Longitudinal Micro-Data Outlier Detection Techniques December 2006

Eric Simants
U.S. Bureau of Labor Statistics
2 Massachusetts Ave, NE
Washington, DC 20212

Abstract

Detecting outliers in longitudinal micro-data is a very involved task. First, the micro-data must be generated and stored in a manner that enables them to be linked through time. Additionally, since it is impractical to review each micro-data record, especially if the data series contains many millions of micro-data records, the data must be aggregated properly. Too high a level of aggregation may have the affect of removing variation in the time-series, thus eliminating observable outliers in the data. This paper presents techniques for identifying outliers in a time-series comprised of cross-sectional micro-data.

Keywords: Outliers, Longitudinal Database, Business Employment Dynamics

1. Introduction

For over half a century, the Bureau of Labor Statistics (BLS) has produced monthly net change data on nonfarm payroll employment. Even though these data are one of the most closely watched economic indicators in the United States, it does not provide a detailed breakdown of the factors that lead to the overall change. A net increase in employment, for instance, can be attributed to one of many factors, such as more business establishments expanding or opening. Declines in the number of business establishments that contract or close is an alternative possibility that can generate the same net result. From one quarter to the next, millions of business establishments make the decision to either expand, open, contract or close, all of which factor in the demand for labor in the job market. Combining this with the fact that many more millions of individuals make labor market decisions affecting the supply-side, a tremendous amount of job churn can be illustrated by tracking the number of gross job gains and gross job losses on an over-the-quarter basis. In an effort to gather and report more detailed data on the gross job flows that underlie the net employment change figure, the BLS released, in September 2003, its initial publication on the newly constructed Business Employment Dynamics program.

Derived from quarterly cross-sectional data that are gathered by each of the 50 states and the District of Columbia, the Business Employment Dynamics data present a new and unique method of analyzing the evolvement of the labor market by utilizing advanced techniques to track single business establishments across time. In order to compute the Business Employment Dynamics gross job gains and gross job losses, establishments that are found to be continuous from one quarter to the next are placed into one of four categories of data elements. Continuous establishments that report a higher level of employment in the third month of the current quarter than the same month in the previous quarter are classified as expanding. Conversely, contracting establishments are those that show a lower level of third month employment in the current quarter as compared to the prior quarter. Opening establishments consist of either those units that cannot be matched with any record existing in the database during the previous quarter, or are establishments that report a positive level of employment during the third month of the current quarter after having reported zero employment one quarter earlier. Similarly, establishments that either report positive third month employment in the prior quarter and report zero employment in the current quarter, as well as establishments that no longer exist in the database, are placed in the closings category. From these definitions, gross job gains and gross job losses can be calculated. The sum of the quarterly net employment change of all expanding and opening establishments equals the number of gross job gains, whereas gross job losses are derived by summing the quarterly net employment change of all contracting and closing establishments. The mathematical difference of the gross job gains and gross job losses is equal to the quarterly net change in employment.

Not only is it possible to draw revealing conclusions regarding labor market dynamics by comparing the gross job gains and gross job losses figures for one quarter because of its time series properties, the Business Employment Dynamics data provide further insight by tracking these figures over time. For example, the data reveal that the recession of 2001 was characterized by a temporary spike in gross job losses accompanied by a decline in gross job gains (Clayton and Spletzer, 2005).

Furthermore, the time series nature of the data allows for the detection of outliers from forecasted trends in the data by utilizing time series modeling techniques. Although it is not evident at levels of higher aggregation, specifically at the National level for all industry classifications, outliers are very apparent with further disaggregation.

Due to the fact that all of the tabulations used to calculate the Business Employment Dynamics statistics are based on micro-level establishment data, it becomes imperative that establishments reporting large levels of over-the-quarter employment change are reviewed for accuracy. Also of great importance is linking establishments that appear to be discontinuous, but rather are involved in a merger, acquisition, or some other form of administrative change. Not linking establishments properly across quarters can lead to an overstatement of one or more of the four data elements in the Business Employment Dynamics.

2. Business Employment Dynamics

The Business Employment Dynamics is a longitudinal data series that uses the establishment as the primary unit of analysis. An establishment is defined as an economic unit that produces goods and services, usually a physical location engaged in predominantly one type of economic activity (Spletzer et al, 2004). This differs from a firm in that a firm may be comprised of more than one establishment. Each establishment is tracked over time and is placed into one of four data element categories for a given quarter. These four data elements (expansions, openings, contractions and closings) can be summed to compute the more commonly known over-the-quarter net employment change figure.

