Appendix E

Statistical Forecasting Techniques

Appendix E. Statistical Forecasting Techniques

Frequently, transportation planners need forecasts of freight data as a basis for a whole range of short-term investment decisions. Indeed, starting a new highway construction project versus a project for the construction of a new intermodal freight terminal may depend upon whether the planner projects truck or rail traffic to growth at a faster pace during the next five years. The problem confronting planners is to take available time-series data on the freight traffic in question and develop projections of future volumes or flows. Indeed, the solution to a whole class of practical transportation planning problems involves assessment of future freight traffic demand based on time-series data. Since time-series freight data exist for a number of different types of freight movements, a number of specific transportation planning issues can be answered by using that data.

For example, there are time-series data on the volume of traffic (by commodity) moving on the inland waterway system. These data can be disaggregated to show traffic volumes on particular segments of the system. Planners are frequently confronted with the problem of projecting future traffic volumes on each segment in order to determine whether existing facilities need to be expanded or whether new facilities are, in fact, required in order to meet demand. Planners also have time-series data for truck traffic by highway segment. Again, the issue confronting the planner is to project that traffic into the future in order to decide whether or not existing facilities need expansion or whether new facilities are needed. Airport planners are faced with critical decisions regarding the mix of air freight versus air passenger facilities on their property. In order to assist in making that decision, they need to use time-series data on air freight shipments in order to project future needs. Planners may confront a decision regarding a need for an expansion of an urban intermodal freight terminal. They could use time-series data from the rail waybill database to project future intermodal shipments in their metropolitan area. These represent selected examples of the type of problems facing transportation planners at the state, local, and even national level whose solution can benefit from projections generated from the use of available time-series freight data.

■ E.1 Regression Analysis

Regression analysis is widely used by analysts for empirical estimation and forecasting. Regression analysis involves identifying one or more independent variables (the explanatory variables) which are believed to influence or determine the value of the dependent variable (the variable to be explained) and calculating a set of parameters which characterize the relationship between the independent and dependent variables.

Assume a variable, y, is linearly dependent upon three independent variables, x^1, x^2 and x^3 plus some unknown, unmeasurable influence, ε:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Given a sufficient number of observations[1] of y, x_1, x_2, and x_3, the regression will use ordinary least squares (OLS) to estimate values of the true parameters β_0, β_1, β_2, and β_3 and use these estimates, b_0, b_1, b_2, and b_3, to calculate an estimated value of each observation of y. This estimate is denoted as $\hat{y}$:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$$

The regression assumes the unknown term, ε, has a mean value of zero. The true value of ε may be non-zero for any given observation. The difference between the observed y and its estimate, $\hat{y}$, is the "error", or "residual" of the estimate. The regression chooses the values of the parameter estimates to minimize the sum of the squared errors[2] of the estimate to produce the "best" fit.

It must be emphasized that, although regression analysis provides the best fit between the independent and dependent variables, this does not mean that the estimated $\hat{y}$ will be a good estimate of y. Regression simply

---

[1]Observations of dependent and independent variables may be time-series or cross sectional in nature. Time-series data contain a single observation for each variable for each of several time periods or points in time. Examples might be annual volume of freight shipments, annual output per employee, and annual real gross domestic product (GDP). If twenty years' worth of data are collected, there would be twenty observations of shipments, output per employee and GDP. Cross sectional data contain observations across the population at a point in time or during a single time interval. Examples include volume of shipments by state, industry, or firm, in a given year.

[2]The sum of the squared errors = $\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$ for n observations.

guarantees that there is no better estimate of y based on the given independent variables. If different independent variables are used, the estimate of y may change significantly. Regression analysis provides little guidance as to which variable or variables should be used to estimate y.[3]

Ordinary Least Squares Regression

The simplest and most commonly used form of regression analysis is the "Ordinary Least Squares" (OLS) approach. OLS is a single equation estimation technique in which each observation is given equal "weight" or importance in estimating the parameters described above. Advanced forms of regression analysis include weighted least squares, two-stage least squares, and step wise regression.[4] Most packages provide capabilities for one or more of these advanced regression techniques.

