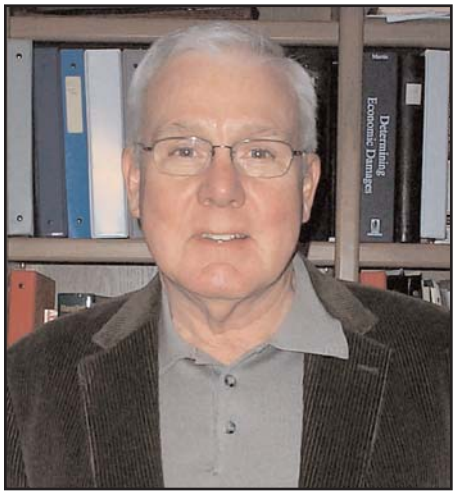
HOW TO SELECT A MOST EFFICIENT OLS MODEL FOR A TIME SERIES DATA

By John C. Pickett, David P. Reilly and Robert M. McIntyre

Ordinary Least Square (OLS) models are often used for time series data, though they are most appropriated for cross-sectional data ... provides a check list of conditions that must be satisfied for an OLS model to be most efficient ... also, gives sufficiency variables that can be used to overcome various problems in the model.

Practicing forecasters seek techniques that maximize forecasting accuracy and minimize forecast error. Their usual challenge is to make forecasts of the next period on the basis of time series data, which has a monthly, quarterly or annual data. Despite the tomes of econometricians' ponderous recommendations residing in (sometimes dusty) university libraries, it is not unusual for many practitioners to resort to an "old friend" — the ordinary least squares (OLS) model. In this article, we will show how to use our old OLS friend for optimum results. We also provide ways of identifying and estimating the most efficient OLS model, the model that minimizes the forecasting error.

OLS models, developed in the early 20th Century, were designed to analyze cross-sectional data, not time series data. A fundamental assumption required of the user of OLS models is that target data are randomly sampled from a population, and as such are independent of each other. There are occasions when practicing forecasters may use data that meet this critical assumption. For example, they may examine cross-sectional data collected from across geographical regions, age groups, income levels in a given time period. In a cross-sectional dataset, knowledge of a value for one observation would tell us nothing about the value of another observation. However, when forecast practitioners use time series data to make their predictions, they face something quite different from cross-sectional data properties. Time series data are collected at equally spaced intervals through time. In contrast to cross-sectional data, knowledge of an observation collected at time 1 may well provide information with regard to the value of another observation collected at time 1 plus 1. In other words, time series data cannot be treated as randomly selected observations from a population. Time series observations collected at proximal time periods tend to be more similar than observations collected from two distant time periods. It is as though the data have time-driven "memory"! The OLS user in this circumstance must develop an efficient OLS model—a modification of the OLS model—that circumvents the violations of the assumption of independence.



JOHN C. PICKETT

Dr. Pickett is a professor of economics at the University of Arkansas at Little Rock. His areas of interest include applied microeconomics, statistics, and forecasting. He maintains an active forecasting agenda using time series techniques. He has made numerous appearances at the Institute of Business Forecasting conferences.

ROBERT M. MCINTYRE

Dr. McIntyre is the Graduate Program Director of the Industrial-Organizational Psychology program at Old Dominion University in Norfolk, Virginia. He specializes in the application of multivariate statistics and quantitative methods. He has served as a consultant and applied researcher for various private and public organizations.

SUFFICIENCY VARIABLES

There are eight assumptions, described in the next section, which must hold for a forecast model to be sufficient. If one of the underlying sufficiency assumptions is not satisfied, then a deficiency exists within the model that may be remedied by incorporating the corresponding "sufficiency variable" in the model. Table 1 gives a list of sufficiency variables along with their description. It is conceivable that a model may fail to meet more than one of the sufficiency assumptions. In such cases, more than one sufficiency variables may be required.

