Outlier Detection by Forecasting

Nathan McDermott and
Brendan Livingston

The Consumer Expenditure Quarterly Interview Survey collects data from consumer units (CUs) about their expenses during the previous 3 months. The purpose of the survey is to gather information about large purchases, such as those of vehicles and appliances, and expenditures that are made on a regular basis, such as rent and utility payments. These data are collected by the U.S. Census Bureau and then transferred to the U.S. Bureau of Labor Statistics (BLS), Division of Consumer Expenditure Surveys (CE). The branch of Production and Control (P&C) screens and processes the raw data for their eventual use in publications and in the weighting of the BLS Consumer Price Index.

P&C's final data-editing procedure for the Interview Survey is the Monthly Tabulation of Expenditures (MTAB), which maps or assigns expenditures to a specific month and a Universal Classification Code (UCC).[1] The MTAB Review procedure then evaluates the created data for suspicious values. To improve the existing review procedure, P&C initiated a research project in August 2005. The goals of this project were to make the MTAB Review more efficient, focus analysts' attention on outliers, create more informative reports, and provide more accurate data to end users.

Three techniques for improving the process of selecting outliers were investigated during the modernization of the MTAB Review. The method that was chosen, which compared forecasted with reported values, was implemented in February 2006. With this technique, the analyst detects outliers by using forecasted prediction intervals created by SAS and comparing them with current means.[2] This article summarizes the forecasting technique adopted for the MTAB Review.

Background

After all quarterly data have been reviewed and deemed complete, the MTAB edit program produces a data set containing the monthly expenditure values. This data set contains approximately 450,000 observations per quarter, categorized into one of 600 UCC codes. A timing variable indicates whether the collected expenditure constitutes a continuous expense, with the same amount every month, or whether it represents a single monthly value. For continuous expenses, the MTAB edit creates three expenditure records, one for each month in the quarter. For all records, the amounts are assigned

Nathan McDermott is an economist formerly working in the Branch of Production and Control, Division of Consumer Expenditure Surveys, U.S. Bureau of Labor Statistics.

Brendan Livingston is an economist formerly working in the Branch of Production and Control, Division of Consumer Expenditure Surveys, U.S. Bureau of Labor Statistics.

[1] The Universal Classification Code, or UCC, is the lowest level of aggregation for consumer expenditures. For example, camping equipment, admission to sporting events, and men's shirts are categorized into different UCCs.

[2] Created by the SAS Institute, SAS, a statistical analysis software package, is widely used throughout the Bureau of Labor Statistics. More information about SAS can be found on the Internet at **www.sas.com**.

to a month of purchase and to the UCC for the appropriate category. After categorizing all of the expenditure variables, analysts review the MTAB data set for suspicious values.

Expenditure data typically have a skewed distribution, with a few extreme observations.[3] Only large expenditure values qualify for review. Extremely small values are usually considered legitimate. They are also too numerous and do not have a substantial enough effect on the mean to warrant investigation; therefore, they are not reviewed. Outliers can arise from unusually high reported expenditures, from incorrectly entered values or codes, or from other data-editing processes that estimate missing values. The MTAB Review procedure attempts to find, document, and manually fix these outliers.

The Interview Survey has several data-editing procedures, used throughout the production cycle, for identifying suspicious data. Screening at different classification levels ensures clean data. Current outlier detection techniques, besides those used for MTAB Review, consist of gap tests, *z*-scores, and mean comparisons. A gap test takes all values above the mean and sorts them in descending order. Then the difference between the expenditure and the value immediately below it is calculated. The largest gap is determined, and every value above the largest gap is flagged for further review. The Priority Index (PINDX), one kind of gap test, scores the difference between each value against the point immediately below it for all observations above the largest gap.[4] Any observation with a PINDX greater than 2.0, where the suspect value is 3 times larger than the value below it, is selected for manual review.

