12 Chapter Statistical Techniques -- II

Chair: Linda Stinson, Bureau of Labor Statistics

David A. Pierce ◆ Laura Bauer Gillis

Peter Ochshorn

James Kennedy

12

Chapter

Time Series and Cross Section Edits

David A. Pierce and Laura Bauer Gillis
Federal Reserve Board

Abstract

Much editing of data from repeated surveys and reports is based on comparing the current or incoming value for a variable or item to that variable's value for the previous week, using a set of published **tolerances**. The previous value represents an estimate or forecast of what the current value would be in the absence of error or unusual circumstance. This paper investigates two generalizations of this editing method, which both involve incorporating information beyond that contained in the previous week's value. One of these is to base this estimate on the item values from a **cross section** of similar institutions in the current time period which have already reported, and the other is to calculate a forecast based on the **time series** of past values of the item. A composite estimate combining these two methods is also presented.

These methods are applied to data from the major deposit reports submitted by commercial banks to the Federal Reserve System. Edit simulations are performed to measure the improvement from this approach (in terms of fewer edit exceptions which are correct and/or increased detection of errors), which is found to be substantial for some items and size groups. Efforts thus far to implement these enhancements are described, and possible further generalizations are mentioned.

Time Series and Cross Section Edits

David A. Pierce and Laura Bauer Gillis
Federal Reserve Board

Background and Introduction

Data for the U.S. Money Supply are regularly transmitted to the Federal Reserve System by commercial banks and other financial institutions at weekly and other intervals. A major vehicle for this transmission is the "Report of Transactions Accounts, Other Deposits and Vault Cash," or simply the "Report of Deposits," on which banks and other financial institutions report weekly data for 25 deposit categories and related items. Based on these data and on similar information contained in other reports, the money supply measures are constructed and reserve requirements are maintained.

The money and reserves figures are important both as barometers of economic activity and in enabling the Federal Reserve to perform its economic stabilization and bank regulatory functions, and it is essential that the data submitted on the Report of Deposits and other reports be reliable and of high quality. To ensure their accuracy, all such data are subjected to numerical edits to detect unusual or deviant values. These edits are to two general types, **validity** edits to ensure that adding-up and other logical constraints are satisfied, and **quality** edits based on statistical or distributional aspects of the data.

The most commonly used quality edit involves the comparison of an incoming weekly figure to the previous value of that variable (in both dollar and percentage terms), using a tolerance band constructed about that value. The **tolerances**, or half-widths of the tolerance bands, are determined from previous estimates of the variable's distribution, in particular measures of spread, and are published in aTechnical Memorandum or "Tech Memo" (Federal Reserve Board, 1993). An edit "exception" occurs if the incoming value falls outside this tolerance band; when this happens, the reporting bank or other institution may be contacted for verification or correction. All tolerance-table comparisons are made (and edit exceptions generated) by machine, whereas the decision to contact the respondent is made by data analysts. The editing is done at both the Federal Reserve Board and the 12 Federal Reserve Banks.

Edits are in essence hypothesis tests, and both Type I and Type II errors can occur. A major task in setting edit tolerances is to ensure adequate sensitivity without generating unnecessarily large quantities of "false positive" edit exceptions. It is because of the large number of these exceptions that editing at both the Reserve Banks and the Board is currently quite labor intensive. All exceptions are reviewed by data analysts who must decide which are to be referred to the respondent institution for verification or revision. At the same time, a large majority of the data errors are not caught by these edits, based on the historical record of revisions submitted by respondents (they may be detected by other edits at a later date). There is consequently a need both to increase the sensitivity of the edits and to streamline the data editing process.

The value to which the tolerances are applied is in effect an estimate or forecast of the incoming figure that is being edited in the absence of error or unusual circumstance. By basing this forecast or estimate on information beyond that contained in the previous week's value, we obtain the generalizations of the current editing method that are investigated in this paper. One generalization is to base this estimate on the item values from a **cross section** of similar institutions in the current time period which have already reported, intending to capture economic, institutional or calendar movements which tend to affect similar respondents in a similar manner. The other is to calculate a forecast based on the **time series** of past values of the item for that respondent, including possibly last month's or last year's figures in addition to the one for last week as in the current procedure. A composite estimate combining these two methods is also investigated, the idea being that each method may incorporate information not captured by the other. (We also generated a composite of the cross section and current edits).

The paper's focus is on the data submitted on the Report of Deposits, also known as the Edited Data Deposits System (EDDS) Report. We investigated four of the more important items on this report, total transactions deposits, total savings deposits, and large and small time deposits. The study was motivated by the desire for greater automation in the Federal Reserve Board's Division of Information Resources Management, which carries out the edits. The improvements resulting from the study are being incorporated into a new software package called DEEP (Distributed EDDS Editing Project), for interactive editing on the PC. (For more detail see, Pierce and Gillis, 1995.)

Our results vary greatly according to item, entity type (e.g., commercial bank, credit union, etc.), and the amount of data in an institution group -- the latter being important for reliable cross-section estimates. In most cases we find that, with sufficient data, the cross section approach is as reliable as the current editing procedure. For total transactions deposits almost uniformly, and for total savings deposits for most commercial bank categories, time series modelling plays a significant role in the edits.

The following section of the paper discusses in greater detail the methodology underlying the different data editing approaches investigated. The third section then describes a set of edit simulations we performed with each of the five types of edits studied, and presents the results of these. Based on the simulation results, we provided a set of recommendations for experimental edits for DEEP, for each entity type and item, which have recently become operational.

