CHAPTER ONE

Introduction to Statistical Data Editing and Imputation

1.1 Introduction

It is the task of National Statistical Institutes (NSIs) and other official statistical institutes to provide high-quality statistical information on many aspects of society, as up-to-date and as accurately as possible. One of the difficulties in performing this task arises from the fact that the data sources that are used for the production of statistical output, both traditional surveys as well as administrative data, inevitably contain errors that may influence the estimates of publication figures. In order to prevent substantial bias and inconsistencies in publication figures, NSIs therefore carry out an extensive process of checking the collected data and correcting them if necessary. This process of improving the data quality by detecting and correcting errors encompasses a variety of procedures, both manual and automatic, that are referred to as *statistical data editing*. The effects of statistical data editing on the errors have been examined since the mid-1950s [see Nordbotten (1955)].

Besides errors in the data, another complicating factor in order to fulfill the task of NSIs and other statistical institutes successfully is that data are often missing. This can be seen as a simple form of erroneous data, simple in the sense that missing values are easy to identify; estimating good values for these missing values may, however, be hard.

Handbook of Statistical Data Editing and Imputation, by T. de Waal, J. Pannekoek, and S. Scholtus

Errors and missing data can arise during the measurement process. Errors arise during the measurement process when the reported values differ from the true values. A possible reason for a measurement error can be that the true value is unknown to the respondent or difficult to obtain. Another reason could be that questions are misinterpreted or misread by the respondent. An example is the so-called unity measurement error that occurs if the respondent reports in euros when it was required to report in thousands of euros. Another example is a respondent reporting his own income when asked for the household income. For business surveys, errors also occur due to differences in definitions used by the statistical office and the accounting system of the responding unit. There may, for instance, be differences in the reference period used by the business and the requested period (financial year versus calender year is an example). After the data have been collected, they will pass through several other processes, such as keying, coding, editing, and imputation. Errors that arise during this further processing are referred to as processing errors. Note that although the purpose of editing is to correct errors, it is mentioned here also as a process that may occasionally introduce errors. This undesirable situation arises if an item value is adjusted because it appeared to be in error but it is actually correct. Missing data can arise when a respondent does not know the answer to a question or refuses to give the answer to a certain question.

Traditionally, NSIs have always put a lot of effort and resources into statistical data editing, because they considered it a prerequisite for publishing accurate statistics. In traditional survey processing, statistical data editing was mainly an interactive activity intended to correct all data in every detail. Detected errors or inconsistencies were reported and explained on a computer screen and corrected after consulting the questionnaire or contacting respondents, which are time- and labor-intensive procedures. In this book we examine more efficient statistical data editing methods.

It has long been recognized that it is not necessary to correct all data in every detail. Several studies [see, for example, Granquist (1984, 1997) and Granquist and Kovar (1997)] have shown that in general it is not necessary to remove all errors from a data set in order to obtain reliable publication figures. The main products of statistical offices are tables containing aggregate data, which are often based on samples of the population. This implies that small errors in individual records are acceptable. First, because small errors in individual records tend to cancel out when aggregated. Second, because if the data are obtained from a sample of the population, there will always be a sampling error in the published figures, even when all collected data are completely correct. In this case an error in the results caused by incorrect data is acceptable as long as it is small in comparison to the sampling error. In order to obtain data of sufficiently high quality, it is usually enough to remove only the most influential errors. The above-mentioned studies have been confirmed by many years of practical experience at several statistical offices.

In the past, and often even in the present, too much effort was spent on correcting errors that did not have a noticeable impact on the ultimately published figures. This has been referred to as ''over-editing.'' Over-editing not only costs

money, but also takes a considerable amount of time, making the period between data collection and publication unnecessarily long. Sometimes over-editing even becomes ''creative editing''; the editing process is then continued for such a length of time that unlikely, but correct, data are ''corrected.'' Such unjustified alterations can be detrimental for data quality. For more about the danger of over-editing and creative editing, see, for example, Granquist (1995, 1997) and Granquist and Kovar (1997).

It has been argued that the role of statistical data editing should be broader than only error localization and correction. Granquist (1995) identifies the following main objectives:

1. Identify error sources in order to provide feedback on the entire survey process.
2. Provide information about the quality of the incoming and outgoing data.
3. Identify and treat influential errors and outliers in individual data.
4. When needed, provide complete and consistent individual data.

During the last few years, the first two goals—providing feedback on the other survey phases, such as the data collection phase, and providing information on the quality of the final results—have gained in importance. The feedback on other survey phases can be used to improve those phases and reduce the amount of errors arising in these phases. In the next few years the first two goals of data editing are likely to become even more important. The main focus in this book is, however, on the latter, more traditional, two goals of statistical data editing. Statistical data editing is examined in Chapters 2 to 6.

Missing data is a well-known problem that has to be faced by basically all institutes that collect data on persons or enterprises. In the statistical literature, ample attention is hence paid to missing data. The most common solution to handle missing data in data sets is imputation, where missing values are estimated and filled in. An important problem of imputation is to preserve the statistical distribution of the data set. This is a complicated problem, especially for high-dimensional data. Chapters 7 and 8 examine this aspect of imputation.

At NSIs the imputation problem is further complicated owing to the existence of constraints in the form of edit restrictions, or edits for short, that have to be satisfied by the data. Examples of such edits are that the profit and the costs of an enterprise have to sum up to its turnover and that the turnover of an enterprise should be at least zero. Records that do not satisfy these edits are inconsistent and are hence considered incorrect. Details about imputation and adjustment techniques that ensure that edits are satisfied can be found in Chapters 9 and 10.

The rest of this chapter is organized as follows. In Section 1.2 we examine the statistical process at NSIs and other statistical organizations, and especially the role that statistical data editing and imputation play in this process. In Section 1.3 we examine (kinds of) data, errors, missing data, and edits. Section 1.4 briefly describes the editing methods that will be explored in more detail later in this book. Finally, Section 1.5 concludes this chapter by describing a basic editing strategy.