OLS is based on several key assumptions. If one or more of these assumptions is violated, it may be necessary to use an advanced estimation procedure to obtain a satisfactory model. OLS assumes:

1. A "one-way causality" exists between the independent and dependent variables;

2. The regression includes all relevant independent variables and excludes irrelevant ones (i.e., the "right regressors" are chosen);

3. The dependent variable can be calculated as a linear function of a specific set of independent variables plus an error term;

4. The expected value of the error term is zero;

5. The error terms have the same variance and are independent of each other (uncorrelated errors);

---

[3]There are statistical tests available to assist in determining the "significance" of an independent variable in explaining variation in the dependent variable (the *t*-test is most commonly used). However, these tests do not ascertain whether there is a meaningful relationship between the variables or a coincidental correlation. If it is the latter, the model is unlikely to be a reliable forecasting tool, even if it shows a good statistical "fit."

[4]For further discussion of these and other advanced regression techniques the reader may consult any standard text on econometrics. Particularly good intuitive discussions are contained in Peter Kennedy, *A Guide to Econometrics*, Third Edition, MIT Press, Cambridge, Mass., 1992. For more technical specifications see J. Johnston, *Econometric Methods*, Third Edition, McGraw-Hill, New York, 1984.

6. Observations of independent variables do not depend on the sample chosen;

7. No independent variables are linear combinations of other independent variables (no perfect multicollinearity); and

8. The number of observations exceeds the number of independent variables.

Some of these assumptions are essential for using OLS, either because of the implied underlying theoretical relationships (e.g., Assumption 1) or because of the pure mathematical properties involved in minimizing the sum of squared errors (e.g., Assumption 7). Others represent nice, desirable, properties but may be set aside without invalidating the regression results. Methods for identifying violations of these assumptions in an OLS model and the consequences of these violations are discussed in the second subsection below.

Model Specification and Testing

The key to a good OLS model lies in the theoretical relationship between independent and dependent variables, as described in assumptions one and two above. Johnston[5] provides a very useful description of the process of building and evaluating a regression model. He emphasizes the following steps:

1. Talk with experts to become knowledgeable about the problem being modeled;

2. Become familiar with the relevant institutions and the constraints they impose on the problem;

3. Look at the data to gain a better understanding of the problem or process being modeled and the limitations of the data;

4. Base the model on sound economic theory;

5. Avoid "data mining" in which models are selected on the basis of high R^2 or high *t*-values while ignoring other, more fundamental, relationships; and

6. Use the judgment of an "experienced critic" to shape the model.

[5]*Op. cit.*, pp. 498-510.

Building a good regression model requires good data, a thorough understanding of the expected relationships among variables, and the time to test the various outcomes against a range of criteria. Among the most commonly used tests are:

1. The *t-test* which assesses the likelihood that the estimated parameter, b_n is significantly different from zero. If the parameter equals zero, the corresponding independent variable, x_n, provides no information in the given specification; i.e., it does not explain any of the variation of the dependent variable. Most software packages provide the *t*-value for each independent variable. The *t*-value can then be compared with the critical value of the Student's *t* Distribution table, found in statistics books.[6] If the absolute value of the computed *t* exceeds the table value for the appropriate number of degrees of freedom[7] at the desired confidence level, the parameter estimate is considered to be significantly different from zero. To be significant at the five percent level,[8] for example, the *t*-value generally must exceed a value close to two. It is often possible to test the significance of the *t*-value without referring to the table of values; absolute values of *t* above three will always be considered significant while those below one will always be considered insignificant. The *t*-test is not valid, however, when autocorrelated errors are present (see discussion below).