Methodology

Given a variable or item of interest, many data editing procedures can be characterized as first generating a forecast (a point estimate) of the incoming value for that item, next applying a tolerance to the forecast to form a tolerance interval (an interval estimate) for the incoming value, and then flagging that value if it is outside the tolerance interval. In the current editing framework, that forecast is taken to be the previous week's item value, and the tolerance is as given by the Tech Memo (Federal Reserve Board, 1993). In this section the two generalizations to the forecast noted above are presented, along with composite procedures, after first describing the data and framework used.

Choice of Items and Statistical Form

The current approach to editing data from financial institutions is to subdivide them into homogeneous "cells," which are combinations of an institution's size group, entity type, and geographic location. There are six size groups for commercial banks and a smaller number of size groups for each of the other entity types, which are credit unions, savings and loans, savings banks, agencies and branches of foreign banks, and Edge and Agreement Corporations. The geographic locations are defined in terms of 12 Federal Reserve districts.

There are thus a great many edit cells, and to make our task manageable, and to achieve comparability with the current edits, we have simplified this study in the following ways:

❒ Staying with the **same cells** of the current EDDS edits. This will facilitate assessing the effects of the cross section estimates, model forecasts, and composite procedures. We recognize that more sophisticated groupings into cells may enhance the performance of the edits and plan to work with these in the future. Also we have eliminated all acquisitions and mergers from the institutions studied and have placed "credit-card banks" in a separate group.

❒ Maintaining the **same tolerance** widths as currently (applied, however, to the time series / cross section estimates that we generate, as well as to the most recent value as currently done). This may at first seem unnecessary, since standard deviations, percentiles, and other aspects of the distribution can be determined from either the cross section data or the historical model. However, such calculations can sometimes be unreliable, especially with cross sections without at least several hundred institutions in a group, as we are working with the extremes of distributions. And as with the cells themselves, keeping the current cell tolerance-interval widths facilitates comparisons among procedures.

We have also confined our attention in this study to the smaller institutions ("Priority-3" or P-3 institutions), where there may be the greatest potential for human resource savings from this approach. (Essentially this excludes the largest three size groups for commercial banks and a portion of the largest size group for other entity types.) For these institutions, we have examined the following items:

Total transactions deposits
Savings deposits
Large time deposits
Small time deposits.

Current EDDS editing is performed with both dollar and percentage changes of the item being edited, with both required to exceed tolerances ("and" condition) for an exception to occur The modifications outlined in this report are only for percentage changes; the Tech Memo tolerances continue to be applied to the dollar changes. There are several reasons for choosing percentage changes as the focus. Since they are used in current edits, the present edit cells and tolerances can be employed, and comparisons with current procedures can be made. They (or their annualized versions, growth rates) are also used in other analyses, such as with the Small Bank Sample of early reporting institutions. They are more homogeneous than dollar changes among different sized institutions, so that fewer edit groupings should eventually be needed. Percentage changes were found to be more sensitive to reporting and other errors than ratios to other items such as total deposits, which change with the denominator as well as the numerator and moreover present difficulty when the denominator was zero.

Cross Section Edits

Period-to-period edits compare an institution's current value for an item to the previous period's value. However, useful additional information may be contained in the current values of that item for other institutions that are similar to the one being edited. For example, if most of the institutions in a group experience a surge in large time deposits in a given week, then it would probably be inaccurate to list them as exceptions simply because they were outside the EDDS tolerances. Conversely, a very small change that week in large time deposits for a particular institution in that group may be suspicious even though current period-to-period tolerances would not be exceeded.

Cross section edits are carried out by examining the distribution of values (here, of percentage changes) for institutions within a homogeneous group, and listing as exceptions any values that were unusual compared to that distribution. Ordinarily one would calculate the mean and standard deviation of the percentage changes and flag those that were farther away from the mean than (say) two or three standard deviations; but in the present study we modified this set-up in two ways. First, because extreme values (the ones we hope to detect) would themselves influence the mean to which they would be compared, we "trimmed" the mean by eliminating the largest and smallest 5 percent of the values before calculating the estimated mean. Second, more observations are required to form a reliable estimate of the standard deviation than of the mean, and since most of the cells or groupings of institutions were too small for this, we chose to use multiples of the current EDDS tolerances as proxies for the standard deviations. As noted earlier, an additional advantage of this practice is to facilitate comparisons with the current edits.

One difficulty in using a cross section edit is that the data for an editing group need to be available in order to calculate such quantities as the average percentage change for that group. But the data for Priority-3 institutions are not due at the Board until nine days after the as-of date; and since timely estimates of the monetary aggregates and required reserves are needed, the editing process cannot be postponed this long. Our solution to this is to wait until a large enough fraction of the institutions have reported, and to form the distributional estimates (the trimmed means in this case) from the data available at that time.

For the EDDS data, more than half of the P-3 institutions' records are received by the Federal Reserve Board on the Thursday night following the as-of date (the previous Monday, on which the statement week ends), with the majority of those outstanding arriving by Friday night and the few remaining ones by the following Wednesday. For this study it was, therefore, decided to start the cross section editing on Friday morning. In either case, the trimmed mean estimates initially formed are not modified when more institutions have reported, in order not to confuse the editing process.

Some of the editing cells contain only a small number of respondents (and an even smaller number reporting by Friday), so that the estimated mean for those cells may not be very reliable. We required a minimum of 50 available observations in order to use the cross section estimate by itself. If the number of available observations is less than 50 but at least 20, a composite of that estimate and the previous week's value for the institution is employed, and with less than 20 the previous week's value alone is used.

The cross section edit is performed by comparing the deviation between the observed and the estimated percentage changes to the current EDDS edit tolerance for the item. As noted earlier, if the percentage-change condition is violated, then a second comparison of the magnitude of the dollar change versus its tolerance is performed, and the item is flagged only if both sets of tolerances are exceeded. An exception to this is that, as is done with the current edits, when the item changes from zero to a nonzero value or vice versa, the current dollar-change edit tolerances are applied without any adjustment.