By definition, the over-the-quarter net employment change is the sum of employment over all establishments in the current quarter less the sum of employment over all establishments in the previous quarter. This can be further decomposed by summing the net employment change of all establishments adding to its workforce from the prior quarter, a positive value, with the net employment change of all establishments showing a decline in its over-the-quarter payroll, which is a negative. The net change of all establishments increasing in employment, defined as gross job gains, is comprised of the total increase in employment at expanding establishments that report a larger positive level of employment in the current quarter compared to the prior quarter, along with the current quarter level of employment at opening establishments that either reported zero employment, or were nonexistent, in the prior quarter. Similarly, gross job losses are the sum of the net change in employment at contracting establishments reporting a lower level of employment in the current quarter than the previous one, combined with prior quarter level of employment of closing establishments that either reported zero employment in the current quarter after having reported a positive value in the prior quarter, or those establishments that are no longer reporting.

Business Employment Dynamics data from fourth quarter 2005 reveal that over 1.5 million establishments expanded, adding nearly 6.3 million new jobs. The number of establishments that opened or reopened was 375,000, showing a total of 1.5 million new jobs. The gross job gains figure of 7.8 million is slightly less than what was reported in the fourth quarter 2001, during the most recent recession. However, because the level of gross job losses, 7.3 million, was at one of the lowest points it has been for the past decade, the over-the-quarter net employment change exceeded 500,000 compared to a net loss of nearly 1 million in the fourth quarter 2001.

Since the Business Employment Dynamics data are based on the over-the-quarter changes in the employment levels at the establishment level, the accuracy of the statistics are directly dependent on the accuracy and timeliness of the reported cross-sectional micro-data. Obviously, data entry errors can skew one of the four data elements used to calculate the gross job gains and gross job loss statistics. Moreover,

misreported and delinquently reported administrative changes can wrongly inflate one or more of the four data elements. For example, establishment A and establishment B merge to create establishment C in the third quarter of a given year. In the process of consolidating its human resource operations, establishment C does not report its quarterly employment for third quarter. This situation would result in overstating the level of closings for third quarter because establishments A and B appear to be no longer in existence. To compound the situation, the level of openings will be falsely inflated when establishment C begins reporting employment in the fourth quarter. To ensure the highest level of data quality, each State Labor Market Information staff, as well as National and Regional BLS staff members, conduct extensive and thorough analytic reviews of the cross-sectional micro-data every quarter.

3. Quarterly Census of Employment and Wages

The Business Employment Dynamics data series is derived from micro-data collected through the BLS Quarterly Census of Employment and Wages (QCEW) program. It is a requirement that all employers subject to state Unemployment Insurance laws submit quarterly contribution reports to the State Employment Security Agency detailing the number of persons on its payroll for each month of the quarter along with the amount paid in wages during the quarter. Employment in the QCEW is defined as the number of workers whose wages are subject to unemployment insurance taxes and earned wages during the pay period that includes the 12^{th} of the month. These data present reliable and timely accounts of the number of workers employed in the United States.

Approximately 98% of employees on nonfarm payrolls are covered under unemployment insurance programs. In the fourth quarter of 2005, there were 8.7 million establishments reporting a total employment level of 134 million. Employers subject to unemployment insurance tax laws have one month from the end of the quarter to file reports to the state. After the data are collected, the State Labor Market Information staff has three months to obtain delinquent reports, comprehensively examine and, if necessary, edit these data before transmitting them to the BLS.

4. Longitudinal Database

Once the data are received at the BLS, they are again thoroughly checked for data entry errors, as well as any administrative changes that may have an adverse impact of the accuracy of the published data. Since these data are compiled from a virtual census, they are not subject to sampling error. Data entry errors are generally rare, and those that do occur are quite easy to detect and correct. The type of error that is most troublesome is caused by changes in administrative, rather than economic, data. Administrative changes that have the most significant impact on Business Employment Dynamics data stem from establishments becoming discontinuous in the data series because of a missed predecessor/successor relationship.