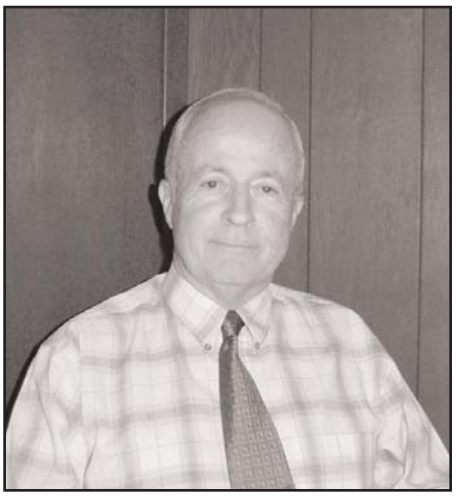
OLS MODEL SUFFICIENCY CONDITIONS

An OLS model must satisfy the following conditions (assumptions):

1. All the variables contained in a model are statistically significant. It means that each of the model parameter estimates is statistically significant and is greater than zero. Furthermore, the model does not omit any essential variables. These conditions are so basic that they cannot be addressed by means of sufficiency variables listed in Table 1.

2. This condition pertains to residuals. Residuals represent the difference between the actual and fitted value of each observation in the data series. The mean of residuals should not deviate significantly from zero in any subset of the time series. In other words, if the data comprising the time series data are broken down into different subsets, the mean of residuals of each subset should not deviate significantly from zero. If the mean of a subset of the residuals deviates significantly from zero, then sufficiency variables 4-7 may be required. Figure 1 portrays an example of this insufficiency. In the sequential graph, given in Figure 1, the mean of the residuals begins to increase at observation 30. The difference in the means of the two subsets is significantly different from zero.

3. The residuals have a constant variance throughout the series. Here again, the analyst must test all possible pairs of subsets. If the variance of the residuals is not constant, then a number of remedies can be used. A weighted regression, a power transformation, or generalized autoregressive conditional heteroschedasticity (GARCH) techniques may be applied to the data to insure the variance of the residuals is constant. Weighted regression is a technique that divides all series by the standard deviation of the error terms. A power transformation searches for the exponent of the series between +1 to -1 that result in a constant variance of the residuals. GARCH is a time series technique that allows users to model the serial dependence of the errors. Figure 2 shows a plot of the residuals, where the variance increases, beginning with observation 60. In this case, in addition to a difference in the variance across subsets of data, the residuals are related to one another, i.e., they are "autocorrelated." (One cue to identifying autocorrelation in a sequential plot of residuals is the relative similarity in values in points close in time.) Oftentimes, a violation manifests different symptoms.



DAVID P. REILLY

Mr. Reilly, one of the founders of Automatic Forecasting Systems, has a distinguished career as a developer of creative time-series applications. He is a former GE executive and a Senior Manager at the American Stock Exchange. He has been a consultant to many companies including Pepsi Cola, General Mills and Anheuser-Busch.

4. The residuals are free from autocorrelation for all lags. With regard to tests for autocorrelation, readers should be aware that the Durbin-Watson statistic tests only for the existence of first-order autocorrelation, that is, lag of 1. The analyst must, therefore, examine autocorrelation of residuals for all possible lags. If the residuals are autocorrelated, then the model builder should consider adding sufficiency variables 1, 2, or 7. Figure 3 shows a plot of residuals. Here the pattern clearly shows the presence of autocorrelation among residuals, because it forms a wave like pattern.

5. This condition states that the residuals are normally and independently distributed. Failure of this assumption is closely related to the failure of condition 4. If the residuals are not normally distributed, then either a power transform of the dependent variable or a type 1, 2, or 7 sufficiency variables may be needed to "cure" the deficiency in the model. Figure 4 shows a plot of the residuals that fail to meet this condition. Here the histogram of the first 16 observations would have a different shape than the histogram of the last 24 observations.