Time Series Edits

These edits are based on time series **models**, which predict or explain an item's present value in terms of its past history. This usually involves the immediately previous value, on which the current edits are based, and often additional values as well, such as last year's. To the extent that these more distant values are important in predicting the incoming value, more sensitive edits should result from taking them into account.

Editing using a time series model for generating forecasts of percentage changes implies that a historical relationship exists between the item and its previous values. The "random walk" model is a time series model in which the best forecast of the current value is simply last week's value. Thus, the random walk model is implied by the current period-to-period change edits, which take last week's value as the current-period forecast around which the tolerances are applied. More complicated time series models yield forecasts which are weighted averages of several past values of the percentage change.

We first investigated the fitting of time series models for each institution separately. Some institutions' data fit the models quite well, with reductions in the standard deviation of the forecast errors (a key to the effectiveness of tolerances of a given width) of 50 percent or more, while other institutions exhibited only weak fits, or only the random walk behavior that the current editing framework already captures. Although fitting individual models is the preferable method for forecasting, it was not feasible to maintain over 7,000 models for each item edited within the DEEP framework -- at least not at this time. Thus, at this stage and for the P-3 institutions, a single time series model was fit to each editing cell's aggregate, and the coefficients from that estimated model were used to obtain an individual bank's forecast using its own previous values. While the benefits of time series modelling are reduced by doing this, the method can be easily implemented, and updated when necessary. Another constraint at present is that, because of data storage limitations, we only utilized terms in the model at lags of 1, 2, 3, 52 and 53 weeks, thus capturing nearby effects and annual seasonal influences but not, say, monthly or quarterly effects.

As an example of the model-fitting results, Table 1 provides information on time series models fit to cell aggregates of Total Transactions Deposits for three of the editing cells. Notice the highly statistically significant seasonal effect (lag 52, and in some cases lag 53). The strength of the fit declines going down the page, with the third one (Edges & Agreements, a root MSE reduction of only 9.2 percent) being not much different from the random walk model underlying current edits. On the other hand the results suggest that model-based editing may be valuable for certain commercial bank cells, for total transactions.

As with cross section edits, the deviation between the actual percentage change and the forecasted change from the time series model is compared to the edit tolerances. A tolerance exceedance both here and on the dollar change (also using current EDDS tolerances) triggers an edit exception for the record.

Composite Edits

The cross section and time series edits are based on different sets of information, past values of the institution being edited and present values of similar institutions. Thus a forecast which combined these two estimates, thereby utilizing both sources of information, may be more accurate than either one separately, and edits derived from such forecasts correspondingly more sensitive.

Table 1. --Percentage Change Models for Total Transactions Aggregates, Selected Editing Cells

------------------Cell = CB, Size Group 4, Region I------------------

Root MSE(orig.) = 0.0383 Root MSE(model) = 0.0211
Reduction in Root MSE = 44.9 percent

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.4349	0.0483	-9.005	0.0001
TRN_{t-2}	-0.0341	0.0329	-1.039	0.2996
TRN_{t-3}	-0.1510	0.0338	-4.467	0.0001
TRN_{t-52}	0.6494	0.0318	20.391	0.0001
TRN_{t-53}	0.4668	0.0440	10.606	0.0001

---------------Cell = CU, Size Group 2, Regions II&III---------------

Root MSE(orig.) = 0.1067 Root MSE(model)=0.0809
Reduction in Root MSE = 24.2 percent

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.2450	0.0546	-4.486	0.0001
TRN_{t-2}	-0.1160	0.0474	-2.444	0.0151
TRN_{t-3}	-0.2200	0.0486	-4.525	0.0001
$TRNt_{t-52}$	0.4922	0.0477	10.312	0.0001
TRN_{t-53}	0.1866	0.0533	3.498	0.0005

----------------------------Cell = EA, All---------------------------

Root MSE(orig.) = 0.0564 Root MSE(model)=0.0512
Reduction in Root MSE = 9.2 percent

Variable	Parameter Estimate	Standard Error	T-stat	p-value
TRN_{t-1}	-0.3776	0.0569	-6.632	0.0001
TRN_{t-2}	-0.1547	0.0586	-2.642	0.0087
TRN_{t-3}	-0.0449	0.0553	-0.815	0.4181
TRN_{t-52}	0.2432	0.0524	4.638	0.0001
TRN_{t-53}	0.1057	0.0540	1.955	0.0514

For a given institution (e.g., bank) and a given item, if T denotes a time series estimate (forecast) for a given week, C represents a cross section estimate, and A the actual value that is reported, then the **composite estimate** is a weighted average of T and C which is of the form

$$wT + (1-w)C.$$

The weights w and 1-w depend on the relative sizes and the correlation between the estimation / forecast errors of T and C. If these errors are given by

$$ET = A - T \text{ and } EC = A - C ,$$

then

$$w = [Var(EC) - Cov(ET,EC)] / Var(ET-EC) .$$

A composite forecast is thus a weighted average of individual component forecasts where the relative weights are chosen to minimize the sum of the squared forecast or estimation errors, and where the sum of the weights is one.

Using past data, we investigated a composite estimate of the cross section (CS) and the time series (TS) forecasts, denoted "CSTS", for each editing cell and each variable or item. The composite forecast defaults to the time series forecast with fewer than 20 available observations in the cell average. (With exactly 20 and using the 90 percent trim, 18 observed changes would be used in the cell estimate.)

