A FLEXIBLE AND INTERACTIVE EDIT AND IMPUTATION SYSTEM FOR RATIO EDITS

Brian Greenberg and Rita Surdi, U.S. Bureau of the Census

All survey and census programs are subject to nonresponse and erroneous reporting, whereas data users demand complete and accurate data to be used for a variety of statistical purposes. Although the implementation of an edit and imputation system is highly survey specific, coherent methodologies can be developed that integrate diverse features and needs into a structured framework. Various imputation strategies, subject-matter expertise, and auxiliary information can be incorporated within such a framework.

A widely used criterion for economic data requires that the ratio of two responses lie between prescribed bounds. The upper and lower bounds are determined by historical information, subject-matter expertise, and when feasible, by a sample of responses. In addition to comparing two fields on the report form, ratio edits can incorporate data from an earlier time frame as well as information from an external data file. A system to edit data under ratio edits has been developed at the Bureau of the Census and a prototype model has been developed for the Annual Survey of Manufactures. A modification of this prototype system was designed and used to process two segments of the 1982 Economic Census. An interactive version of this system has been developed for use by subject-matter analysts for on-line processing of referral cases.

1. INTRODUCTION

All survey and census programs are subject to nonresponse and erroneous reporting, whereas data users demand complete and accurate data to be used for a variety of statistical purposes. It is well-recognized that the data collection agency has the optimal vantage point and attendant obligation to provide valid allocations for missing values and to adjust spurious responses. The development of statistically precise and mathematically rigorous edit and imputation systems is essential in meeting this objective and is vital in providing users with high quality data products.

Although the implementation of an edit and imputation system is highly survey-specific, coherent methodologies can be developed that integrate diverse features and needs into a structured framework. Within such a framework, various imputation strategies, subject-matter expertise, and auxiliary information can be incorporated. State-of-the-art edit systems draw upon operations research optimization techniques, mathematics, and statistical analysis to incorporate prior knowledge and concurrent information. Development and implementation of such systems require that mathematical and statistical investigators work jointly with subject-matter specialists familiar with the survey environment.

The role of the edit process is to alter erroneous responses and not to alter valid ones. In most discussions of editing the focus is usually on altering erroneous fields; however, we should beware of overzealousness and take precautions against changing correctly reported values. One should endeavor to assert that a record is acceptable, even in the face of several failed statistical edits if information can be garnered from ancillary sources or from the record itself to support its validity.

One imputes because of item nonresponse and because fields have been targeted for change based on patterns of edit failure. The role of the imputation process is not simply to create a consistent record nor to allocate values based on a random generation from a presumed underlying distribution. The ideal goal (though generally not practicable) is to create a revised record close to what a respondent would have reported were there no errors. In particular, when one imputes in a field deleted due to edit failures the imputation strategy should take into account the reported value (albeit incorrect) whenever possible, and the imputation for edit failures might be different from that for item nonresponse. For example, in some surveys, a frequent reporting (or keying) problem is that a field is in error by a multiple of one thousand. For the fields susceptible to this sort of error, one should attempt to detect it and divide the recorded response by one thousand.

The relation between editing and imputation is fundamental, and it is crucial to integrate these two features when designing an error correction system. One aspect of the relation is technical: imputed values should not fail edits except in prespecified special cases. Accordingly, an important aspect of the imputation process is the editing of imputed values—assuming that non-imputed variables all pass edit checks. An imputation procedure based on an estimation process, especially one involving a stochastic component, can yield specious imputations. For example, due to the contribution of a residual, an estimate of a missing value may be negative—usually proscribed. But more generally, interrelated data items often must conform to edit constraints, and to ensure that one does not impute a value that would be rejected if it were reported, the candidates for imputation have to be checked for feasibility. Those not feasible have to be either reimputed or adjusted. If a non-feasible or suspicious imputation occurs in a set of fields that were targeted for change due to edit failures, an alternate set of fields to adjust may be indicated. Of course, if the imputation strategy can ensure feasibility, so much the better.

Another aspect of the relation between editing and imputation is far more intimate and must run throughout a coherent system. Simply stated, the variables and criteria that contribute to the editing of reported data and are embedded in the edit constraints should play a role in determining a valid and meaningful imputation. For example, if the imputation is to be based on matching to records from other respondents (e.g., hot deck, statistical matching) the connection between the edit step and the imputation is that the matching be based on variables that enter edits for missing fields. If the imputation is based on other reported values on the same record (as in a regression procedure), once again, the variables most prominently contributing to the impute should be those in edits for that field. By utilizing variables most closely related to the field to be corrected in both editing and imputation, one endeavors to guarantee that imputed values pass all edits.

The seminal paper relating editing and imputation is by Fellegi and Holt, [2]. In that paper, the