrelevant extra records: Sometimes it is found that the data-file contains some records which are not relevant to the context. For example, suppose information for a particular block is supposed to be collected for regular-salaried persons. But it is found that some of these records are for persons who are employed but not regular-salaried. These records should be deleted from the data-file. Otherwise the estimates generated from the records of this block will be erroneous.
5. Creation of dummy records: Sometimes the situation may be just opposite of above situation i.e., some expected records are missing. In above example, no record is available for a regular-salaried person. In such a case, a dummy record is created for that persons and information for various fields of that dummy record is allocated as far as possible. This may be done either by transferring the information available for that person in other records or by logical editing or by hot-deck/cold-deck imputation.

NSSO follows cold-deck approach in imputation of quantitative fields. In the cases of qualitative fields (codes), the edit programme either corrects the entry by logical editing or makes it blank. Hot-deck approach is hardly used. A blank field is treated as 'not recorded' for the purpose of tabulation. For quantitative fields, it is essential that a 'zero' value is distinguished from a 'blank' entry.

In NSS schedules, the coverage of persons in different blocks is interrelated. To make the coverage complete and consistent throughout all the blocks, in some cases, records in certain blocks may have to be created or corrected. The procedures for such situation are laid down in the edit software so that in the edited data the coverage of persons in different blocks remain same as envisaged in the design of the schedule.

Few important edit rules followed at NSSO are given below:

1. **Generation of totals and sub-totals**: In NSS schedules, there are entries which are derived from other individual entries. For example, total and sub-total entries are obtained by adding the concerned set of entries. During computer edit, all the totals and sub-totals are generated from the individual entries and they are replaced in the data file wherever necessary. This is mostly done in consumer expenditure schedule.
2. **Generation or Dropping of records**: The number of person records in certain blocks is to be consistent with that of parent block with same identification and age. In case of mismatch of records, action may be required to generate new records or drop some

existing records on the basis of essential demographic blocks. This is mostly done in employment and unemployment schedule.

3. **Dropping of Sample Household records**: In each schedule of NSS, certain blocks are considered to be essential blocks. Usually, demographic block is considered to be an essential block. Without the presence of those blocks, the entries recorded in the other blocks are of no use. In case this block is missing, all the data of this household is dropped (i.e. rejected).
4. **Imputation of Quantity/Value figures**: In spite of all the efforts made in the previous stages of data validation, there may be some items having only quantity figures recorded but no value or vice versa. In such situations the value/quantity figures are imputed based on the suggested principle.
5. **Correction of invalid NIC-5digit codes:** Depending upon activity status some of NIC codes can be corrected say casual labour working in construction. If NIC code is valid at 4-digit level then the last digit could be replaced based on last code available in the NIC master list.
6. **Distribution of total across components:** Distribution of total cost of construction to material, labour and other in proportion to cold deck ratios in case components are missing or residual amount to others.

Finalisation of edit software: The computer edit programme is initially developed as a provisional one and after trial running on validated data and thorough examination of output, edit rules are finalised before generation of trial tables. While giving trial runs the following points are kept in view.

(i) The frequency of edited values against each check point
(ii) The actual entry recorded in the data file after editing for each check point.
(iii) A suitable print-file is generated for proper examination of all types of data editing (i.e., deleting records, creating new records, edited figures etc.)
(iv) Number of households for which each such check has been run.

After finalizing the computer edit software and running it on the validated data, we get the autocorrected data which is used for generating trial tables. The pre-autocorrected data are also kept. Corrections, if any, during the examination of trial tables are made in the pre-autocorrected data. After incorporating the corrections in the data-file, again auto correction software is run before generation of final tables.

****