The other type of composite edit we considered combines the cross section and the random walk forecasts (CSRW). We employed this edit when a CS edit was indicated but the sample size -- the number of observations available on Friday morning when the cell means are formed -- was insufficient (less than 50) to obtain an adequately reliable cross section estimate. For very small sample sizes (less than 20), our procedure is to revert to the use of only the RW edit.

Modelling and Simulation

To examine the relative performance of different types of edits, we conducted simulations of these edits over the 1991-1992 time period. For each cell (choice of item, entity type, size group and geographic region), we performed five sets of simulations, corresponding to the different types of edits under consideration: current (random walk), cross section, time series, cross section/time series composite, and cross section/random walk composite.

Simulation Procedure

Data preparation was a time consuming task. First, all Priority-3 reporters' weekly average data were compiled for the period from January 1986 through December 1992. While the edits were simulated only for the most recent two years, the additional data were used for fitting time series models with potential annual patterns. To avoid distortions, we eliminated all banks involved in mergers during this period. We next partitioned the data set into the editing groups or cells. We found that not all cells had a sufficient number of reporters to fit a model or to obtain reliable cross section estimates, and so some of them were combined. For commercial banks of size group 3 (total deposits between $1B+ and $3B), there were too few P-3 reporters to employ any of the new approaches. In addition, we added an editing category for known credit card banks. In total there were 40 edit cells, 37 of which were involved in the simulations.

Once the data were prepared, time series models were fit to the percentage changes in each cell's aggregate, as described above. Using the fitted model for a cell, predicted values for the last two years were generated for each institution in the cell. (Although forecasted values of the percentage change were generated for all periods, those in which a change of zero to a value or a value to zero were edited using special tolerances). Both the model-based and the zero-valued random walk forecasts were assigned to each observation in the cell. The 10 percent trimmed mean of the percentage changes was also calculated for each cell and each week of the two year simulation period, for use in the cross section edits. Since the cross section simulation employed all the data within a cell to calculate the current-period forecast, rather than the available data as of Friday morning when editing begins, the simulated results will differ from those in practice. In order to generate the two composite forecasts, the prediction errors from the original three forecasts were computed and the formulas in the section on methodology applied by institution. A cell root mean square prediction error (RMSE) was also computed.

Since the composite forecast combines the component forecasts in such a way as to minimize the sum of the squared prediction errors, we chose to estimate the appropriate weights for each bank in a cell and then to average those weights over the cell in order to obtain the composite for editing. Since the composite is a weighted average of the individual forecasts, the sum of the weights must equal one. For some institutions, where the prediction errors were very highly correlated between methods, we obtained pairs of weights with one value less than zero and the other greater than one. Evidently it only requires a small number of observations away from that correlation structure to cause such disproportionate weights. In calculating the average pair of composite weights for each cell, therefore, we first screened out those sets of weights not within the (0,1) range. After the two composite weighting schemes were determined for each cell, the mean square prediction errors were computed for these two forecasts as well.

For each of the five edit methods, Table 2 presents the root mean square prediction errors and composite weights for the commercial bank cells, for total transactions and total savings deposits. Similar calculations for other entities (savings and loans, etc.) and other items were also made. We anticipate the method with the smallest forecasting error to have the best potential as an edit, but until our tolerances are better tailored to the actual editing method, this potential may not be realized.

To apply the edits, we first looked for percentage changes that differed from the forecasted percentage changes by more than the appropriate tolerance (whether taken from the Tech Memo or generated as described in this paragraph), and for those ascertaining whether the dollar change tolerance was also exceeded. Since total savings and large time deposits are currently edited items, their current tolerances can be used. However, for total transactions and small time deposits, current tolerances do not exist. We therefore generated tolerances in a manner similar to that used for the creation of the current ones. This involved iterative steps with the intent of flagging approximately 0.3 percent of the observations per cell on average (the maximum percentage of observations flagged using current editing methods for other items for the year 1991). Using the components of total transactions and items that were related to small time, such as total and large time, we first compiled a range of feasible values for the tolerances. We then examined where these values occurred on the distribution of percentage changes over each cell for the two-year period. Given a reasonable proportion of the changes exceeding the initial values, we then examined the dollar change distribution for the subset of percentage change exceptions. Percentiles of this distribution were then determined in order to obtain the expected 0.3 percent edit failures under the current random walk model. These percentiles became the dollar change tolerances.

Table 2. --Root Mean Square Errors for Forecasts, Commercial Bank Cells

A. *Total Transactions*

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.077	0.073	0.077	0.074	0.071	0.72	0.51
-Size 5	0.096	0.094	0.097	0.094	0.089	0.73	0.55
-Size 6	1.190	1.190	1.276	1.190	1.204	0.70	0.58
Region 2							
-Size 4	0.064	0.059	0.236	0.060	0.121	0.77	0.58
-Size 5	0.210	0.209	0.223	0.209	0.212	0.62	0.55
-Size 6	0.331	0.330	0.344	0.330	0.333	0.68	0.57
Region 3							
-Size 4	0.102	0.099	0.108	0.100	0.100	0.75	0.51
-Size 5	0.054	0.048	0.051	0.050	0.046	0.74	0.58
-Size 6	0.067	0.063	0.071	0.064	0.062	0.70	0.60

B. *Total Savings*

	Root Mean Square Error					Weight of CS in Composite	
Cell	RW	CS	TS	CSRW	CSTS	CSRW	CSTS
Region 1							
-Size 4	0.042	0.042	0.045	0.042	0.042	0.64	0.73
-Size 5	0.054	0.054	0.056	0.054	0.054	0.64	0.67
-Size 6	0.048	0.048	0.055	0.048	0.048	0.60	0.72
Region 2							
-Size 4	0.038	0.038	0.099	0.038	0.043	0.65	0.76
-Size 5	0.235	0.234	0.244	0.234	0.236	0.64	0.64
-Size 6	0.055	0.055	0.067	0.055	0.055	0.64	0.66
Region 3							
-Size 4	0.051	0.051	0.998	0.051	0.274	0.68	0.74
-Size 5	0.041	0.040	0.041	0.040	0.040	0.63	0.66
-Size 6	0.055	0.055	0.065	0.055	0.055	0.61	0.75

