The Problem

Consider a case where we only have 4 readings with each one taken an hour apart. By using data at each minute we are able to increase our sample size to 240. We are not increasing the number of samples, but the statistical calculation is done as if we have, and so the number of degrees of freedom for the significance test is incorrectly increased and a spurious conclusion is reached. This is one of primary causes of "spurious correlation".

Business to business tracking systems can now deliver significant amount of data but data is not information Sound statistical analysis is required to convert data to information leading to understanding and subsequently control. The additional dimensions that data can be classified into, literally scream for better analysis.

Autobox's Methodology

By taking observations at closer intervals we create series with higher and higher autocorrelation. The task for AUTOBOX is to somehow adjust for the intra-relationship and its effects on test statistics.We do not observe time series or rather collect time series data. We observe transactions or events that occur at specific points in time. For example, to study web hits or web traffic we might observe and record the following log information

User x002bn21 logs in at 00:01:05

User x022bn24 logs in at 00:01:08

User x132bn21 logs in at 00:01:09

User x002tc214 logs in at 00:02:01

…

User x04bxxx9 logs in at 00:03:05

User xwmcp02 logs in at 00:03:06

Thus if we aggregate these transactions into slices, we can analyze data sets that have intervals of 1 minute. These temporal aggregates are then referred to as time series.

MIN 1 3

MIN 2 1

…..

MIN ? ?

3,1,,,,,,,,,,?

It is quite natural to perform an analysis of this time series. Two kinds of analyses are possible

1. **Descriptive:** tables, graphs, means/standard deviation, percentages etc..

2. **Inferential:** predictive equation, data cleansing , regime change identification, cluster analysis , impact analysis, outlier identification, etc..

AUTOBOX was developed in 1975 and was originally distributed to both mainframe users and time-sharing network systems such as IDC, CSC and Compuserv. The new generation of AUTOBOX emphasizes the inferential aspects of time series and provides a computational aid to modeling. Modeling is the process of comprehensive statistical analysis often referred to as time series analysis. Rather than dumb-down the approach to modeling and analyzing time series data, AUTOBOX has taken the high road in making sophisticated approaches easy to use. Autobox marries the concepts of time series (Box-Jenkins) , multiple regression and outlier detection and creates a symphony.

Time series analysis consists of using information or data from the following three categories.

1. The history of the series being forecasted/modeled

2. User-specified auxiliary or helping variables that have had a statistically proven effect on the behaviour/pattern of the series being analyzed. This evidented pattern is then used to predict the impact of these auxiliary or helping variables.

3. Statistically identified events such as pulses , seasonal pulses , level shifts and local time trends which have had a significant impact on the past and whose effect must be taken into account.

These studies can lead directly to:

1. Development of Early Warning Systems not based simply on a percent change or a specified deviation but rather on the statistical significance of recent actual observations. Autobox is a pro-active tool that actually pushes out the identification of anomalies and leads directly to ranking schemes that order series based on their degree of exception This is in fact a form of cluster analysis and market segmentation based on actual performance data.

2. Statistically rigorous predictions that are based on the signal or equation found to be optimal for each time series. This stands in marked contrast to standard analytical applications which rely on simple equations within calculation scripts to generate forecasts.

3. Declarative statements such as “the series had a significant level change at February of 1999” or “a statistically significant change in trend occurred in March of 1998” or “last’s month sales were statistically significant at the .005 level of probability”. The ability to mine through databases and then provide this type of subjective narrative to the end user, without forcing them to view heaps of extended spreadsheet reports, takes analytical capabilities to an entirely new level.

4. If you reduce your price by z percent your sales are expected to increase y x percent. Imagine this being drawn from the data rather than simply guessed at. This is due to the fact that Autobox equations are defined and predicated upon the trends within the database.

5. Forecasting or prediction equations employ not the ordinary user-specified specification but alternatively statistically significant coefficients, yielding true forecasting not simply “let’s assume this and so” which has been the genre for many years

6. Data Cleansing. For example the sales series 1,9,1,9,1,9,5,9 has an anomaly . Can you find it ? Can your statistical tools find it ? If not why not ? Are they not as good as your eye ? How can the mean be unusual ?

7. Cluster Analysis based upon autoprojective models identifying different patterns.

8. Tests of equivalents over groups . Are sales in New jersey growing at the same rate as a Pennsylynia and New York ?

9. What has been the effect of a public statement or law change. Which series responded and which ones did not ?

10. Rejection of silly models like the number of bars is related to the number of churches or that the amount of damage at a fire is related to the number of fireman. More firemen ….more damage!

11. If we hadn't merged what would our sales have been ?

12. So you have 5 years of weekly data. Who said one model or set of parameters should be used for all 5 years? Maybe we have too much data ? Find out where the model changed and identify the cause of that change.

13. Autobox delivers "actionable items" leading to effective business contro without requiring the user to deal with the details of modeling and forecasting.