1.2 Statistical Data Editing and Imputation in the Statistical Process

1.2.1 OVERVIEW OF THE STATISTICAL PROCESS

The processes of detecting and correcting errors and handling missing data form a part of the process of producing statistical information as practiced at NSIs. This process of producing statistical information can be broken down into a number of steps. Willeboordse (1998) distinguishes the following phases in the statistical process for business surveys:

- Setting survey objectives.
- Questionnaire design and sampling design.
- Data collection and data entry.
- Data processing and data analysis.
- Publication and data dissemination.

A similar division can be made for social surveys. Each phase itself can be subdivided into several steps.

Setting Survey Objectives. In the first phase, user groups for the statistical information under consideration are identified, user needs are assessed, available data sources are explored, potential respondents are consulted about their willingness to cooperate, the survey is embedded in the general framework for business surveys, the target population and the target variables of the intended output are specified, and the output table is designed.

Questionnaire Design and Sampling Design. In the second phase the potential usefulness of available administrative registers is determined, the frame population in the so-called Statistical Business Register is compared with the target population, the sampling frame is defined, the sampling design and estimation method are selected, and the questionnaire is designed. There is a decision process on how to collect the data: paper questionnaires, personal interviews, telephone interviews, or electronic data interchange.

Data Collection and Data Entry. In the third phase the sample is drawn, data are collected from the sampled units and entered into the computer system at the statistical office. During this phase the statistical office tries to minimize the response burden for businesses and to minimize nonresponse.

Data Processing and Data Analysis. In the fourth phase the collected data are edited, missing and erroneous data are imputed, raising weights are determined, population figures are estimated, the data are incorporated in the integration framework, and the data are analysed (for example, to adjust for seasonal effects). The process of detecting and correcting

errors and handling missing data forms an important part of this phase. The bulk of the work at NSIs and other statistical agencies that collect and process data is spent on this phase, especially on statistical data editing.

Publication and Data Dissemination. The final phase includes setting out a publication and dissemination strategy, protecting the final data (both tabular data and microdata, i.e. the data of individual respondents) against disclosure of sensitive information, and lastly publication of the protected data.

1.2.2 THE EDIT AND IMPUTATION PROCESS

During statistical data editing and the imputation process, erroneous records—and erroneous values within these records—are localized and new values are estimated for the erroneous values and values missing in the data set. To edit an erroneous record, two steps have to be carried out. First, the incorrect values in such a record have to be localized. This is often called *error localization*. Second, after the faulty fields in an erroneous record have been identified, these faulty fields have to be *imputed*; that is, the values of these fields have to be replaced by better, preferably the correct, values.

For erroneous records, error localization and imputation are closely related. Often it is hard to distinguish where the error localization phase ends and where the imputation phase starts. For instance, when humans edit data, they frequently look at possible ways of imputing a record before completing the error localization phase. Another example is that the localization of erroneous values might be based on estimating values first and then determining the deviation between the observed and estimated values. The observed values that differ most from their estimated counterparts are then considered erroneous, or in any case suspicious. In this approach the detection of erroneous values and the estimation of better values are highly intertwined. A third example is that during manual review (see Chapter 6) the detection of erroneous values and the ''estimation'' of better values are highly intertwined. This ''estimation'' often simply consists of filling in correct answers obtained by recontacting the respondents. Despite the fact that error localization and imputation can be closely related, we will treat them as two separate processes throughout most of this book. This is a simplification of the edit and imputation problem, but one that has shown to work well for most cases arising in practice.

In principle, it is not necessary to impute missing or erroneous values in order to obtain valid estimates for the target variables. Instead, one can estimate the target variables directly during an estimation phase, without imputing the missing and erroneous values first. However, this approach would in most practical cases become extremely complex and very demanding from a computational point of view. By first imputing the missing and erroneous values, a complete data set is obtained. From this complete data set, estimates can be obtained by standard estimation methods. In other words, imputation is often applied to simplify the estimation process.

1.3 Data, Errors, Missing Data, and Edits

1.3.1 KINDS OF DATA

Edit and imputation techniques can be divided into two main classes, depending on the kind of data to be edited or imputed: techniques for numerical data and techniques for categorical data. Generally, there are major differences between techniques for these kinds of data. At NSIs and other statistical institutes, numerical data occur mainly in surveys on businesses whereas categorical data occur mainly in social surveys—for instance, surveys on persons or households.

At Statistics Netherlands and other NSIs, editing of business surveys is a much bigger problem than editing of most social surveys on households and persons. The main reason is that for business surveys, generally much more edit rules (see below) are defined than for social surveys, and business surveys generally contain much more errors than social surveys. Typically, business surveys are not very large. Large and complicated business surveys may have somewhat over 100 variables and 100 edits. In a small country such as the Netherlands, the number of records in a business survey is usually a few thousand.

Population censuses form an important exception to the general rule that editing is easier for social surveys than for business surveys. Census data do not contain a high percentage of errors, but the number of edits (a few hundred), the number of variables (a few hundred), and especially the number of records can be high (several millions). Due to the sheer volume of the data, editing of data from a population census forms a major problem. The only efficient way to edit such volumes of data is to edit these data in an automatic manner whenever possible.

A recent development at NSIs is the increasing use of administrative (or register-based) data, as opposed to the more traditional data collection by means of sample surveys. The editing and imputation of administrative data for statistical purposes has certain specific features not shared by sample surveys. For instance, if data from several registers are combined, apart from the errors that are present in the individual registers, additional inconsistencies may occur between data from different registers due to matching errors or diverging metadata definitions. Because this is a relatively new topic, suitable methodology for the statistical data editing and imputation of administrative data has not yet been fully developed. We refer to Wallgren and Wallgren (2007) for an overview of current methods for register-based statistics.

1.3.2 KINDS OF ERRORS

One of the important goals of statistical data editing is the detection and correction of errors. Errors can be subdivided in several ways. A first important distinction we shall make is between systematic and random errors. A second important distinction we shall make is between influential errors and noninfluential errors. The final distinction is between outliers and nonoutliers.

Systematic Errors. A systematic error is an error that occurs frequently between responding units. This type of error can occur when a respondent misunderstands or misreads a survey question. A well-known type of systematic error is the so-called unity measure error, which is the error of, for example, reporting financial amounts in euros instead of the requested thousands of euros. Systematic errors can lead to substantial bias in aggregates. Once detected, systematic errors can easily be corrected because the underlying error mechanism is known.