6. This condition states that the model's residuals are not a function of the lagged values of each of the independent variables. Failure to meet

TABLE 1
LIST OF SUFFICIENCY VARIABLES

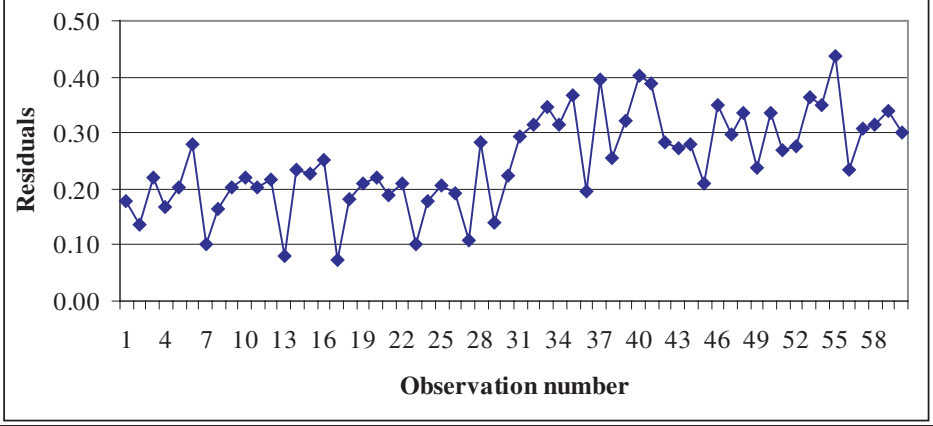
Sufficiency Variable Type	Sufficiency Variable Description
1	Lagged values of Y
2	Lagged values of each X & lead values for each X, where applicable
3	Intervention variable(s) representing a pulse(s)
4	Intervention variable(s) representing a seasonal pulse(s)
5	Intervention variable(s) representing a level shift(s)
6	Intervention variable(s) representing local time trend(s)
7	Moving average term(s) representing lagged values of the error terms

Notes:
i. Intervention variable are dummy variables included to account for qualitative phenomenon.
ii. A pulse implies a one time only effect or may re-occur at fixed intervals of time as a seasonal pulse.
iii. A level shift implies the phenomenon occurs in contiguous periods of time in a common way, so that the series shifts up or down during the period.

this condition implies that the model builder did not use prior lags of one or more of the independent variables. In other words, a necessary sufficiency variable 2 was omitted. Figure 5 shows a characteristic pattern of data that fail to meet this condition. A simple regression of the residuals on the lagged values of X would show the estimated parameter is statistically significant.

7. This condition states that the X values in a series are not a function of the lagged residuals. If the Xs are a function of the lagged residuals, then the one-way causal model is the wrong functional form. Instead of estimating a single equation model, the relationship should be modeled using simultaneous equations. Failure of this assumption is frequently observed when modeling large macroeconomic systems. In these models, the dependent and independent variables are all interdependent, which requires multivariate time series techniques (vector time series). Figure 6 shows a characteristic pattern where this condition is not met. The parameter estimated from a simple regression between X and the lagged residuals would be statistically significant. However, there are many cases where a series is affected by "future values" such as when customers forestall purchasing products because they are aware of price changes that will occur. This is a "lead effect" and can be easily modeled.

FIGURE 1
RESIDUALS WITH A NON-CONSTANT MEAN



8. The distribution of the residuals is invariant over time, that is, one subset of the series data should have the same covariance structure as another subset. With regard to Figure 7, the autocorrelation function (the series of autocorrelations for all lags) for first 10 observations would be different from that for the last nine observations. One way of depicting this would be in terms of the "autoregressive process." (An autoregressive process is the one in which the value of some present measure is a function of previous measurements in time.) The autoregressive parameter estimate for the first 10 observations would take on a positive value while it would take on a negative value for the last nine. Hence, the parameter would not be consistent over time. If the covariance structure is not the same, then the data exhibit time-variant model parameters. In such a case, one should focus on the most recent homogenous set. It is fair to say that modeling time series data comprising more than one underlying "generating process" is technically beyond the state of the art at this time. If this condition is not met, then the analyst should identify when the change occurs and then estimate the model with the set of most recent observations. Figure 7 shows the data pattern of a series with such a problem. Clearly the slope of the line fitted for