Z-scores use distributional statistics, such as the standard deviation or the interquartile range, to compare individual points against the population mean in relative terms. The standardized *z*-score is equal to the observation, minus the mean, divided by the standard deviation. For a two-tailed test, a *z*-score of 3 is in the 99th percentile. The CE uses a modified, more robust version of this test, in which the observation is divided by the interquartile range.[5] A "robust" *z*-score of 25 is considered large enough for the observation to be an outlier and is equal to approximately the 99.9th percentile.[6]

Unlike *z*-scores, mean comparisons consist of *t*-tests and other descriptive statistics that compare means between groups. Mean comparisons are useful because the mean is sensitive to extremely large values. Although *t*-tests, which use the standard deviation, are the most common type of mean comparison, a simple percent change also can be used. However, without any normalization, percent changes between means have no scale for comparison. Therefore, each record must be manually examined to determine whether it contains an outlier.

These different techniques continue to be used in statistical investigations. However, analysts believed that improvements could be made to the method used in the MTAB Review. A description of the old procedure and the new procedure that was adopted follows, along with a discussion of other methods that were considered.

Previous MTAB Review Procedure

The old MTAB Review procedure was based on comparing changes in mean values. Analysts received two worksheets to be used in detecting outliers. One worksheet compared the percent change from the current quarter with the percent changes from each of the previous three quarters; the second worksheet compared the percent change from the current quarter with that from the same quarter for the previous 3 years. (See example 1.) The comparison with the previous three quarters facilitated the detection of large single-quarter shifts, while the comparison with the same quarter for the previous 3 years looked for spikes in the yearly trends. Analysts then searched for particularly large percent changes in UCCs, where the percent changes were based on the categorical type of UCC.

MTAB Review worksheet

The old review procedure was particularly cumbersome for several reasons:

- The review consisted of manually comparing percentages for a very large number of groups.
- Each UCC appeared in both worksheets, together with the changes for the three respective quarters.
- Analysts reviewed every UCC, because there was no standardized method for identifying suspicious UCCs for further outlier review.

Methods Investigated

A number of outlier detection techniques were considered in the investigation of a new methodology for the MTAB Review procedure. One method compared histograms in order to identify distributional differences. Tests of the distribution of the current quarter against the previous quarter's distribution produced no reliable results, because outliers do not necessarily change the underlying distribution and single values are too hard to detect on a large scale.

A second method used *t*-tests to determine whether there was a statistical difference in the means. Because the skewed distributional pattern of the CE data did not meet all of the requirements of a regular *t*-test, the Wilcoxon rank-sum test, a nonparametric *t*-test,

[3] Expenditures are recorded as positive real numbers. Reimbursements can be recorded as negative values. The distribution is generally skewed to the right.

[4] For example, if the top values for a UCC were 150, 50, 45, 40, and 35, then the PINDX for the top observation would be ((150 – 50)/50) = 2.0.

[5] See Appendix A.

[6] The exact distribution of the "robust" *z*-score is unknown. The percentile approximation for a z-score of 25 is equal to 99.88. This is calculated by using income and expendi-ture data from 2004 through 2006.

Example 1. **Previous MTAB Review worksheet**

UCC	RTYPE	EXPNAME	PC_Q041	PC_Q042
220612 DWASH/DISP/HOOD CAP IMP	CRB	QADEQPX5	27400.00	27400.00
320522 PORTABLE HEATING/ COOLING EQUIP	CRB	QADEQPX1	520.69	27229.22
600210 GENERAL SPORT/EXERCISE EQUIP	FRA	FURNPURX	99.77	1610.33
790600 MAINT/REP/UTIL OTH PROP	CRB	QADPSPLX	.	1522.02
220615 CAP IMPROVE LABOR/MAT OWNV	CRB	QADLAB3X	741.76	1408.59
240321 ELEC SUPP, HEAT/COOL EQUIP RNTR	CRB	QADPSP3X	114.29	1400.00
790600 MAINT/REP/UTIL OTH PROP	CRB	QADLAB1X	.	1319.10
600121 BOAT W/O MOTOR/BOAT TRAILERS	OVB	QTRADEX	381.01	1285.41
870401 BOAT/TRAILERS, NOT FIN.	OVB	QTRADEX	381.01	1285.41
300322 MICROWAVE OVENS OWND	CRB	QADEQPX2	66.67	1007.26
450312 TRADE-IN ALLOWANCE/CAR LEASE	LSD	TRADEEXP	400.00	829.13
240311 PLUMBING SUPP/EQUIP RNTR	CRB	QADPSPLX	9556.98	777.91
220615 CAP IMPROVE LABOR/MAT OWNV	CRB	QADLAB2X	4560.39	616.02
240213 MAT/EQUIP FOR ROOF/GUTTER OWND	CRB	QADPSP2X	-79.74	606.58
230150 REP/MAINT LABOR/MAT RNTR	CRB	QADLAB1X	784.32	600.35