Systematic errors, such as unity measure errors, can often be detected by comparing a respondent's present values with those from previous years, by comparing the responses to questionnaire variables with values of register variables, or by using subject-matter knowledge. Other systematic errors, such as transpositions of returns and costs and redundant minus signs, can be detected and corrected by systematically exploring all possible transpositions and inclusions/omissions of minus signs. Rounding errors—a class of systematic errors where balance edits (see Section 1.3.4) are violated because the values of the involved variables have been rounded—can be detected by testing whether failed balance edits can be satisfied by slightly changing the values of the involved variables. We treat systematic errors in more detail in Chapter 2.

Random Errors. Random errors are not caused by a systematic deficiency, but by accident. An example is an observed value where a respondent by mistake typed in a digit too many. In general statistics, the expectation of a random error is typically zero. In our case, however, the expectation of a random error may also differ from zero. This is, for instance, the case in the above-mentioned example.

Random errors can result in outlying values. In such a case they can be detected by outlier detection techniques or by selective editing techniques (see Chapter 6). Random errors can also be influential (see below), in which case they may again be detected by selective editing techniques. In many cases, random errors do not lead to outlying values or influential errors. In such cases, random errors can often be corrected automatically, assuming that they do lead to violated edit restrictions. Automatic editing of random errors is treated in detail in Chapters 3 to 5.

Influential Errors. Errors that have a substantial influence on publication figures are called influential errors. They may be detected by selective editing techniques (see Chapter 6).

The fact that a value has a substantial influence on publication figures does not necessarily imply that this value is erroneous. It may also be a correct value. In fact, in business surveys, influential observations are quite common, because many variables of businesses, such as turnover and costs, are often highly skewed.

Outliers. A value, or a record, is called an outlier if it is not fitted well by a model that is posited for the observed data. If a single value is an outlier, this is called a univariate outlier. If an entire record, or at least a subset consisting of several values

in a record, is an outlier when the values are considered simultaneously—that is, if they do not fit the posited model well when considered simultaneously—this is called a multivariate outlier. Again we have that the mere fact that a value (or a record) is an outlier does not necessarily imply that this value (set of values) contains an error. It may also be a correct value (set of values).

Outliers are related to influential values. An influential value is often also an outlier, and vice versa. However, an outlier may also be a noninfluential value and an influential value may also be a nonoutlying value. In the statistical editing process, outliers are often detected during so-called macro-editing (see Chapter 6). In this book we do not examine general outlier detection techniques, except for a very brief discussion in Chapter 3. We refer to the literature for descriptions of these techniques [see, e.g., Rousseeuw and Leroy (1987), Barnett and Lewis (1994), Rocke and Woodruff (1996), Chambers, Hentges and Zhao (2004), and Todorov, Templ, and Filzmoser (2009)].

1.3.3 KINDS OF MISSING DATA

The occurrence of missing data implies a reduction of the effective sample size and consequently an increase in the standard error of parameter estimates. This loss of precision is often not the main problem with nonresponse. Survey organizations can anticipate the occurrence of nonresponse by oversampling, and moreover the loss of precision can be quantified when standard errors are estimated. A more problematic effect, which cannot be measured easily, is that nonresponse may result in biased estimates.

Missing data can be subdivided in several ways according to the underlying nonresponse mechanism. Whether the problem of biased estimates due to nonresponse actually occurs will depend on the nonresponse mechanism. Informally speaking, if the nonresponse mechanism does not depend on unobserved data (conditionally on the observed data), imputation may lead to unbiased estimates without making further assumptions. If the nonresponse mechanism does depend on unobserved data, then further—unverifiable—assumptions are necessary to reduce bias by means of imputation. We shall now make these statements more precise.

A well-known and very often used classification of nonresponse mechanisms is: ''missing completely at random'' (MCAR), ''missing at random'' (MAR), and ''not missing at random'' (NMAR); see Rubin (1987), Schafer (1997), and Little and Rubin (2002).

MCAR. When missing data are MCAR, the probability that a value is missing does not depend on the value(s) of the target variable(s) to be imputed or on the values of auxiliary variables. This situation can occur when a respondent forgets to answer a question or when a random part of the data is lost while processing it. MCAR is the simplest nonresponse mechanism, because the item nonrespondents (i.e., the units that did not respond to the target variable) are similar to the item respondents (i.e., the units that did respond to the target

variable). Under MCAR, the observed data may simply be regarded as a random subset of the complete data. Unfortunately, MCAR rarely occurs in practice.

More formally, a nonresponse mechanism is called MCAR if

$$\mathrm{P}(r_j \mid y_j, \mathbf{x}, \xi) = \mathrm{P}(r_j \mid \xi). \tag{1.1}$$

In this notation, r_j is the response indicator of target variable y_j, where $r_{ij} = 1$ means that record i contains a response for variable y_j, and $r_{ij} = 0$ that the value of variable y_j is missing for record i, $\mathbf{x}$ is a vector of always observed auxiliary variables, and ξ is a parameter of the nonresponse mechanism.

MAR. When missing data are MAR, the probability that a value is missing does depend on the values of auxiliary variables, but not on the value(s) of the target variable(s) to be imputed. Within appropriately defined groups of population units, the nonresponse mechanism is again MCAR. This situation can occur, for instance, when the nonresponse mechanism of elderly people differs from that of younger people, but within the group of elderly people and the group of younger people the probability that a value is missing does not depend on the value(s) of the target variable(s) or on the values of other auxiliary variables. Similarly, for business surveys, larger businesses may exhibit a different nonresponse mechanism than small businesses, but within each group of larger businesses, respectively small businesses, the nonresponse mechanism may be MCAR.

In more formal terms, a nonresponse mechanism is called MAR if

$$\mathrm{P}(r_j \mid y_j, \mathbf{x}, \xi) = \mathrm{P}(r_j \mid \mathbf{x}, \xi), \tag{1.2}$$

using the same notation as in (1.1).