Summary

With modeling capabilities, users can run "what if?" scenarios using estimated coefficients and lagged/lead dependencies even incorporating the effects of missing variables. For example a series may be effected by weather and as is the usual case where weather data is not available statistically significant seasonal adjustments and predictions based upon previous values of Y (the series being predicted) can be automatically found and used. Dynamic aspects such as a price change having a "ripple effect" over time are all part of the "art of the possible" . No longer any need to assume a static relationship between the series or a particular guess as to the effect of one series on another.

How does it work?

AUTOBOX handles a single endogenous equation incorporating either pre-identified causal series or empirically identified dummy series which are found to be statistically significant. The set of pre-identified series can be either stochastic or deterministic (dummy) in form. In its search for the most appropriate model form and the optimal set of parameters the final model can either be:

- Purely empirical or
- A starting model could be used.

A final model may require one or more of the following structures:

- Power transforms like Log, Square Root, Reciprocal etc.
- Variance stabilization due to deterministic changes in the background error variance.
- Data segmentation or splitting as evidenced by a statistically significant change in either model form or parameters.

Enroute to its tour de force AUTOBOX will evaluate numerous possible models/parameters that have been suggested by the data itself. In practice, a realistic limit is set on the maximum number of model form iterations. The exact specifics of each tentative model is not pre-set thus the power of AUTOBOX emerges. The kind and form of the tentative models may never before been tried. Each dataset speaks for itself and suggests the iterative process.

The Final Model could be as simple as:

- A simple trend model or a simple ordinary least squares model.
- An exponential smoothing model.
- A simple weighted average where the weights are either equal or unequal.
- A Cochrane-Orcutt or ordinary least squares with a first order fixup.
- A simple ordinary least squares model in differences containing some needed lags.
- A spline-like set of local trends superimposed with an arbitrary ARIMA model and perhaps a pulse or two.

The number of possible final models that AUTOBOX could find is infinite and only discoverable via a true expert system like AUTOBOX. A final model may require one or more of the following seasonal structures:

1. Seasonal ARIMA structure where the prediction depends on some previous reading S periods ago.

2. Seasonal structure via a complete set of seasonal dummies reflecting a fixed response based upon the particular period.

3. Seasonal structure via a partial set of seasonal dummies reflecting a fixed response based upon the particular period.

The Final Model will satisfy both:

1. Necessity tests that guarantee the estimated coefficient is statistically significant.

2. Sufficiency tests that guarantee that the error process is:

- ✓ unpredictable on itself.
- ✓ not predictable from the set of causals.
- ✓ has a constant mean of zero.

The Final model will contain one or more of the following structures:

1. CAUSAL with correct lead/lag specification.

2. MEMORY with correct "autoregressive memory".

3. DUMMY with correct pulses, level shifts or spline time trends

AUTOBOX provides both automated, semi-automated and manual capabilities. AUTOBOX has a complete set of forecasting features that will appeal to both novice and expert forecasters. Autobox's automatic features are unparalleled in breadth and depth of implementation. Autobox is truly the power forecasters dream tool with a pallette of tools that allows the forecaster to build models that work.

AFS was the first company to automate the BJ model building process. Our approach is to program the model identification, estimation and diagnostic feedback loop as originally described by Box and Jenkins. This is implemented for both ARIMA (univariate) modeling and Transfer Function (multivariate or regression) modeling. What this means is that the user, from novice to expert, can feed Autobox any number of series and the programs powerful modeling heuristic can do the work for you. This option is implemented in a such that it can be turned on at any stage of the modeling process. There is complete control over the statistical sensitivities for the inclusion/exclusion of model parameters and structures. These features allow the user complete control over the modeling process. The user can let Autobox do as much or as little of the model building process as you or the complexity of the problem dictates.

Autobox comes with a complete set of indentification and modeling tools for use in the BJ framework. This means that you have the ability to transform or prewhiten the chosen series for identification purposes. Autobox handles both ARIMA (univariate) modeling and Transfer Function (multivariate) modeling allowing for the inclusion of interventions (see below for more information). Tests for interventions, need for transformations, need to add or delete model parameters are all available. Autocorrelation (both traditional and robust), partial autocorrelation and cross-correlation

functions and their respectives tests of significance are calculated as needed. Model fit statistics, including R^2, SSE, variance of errors, adjusted variance of errors all reported. Information criteria statistics for alternate model identification approaches are provided.

One of the most powerful features of Autobox is the inclusion of Automatic Intervention detection capabilities in both ARIMA and Transfer Function models. Almost all forecasting packages allow for interventions to be included in a regression model. What these packages don't tell you is how sensitive all forecasting methodologies are to the impact of interventions or missing variables. These packages don't tell you if your series may be influenced by missing variables or changes that are outside the current model. If a data series is impacted by changes in the underlying process at discrete points in time, both ARIMA models and Transfer Function models will produce poor results. For example, a competitors price change changes the level of demand for your product. Without a variable to account for this change you forecast model will perform poorly. Autobox implements ground breaking techniques which quickly and accurately identify potential interventions (level shifts, season pulses, single point outliers and changes in the variance of the series). These variables can then be included in your model at your discretion. The result is more robust models and greater forecast accuracy.