2. R^2, the *coefficient of determination*, measures the "goodness of fit," i.e., the amount of the variation in the dependent variable explained by the independent variables.[9] Many researchers search for the highest possible value of R^2, regardless of the sensibility of the model they have developed. This is the wrong approach to model building. One glaring problem with this approach is that the mathematical nature of the R^2 is such that R^2 cannot fall when an additional variable is added to a regression and may rise, regardless of the quality of the variable. The analyst who wants a high R^2 needs only to add more variables, whether they are sensible or not. A further problem with R^2 is that

---

[6]*E.g.*, see Stephen Kokoska and Christopher Nevison, *Statistical Tables and Formulae*, Springer-Verlag, Inc., New York, 1989.

[7]The degrees of freedom for a particular equation equal the number of observations (sample size) minus the number of parameters estimated.

[8]Significance testing must be based on a level of confidence. The five percent level represents 95 percent confidence that the result is true.

[9]R^2, also known as the "coefficient of determination," is calculated as the ratio of the regression sum of squares to the total sum of squares and its value ranges between zero and one. The total sum of squares is the total variation of the dependent variable around its mean, $\sum(y - \bar{y})^2$; the regression sum of squares is the variation of the dependent variable "explained" by the regression, $\sum(y - \hat{y})^2$.

time-series regressions typically produce high values of R^2 because the time trend is a strong determinant of both independent and dependent variables.

3. The adjusted[10] measure of R^2 corrects for the number of observations and the number of independent variables and may fall when a new, meaningless variable is added to the regression. Although a better measure than R^2, the adjusted R^2 still cannot distinguish between a model that fortuitously fits the data well and one which has identified the true underlying relationships. Analysts are advised to ignore the R^2, adjusted or not, when building their models.

4. The *standard error of the estimate* (SEE), which provides a better indication of how well the independent variables explain the variation in the dependent variable. Adding more variables may reduce or increase the SEE.

For a more detailed discussion of the range of tests and how to interpret their results, consult any econometrics textbook.[11]

Violating the OLS Assumptions

The assumptions identified at the beginning of this section are often violated in practice, and the tests identified above are often inadequate to identify the problem. Exhibit E.1 summarizes the effect(s) of each violation on the ability of OLS to calculate parameter estimates and the usefulness of the model's results.

Although Exhibit E.1 shows that only two violations interfere with OLS' ability to perform the calculations, there are several situations in which the parameter estimates or the estimate of the dependent variable (or both) are less than ideal. If the parameter estimates are questionable, the analyst would have little confidence in describing the influence that a change in a particular independent variable's value would have on the value of the dependent variable. However, the estimate of the dependent variable

---

[10]Adjusted R^2, also known as "$\overline{R}^2$," is calculated as:

$$R^2 - \frac{K-1}{T-K}(1-R^2)$$

where K is the number of independent variables and T is the number of observations.

[11]One of the less technical textbooks is Harry H. Kelejian and Wallace E. Oates, *Introduction to Econometrics: Principles and Applications*, Third Edition, Harper & Row, New York, 1989.

Exhibit E.1 Consequences of Violating the Basic OLS Assumptions

Assumption	Effect on Ability to Calculate Parameters	Effect on Quality of Parameter Estimates	Effect on Quality of the Estimate of the Dependent Variable
1. One-way causality	None	None	Not suitable for forecasting because no meaningful relationship has been identified
2. The "right" regressors are used	None	May cause bias	May produce larger errors; less reliable
3. The dependent variable is a linear function of the independent variables	None	Biased	Poor estimate unless the particular sample used for estimation is nearly linear; unsuitable for forecasting
4. The expected value of the error is zero	None	Estimate of the intercept is biased; other parameter estimates may be biased or unbiased	May be adequate for estimation and forecasting
5a. Error terms have some variances (homoscedasticity)	None	Parameter estimates unbiased but no longer have minimum variance	May be adequate but can be distorted by the undue influence of some of the observations
5b. Errors are uncorrelated with each other	None	Parameter estimates unbiased but *t*-test invalid to determine parameter significance	May be adequate for estimation and forecasting but using an autocorrelation correction technique is advised
6. Observations of the independent variable are fixed even with repeated sampling	None	Biased, especially if autoregressive model	May still provide good estimation and forecasts
7a. No variables are linear combinations of other independent variables	OLS cannot perform calculation; matrix cannot be inverted		
7b. No collinearity	None	Parameter estimates unbiased but large variances make them unreliable	Estimation and forecasting may be reliable
8. Number of observations	OLS cannot perform calculation		