Once all the forecasts and tolerances were in place, the editing experience for the 1991-1992 period was simulated for each of the five forecast methods. For each method we observed which observations were flagged as edit exceptions. Then, based on a history of weekly revisions to the EDDS file maintained by the Federal Reserve's Statistical Services branch, we were able to determine the rate of type I and type II errors for each method. (A type I error or "false positive" refers to an item that was flagged but not in error, or at least not revised. A type II error occurs when an item is not flagged but is erroneous -- as evidenced by a later revision.)

Simulation Results

For reference in this section, Table 3 shows our recommended edits based on these simulations. As mentioned in Section 1, these have been implemented as part of the Federal Reserve Board's DEEP editing software. In Table 3, the left column lists the entities (with the included size groups in parentheses), followed by the chosen edit for each item.

Details of the results on which this table is based are contained in earlier reports available on request. To give the flavor of the analysis, Table 4 summarizes the editing simulations for commercial banks' total transactions deposits; those for other entity types and other variables were similar. To assess the magnitudes and the implications of errors caught and errors missed by the editing schemes, Table 4 breaks down these errors in terms of their size (i.e., the size of the revision -- we assume, however accurately, that revised data are correct and the revision is the error in the unrevised data). It is clear from these simulations that there is room for improvement, especially regarding the type II error probabilities, which range from 98 percent to 99 percent. And although the type I error probabilities appear small, the number of flagged items that are not in error is quite large (between 87 percent and 94 percent).

Wherever the fitted time series model indicated a potentially substantial payoff relative to the random walk model (as in the first model in Table 1), the time series edit tended to be the most accurate, yielding the smallest number of edit exceptions and with fewer errors missed that were captured by other methods than vice versa. The reduction in the number of edit exceptions was not as great for the CS and CSTS composite methods, but often the composite method caused less of an increase in the type II error probability. The CS and the CSRW composite often mimicked the current RW results. Where there was doubt regarding the preferable edit method, we tended to favor the CS or CSRW -- even when the reduction in RMSE and the number of edit exceptions was small relative to the current (RW) method -- since cross section edits would allow possibly large shifts in behavior for a given week to be incorporated into the editing norm, and the DEEP software is well-suited to this type of edit. Also, we gave some preference to a uniformity of editing method across related cells (e.g., adjacent size groups within an FR region, or like size groups between regions).

For commercial banks, the alternative edits on the whole did quite well. The time series edits for total transactions and total savings were effective in reducing the total number of exceptions while missing only 3 small revisions and actually finding an additional error of over $25M. (This revision was generated either by an outside source or by an edit of another report that is not being considered here. This occurrence brings to light that some errors are detected by other sources - not the Reserve Banks or the Board. What we gain from this additional edit exception an earlier detection of the error; it would not necessarily go undetected permanently.) For the other entity types, total transactions was the only item that allowed for an alternative other than the CSRW method (CSRW was selected for

Table 3. --Experimental Edits for DEEP

Institution	Total Transactions	Total Savings	Large Time	Small Time
Commercial Banks (3,Ccd)	RW	RW	RW	RW
Commercial Banks (4,5,6)	CSTS	TS	CS	CS
Credit Unions (1,2,3,4)	TS	CSRW	CSRW	CSRW
S&Ls, Coops, Sbs (1,2,3,4)	ℜI ℜII-IV TS CSTS	CSRW	CSRW	CSRW
Agencies & Brs.(1,2,3)	CSTS	CSRW	CSRW	CSRW
Edges & Agr. (1,2)	CSRW	CSRW	CSRW	CSRW

The numbers in parentheses are the size groups, with "Ccd" denoting credit card banks. CB size groups 1 and 2 are omitted, as they are priority 1 and 2 institutions. ℜ denotes the FR Region, as in TM#16. The other entries in this table have the following explanations:

TS: The time-series model-based forecast, utilizing the institution's past percentage changes (of 1, 2, 3, 52, and 53 weeks ago).

CS: The cross-section forecast, or estimate of the average percentage change over all the institutions in the editing group or cell. Uses only the data received by the Friday after the as-of date and is calculated as the 90 percemt trimmed mean of the individual percentage changes in the cell.

CSTS: A weighted average of the TS and CS percentage-change forecasts, with statistically determined weights. When the number (n) of institutions in the group available on Friday for calculating the mean is less than 20, the weights are 1 and 0 (only the TS forecast is used).

RW: The forecast based on the "random walk" model, or the time series model giving a zero period-to-period change as the best forecast -- and is thus the implicit model underlying the current edits. This translates into a percentage-change forecast of zero.

CSRW: The forecast based on a composite of the CS and RW estimates of the percentage change, again depending on the number n of available observations in the cell. Thus:
if $n \geq 50$, use CS only;
if $20 \leq n < 50$, use weighted average of the CS and RW estimates;
if $n < 20$, use the RW estimate (zero percentage change forecast).