UCC	SC_Q034	SC_Q041	SC_Q042
220612 DWASH/DISP/HOOD CAP IMP	27400.00	.	14411.05
320522 PORTABLE HEATING/ COOLING EQUIP	27229.22	41.91	272.84
600210 GENERAL SPORT/EXERCISE EQUIP	1610.33	83.76	83.91
790600 MAINT/REP/UTIL OTH PROP	1522.02	.	.
220615 CAP IMPROVE LABOR/MAT OWNV	1408.59	918.70	420.13
240321 ELEC SUPP, HEAT/COOL EQUIP RNTR	1400.00	.	59.83
790600 MAINT/REP/UTIL OTH PROP	1319.10	.	.
600121 BOAT W/O MOTOR/BOAT TRAILERS	1285.41	512.94	269.26
870401 BOAT/TRAILERS, NOT FIN.	1285.41	512.94	269.26
300322 MICROWAVE OVENS OWND	1007.26	22.21	37.01
450312 TRADE-IN ALLOWANCE/CAR LEASE	829.13	421.16	208.21
240311 PLUMBING SUPP/EQUIP RNTR	777.91	765.55	5062.77
220615 CAP IMPROVE LABOR/MAT OWNV	616.02	2165.42	2513.58
240213 MAT/EQUIP FOR ROOF/GUTTER OWND	606.58	8.84	258.94
230150 REP/MAINT LABOR/MAT RNTR	600.35	751.57	612.17

was used. This test examines the distributional differences between two samples. A disadvantage of the test is that extremely large values do not have a significant impact on the ordinal ranking of observations and thus cannot be identified. Another disadvantage is that the test can compare the current quarter only with a single previous quarter; it cannot identify trends or seasonality sometimes found in CE data.

The final method investigated to detect outliers used a forecasting model to predict UCC means for the current quarter. The forecasting procedure, accounting for trend and seasonality, creates a prediction interval that is then compared against the actual mean. This method was determined to be the most effective, and it replaced the previous method beginning with the second quarter of 2005.

Adopted MTAB Review procedure

The new procedure uses forecasting to create a prediction for the current quarter of data and then compares the predicted value against the mean of actual data value collected in the current quarter. Let μ_t denote the collected mean of the current quarter. The input time-series data consist of quarterly means taken from the previous 10 years of data (from μ_{t-41} through the preceding quarter, μ_{t-1}). The procedure then forecasts a mean $\hat{\mu}_t$ and compares this predicted mean against the collected μ_t. The width of the confidence interval is calculated on the basis of the average number of observations from each quarter, and any collected μ_t that is greater than the upper bound of the confidence interval for the predicted μ_t will be output for the analyst to review.

Before forecasting, a check is run to ensure that there are enough observations for an accurate prediction. Any UCC that does not have at least 10 quarters of historical means, either because it was recently added or it was rarely collected, cannot be accurately forecasted and is output for manual review. This minimum requirement is satisfied for the majority of UCCs, including those collected annually, by using 10 previous years of data as the starting date for the collection. After making certain that the UCC has a sufficient number of observations, analysts test whether a logarithmic transformation is appropriate.

The LOGTEST macro applies a logarithmic test to each UCC that has 10 or more observations.[7] If the log-trans-

[7] The LOGTEST macro is included in the SAS/ETS software package. Details of the macro can be found on the Internet at **v8doc.sas.com/sashtml/ets/chap4/sect17.htm**.

formed model has a larger log likelihood than that of the untransformed model, then the log transformation is run on the UCC.[8] This transformation smoothes the data, thereby correcting for exponential growth (exhibited, for example, by expenditures on cellular phones) and exponential decline (demonstrated, for instance by spending on pagers).

