Advisory Group Paper (00)06

EDIT AND IMPUTATION FOR 2001

This paper summarises the plans to edit and impute data in 2001.

Advisory group members are asked to:-

(a) note the contents.

A more detailed document will be circulated to the Advisory Groups towards the end of March.

Paul Vickers
Head of Edit, Imputation and Disclosure Control
Census Division, Office for National Statistics
Room 4200N, Segensworth Road, Titchfield, Hampshire, PO15 5RR
Tel: 01329 813685, e-mail paul.vickers@ons.gov.uk

February 2000

Edit and Imputation for 2001

Advisory group members are asked to note our plans to edit and impute data in 2001. This paper summarises our plans. A more detailed document will be circulated to the Advisory Groups towards the end of March.

Introduction

Census data will contain some invalid, inconsistent or missing data. This paper explains how the Census Offices plan to correct these errors when they occur within household data. In particular, it deals with missing data arising from a partially completed form.

We edit and impute data so that as far as possible we can provide a complete and consistent database to our users. This removes the danger of inconsistent interpretations of the data being made as users employ different methods of analysing non-response and dealing with inconsistent data. We are also best placed to carry out imputation because we have access to all micro-data which improves the quality of the imputation.

Missing data, however, can also arise when we know a household exists or we estimate it exists but we do not have any information for it. These missing households and missing people are dealt with by the One Number Census (ONC) process and the imputation methodology in this instance is covered in 'A guide to the ONC' which was produced for the ONC Roadshows in 1999.

Invalid data

This consists of errors that arise because:

- The values given are outside a realistic range, such as an age of 140 years,
- Multi-ticking of responses when only one tick was expected, such as someone who ticks the single and married boxes for marital status, and
- Filters on the forms not being followed such as someone answering the economic activity questions when they were less than 16 years old.

These will be dealt with by a series of rules that will be invoked during the data capture and coding phase. The rules will either set the value to missing or give a correct value for the error. These rules are being reviewed using the 1999 rehearsal data.

Inconsistent Responses between Questions

We have devised a set of consistency rules that will identify certain combinations of responses that cannot occur. Generally, these involve the key demographic variables including age, sex, marital status and relationship. Examples of these are:

- No-one under sixteen can be married; and
- The difference in age between a parent and a child must be at least 13 years.

We will not, however, identify every possible inconsistency such as those involving occupation and age or occupation and location. So the database may contain a coal miner who works in London or a 16 year old doctor.

We will correct inconsistencies in a similar way to 1991 by the use of an 'edit matrix'. This will take place after the data has been captured and coded. This matrix considers all combinations of consistency rule failures that involve age and proposes an action that will either change the value of one variable or set the value of one or more variables to missing. The matrix will attempt to keep the number of changes to minimum.

There are a series of other non-age related checks and checks between persons in the household which are each dealt with in turn.

Missing data

As a result of the previous 2 processes the data will contain a 'consistent' database with missing data.

We will impute missing data using information from similar households that have complete data for the data that is missing. This is similar to the 1991 method. The

main difference between this method and the one used in 1991 is that we aim to impute all the missing data from one 'donor' household. In 1991 several 'donors' may have been used to impute data into a household with more than one item missing.

Another difference between 2001 and 1991 is that we aim to impute values for all variables. There may, however, be some variables for which we cannot remove the bias using our imputation methodology. These are variables that have a large number of possible values such as occupation, industry, postcode of address one year ago and postcode of workplace address. These variables will be subject to evaluation using Census Rehearsal or 1991 data.

Future Paper

This paper has provided a high level view of the edit and imputation process planned for 2001. A more detailed paper will be circulated to advisory group members towards the end of March.

Paul Vickers
Head of Edit, Imputation and Disclosure Control
Census Division, Office for National Statistics
Room 4200N, Segensworth Road, Titchfield, Hampshire, PO15 5RR
Tel: 01329 813685, e-mail paul.vickers@ons.gov.uk