Table 4.--Selected Editing Simulation Results for Commercial Banks

A. *Total Transactions*

1. Random Walk (Standard Edit)

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M >$25M	< $25M	Total
Not flagged	557,166 97.76	9,732 1.71	791 0.14	508 0.09	168 0.03	568,365 99.73
Flagged	1,444 0.25	75 0.01	17 0.00	12 0.00	6 0.00	1,554 0.27
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.26 percent
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.0 percent
Pr(Item not in error | Item Flagged) = 92.9 percent

2. Cross Section -- Time Series Composite

Frequency/ Percent	Not Revised	<$5M	$5M <$10M	$10M >$25M	<$25M	Total
Not flagged	557,326 97.78	9,743 1.71	792 0.14	509 0.09	167 0.03	568,537 99.76
Flagged	1,444 0.23	75 0.01	17 0.00	12 0.00	6 0.00	1,554 0.24
Total	558,610 98.01	9,807 1.72	808 0.14	520 0.09	174 0.03	569,919 100.00

Pr(type I error) = Pr(Flag Item | Item not in error) = 0.23 percent
Pr(type II error) = Pr(Do not Flag Item | Item in error) = 99.1 percent
Pr(Item not in error | Item Flagged) = 92.9 percent

Reduction in edit exceptions = 11.1 percent
Reduction in type I error probability = 11.5 percent
Increase in type II error probability = 0.1 percent

these entity types in place of CS in order to accommodate smaller sample sizes in the preliminary data). Those credit unions and savings institutions which would have more activity in transactions accounts than the other entity types, do exhibit cyclical patterns which the time series model was able to capture. Agencies and branches also exhibited improved editing results with the CSTS method. As mentioned, this combination of alternative strategies yielded an 11 percent reduction in both the type I error probability and the number of edit exceptions, with only a very slight increase in the type II error likelihood (about 0.1 percent).

All of these results are based on simulations using data reported to the Board by the Reserve Banks. Thus, any errors caught at an earlier stage are not reflected in these data, nor are errors undetected by Banks or Board that do not show up in the revision files. And as previously mentioned, the other factor to be monitored is the use of preliminary data in cross section estimates of the mean percentage change. Depending on where the preliminary data fall in the distribution of all percentage changes for an item, the operational results based on the CS, CSTS, or CSRW methods may differ significantly from what is expected based on the simulation results.

This investigation is still in progress, and further generalizations of the work are underway or planned. Among these are examining time series models with regression components to account for such phenomena as tax dates, calendar effects or related variables, alternative groupings of the data according to size or geographic region, modelling larger banks individually, and examining additional items or variables.

References

Federal Reserve Board (1993). "Processing Procedures for the Report of Transaction Accounts, Other Deposits and Vault Cash (FR2900)," Technical Memorandum No. 16, Publications Section, (December).

Pierce, David A. and Laura Bauer Gillis (1995). "Time Series and Cross Section Edits with Applications to Federal Reserve Deposit Reports," *Seminar on New Directions in Statistical Methodology*, Statistical Policy Working Paper No. 23, Part 1, pp.152-171 (June). ■

12

Chapter

Inflation Factors for Stratified Samples with Control Information

Peter Ochshorn, University at Albany, SUNY

Abstract

This paper looks at the "inflation factor" or "weight" for the members of a stratified sample. As the inverse of the sampling ratio, it is often interpreted as the number of members of the population "represented" by each sample member. In the first section, standard estimators are reexamined to construct corresponding "implicit" inflation factors. Then, inflation factors are chosen to minimize error functions. Two examples are discussed, both of which are easily computed by linear programs. An extract of the data for the problem will also be displayed.

Inflation Factors for Stratified Samples with Control Information

Peter Ochshorn, University at Albany, SUNY

Introduction

The quantity $w_h = N_h/n_h$ is sometimes referred to as an "inflation factor" or "weight" for the members of a stratified sample survey in stratum h. As the inverse of the sampling ratio, it is often interpreted as the number of members of the population "represented" by each sample member. By applying the factors to the sample, sums of variates are "inflated" from the sample to the population, providing standard estimators such as $\sum_{h,i} w_h y_{hi}$ for the sum of variate y over the population. When covariate parameters or "controls" such as μ_{xh} are known, it is usually possible to construct more efficient estimators, such as ratio and regression estimators. For the theory, compare any sample survey textbook, e.g., (Cochran, 1977). These in turn require a degree of statistical sophistication not always present in the user of the sample, especially if the sample is made available to "outside" users. Additionally, tables of control information are not always available to these "outside" data users. Thus, there is some motivation to adjust the inflation factors, taking account of control information in some way to force better estimation.

Inflation Factor Form of Control-Based Estimators

In this section standard estimators are reexamined to construct corresponding "implicit" inflation factors. This can be thought of as "a priori" construction because the estimators usually are chosen according to their average behavior, i.e., prior to sampling. The advantages of the different choices are discussed in the textbooks. Equation numbers below beginning with "c" refer to (Cochran, 1977) (e.g., "c6.44/p164" is Cochran's equation 6.44 found on page 164).

The Separate Ratio Estimator

The inflation factor corresponding to the separate ratio estimator can be derived starting with equation c6.44/p. 164:

$$\sum_h \frac{y_h}{x_h} X_h = \sum_h \frac{N_h}{n_h} \frac{\mu_{xh}}{m_{xh}} \sum_i y_{hi} = \sum_h w_h \frac{\mu_{xh}}{m_{xh}} \sum_i y_{hi}$$

so that the implicit inflation factors are just $w_h(\mu_{xh}/m_{xh})$, the adjustment being the ratio of population to sample mean of x in each strata. In other words the sample is inflated by the ratio of x totals rather than counts.

The Combined Ratio Estimator

To derive the inflation factor corresponding to the combined ratio estimator, one can start with equation c6.48/p165:

$$\frac{\hat{Y}_{st}}{\hat{X}_{st}}X = \frac{N\mu_x}{\sum_h N_h m_{xh}} \sum_h \frac{N_h}{n_h} \sum_i y_{hi} = \frac{N\mu_x}{\sum_h N_h m_{xh}} \sum_h w_h \sum_i y_{hi}$$

so that in this case the adjustment to w_h is constant across strata, being the ratio of population to count-inflated sample totals of x. Another view of the adjustment factor is given by:

$$\left[\sum_h \frac{N_h}{N} (\frac{\mu_x}{m_{xh}})^{-1}\right]^{-1}$$

which is a harmonic mean of the strata ratios of population to sample mean of x, weighted by population count shares.