may still be adequate, allowing for the interactions among the independent variables and the error term. The model may be a good forecasting tool. If the model does a poor job of estimating the dependent variable but is adequate in its parameter estimation, the model may be useful for simulation studies. It is important, therefore, to know the intended purpose of the model before deciding whether a violation of the basic assumptions renders the model unfit for that purpose.

Perhaps the most difficult assumptions to satisfy are the first two, namely that the model is using the "right" independent variables to explain the variation in the dependent variable. One can never be sure that there are no other pertinent influences on the dependent variable. Furthermore, even if statistical testing points to the need for an additional variable, there is no standard procedure for identifying the missing variable. It is easier to reject potential existing variables than to find and incorporate new ones. The model builder is advised to choose independent variables that are consistent with economic or other appropriate theory as a first step.

Regression analysis assumes a "one-way causality" among the variables: the independent variables must affect the value of the dependent variable but the dependent variable cannot affect the values of the independent variables. In some situations this is clearly the case, but in other situations the relationships may be tangled. Consider the following examples:

1. The number of umbrellas carried in a city on a given day depends on the region's population and the expected probability of rain. Population does not depend upon the number of people carrying umbrellas nor does the probability of rain. One-way causality is well established in this case.

2. The demand for a new car depends in part on the vehicle's price, but the vehicle's price is determined in part by the aggregate demand for cars. In this case, there is a simultaneity between demand and price. OLS would not provide a good basis for forecasting demand if vehicle price is used as an independent variable. More advanced techniques[12] might be used in this case.

OLS is a linear estimator and assumes that the variables are linearly related. OLS will still calculate parameter estimates if the true relationship is nonlinear but the estimates will not be useful for forecasting. It is critical that the model builder using OLS specify a linear relationship.

---

[12]See Kennedy (*op. cit.*, pp. 157-163) for a good discussion of indirect least squares, instrumental variables, two-stage least squares, and limited information, maximum likelihood techniques which might be used to overcome a simultaneous equation problem.

The linearity of the relationship can often be examined through simple scatter plots of the data and a nonlinear relationship can sometimes be "linearized" by transforming the variables in a specific manner. The logarithmic transformation is probably the most commonly used. It is appropriate when the growth rates of the variables are related in a linear manner. When Assumption 4 is violated and the expected value of the error term is not zero, the estimation of the intercept, b_0, will be biased. Omitting a key, relevant independent variable from the regression is often responsible for violating Assumption 4. The error term will reflect the variation in this missing variable and the mean of the error term will likely not be zero. This is, however, more properly viewed as a violation of Assumption 2.

Assumption 4 can also be violated if the dependent variable is restricted to a limited range of values, thereby limiting the potential size of the error. The nature of the study may make this truncation unavoidable. For example, a study of low-volume roads would exclude observations of AADT above a cutoff value. This truncation ensures that the error terms would not be large enough to cause the dependent variable to be less than the AADT cutoff, leading to a truncation of the upper end of the error distribution. The expected value of the truncated error distribution is negative, not zero.

The two major problems resulting from violating Assumption 5 are heteroscedasticity (errors with different variances) and autocorrelated errors. Heteroscedasticity often occurs when higher values of the independent variable are associated with larger variances of the error. This may be an entirely logical outcome. For example, if personal VMT depends on income, at higher levels of income there is more opportunity for spontaneous discretionary travel. This spontaneity would produce a larger variability in observed VMT, and, consequently, a greater variability in the error of the estimate. Several large values of the independent variable could shift the regression line, weakening its predictive value. A weighted least squares approach ("generalized least squares") is often used instead of OLS when heteroscedasticity is present to reduce the influence of the observations that are expected to have large errors.