All forecasting packages allow for you to produce forecasts using the models you have constructed. Autobox presents the critical information you need to determine if those forecasts are acceptable. Autobox has options that allow you to analyze the stability and forecasting ability of your forecast model. This is achieved through a series of ex-poste forecast analyses. You can automatically withhold any number of observations, re-estimate the model form and forecast. Observations are then added back one at a time and the model is re-estimated and reforecast. Forecast accuracy statistics, including Mean Absolute Percent Error (MAPE) and Bias, are calculated at each forecast end point. Thus the stability of the model and its ability to forecast from various end points can be analyzed. Finally, you can optionally allow Autobox to actually re-identify the model form at each level of withheld data to see if the model form is unduly influenced by recent observations.

AUTOBOX provides a very comprehensive range of causal models , including but not limited to incorporating lead effects as well as contemporaneous and lag effects. It can detect and compensate for changes in variance , changes in model form and changes in parameters Multiple Regression was originally developed for cross-sectional data but Statisticians/Economists have been applying it (mostly incorrectly) to chronological or longitudinal data with little regard for the Gaussian assumptions of constant mean of the errors, constant variance, identical distribution of the errors and independence of the errors. AUTOBOX tests for and remedies any proven violations.

A brief introduction to time series analysis

Time series = a sequence of observations taken on a variable or multiple variables at successive points in time.

Objectives of time series analysis:

1. To understand the structure of the time series (how it depends on time, itself, and other time series variables)

2. To forecast/predict future values of the time series

What is wrong with using regression for modeling time series?

- Perhaps nothing. The test is whether the residuals satisfy the regression assumptions: linearity, homoscedasticity, independence, and (if necessary) normality.

  It is important to test for Pulses or one-time unusual values and to either adjust the data or to incorporate a Pulse Intervention variable to account for the identified anomaly. Unusual values can often arise Seasonally , thus one has to identify and incorporate Seasonal Intervention variables. Unusual values can often arise at successive points in time earmarking the need for either a Level Shift Intervention to deal with the proven mean shift in the residuals.

- Often, time series analyzed by regression suffer from autocorrelated residuals. In practice, positive autocorrelation seems to occur much more frequently than negative.

- Positively autocorrelated residuals make regression tests more significant than they should be and confidence intervals too narrow; negatively autocorrelated residuals do the reverse.

- In some time series regression models, autocorrelation makes biased estimates, where the bias cannot be fixed no matter how many data points or observations that you have. To use regression methods on time series data, first plot the data over time. Study the plot for evidence of trend and seasonality. Use numerical tests for autocorrelation, if not apparent from the plot.

- Trend can be dealt with by using functions of time as predictors. Sometimes we have multiple trends and the trick is to identify the beginning and end periods for each of the trends.

- Seasonality can be dealt with by using seasonal indicators (Seasonal Pulses) as predictors or by allowing specic auto-dependence or auto-projection such that the historical values (Y(t-s)) are used to predict Y(t)

- Autocorrelation can be dealt with by using lags of the response variable Y as predictors.

- Run the regression and diagnose how well the regression assumptions are met.

- the residuals should have approximately the same variance (homoscedasticity) otherwise some form of "weighted" analysis might be

needed.

- the model form/parameters should be invariant i.e. unchanging over time. If not then we perhaps have too much data and need to determine at what points in time the model form or parameters changed.

Time series data presents a number of problems/opportunities that standard statistical packages either avoid or ignore.

1. How to determine the temporal relationship for each input series ,i.e. is the relationship contemporaneous, lead or lag or some combination ? (How to identify the form of a multi-input transfer function without assuming independence of the inputs .)

2. How to determine the arima model for the noise structure reflecting omitted variables.

3. How to do this in a ROBUST MANNER where pulses, seasonal pulses , level shifts and local time trends are identified and incorporated.

4. How to test for and include specific structure to deal with non-constant variance of the error process.

5 How to test for and treat non-constancy of parameters or model form.

6. Do we model the original series or the differenced series ? AUTOBOX deals with these issues and more .

A very natural question arises in the selection and utilization of models. One asks,"Why not use simple models that provide uncomplicated solutions?" The answer is very straightforward, "Use enough complexity to deal with the problem and not an ounce more". Restated, let the data speak and validate all assumptions underlying the model. Don't assume a simple model will adequately describe the data. Use identification/validation schemes to identify the model.

A transfer function can be expressed as a lagged autoregression in all variables in the model. AUTOBOX reports this form so users can go directly to spreadsheets for the purposes that you require. Care should be taken to deal with Gaussian violations such as Outliers (pulses) , Level Shifts , Seasonal Pulses , Local time trends , changes in variance , changes in parameters , changes in models ..... just to name a few ..