Another important responsibility of the State Labor Market Information staff is the assignment of a unique identifier to each reporting establishment. Every establishment is given a ten digit unemployment insurance (UI) account number. Tracking this unique identifying number from one quarter to the next is the primary method in determining the continuity of a given establishment.

In order to link establishments across quarters, the BLS designed the Longitudinal Database (LDB). The LDB currently contains over 400 million observations of establishment level micro-data from the first quarter of 1990 through the fourth quarter of 2005. Approximately 97% of all records processed in the current quarter are directly matched with the UI account number assigned to them from the preceding quarter. The number of all records that are not continuous, i.e. new openings, is around two percent each quarter. This leaves about one percent of establishments that are linked from a predecessor UI account number to a successor UI account number. The methods for detecting predecessor/successor relationships are; 1) systematically generated matches, 2) probability weighted matches, and 3) matches found by analysts during the data review process. A more detailed description of the LDB record linkage process can be found in Robertson, Huff, Mikkelson, Pivetz, and Winkler (1997).

It was with the initial release of the Business Employment Dynamics data in 2003 that the process of an analyst review of record linkages

began. However, it has only been within the past year that the entire LDB has been reviewed for missed, or false, linkages. This may be due to the fact that large, unusual variations in the time series are unobservable at the levels of aggregation originally published. For example, a missed link of a health services establishment in Alaska would have very little impact on the National data, even within the health and educational services industry super-sector. In preparation for the release of state-level Business Employment Dynamics data, over the past year the BLS has conducted a thorough and comprehensive analytical review of the entire LDB, broken down at the state-level.

5. Outliers in the State-Level Business Employment Dynamics Time Series

By definition, an outlier is some point in a data set that is very unique from the rest of the data based on certain criteria (Aggarwal and Yu, 2001). To be sure, any definition of outliers is broad, if not vague, thus making their detection even that more daunting of a task. Complicating matters further is the fact that there are two types of outliers, those that are caused by real events and those that result from gross errors in the data (Tolvi, 1998).

The process of identifying and correcting accounting and administrative errors in the state-level Business Employment Dynamics series was begun by carrying out direct seasonal adjustment runs on the four data elements for each state. The seasonal adjustment was processed using the U.S. Census Bureau's X-12-ARIMA program with an "airline" model, which is an ARIMA model to the order (0,1,1)(0,1,1) as defined by Box and Jenkins (1976). Detailed information regarding the X-12-ARIMA program can be obtained in Findley, Monsell, Bell, Otto, and Chen (1998).

Charts of the seasonally adjusted time-series were then created using SAS. These charts served as the initial approach for trying to determine those points in the data series that appeared unusual by visually identifying "spikes" in the graphs of the time-series. Although useful because of its simplicity, this naïve method quickly became unwieldy due to the fact that there were 204 separate time-series, all with disparate scalability. For instance, a "spike" in the time-series of the openings data element for a small state may only be a slight deviation from the normal trend. Conversely, a significant divergence from the trend in the openings data element for a given quarter in larger states may be unnoticeable using only a visual detection technique.

The next step involved running the seasonally adjusted time-series of all four data elements for each state through a macro routine using the SAS ARIMA procedure. Each series was tested for stationarity using augmented Dickey-Fuller unit-root tests, and, as is the case with most economic time-series, most of these series were non-stationary. In order to filter out the trends of the time-series, an ARIMA model of order (1,1,1) was finally selected.

Outliers were then found at points in the time-series where the observed seasonally adjusted value differed, positively or negatively, from the forecasted value by at least two standard deviations. In total, there were 392 points found to be outliers out of a possible 8,568, or less than 5% of all possible points. Many of the outliers were the result of true economic events. This was particularly true in the first quarter of 2002, the first point in the time-series following the 2001 recession. At this point the program was expecting a continuation of the high levels in the contractions and closings data elements in most states. Nevertheless, it was a fairly uncomplicated task to determine the points in the time-series where errors in administrative data occurred. Almost all of these errors took place prior to the initiation of the analyst review process.