After testing whether a log transformation is appropriate, the width of the confidence interval is determined. This calculation is based on the mean number of observations in the historical quarters. For UCCs for which there are a large number of observations, the mean is less vulnerable to a single large value; thus, it becomes more difficult to find outliers. In order to offset the reduced effect of the outlier, the width of the confidence interval is decreased. For example, UCCs with an average of 1,000 or more observations each quarter are assigned a confidence interval of 85 percent, while less common UCCs are tested at a wider confidence interval of 97 percent.

After an appropriate width has been established, the Proc Forecast procedure in SAS is used to predict the current quarter's mean. This procedure employs a user-specified method—the Holts-Winter exponentially smoothed trend-seasonal method—to decompose the data into trend, seasonal, and irregular components. Exponential smoothing weights previous data points according to how important they are in predicting future quarters' values. The Proc Forecast procedure allows the user to specify the weights given to previous quarters, from zero to unity. A weight closer to zero makes the forecast less sensitive to recent trends. A weight near zero is used with time-series data that are not volatile. A weight closer to unity makes the forecast *more* responsive to recent trends. A weight of 0.3 was chosen for the project because CE data, while volatile, still follow long-term trends. This weight is on the upper end of a reasonable bound and compensates for the lack of stability in some of the UCC predictions.[9] The Holts-Winter method was chosen for the project because of its ability to adjust for seasonal fluctuations in the series. The forecast provides mean and interval predictions for four quarters into the future. These predictions are then used to test for outliers.

Proc Forecast creates an output data set containing the actual and predicted points, along with the upper and lower bounds of the confidence interval. This data set is then used by Proc Gplot to create a graphical representation of the UCC's life cycle.[10] For documentation purposes, a graph is created for every UCC. The resulting graphs allow the analyst to visually compare the current quarter's mean with previous means and predicted means. For example, the plot of UCC 270102, *Cellular phone service* (see chart 1), shows that the mean has been relatively steady and increasing gradually over time. The mean for the first quarter of 2006 is within the confidence interval and very close to the predicted mean; therefore, this UCC would not require any further investigation.

The plot of UCC 310220, *Video cassettes, tapes, and discs* (see chart 2), shows an increasing trend over time, with strong seasonal spikes in the first quarter of each year. The previous methodology for the MTAB Review, which involved the percent change between quarters, could not identify seasonality. Once again, the actual mean is within the prediction intervals, so this UCC would not be reviewed by the analyst.

Finally, the plot of UCC 450310, *Car lease payments* (see chart 3), reveals a break in the trend, with the mean starting to decrease in the third quarter of 2003. The prediction model quickly adapts and corrects itself, showing a downward trend. From the graph, an analyst can see that the current quarter's mean not only is above the upper bound of the confidence interval, but also is equal to the maximum of the past means. This UCC would be considered an outlier and would thus be investigated.

A report is generated for a UCC when its current-quarter mean, μ_t, is greater than the upper bound of the prediction interval. The report consists of summary statistics, the forecasted model's graph, a plot of the number of observations by quarter, and the highest 20 expenditures for the UCC. For the highest 3 expenditures, the report also includes CU characteristics and income—additional information that aids the analyst in deciding whether the outlying expenditure is valid.

Implementation of the Forecasting Technique

The use of a forecasting model to detect outliers offers several advantages over the previous review procedure:

- Forty quarters of data are used to forecast trends and seasonality. In contrast, in the previous review procedure, analysts could compare only 7 quarters of data.
- Reviewer burden is reduced. The number of UCCs reviewed per quarter has decreased from more than 600 to approximately 35, which are selected by the forecasting model.
- Reviewers have visual summaries of the data they are reviewing. The new program creates graphs that plot the actual and forecasted means and the number of observations per quarter. In contrast, in the old review procedure, reviewers had access only to a tabular presentation of the data.
- The nonnormal distribution of the expenditure data does not invalidate statistical results.

Forecasting as a means of detecting outliers yields better results than the previous method or any of the other methods investigated. The updated MTAB Review process now uses 40

[8] See Appendix B.

[9] *Forecasting Methods*, on the Internet at **v8doc.sas.com/sashtml/ets/chap12/sect13.htm**.