Autocorrelated errors exist when the errors are not independent of each other. The Durbin-Watson test is commonly used to detect the presence of autocorrelated errors although the test is not reliable when lagged values of the dependent variable are used as independent variables. The presence of autocorrelated errors reduces the reliability of the OLS estimate.

Autocorrelated errors are often found in time-series data because the effect of a disturbance usually persists beyond the period in which it occurs. For example, the Mississippi River flooding affected travel when it occurred and in the months following. If a model overestimated barge traffic during the flood, (i.e., its error was positive), it likely would have underestimated barge traffic during the several months following. The errors would all be related to the flood and would be correlated with each other.

More generally, there are almost always some exogenous influences that have been omitted from a model which tend to increase (or decrease) the dependent variable for several consecutive time periods, thus producing a series of negative (or positive) errors for these time periods.

Autocorrelated errors may also result from model misspecification, especially the omission of a relevant variable, a violation of Assumption 2. If the model appears to be specified correctly, techniques such as the Cochrane-Orcutt method can be used to reduce or eliminate the autocorrelation.

Assumption 6 specifies that observations of the independent variable are fixed even when the sampling is repeated. This assures that the independent variables are uncorrelated with the error terms. When this assumption is violated, the OLS estimate will be biased. Assumption 6 will be violated when the independent variables are improperly measured or when a model is autoregressive. In the latter case, the current value of the dependent variable is influenced by its own past values which were, in part, determined by the error term in those periods. Despite their bias, autoregressive models can still be useful for estimation.

Although not technically a violation of Assumption 7, strong collinearity among the independent variables may still weaken a model. If there is an approximate linear relationship among the independent variables (strong but not perfect multicollinearity), OLS will run but the variances of the parameter estimates will be large, reducing the confidence one should place in the resulting estimates of the dependent variable. Various tests exist to determine whether multicollinearity is present.[13] Even if multicollinearity is found, the analyst may choose to do nothing if the model appears to be satisfactory.[14] Other approaches to multicollinearity include obtaining more observations since a larger sample size helps to reduce variance by providing additional information to the regression. The analyst may also want to consider dropping one of the collinear variables although this may result in a specification error and biased estimates of the remaining parameters if, in fact, the true coefficient of the dropped variable is not zero.

[13]These include an analysis of the correlation matrix and the use of condition indices.

[14]Kennedy (*op. cit.*, p. 181) discusses two "rules of thumb" which suggest doing nothing: all t statistics greater than 2; or, the R^2 from the regression exceeds the R^2 of any independent variable regressed on the other independent variables.

Forecasting with an OLS Model

Although much of the emphasis in this discussion has been on building a satisfactory model to explain the variation in a variable of interest, the purpose of many models is to provide decision-makers with useful forecasts. A well-built econometric model may or may not produce "good" forecasts. This section discusses some of the pitfalls of forecasting.

In order to use an OLS model for forecasting, it is necessary to provide future values (forecasts) of each independent variable. Developing good forecasts of the independent variables may require additional model building, extrapolating past trends, or acquiring forecasts from outside firms or agencies. When the data are trendless, the "naive" forecast that the next period's value will equal the current period's value, may prove satisfactory. To the extent that estimated or forecasted future values of the independent variables contain errors, the forecast of the dependent variable will be weakened.

A second problem in forecasting involves the stability of the parameter estimates. If the parameter estimates are extremely sensitive to the data sample used in the regression, the model's structure may change over time. Forecasting models implicitly assume that the parameter estimates identified by OLS will be invariant over time. This is rarely true. Statistical tests, such as the Chow test, are helpful in analyzing the structural stability of a model. Validation techniques such as estimating the regression over a portion of the sample and allowing it to "forecast" the remaining values of the dependent variable are also helpful in assessing the usefulness of the model for forecasting.