MAR is a more complicated situation than MCAR. In the case of MAR, one needs to find appropriate groups of population units to reduce MAR to MCAR for these groups. Once these groups of population units have been found, it is simple to correct for missing data because within these groups all units may be assumed to have the same probability to respond.

In practice, one usually assumes the nonresponse mechanism to be MAR and tries to construct appropriate groups of population units. These groups are then used to correct for missing data.

NMAR. When missing data are NMAR, the probability that a value is missing does depend on the value(s) of the target variable(s) to be imputed, and possibly also on the values of auxiliary variables. This situation can occur, for instance, when reported values of income are more likely to be missing for persons with a high income, when the value of ethnicity is more likely to be missing for certain ethnic groups, or—for business surveys—when the probability that the value of turnover is missing depends on the value of turnover itself.

In more formal terms, a nonresponse mechanism is called NMAR if

$$\mathrm{P}(r_j \mid y_j, \mathbf{x}, \xi)$$

cannot be simplified, that is, if both (1.1) and (1.2) do not hold.

NMAR is the most complicated case. In order to correct for NMAR, one cannot use only the observed data. Instead, one also has to make model assumptions in order to model the dependence of the nonresponse mechanism on the value(s) of the target variable(s).

A related classification of nonresponse mechanisms is: "ignorable" and "nonignorable."

Ignorable. A nonresponse mechanism is called ignorable if it is MAR (or MCAR) and the parameters to be estimated are distinct from the parameter ξ. Here, distinctness means that knowledge of ξ (which could be inferred from the response indicator r_j that is available for all units, responding or not) is not helpful in estimating the parameters of interest. However, as noted by Little and Rubin (2002, Chapter 6), the MAR condition is more important than distinctness, because MAR alone is sufficient to make valid inference possible. If the parameters are not distinct, this can merely result in some loss of efficiency.

Nonignorable. A nonresponse mechanism is called nonignorable if the conditions for ignorability do not hold. That is, either the nonresponse mechanism is NMAR or the parameter ξ is not distinct from the parameters of interest, or both.

For more details on MCAR, MAR, NMAR, and (non-)ignorability we refer to Rubin (1987), Schafer (1997), and Little and Rubin (2002).

1.3.4 EDIT RULES

Errors are most often detected by edit rules. Edit rules, or edits for short, define the admissible (or plausible) values and combinations of values of the variables in each record. Errors are detected by verifying whether the values are admissible according to the edits—that is, by checking whether the edits are violated or satisfied. An edit e can be formulated as

$$e : x \in S_x,$$

with S_x the set of admissible values of x. As we shall see below, x can refer to a single variable as well as multiple variables. If e is false, the edit is violated and otherwise the edit is satisfied.

Edits can be divided into *hard* (or *fatal*) edits and *soft* (or *query*) edits. Hard edits are edits that must be satisfied in order for a record to qualify as a valid record. As an example, a hard edit for a business survey specifies that the variable *Total costs* needs to be equal to the sum of the variables *Employee costs*, *Capital costs*, *Transport costs*, and *Other costs*. Records that violate one or more hard edits

are considered to be inconsistent and it is deduced that some variable(s) in such a record must be in error. Soft edits are used to identify unlikely or deviating values that are suspected to be in error, although this is not a logical necessity. Examples are (a) an edit specifying that the yearly income of employees must be less than 10 million euros or (b) an edit specifying that the turnover per employee of a firm may not be larger than 10 times the value of the previous year. The violation of soft edits can be a trigger for further investigation of these edit failures, to either confirm or reject the suspected values.

To illustrate the kind of edits that are often applied in practice, examples of a number of typical classes of edits are given below.

Univariate Edits or Range Restrictions. An edit describing the admissible values of a single variable is called a univariate edit or a range restriction. For categorical variables, a range restriction simply verifies whether the observed category codes for the variable belong to the specified set of codes. The set of allowable values S_x is

$$S_x = \{x_1, x_2, \ldots, x_C\}$$

and consists of an enumeration of the C allowed codes. For instance, for the variable *Sex* we could have $S_x = \{0, 1\}$ and for a date variable in the conventional *yyyy-mm-dd* notation the set S_x would consist of all allowed integer combinations describing the year, month, and day. Range restrictions for continuous variables are usually specified using inequalities. The simplest, but often encountered, range restrictions of this type are nonnegativity constraints, that is,

$$S_x = \{x \mid x \geq 0\}.$$

Examples are *Age*, *Rent*, and many of the financial variables in business surveys (costs of various types, turnover and revenues of various activities and so on). Range restrictions describing an interval as

$$S_x = \{x \mid l \leq x \leq u\}$$

are also common. Examples are setting lower (l) and upper (u) bounds on the allowable values of age, income, or working hours per week. Range restrictions can be hard edits (for instance, if S_x is an enumeration of allowable codes), but they can also be soft edits if the bounds set on the allowable range are not a logical necessity (for instance, if the maximum number of weekly working hours is set to 100).

Bivariate Edits. In this case the set of admissible values of a variable x depends on the value of another variable, say y, observed on the same unit. The set of admissible values is then the set of admissible pairs of values (x, y). For instance, if x is *Marital status* with values 0 (never married), 1 (married) and 2 (previously married) and y is *Age*, we may have that

$$S_{xy} = \{(x, y) \mid x = 0 \wedge y < 15\} \cup \{(x, y) \mid y \geq 15\},$$

reflecting the rule that persons younger than 15 are not allowed to be married, while for persons of 15 years or more all marital states are allowed. Another example of a bivariate edit is

$$S_{xy} = \{(x, y) \mid x - y > 14\},$$

with x the age, in years, of a mother and y the age of her child. This example reflects the perhaps not "hard" edit that a mother must be at least 14 years older than her child. Finally, an important and often encountered class of bivariate edits is the so-called *ratio edit* which sets bounds on the allowable range of a ratio between two variables and is defined by

$$S_{xy} = \{(x, y) \mid l \leq \frac{x}{y} \leq u\}.$$

A ratio edit could, for example, specify bounds on the ratio of the turnover and the number of employees of firms in a certain branch of industry. Ratio edits are often used to compare data on the same units from different sources, such as values reported in the current survey (x) with values for the same variables reported in last year's survey (y) or values of variables from a tax register with similarly defined variables from a survey.