[10] *The GPLOT Procedure*, on the Internet at **v8doc.sas.com/sashtml/gref/zlotchap.htm**.

Chart 1. **Forecast of UCC 270102 cellular phone service**

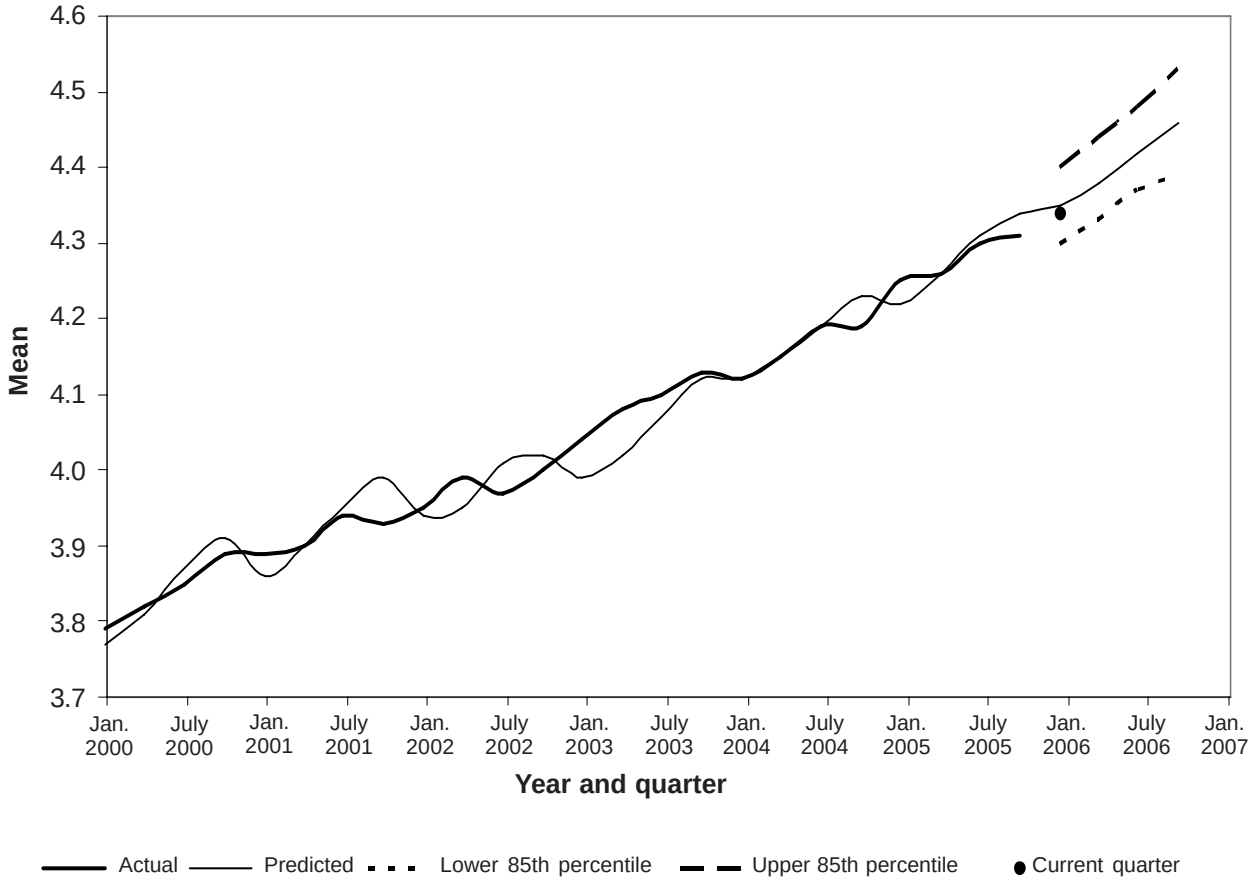