A third problem in forecasting involves the unforeseen disturbances which can cause any forecast to miss its mark. Examples are found in natural disasters (earthquakes, fires, floods), supply shocks (e.g., petroleum), international disturbances, and significant policy changes. The estimated parameters have no knowledge of these events and the manner in which they alter the relationship between independent and dependent variables.

A fourth problem in forecasting involves the range of the independent variables' future values. If the values of the independent variables move outside the range from which the model established its parameter estimates, there is an increasing likelihood that the forecast will have a large error.

Forecasts which fail to predict the level of the dependent variable can still be useful if they forecast the direction of change. Regression models are more likely to forecast "turning points" than the simple ARIMA models which extrapolate past trends.

Regression models can also be used to assess the sensitivity of the dependent variable to possible changes in one or more independent variables. These simulations, sometimes called "what if?" analyses, are not true forecasts but provide a range of outcomes to consider under different input assumptions. By assigning probabilities to the potential values of the independent variables, an expected future value can be derived.

■ E.2 Exponential Smoothing

Time-series data frequently involve some short-term fluctuations, or up and down movements in the data, that seem to deviate from an established pattern. The first type of time-series forecasting technique involves the "smoothing" out of these short-term fluctuations by identifying the underlying pattern in the data and extrapolating the underlying "smooth" pattern into the future. Exponential smoothing techniques remove random fluctuations and establish the underlying pattern in the time-series. Indeed, all "exponential smoothing" techniques use some form of weighted average of past observations to smooth out data fluctuations. The differences in methods involve how much weight should be given to the most recent observation versus the more distant data in generating the smoothing effect.

Specifically, the "exponential smoothing" methods answer the following questions: (1) what weight should be given to the most recent value in the series (in most time-series data, each value is positively correlated with its preceding value – i.e., positive autocorrelation)?; (2) do the data lack any pattern such that the best value to use in developing projections is the overall average of the entire series with no special consideration given to the more recent data?; (3) what is the general trend in the data?; and (4) do the data reflect any seasonal pattern?

The "exponential smoothing" procedure in SPSS Trends estimates four parameters to control for the relative importance of recent observations in developing predictions. One parameter is used in all applications of the procedure, while the researcher selects among the other three parameters depending upon whether the data shows evidence of trends or seasonality. The four parameters are:

1. The alpha parameter – controls the weight given to the most recent observation in determining the overall level and is used in all time-series estimations. (When alpha is one, the single most recent observations is used exclusively in the smoothing process; when alpha is zero, old observations count just as heavily as more recent ones in the process.) This parameter is referred to as the smoothing constant.

2. The gamma or trend parameter – used only when the series shows a trend. (When the gamma is high, forecasts are based on trends estimated from most recent points in the series; when the gamma is low, forecast uses trend based on entire series with all points counting equally.)

3. The delta parameter – used when the data show a seasonal pattern. (When delta is high, the seasonality adjustment is based on the more recent time periods; when delta is low, the seasonality adjustment is based on the entire series with all time periods counting equally.)

4. The phi parameter – used in place of gamma when the series shows a trend and that trend is damped, or dying out. (High values of phi provides rapid response in projections when any indication that the trend is dying out is given, while low values of phi estimate damping of the trend from the entire data series.)

In the "exponential smoothing" procedure within SPSS Trends, the researcher can initially generate a simple data smoothing operation through the application of a smoothing constant – i.e., the alpha parameter. SPSS Trends will evaluate the range of alpha values and recommend a value for the model with the lowest "sum of squared errors."