Balance Edits. Balance edits are multivariate edits that state that the admissible values of a number of variables are related by a linear equality. They occur mainly in business statistics where they are linear equations that should be satisfied according to accounting rules. Two examples are

$$Profit = Turnover - Total\ costs \tag{1.3}$$

and

$$Total\ costs = Employee\ costs + Other\ costs. \tag{1.4}$$

These rules are related because they have the variable *Total costs* in common. If the first rule is satisfied but the second is not, it may seem more likely that *Employee costs* or *Other costs* are in error than *Total costs*, because in the last case the first rule should probably also be violated. Balance edits are of great importance for editing economic surveys where there are often a large number of such edits. For instance, in the yearly structural business statistics there are typically about 100 variables with 30 or more balance edits. These interrelated systems of linear relations that the values must satisfy provide much information about possible errors and missing values.

Since balance edits describe relations between many variables, they are multivariate edits. Moreover, since they are often connected by common variables, they should be treated as a system of linear equations. It is convenient to express such a system in matrix notation. Denoting the five variables in (1.3) and (1.4)

by x_1 (*Profit*), x_2 (*Turnover*), x_3 (*Total costs*), x_4 (*Employee costs*), and x_5 (*Other costs*), the system can be written as

$$\begin{pmatrix} 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

or $\mathbf{Ax} = \mathbf{0}$. The admissible values of a vector $\mathbf{x}$ subject to a system of balance edits, defined by a restriction matrix $\mathbf{A}$, can then be written as

$$S_{\mathbf{x}} = \{\mathbf{x} \mid \mathbf{Ax} = \mathbf{0}\}.$$

1.4 Basic Methods for Statistical Data Editing and Imputation

In this section we have a first brief look at various methods that can be used to edit and impute data. Before we sketch these methods, we first look back to see why these methods were developed.

Computers have been used in the editing process for a long time [see, e.g., Nordbotten (1963)]. In the early years their role was, however, restricted to checking which edits were violated. Subject-matter specialists entered data into a mainframe computer. Subsequently, the computer checked whether these data satisfied all specified edits. For each record all violated edits were listed. Subject-matter specialists then used these lists to correct the records. That is, they retrieved all paper questionnaires that did not pass all edits and corrected these questionnaires. After they had corrected the data, these data were again entered into the mainframe computer, and the computer again checked whether the data satisfied all edits. This iterative process was continued until (nearly) all records passed all edits.

A major problem of this approach was that during the manual correction process the records were not checked for consistency. As a result, a record that was ''corrected'' could still fail one or more specified edits. Such a record hence required more correction. It was not exceptional that some records had to be corrected several times. It is therefore not surprising that editing in this way was very costly, both in terms of money as well as in terms of time. In the literature it was estimated that 25% to 40% of the total budget was spent on editing [see e.g. Federal Committee on Statistical Methodology (1990) and Granquist and Kovar (1997)].

1.4.1 EDITING DURING THE DATA COLLECTION PHASE

The most efficient editing technique of all is no editing at all, but instead ensuring that correct data are obtained during the data collection phase. In this section we briefly discuss ways to obtain data with no or only few errors at data collection.

When one aims to collect correct data at data collection, one generally uses a computer to record the data. This is called computer-assisted data collection. With computer-assisted data collection the computer can immediately check the recorded data for violations of edits. Below we discuss four modes for computer-assisted data collection: CAPI (Computer-Assisted Personal Interviewing), CATI (Computer-Assisted Telephone Interviewing), CASI (Computer-Assisted Self-Interviewing), and CAWI (Computer-Assisted Web Interviewing). For more information on computer-assisted data collection in general, we refer to Couper et al. (1998).

When CAPI is used to collect the data, an interviewer visits the respondent and enters the answers directly into a laptop. When CATI is used to collect the data, the interview is carried out during a telephone call. When CASI or CAWI is used to collect the data, the respondent fills in an electronic questionnaire himself. The difference between these two modes is that for CAWI an electronic questionnaire on the Internet has to be filled in, whereas for CASI an off-line electronic questionnaire is used. When an invalid response is given to a question or an inconsistency between two or more answers is noted during any of these data collection modes, this can be immediately reported by the computer that is used for data entry. The error can then be resolved by asking the respondent the relevant question(s) again. For CASI and CAWI, generally not all edits that could be specified are actually implemented, since the respondent may get annoyed and might refuse to complete the questionnaire when the edits keep on reporting that his/her answers are inconsistent.

Computer-assisted data collection removes the need for data entry by typists, since the data arrive at the statistical office already in digital form. This eliminates one source of potential errors. In many cases, data collected by means of CAPI, CATI, CASI or CAWI also contain fewer errors than data collected by means of paper questionnaires because random errors that affect paper questionnaires are detected and avoided at collection. For face-to-face interviewing CAPI has in fact become the standard. CAPI, CATI, CASI, and CAWI may hence seem to be ideal ways to collect data, but—unfortunately—they too have their disadvantages.

A first disadvantage of CATI and CAPI is that CATI and, especially, CAPI are very expensive. A second disadvantage of CATI and CAPI is that a prerequisite for these two data collection modes is that the respondent is able to answer the questions during the interview. For a survey on persons and households, this is often the case. The respondent often knows (good proxies of) the answers to the questions, or is able to retrieve the answers quickly. For a survey on enterprises the situation is quite different. Often it is impossible to retrieve the correct answers quickly, and often the answers are not even known by one person or one department of an enterprise. Finally, even in the exceptional case that one person knew all answers to the questions, the NSI would generally not know the identity of this person. For the above-mentioned reasons, many NSIs frequently use CAPI and CATI to collect data on persons and households but only rarely for data on enterprises.