Chart 2. **Forecast of UCC 310220 video cassettes, tapes, and discs**

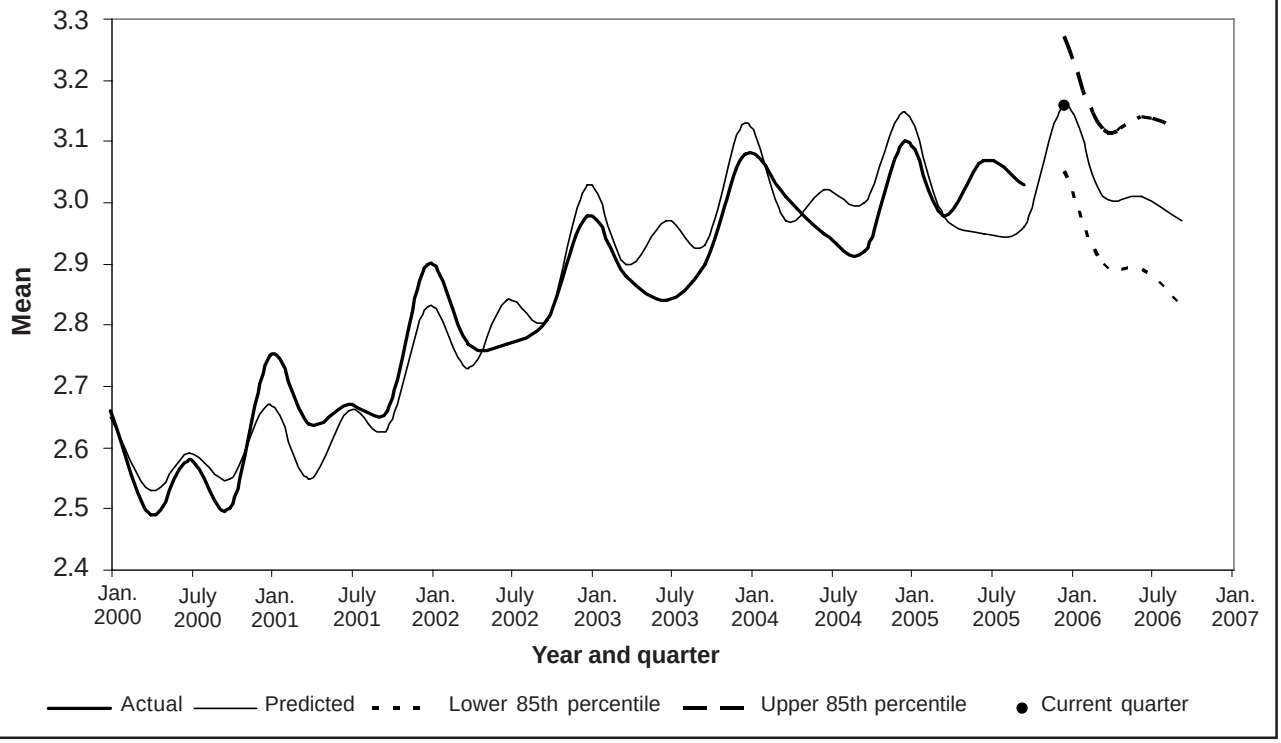
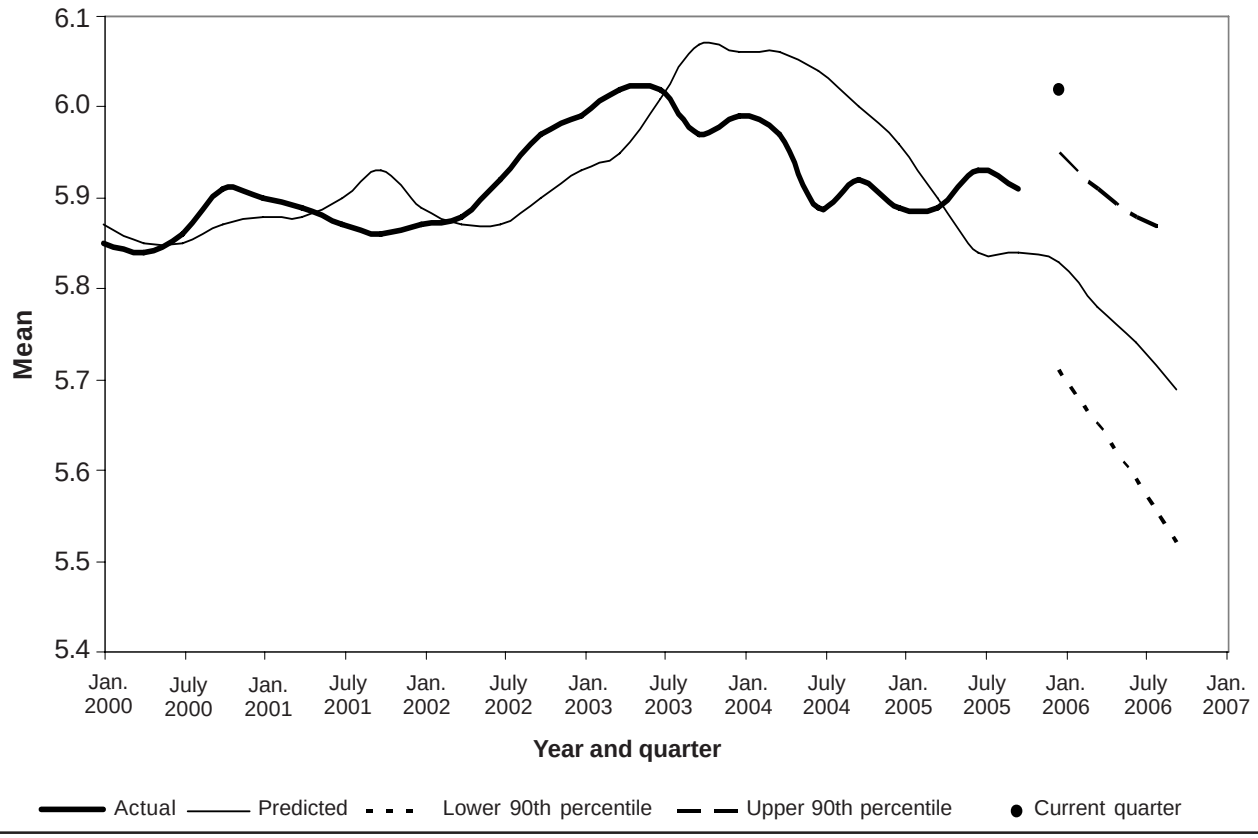