Separate Regression Estimator

Regression estimators can also be the starting point for control-based inflation. For the separate regression estimator the starting point can be equation c7.49/p200 (multiplied by N):

$$\sum_h N_h[m_{yh} + b_h(\mu_{xh} - m_{xh})] = \sum_h w_h \sum_i [1 + n_h(\mu_{xh} - m_{xh})a_{hi}]y_{hi}$$

where a_{hi} are the linear least squares transforms from $b_h = \sum_i a_{hi}y_{hi}$ and are equal to:

$$a_{hi} = (x_{hi} - m_{xh}) / \sum_i (x_{hi} - m_{xh})^2.$$

In this case the inflation factors are adjusted not only according to the difference (not ratio) of population and sample mean of x by stratum, but also to the distance of the covariate from its sample stratum mean. Thus each sample member receives an individual inflation factor, which varies within as well as among the strata.

Combined Regression Estimator

The final "a priori" inflation factor of this section is derived from the combined regression estimator. Starting with c7.50/p200 (multiplied by N):

$$(N_h/n_h)y_{hi}+b_c(N\mu_x-m_{xh}) = \sum_h w_h y_{hi}+b_c N_h(\mu_{xh}-m_{xh}) = \sum_h \sum_i [w_h+N_h(\mu_{xh}-m_{xh})a'_{hi}]y_{hi},$$

where $b_c = \sum_{h,i} a'_{hi} y_{hi}$ is given by equation c(mid-page)/p202 weighted combined regression. As in the separate regression estimator, the implicit inflation factor depends both on the difference between population and sample means in the strata, and on the distance between the covariate and its sample stratum mean.

Inflation Factors as Solutions to Ex-Post Optimal Programs

In this section inflation factors are chosen to minimize error functions. Instead of starting with an estimator to construct inflation factors, one can start with a comparison between control totals and inflated sample totals, minimizing some objective function of the discrepancies by choosing "optimal" inflation factors. An advantage of this approach is the flexibility in choosing the objective, allowing for various criteria to assume their respective importance in determining the solution. Two examples are discussed below, both of which are easily computed by linear programs.

Example 1

Suppose $\{z_j\}$ are a set of variables for which the population totals by strata Z_{jh} are known. Usually the unit variable whose totals are N_h and n_h for the population and sample respectively will be included in this set. Construct the ratios $w_{jh} = Z_{jh}/z_{jh}$ (again usually including $w_h = N_h/n_h$. Let $\{w^*_h\}$ be the set of inflation factors to be determined. Then for each stratum the discrepancy between sample and population for variable z_j is $w^*_h z_{jh} - Z_{jh}$. One possibility for an optimality criterion is the sum of the absolute percentage errors, over variables and strata, or

$$\sum_{j,h} |1 - w^*_h/w_{jh}|.$$

This objective is minimized by minimizing separately for each stratum $\sum_j |1 - w^*_h/w_{jh}|$, summing over the variables z_j. As a function of w^*_h this is a sum of piecewise-linear convex functions, and so inherits these traits. The optimum is therefore one among the $\{w_{hj}\}$. In many applications the solution will be close to the median of these values, but in any case never greater than the median.

To show this result, assume generically that $1 \leq w_{jh} < w_{j+1,h}$. Then the slope of the objective function between is just $w_{j'h}$ and $w_{(j'+1,h)}$ is just

$$\sum_{j\leq j'} -1/w_{jh} + \sum_{j>j'} +1/w_{jh}$$

which, since the individual terms $1 / w_{jh}$ are decreasing over j, must be positive whenever j' is such that the number of terms in the first sum is equal to greater than the number in the second sum. The objective minimum must therefore be less than or equal to all such j'.

In cases where, for each stratum h, the $\{w_{jh}\}_j$ are roughly of the same order of magnitude (a "nice" sampling outcome) then the median value is a likely solution (for this problem, for an even number of values, the lesser of the two middle values is taken as the median.) For example with three ratios, the slope between w_{h1} and w_{h2} is given by

$$+1/w_{h1} - 1/w_{h2} - \ 1/w_{h3},$$

and sufficient conditions for the negativity of this slope are given by $w_{h2} < 2w_{h1}$ and $w_{h3} < 2w_{h1}$.

Example 2

The final example is an actual application to an annual microdata file maintained by the New York State Department of Taxation and Finance, Office of Tax Policy Analysis. The data consist of approximately 90,000 records randomly sampled from a stratified population of about 8,000,000 tax filers. Stratification is by type of tax return (long form, short form, etc.) and by income class. Control information consists of strata totals of return counts, income, and tax liability.

Part of the art of constructing mathematical programs is in eliciting a hierarchy of preferences from users of the final "product." In the present instance it has been determined that it is important to reduce discrepancies in income totaled by income class, in counts totaled by return type, and in overall total tax liability. An implementation has been made using the AMPL (Fourer et al., 1983) algebraic modeling language with the MINOS (Murtaugh and Saunders, 1993) program solver. The objective function has been implemented as a weighted sum of absolute values of the discrepancies listed above. Such a minimization is inherently linear, and can be solved by the usual linear programming techniques. The appendix displays an AMPL model for this problem.

Acknowledgment

The work reported here was begun while the author was employed at the NYS Department of Taxation and Finance. The suggestion to explore the use of linear programs to inflate the personal income tax sample came from Frans Seastrand, Director of the Bureau of Revenue Analysis and Data.