Pilot studies and actual applications have revealed that CASI and CAWI are indeed viable data collection modes, but also that several problems arise when

these modes are used. Besides IT problems, such as that the software—and the Internet connection—should be fast and reliable and the security of the transmitted data should be guaranteed, there are a number of practical and statistical problems. We have already mentioned the practical problem that if the edits keep on reporting that the answers are inconsistent, the respondent may refuse to fill in the rest of the questionnaire. An example of a statistical problem is that the group of people responding to a web survey may be selective, since Internet usage is not uniformly distributed over the population [see, e.g., Bethlehem (2007)].

Another important problem for CAWI and CASI is that data collected by either of these data collection modes may appear to be of higher statistical quality than data collected by means of paper questionnaires, but in fact are not. When data are collected by means of CASI and CAWI, one can enforce that the respondents supply data that satisfy built-in edits, or one can avoid balance edits by automatically calculating total amounts from the reported components. Because less edits are failed by the collected data, these data may appear to be of higher statistical quality. This need not be the case, however. Each edit that is built into the electronic questionnaire will be automatically satisfied by the collected data, and hence cannot be used to check for errors later on. Therefore, the collected data may appear to contain only few errors, but this might be due to a lack of relevant edits. There are indications that respondents can be less accurate when filling in an electronic questionnaire, especially if totals are computed automatically [see Børke (2008) and Hoogland and Smit (2008)].

NSIs seem to be moving toward the use of mixed-mode data collection, where data are collected by a mix of several data collection modes. This obviously has consequences for statistical data editing. Some of the potential consequences have been examined by Børke (2008), Hoogland and Smit (2008), and Van der Loo (2008).

1.4.2 MODERN EDITING METHODS

Below we briefly mention editing methods that are used in modern practice. The editing techniques are examined in detail in other chapters of this book.

Interactive Editing. Subject-matter specialists have extensive knowledge on their area of expertise. This knowledge should be used as well as possible. This aim can be achieved by providing subject-matter specialists with efficient and effective interactive data editing tools. Most interactive data editing tools applied at NSIs allow one to check the specified edits during or after data entry, and—if necessary—to correct erroneous data immediately. This is referred to as interactive or computer-assisted editing. To correct erroneous data, several approaches can be followed: The respondent can be recontacted, the respondent's data can be compared to his data from previous years, the respondent's data can be compared to data from similar respondents, and subject-matter knowledge of the human editor can be used. Interactive editing is nowadays a standard way to edit data. It can be used to edit both categorical and numerical data. The number

of variables, edits, and records may, in principle, be high. Generally, the quality of data editing in a computer-assisted manner is considered high. Interactive editing is examined in more detail in Section 6.5.

Selective Editing. Selective editing is an umbrella term for several methods to identify the influential errors (i.e., the errors that have a substantial impact on the publication figures) and outliers (i.e., values that do not fit a model of the data well). Selective editing techniques aim to apply interactive editing to a well-chosen subset of the records, such that the limited time and resources available for interactive editing are allocated to those records where it has the most effect on the quality of the final estimates of publication figures. Selective editing techniques try to achieve this aim by splitting the data into two streams: the critical stream and the noncritical stream. The critical stream consists of records that are the most likely ones to contain influential errors; the noncritical stream consists of records that are unlikely to contain influential errors. The records in the critical stream, the critical records, are edited in a traditional interactive manner. The records in the noncritical stream, the noncritical records, are not edited in a computer-assisted manner. They may later be edited automatically. Selective editing is examined in Chapter 6.

Macro-editing. We distinguish between two forms of macro-editing. The first form is sometimes called the aggregation method [see, e.g., Granquist (1990)]. It formalizes and systematizes what every statistical agency does before publication: verifying whether figures to be published seem plausible. This is accomplished by comparing quantities in publication tables with the same quantities in previous publications. Only if an unusual value is observed, a micro-editing procedure is applied to the individual records and fields contributing to the suspicious quantity. A second form of macro-editing is the distribution method. The available data are used to characterize the distribution of the variables. Then, all individual values are compared with the distribution. Typically, measures of location and spread are computed. Records containing values that could be considered uncommon (given the distribution) are candidates for further inspection and possibly for editing. Macro-editing, in particular the aggregation method, has always been applied at statistical offices. Macro-editing is again examined further in Chapter 6.

Automatic Editing. When automatic editing is applied, records are edited by a computer without human intervention. In this sense, automatic editing is the opposite of the traditional approach to the editing problem, where each record is edited manually. Automatic editing has already been used in the 1960s and 1970s [see, e.g., Nordbotten (1963)]. Nevertheless, it has never become very popular. For this we point out two reasons. First, in former days, computers were too slow to edit data automatically. Second, development of a system for automatic editing was often considered too complicated and too costly by many statistical offices. In the last two decades, however, a lot of progress has been made with respect to automatic editing: Computers have become faster and algorithms

have been simplified and have also become more efficient. For these reasons we pay more attention to automatic editing than to the other editing techniques in this book. Automatic editing of systematic errors is examined in Chapter 2, and automatic editing of random errors in Chapters 3 to 5.

1.4.3 IMPUTATION METHODS

To estimate a missing value, or a value that was identified as being erroneous during statistical data editing, two main approaches can be used. The first approach is manual imputation or correction, where the corresponding respondent is recontacted or subject-matter knowledge is used to obtain an estimate for the missing or erroneous value. The second approach is automated imputation, which is based on statistical estimation techniques, such as regression models. In this book, we only treat the latter approach.

In imputation, predictions from parametric or nonparametric models are derived for values that are missing or flagged as erroneous. An imputation model predicts a missing value using a function of auxiliary variables, the predictors. The auxiliary variables may be obtained from the current survey or from other sources such as historical information (the value of the missing variable in a previous period) or, increasingly important, administrative data. The most common types of imputation models are variants of regression models with parameters estimated from the observed correct data. However, especially for categorical variables, donor methods are also frequently used. Donor methods replace missing values in a record with the corresponding values from a nearby complete and valid record. Often a donor record is chosen such that it resembles as much as possible the record with missing values. Imputation methods are treated in Chapters 7 to 9 of this book.