In most applications, however, the researcher is confronted with a more complicated problem that would benefit from a specification of one or more additional parameters. The underlying data might have either a growth or trend component or, alternatively, a seasonal component. In SPSS Trends, the exponential smoothing procedure provides the flexibility to handle each of these situations. The routine can estimate both an alpha and a gamma parameter to achieve minimum error and generate a corresponding smooth curve and projections based on the estimated parameters. This procedure is based on Holt's exponential smoothing routine. If, however, the researcher suspects the data involve both a trend and a seasonal component, the routine can estimate three parameters – alpha (the smoothing constant); gamma (the trend parameter); and delta (the seasonal parameter). Again, the model will evaluate a range of values for each of the parameters and recommend values for each based on the achievement of a minimum sum of squared errors.

The "exponential smoothing" procedure establishes the underlying pattern of the databased on the combination of parameters specified by the researcher and uses that pattern to make projections of the time-series data into the future. The exponential smoothing procedure in SPSS Trends includes a number of features to facilitate its use by planners. It adds two new series to the existing time-series data for each application. The first additional series contains the predicted values resulting from the exponential smoothing and the second contains the error terms. These data can be plotted against the actual time-series data to show how the smoothed data compare to the original.

In addition, the package enables the researcher to develop a plot of the residuals for examination in order to establish whether a pattern exist in the residuals. Indeed, the residuals should be randomly distributed. If they display a pattern, then the model is, indeed, inadequate.

Finally, the procedure can provide either one-step or n-step ahead forecasts based on projection of the "smoothed" underlying pattern into the future. The researcher can specify the number of time periods beyond the data for which a projection is requested. Of course, the "exponential smoothing" routine is most appropriate for the short-range forecasting situation.

■ E.3 Leading Indicator Regression

The curve fitting procedure does not make any assumptions about why the time-series curve has the particular modeled shape. Indeed, the curve fitting procedure may indicate that the time-series data best conform to a linear model and that, indeed, the linear model provides very close predictions of the time-series in the validation period. However, there are many instances in which researchers believe that the time-series data, i.e., the modeled variable, is closely related to another time-series variable. In fact, the related data series may lead or provide a good prediction of the time-series variable, i.e., the modeled variable. Thus, if researchers know the value of the lead or indicator variables at the current moment, they will be able to develop predictors for the "modeled variable" at some specified point in the future as indicated by the lead time. In fact, the indicator variables will be of most value if they lead or predict values of the "modeled variable" in the future.

Selecting a Lead Variable

The following example of relevance to a transportation planner will illustrate the point. A need might arise to predict the level of household goods shipments on a national or regional basis. The future levels of such shipments might establish the need for additional drivers or, perhaps, new facilities. While curve fitting procedures might provide future estimates of household goods shipments, there may be reason to believe that other independent variables will provide "leading indications" of household goods shipments in the future. Indeed, a recent investigation showed that sales of existing homes and retail sales of new automobiles lead by four months the number of individual household goods shipments. Thus, the model can predict household goods shipments four months into the future based on sales of existing homes and retail sales of automobiles in the current month.

SPSS Trends provides the researcher the means to evaluate the appropriateness of independent explanatory variables, determine an appropriate lead time for each variable and provide the actual estimation of the variable's effect on the dependent variable – i.e., determine the statistical coefficients specifying the relationship between auto sales, new home sales and household goods movements.

Determine Lead Time

The leading indicator regression depends critically on determining an appropriate leading indicator variable and establishing the appropriate lead time. The leading indicator regression procedure within SPSS Trends provides all the necessary tools to make an appropriate analysis. The establishment of an appropriate lead time between an indicator variable and the time-series variable of interest, the dependent variable, requires an examination of a cross-correlation function – i.e., the correlation between two time-series at the same time and also with each series leading by one or more lags. By analyzing a cross-correlation function between two series, researchers can see the lag at which they are most highly correlated.