References

Cochran, W. G. (1977). *Sampling Techniques* (third edition), Wiley, New York.

Fourer, R.; Gay, D. M.; and Kernighan, B. W. (1993). *AMPL: A Modeling Language for Mathematical Programming*. The Scientific Press, San Francisco.

Murtaugh, B. A. and Saunders, M. A. (1993). *MINOS 5.4 User's Guide*, Stanford University, Stanford CA.

Appendix: AMPL Model

The AMPL model below specifies the minimization of an error function as explained above. The inflation factors to be determined are coded in the statement beginning "var Infl ..."; the objective statement starts with "minimize Errors: ..."; and the constraints follow with the phrase "subject to ..." (the double construction for each constraint is a trick to convert absolute value problems into linear form)

```
set RetType ;
set IncCls ;
set Cells within {RetType,IncCls} ;

param CtrlInc {Cells} ;
param CtrlCnt {Cells} ;
param CtrlTax {Cells} ;
param SmplInc {Cells} ;
param SmplCnt {Cells} ;
param SmplTax {Cells} ;

param ClassCtrlInc {i in IncCls} := sum {(r,i) in Cells} CtrlInc[r,i] ;
param TypeCtrlCnt {r in RetType} := sum {(r,i) in Cells} CtrlCnt[r,i] ;

param TotCtrlInc := sum {(r,i) in Cells} CtrlInc[r,i] ;
param TotCtrlCnt := sum {(r,i) in Cells} CtrlCnt[r,i] ;
param TotCtrlTax := sum {(r,i) in Cells} CtrlTax[r,i] ;

param ClassIncWgt {i in IncCls} := ClassCtrlInc[i]/TotCtrlInc ;
param TypeCntWgt {r in RetType} := TypeCtrlCnt[r]/TotCtrlCnt ;

param IncWgt default 1 >0 ;
param InvCntRatio {(r,i) in Cells} := CtrlCnt[r,i]/SmplCnt[r,i] ;
param InvIncRatio {(r,i) in Cells} := (if not SmplInc[r,i]
    then InvCntRatio[r,i] else CtrlInc[r,i]/SmplInc[r,i]) ;
param InvTaxRatio {(r,i) in Cells} := (if not SmplTax[r,i]
    then InvCntRatio[r,i] else CtrlTax[r,i]/SmplTax[r,i]) ;
param MaxRatio {(r,i) in Cells} :=
    max ( InvIncRatio[r,i], InvCntRatio[r,i], InvTaxRatio[r,i] ) ;
param MinRatio {(r,i) in Cells} :=
    min ( InvIncRatio[r,i], InvCntRatio[r,i], InvTaxRatio[r,i] ) ;

var Infl{(r,i) in Cells}  >=MinRatio[r, i], <=MaxRatio[r,i],
    := InvCntRatio[r,i] ;

var ClassInflSmplInc {i in IncCls} = sum {(r,i) in Cells}      Infl[r,i]*SmplInc[r,i] ;
var TypeInflSmplCnt {r in RetType} = sum {(r,i) in Cells}     Infl[r,i]*SmplCnt[r,i] ;
var TotInflSmplTax = sum {(r,i) in Cells} Infl[r,i]*SmplTax[r,i] ;

var ClassIncErr {i in IncCls} = (if ClassCtrlInc[i]=0 then 0 else
    ClassInflSmplInc[i]/ClassCtrlInc[i] - 1) ;
var TypeCntErr {r in RetType} = TypeInflSmplCnt[r]/TypeCtrlCnt[r] - 1;
var TotTaxErr = TotInflSmplTax/TotCtrlTax - 1;

var AbsClassIncErr {IncCls} ;
var AbsTypeCntErr {RetType} ;
var AbsTotTaxErr;

minimize Errors:
    IncWgt*(sum {i in IncCls} ClassIncWgt[i]*AbsClassIncErr[i]) +
    (sum {r in RetType} TypeCntWgt[r]*AbsTypeCntErr[r]) +
    AbsTotTaxErr ;

subject to IncPos {i in IncCls}: AbsClassIncErr[i] >=  ClassIncErr[i] ;
subject to IncNeg {i in IncCls}: AbsClassIncErr[i] >= -ClassIncErr[i] ;
subject to CntPos {r in RetType}: AbsTypeCntErr[r] >=  TypeCntErr[r] ;
subject to CntNeg {r in RetType}: AbsTypeCntErr[r] >= -TypeCntErr[r] ;
subject to TaxPos: AbsTotTaxErr >=  TotTaxErr;
subject to TaxNeg: AbsTotTaxErr >= -TotTaxErr;
```

■

12

Chapter

Empirical Data Review: Objective Detection of Unusual Patterns of Data

James Kennedy, U. S. Bureau of Labor Statistics

Abstract

Hard edits are often coded into automated editing routines. Illogical and inconsistent responses are flagged; occasionally "soft edits," such as extreme numbers, are flagged as well. These responses are not necessarily incorrect, but require documentation. Often individual data are within normal limits, but the pattern of responses is unusual. The current paper discusses an empirical method for determining when patterns of data fall outside of a normal range.

An observation can be represented as a vector of N variables defining a point in N-dimensional hyperspace. Similar patterns are conceived to be located near one another in this hyperspace; cluster analysis produces the means of multidimensional clusters. The "unusualness" of patterns of data can then be defined in terms of an observation's distance from cluster centers.

The method is demonstrated with data from the COMP2000 generic leveling field test. Data patterns from nine generic leveling factors were analyzed into ten clusters. A Windows program demonstrates the comparison of new data with clusters found in previously-collected data. ■