FIGURE 2
RESIDUALS WITH A NON-CONSTANT VARIABLES

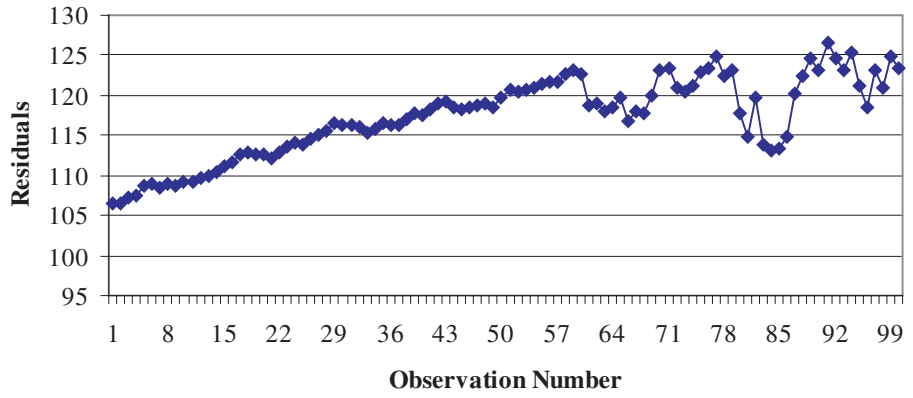


FIGURE 3
RESIDUALS EXHIBITING AUTOCORRELATION

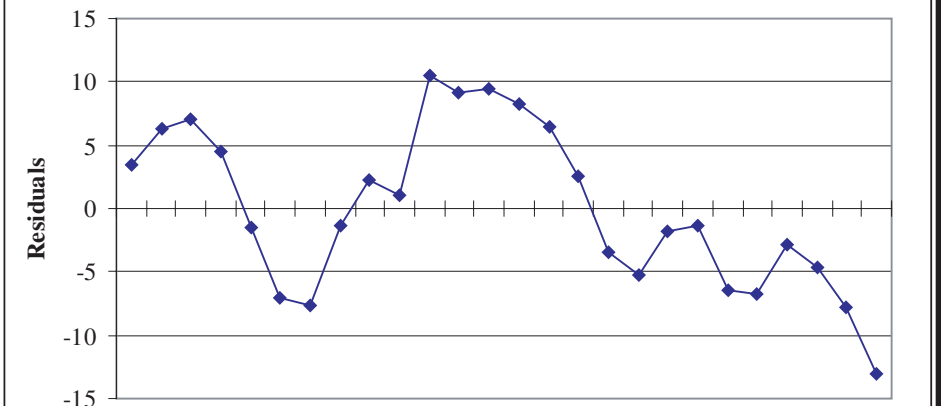
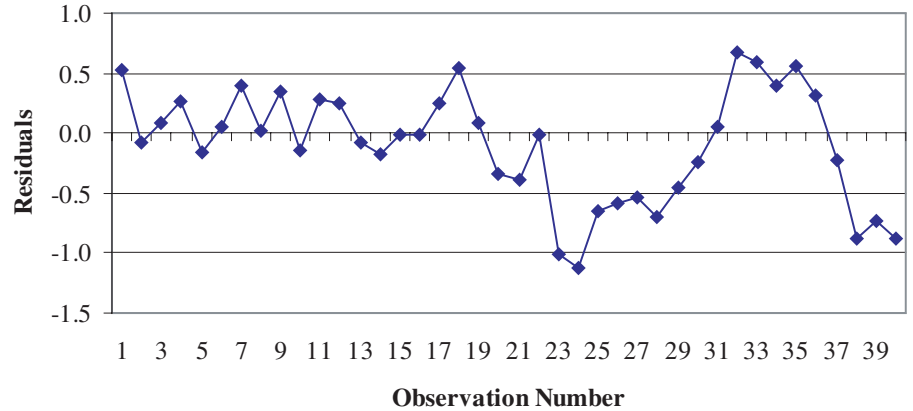


FIGURE 4
RESIDUALS NOT NID



the first 10 observations would be positive while the slope of the line fitted for the last nine observations would be negative.