1.5 An Edit and Imputation Strategy

Data editing is usually performed as a sequence of different detection and/or correction process steps. In this section we give a global description of an editing strategy. This description is general enough to include the situation for many data sets as special cases, and most editing strategies applied in practice will at least include a number of the elements and principles described here. The process steps can be characterized from different points of view—for instance, by the type of errors they try to detect or resolve or by the methods that are used for detection or correction. Another important distinction is between automatic methods that can be executed without human intervention and interactive editing that is performed by editors.

The global editing strategy as depicted in Figure 1.1 consists of the following five steps that are clarified below.

1. *Treatment of Systematic Errors*. Identify and eliminate errors that are evident and easy to treat with sufficient reliability.

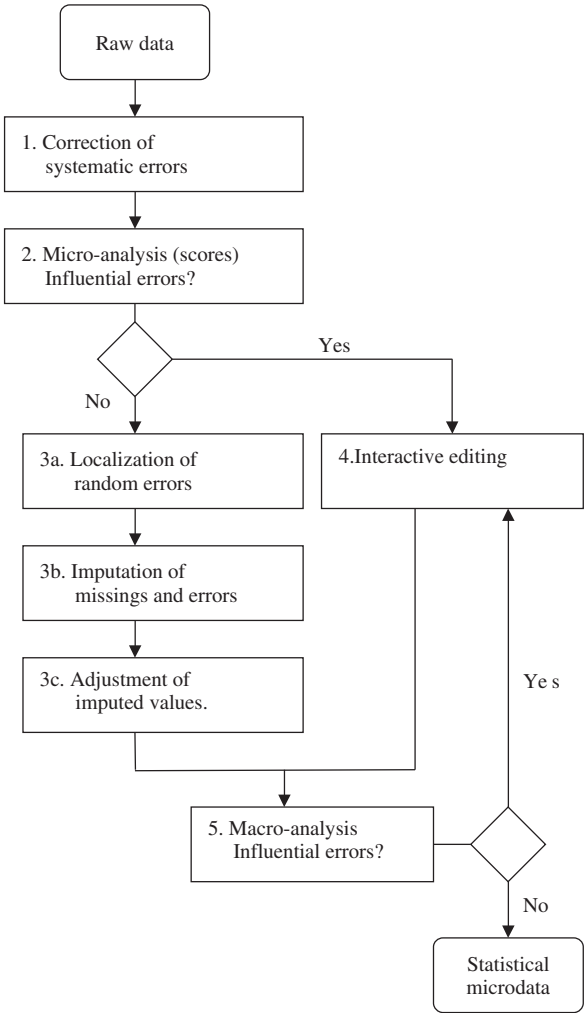

FIGURE 1.1 **Example of a process flow.**

2. *Micro-selection*. Select records for interactive treatment that contain influential errors that cannot be treated automatically with sufficient reliability.
3. *Automatic Editing*. Apply all relevant automatic error detection and correction procedures to the (many) records that are not selected for interactive editing in step 2.
4. *Interactive Editing*. Apply interactive editing to the minority of the records with influential errors.
5. *Macro-selection*. Select records with influential errors by using methods based on outlier detection techniques and other procedures that make use of all or a large fraction of the response.

We distinguish two kinds of process steps: those that localize or treat errors and those that direct the records through the different stages of the process. The processes in step 2 and 5 are of the latter kind; they are "selectors" that do not actually treat errors, but select records for specific kinds of further processing.

Step 1. *Correction of Systematic Errors*. Detection and correction of systematic errors is an important first step in an editing process. It can be done automatically and reliably at virtually no costs and hence will improve both the efficiency and the quality of the editing process. It is in fact a very efficient and probably often underused correction approach. Systematic and other evident errors and algorithms that can automatically resolve these errors are described in Chapter 2 of this book.

Step 2. *Micro-selection*. Errors that cannot be resolved in the previous step will be taken care of either manually (by subject-matter specialists) or automatically (by specialized edit and imputation algorithms). In this step, the data are split into a critical stream and a noncritical stream, using selective editing techniques as mentioned in Section 1.4.2. The extent to which a record potentially contains influential errors can be measured by a score function [cf. Latouche and Berthelot (1992), Lawrence and McKenzie (2000), and Farwell and Rain (2000)]. This function is constructed such that records with high scores likely contain errors that have substantial effects on estimates of target parameters. For this selection step, a threshold value for the score has been set and all records with scores above this threshold are directed to manual reviewers whereas records with scores below the threshold are treated automatically. More details can be found in Chapter 6.

Apart from the score function, which looks at influential errors, another important selection criterion is imputability. For some variables, very accurate imputation models can be developed. If such a variable fails an edit, the erroneous value can safely be replaced by an imputed value, even if it is an influential value. Note that the correction of systematic errors in the previous step can also be an example of automatic treatment of influential errors, if the systematic error is an influential one.

Step 3a. *Localization of erroneous values (random errors)*. The next three steps are automatic detection and correction procedures. In principle, they are designed to solve hard edit failures, including missing values, but they can be applied to soft edits if the soft edit is treated as a hard one. These three steps together represent the vast majority of all edit and imputation methodology. The other chapters of this book are devoted to this methodology for automatic detection and correction of erroneous and missing values.

The first step in the automatic treatment of errors is the localization of errors. Since systematic errors have already been removed, the remaining errors at this stage are random errors. Once the (hard) edits are defined and implemented, it is straightforward to check whether the values in a record are inconsistent in the sense that some of these edits are violated. It is, however, not so obvious how to decide which variables in an inconsistent record are in error. The designation of erroneous values in an inconsistent record is

called the error localization problem, which is treated in Chapters 3 to 5 of this book.

Step 3b. *Imputation*. In this step, missing data are imputed in an automatic manner. The imputation method that is best suited for a particular situation will depend on the characteristics of the data set and the research goals. In Chapters 7 to 9 we examine imputation methods in detail.

Step 3c. *Consistency Adjustment of Imputed Values*. In most cases, the edits are not taken into account by the imputation methods; some exceptions are examined in Chapter 9. As a consequence, the imputed records are in general inconsistent with the edits. This problem can be solved by the introduction of an adjustment step in which adjustments are made to the imputed values such that the record satisfies all edits and the adjustments are as small as possible. This problem can be formulated as a linear or a quadratic programming problem and is treated in Chapter 10.