Chart 3. **Forecast of UCC 450310 car lease payments**



quarters of data and produces output in the form of tables and graphs. Mean comparisons, by contrast, can display only a limited number of historical means and percent changes in the form of a spreadsheet. Graphs displaying the last 40 quarters are easier to understand than numbers on spreadsheets. With the new method, the number of UCCs to investigate has increased from roughly 20 to approximately 35, but the effort of deciding which UCCs are selected is determined by the prediction interval. Using prediction intervals, analysts save time in the detection phase of a review and concentrate on the investigative stage. This method allows analysts more time to determine why the UCC was outside of the confidence interval.

Conclusion

A comparison of the forecasted mean with the reported mean as a technique for detecting outliers is superior to the previous method used for the MTAB Review. The new method accounts for levels, trends, and seasonality and successfully identifies outlying means, whereas traditional techniques do not. The use of a prediction interval to detect outliers reduces reviewer burden by eliminating the need to review every UCC individually, enabling analysts to focus on suspicious expenditure values.

An issue for further investigation is the examination of instances in which insufficient reports on a UCC render the forecasting technique ineffective. These UCCs include rarely collected expenditures, as well as added categories created to capture new technologies and changes in consumer spending. A statistical method for detecting outliers within these UCCs is needed and would save analysts the task of reviewing such UCCs manually. ■

Technical note A

The z-score is defined as

$$Z = \frac{(X_i - \mu)}{\sigma},$$

where X_i = the expenditure value of the individual observation, μ = the mean of the UCC, and σ = the standard deviation of the UCC.

The robust z-score is defined as

$$Z = \frac{X_i}{\theta},$$

where θ = the interquartile range.

Technical note B

The log test macro runs "an autoregressive model to a series and fits the same model to the log of the series. Both models are estimated by the maximum likelihood method, and the maximum log likelihood values for both autoregressive models are computed. These log likelihood values are then expressed in terms of the original data and compared."[11]

[11] *Overview*, on the Internet at **v8doc.sas.com/sashtml/ets/chap4/sect18.htm**.