Points in the time-series where outliers in either expansions or openings are followed, or preceded, by outliers in contractions or closings are obvious indications of missed predecessor/successor linkages. This is also true if the same phenomenon is observed in the same quarter. To be certain, with the exceptions of a recessionary downturn or an expansionary boom, either of which would only impact one side of the gross job flows data, it is highly unlikely for the labor market of any given state to change drastically over one quarter. In one case in particular, many erroneous partial predecessor/successor relationships occurred. This was caused by the fact that the predecessors, which should have remained active in the quarter the relationships took place, skipped one quarter of reporting. To give an example, establishment A reported 500

employees in the quarter prior to where a partial predecessor/successor transaction should have occurred. Establishments B and C are then reported with 100 employees each as successors to establishment A. However, the remaining 300 employees for establishment A that should have been reported in the same quarter the transaction occurred are not reported until the next quarter. The number of contractions in this case would be greater by 300 than it otherwise would have been if the data were correctly accounted for. Furthermore, there would be an additional 300 reported that should not be in the openings data for the subsequent quarter.

6. Outlier Detection of Cross-Sectional Micro-Data

As a measure of ensuring that none of the four Business Employment Dynamics data elements are overstated, an extensive review process of the cross-sectional micro-data is conducted each quarter. Even though the volume of records reported each quarter is staggering, coupled with the fact that the data are in high dimensional space, automated processes allow for a few analysts to review the data from all 50 states, and the District of Columbia, in a relatively short period of time.

The review primarily focuses on those establishments that show a large value of over-the-quarter employment change. Of particular concern are large establishments that are appearing in the LDB for the first time, as well as those that are no longer reporting, in the current quarter. It is uncommon for a single establishment to open or close reporting more than 100 employees.

Again, because of the high dimensionality of the data, the criteria of what constitutes a "large" change in over-the-quarter employment depends upon factors such as geographic location, industry decomposition, and seasonal factors. For example, a school located in a large county showing an increase of 200 or more employees in the third quarter, or a ski resort located in a mountainous state reporting a similar increase for fourth quarter, would not be reviewed insofar as its level of employment in the same quarter one year prior was comparable.

Although there is room for improving the process, filtering the data in this manner significantly reduces the number of establishments necessary to be reviewed each quarter. During the fourth quarter 2005 review process, less than 2,000, or 0.02%, out of 8.5 million establishments were identified as having reported questionable data. Of these approximately 2,000 establishments, 128 were determined to be erroneous and were corrected immediately.

7) Conclusion

In conclusion, the task of identifying outliers, and distinguishing true abnormal events from erroneous data, can be intimidating, especially when working with databases as vast as the LDB. However, the use of advanced statistical techniques and software packages can minimize the burden of highly labor-intensive processes. Once the process of cutting the data into smaller pieces, thus making it possible to detect outliers, it took a small number of analysts a relatively short amount of time to find the root cause of the outliers and make necessary corrections. Moreover, enhancing the review methods of quarterly cross-sectional micro-data will provide analysts the ability to work more effectively and efficiently, and is an essential part of ensuring that the Business Employment Dynamics data continue to be the highest quality data product of its kind.

Disclaimer

All views in this paper are those of the authors and do not necessarily reflect the policies of BLS or the views of its staff members.

References

Aggarwal, C., Philip Yu (2001). "Outlier Detection for High Dimensional Data", *ACM SIGMOD Conference Proceedings.*

Box, G. E. P., Jenkins, G. (1976), *Time Series Analysis: Forecasting and Control.* Holden-Day.

Clayton, Richard L., James R. Spletzer (2005), "Business Employment Dynamics", *NBER-CRIW Producer Dynamics Conference Proceedings.*

Findley, David F., Brian C. Monsell, William R. Bell, Mark C. Otto, Bor-Chung Chen (1998), "New Capabilities and Methods of X-12-ARIMA Seasonal Adjustment Program",

Journal of Business and Economic Statistics, Vol. 16, pp. 127-177.

Robertson, Kenneth, Larry Huff, Gordon Mikkelson, Timothy Pivetz, Alice Winkler (1997), "Improvements in Record Linkage Processes for the Bureau of Labor Statistics' Business Establishment List", *1997 Record Linkage Workshop and Exposition Proceedings*, pp. 212-221.

Spletzer, James R., R. Jason Faberman, Akbar Sadeghi, David M. Talan, Richard L. Clayton (2004), "Business Employment Dynamics", *Monthly Labor Review*, Vol. 127, No. 4, pp. 3-8.

Tolvi, Jussi (1998), "Outliers in Time Series: A Review", *University of Turku, Department of Economics*, Research report No. 76.