SUMMARY

Practitioners devote significant resources to forecasting business time series. They are wasting resources and providing error-prone forecasts if they blindly use traditional OLS methods. The fact is that time series data cannot be treated as random samples from a static population. Rather, these data characteristically contain "memory structure." If users of OLS methods do not pay attention to this fact, their models will contain violations of the underlying assumption of independence. This is not some esoteric, mathematical nuance. It will generate inaccurate and inefficient forecasts. We have described the "sufficiency conditions" required of an OLS forecasting model. The set of conditions serves as a checklist for identifying the variables that must be included in the efficient model. Practically speaking, if followed carefully, the practitioner will avoid potentially important errors in his or her forecasts. ■

FIGURE 5
RESIDUALS DEPENDENT ON INDEPENDENT VARIABLE

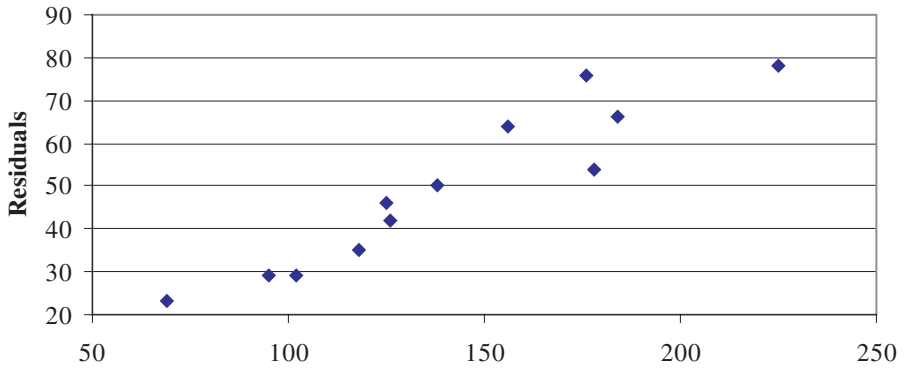


FIGURE 6
INDEPENDENT VARIABLE DEPENDENT ON THE RESIDUALS

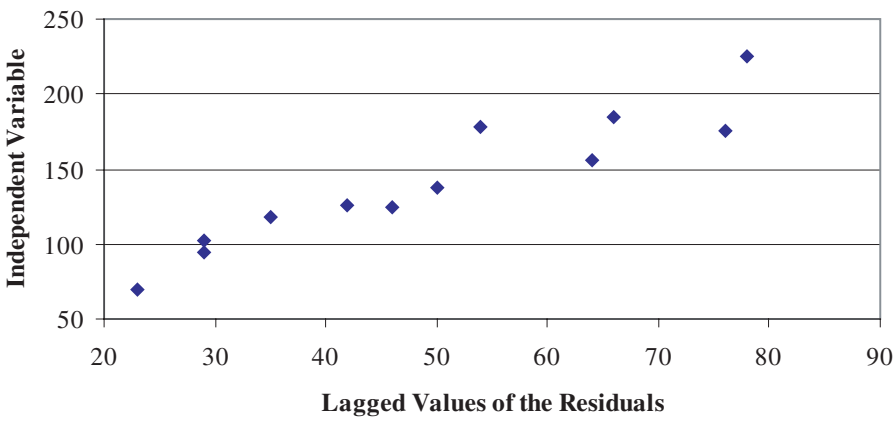


FIGURE 7
TIME VARYING PARAMETERS