Step 4. *Interactive Editing*. Substantial mistakes by somewhat larger enterprises that have an appreciable influence on publication aggregates and for which no accurate imputation model exists are not considered suitable for the generic procedures of automatic editing. These records are treated by subject-matter specialists in a process step called interactive editing; see Section 1.4.2 above and Chapter 6.

Step 5. *Macro-selection*. The steps considered so far all use micro-editing methods—that is, methods that use the data of a single record and related auxiliary information to check and correct it. Micro-editing processes can be conducted from the start of the data collection phase, as soon as records become available. In contrast, macro-selection techniques use information from other records and can only be applied if a substantial part of the data is collected or has been imputed. Macro-selection techniques are also selective editing techniques in the sense that they aim to direct the attention only to possibly influential erroneous values. Macro-editing is treated in Chapter 6 of this book.

The process flow suggested in Figure 1.1 is just one possibility. Depending on the type of survey and the available resources and auxiliary information, the process flow can be different. Not all steps are always carried out, the order of steps may be different, and the particular methods used in each step can differ between types of surveys. For social surveys, for instance, selective editing is not very important because the contributions of individuals to a publication total are not so much different, unlike the contributions of small and large enterprises in business surveys. Consequently, there is less need for manual editing of influential records, and step 4 need not be performed. Often, in social surveys, due to a lack of hard edits, the main type of detectable error is the missing value, and process steps 3a and 3c are not performed either. For administrative data the collection of all records, or a large part of it, is often available at once. This is different from the situation for surveys where the data are collected over a period of time. For administrative data it is therefore possible to form preliminary estimates immediately and to start with macro-editing as a tool for selective editing, and

a process could start with step 1, followed by step 5 and possibly by step 4 and/or step 3.

Although automatic procedures are frequently used for relatively unimportant errors, choosing the most suitable error detection and/or imputation methods is still important. If nonappropriate methods are used, especially for large amounts of random errors and/or missing values, additional bias may be introduced. Furthermore, as the quality of the automatic error localization and imputation methods and models gets better, more records can be entrusted to the automatic treatment in step 3 and less records have to be selected for the time-consuming and costly interactive editing step.

REFERENCES

Barnett, V., and T. Lewis (1994), *Outliers in Statistical Data*. John Wiley & Sons, New York.

Bethlehem, J. (2007), *Reducing the Bias of Web Survey Based Estimates*. Discussion paper 07001, Statistics Netherlands, Voorburg (see also www.cbs.nl).

Børke, S. (2008), *Using "Traditional" Control (Editing) Systems to Reveal Changes when Introducing New Data Collection Instruments*. Working Paper No. 6, UN/ECE Work Session on Statistical Data Editing, Vienna.

Chambers, R., A. Hentges, and X. Zhao (2004), Robust Automatic Methods for Outlier and Error Detection. *Journal of the Royal Statistical Society A 167*, pp. 323–339.

Couper, M. P., R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nichols II, and J. M. O'Reilly (eds.) (1998), *Computer Assisted Survey Information Collection*. John Wiley & Sons, New York.

Farwell, K., and M. Rain (2000), Some Current Approaches to Editing in the ABS. *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, pp. 529–538.

Federal Committee on Statistical Methodology (1990), *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18, U.S. Office of Management and Budget, Washington, D.C.

Granquist, L. (1984), *Data Editing and its Impact on the Further Processing of Statistical Data*. Workshop on Statistical Computing, Budapest.

Granquist, L. (1990), A Review of Some Macro-Editing Methods for Rationalizing the Editing Process. *Proceedings of the Statistics Canada Symposium*, pp. 225–234.

Granquist, L. (1995), Improving the Traditional Editing Process. In: *Business Survey Methods*, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, eds. John Wiley & Sons, New York, pp. 385–401.

Granquist, L. (1997), The New View on Editing. *International Statistical Review 65*, pp. 381–387.

Granquist, L. and J. Kovar (1997), Editing of Survey Data: How Much Is Enough? In: *Survey Measurement and Process Quality*, L.E. Lyberg, P. Biemer, M. Collins, E.D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin, eds. John Wiley & Sons, New York, pp. 415–435.

Hoogland, J., and R. Smit (2008), *Selective Automatic Editing of Mixed Mode Questionnaires for Structural Business Statistics*. Working Paper No. 2, UN/ECE Work Session on Statistical Data Editing, Vienna.

Latouche, M., and J. M. Berthelot (1992), Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. *Journal of Official Statistics 8*, pp. 389–400.

Lawrence, D., and R. McKenzie (2000), The General Application of Significance Editing. *Journal of Official Statistics 16*, pp. 243–253.

Little, R. J. A., and D. B. Rubin (2002), *Statistical Analysis with Missing Data*, second edition. John Wiley & Sons, New York.

Nordbotten, S. (1955), Measuring the Error of Editing the Questionnaires in a Census. *Journal of the American Statistical Association 50*, pp. 364–369.

Nordbotten, S. (1963), Automatic Editing of Individual Statistical Observations. In: *Conference of European Statisticians Statistical Standards and Studies No. 2*, United Nations, New York.

Rocke, D. M., and D. L. Woodruff (1996), Identification of Outliers in Multivariate Data. *Journal of the American Statistical Association 91*, pp. 1047–1061.

Rousseeuw, P. J., and M. L. Leroy (1987), *Robust Regression & Outlier Detection*. John Wiley & Sons, New York.

Rubin, D. B. (1987), *Multiple Imputation for Non-Response in Surveys*. John Wiley & Sons, New York.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Todorov, V., M. Templ, and P. Filzmoser (2009), *Outlier Detection in Survey Data Using Robust Methods*. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing, Neuchâtel.

Van der Loo, M. P. J. (2008), *An Analysis of Editing Strategies for Mixed-Mode Establishment Surveys*. Discussion paper 08004, Statistics Netherlands (see also www.cbs.nl).

Wallgren, A., and B. Wallgren (2007), *Register-Based Statistics—Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester.

Willeboordse, A. (ed.) (1998), *Handbook on the Design and Implementation of Business Surveys*. Office for Official Publications of the European Communities, Luxembourg.