THE WORKPLACE AND EMPLOYEE SURVEY – ERROR DETECTION

Editing of Data

The employer CAPI (Computer Assisted Personal Interviewing) capture vehicle performs validity, range, and inter-field edits. These are the types of edits that are performed during the collection of the first wave data. For subsequent waves a suitable set of historical edits has been developed. The majority of inter-field edits are confined to a single content block. If an edit failure occurs between blocks, then the primary respondent is asked to confirm the information.
An example of a validity edit is that total annual expenditures be positive. The corresponding range edit requires that expenditures not exceed an upper bound. A related inter-field edit for total annual expenditures ensures that the sum of annual gross payroll and non-wage expenditures does not exceed total annual expenditures.

The employee CATI application performs validity, range, inter-field and historical edits. Any edit failures are resolved during the telephone interview.

Outlier Detection

The use of CAPI/CATI for data collection greatly reduces the number of response and typographical errors. The system incorporates basic data validation and verification of known relationships such as full time and part time employment not exceeding total employment. To detect errors that have eluded the CAPI/CATI application, both micro and macro level analysis of questionable responses is performed to protect the coherence of the data.

At the macro level, the top ten contributors to their respective estimates are investigated along with the records comprising an estimate that has undergone a relatively large change from year to year. This change may be positive or negative. The two techniques are related as an unusually large contributor to an estimate may also be the cause for its large change. To make the analysis more efficient, an expected contribution of a unit to an estimate is computed using the reported employment. This is then compared to the corresponding observed contribution. A test is conducted to determine if the difference between the expected and observed contributions is significant. The approach works well for variables well correlated with employment, and is still a good indicator of potential problems even for variables whose correlation with employment is weaker.

When large year-to-year changes are detected in the estimates, all corresponding records are investigated. In many cases the change may be real if a particular sector experiences a period of strong growth or decline. No one record contributes a significant amount to the estimate but the cumulative effect of small changes causes the numbers to change dramatically. The macro analysis is univariate and as such may not detect problems between variables.

At the micro data level, a univariate outlier detection routine is applied to all complete and partial respondents prior to imputation. The outlier detection is performed on individual variables or ratios of variables, cross-sectionally and longitudinally. The method used for outlier detection standardizes the variable(s) of interest by subtracting a location measure and dividing by a scale measure. In WES, the location measure used in the median and the scale measure is the inter quartile range (IQR). This type of outlier detection is performed for workplaces at the micro data level. The sensitivity of the process can be adjusted to suit the survey's needs.

To be able to perform outlier detection successfully with business survey data, one has to satisfy two criteria: (a) data homogeneity, and (b) data symmetry. Achieving data homogeneity obviates the need to use design weights when pooling neighbouring strata to increase the resolution of the outlier routine. Data homogeneity reduces the effect of the design and the complex problem of identifying aberrant observations in a sample drawn from a finite population reduces to a much simpler problem of dealing with outliers in the context of an infinite population. Homogeneity can be achieved by applying an appropriate transformation to one or more variables. The transformed data are then tested for approximate symmetry.