However, the use of the cross-correlation procedure requires that the two time-series variables, the dependent modeled variable and the indicator variable, are stationary – i.e., each variable's mean and variance stay at about the same over length of the series. For variables with a gradually increasing value over the time-series, an effective way to make the series stationary is to difference it. Taking differences means replacing the original time-series by the differences between adjacent values in the series. The leading indicator regression procedure provides for differencing of time-series data (for one or more differences) and the calculation of a cross-correlation function between the differenced variables.

Once the cross-correlation function is examined to select an appropriate lead time indicator, the SPSS routine can automatically alter the database so that each value of the selected lead time indicator variable is matched with the appropriate value of the dependent or modeled variable during both the historical and validation periods. Thus, if the indicator variable leads the dependent variable by three months, then a new variable is created in which the first value of the indicator variable is matched with the fourth value of the dependent variable.

The procedure then enables the researcher to calculate a regression between the dependent time-series variable, the "modeled variable," and the lead-indicator independent variable. The regression establishes a coefficient of impact of the lead variable on the value of the dependent variable. The entire regression equation is used to produce predicted values of the dependent variable during the historical period, the validation period, and a future period as well.

Adjusting for Autocorrelation

One of the assumptions made in regression analysis is that the residuals or errors from regression are uncorrelated among themselves. When important explanatory variables are omitted from a regression analysis, autocorrelated residuals commonly occur. When residuals are strongly autocorrelated, the significance levels reported for the regression coefficients are wrong and the R-squared value does not accurately summarize the explanatory power of the independent variables. Time-series regression frequently violates the assumption of uncorrelated errors, since it is difficult to include all the important explanatory variables in the regression.

One way to explain the problem is to note that the time-series regression involves use of a dependent and independent variables that most probably have trends, either up or down. The two time-series variables with trends will correlate simply because of the trends regardless of whether the two variables are casually related or not. What the researcher wants to know is whether the two variables are related apart from a similarity due to autocorrelation. Thus, it becomes necessary to remove the autocorrelation prior to model estimation.

The leading indicator regression package within SPSS Trends provides information researchers can use to determine the presence of autocorrelated errors in the time-series regression and procedures to correct for these errors. Autocorrelation among errors is most frequently determined by reference to a residual analysis statistic, labeled the Durbin-Watson Statistic, produced as part of the regression output. Values of this statistic range from zero to four, with values less than two indicating positively correlated residuals and values greater than two indicating negatively correlated residuals. Statistical tables indicate whether a given Durbin-Watson statistic is statistically significant given the sample size. Statisticians recommend that researchers review not only a Durbin-Watson statistic and determine its statistical significant, but also examine statistical plots of residuals from a regression against the predicted values and also against each of the predictor variables.

The SPSS procedure provides the researcher with three approaches to removing autocorrelation: two algorithms (Prais-Winsten and Cochrane-Orcutt) transform the regression equation to remove the autocorrelation. The third method uses a maximum likelihood method for removing autocorrelation. Regardless of the removal procedure, the program provides a new estimated model with autocorrelation removed. The new model includes estimates of the impact of each of the predictor variables on the "modeled" or dependent variable as well as predicted values of the dependent variable in the historical and validation period. Finally, the coefficients from the equation can be used to make projections into the future for the "modeled variable."

■ E.4 ARIMA Modeling

In developing regression forecasts based on indicator variables, the planner must have a very clear idea regarding the variables that might be causally linked with the "modeled" variable of interest. However, in many practical situations, the planner lacks such information or, in some instances, does not have adequate time-series data for the indicator variables. While such circumstances might dictate the use of an exponential smoothing procedure or a curve estimation regression, there is a technically sophisticated time-series modeling approach that builds forecasts from more inclusive and simultaneous analysis of complex past patterns in the time-series than is achievable with application of either the exponential smoothing or curve estimation regression approach. This class of models is called the Box-Jenkins ARIMA Models.

ARIMA models process a great deal of information from time-series data, but require the researcher to specify only a minimum number of parameters. ARIMA models are highly flexible and compare a wide variety of alternative models in developing the "best" or "correct" model for the time-series data. Indeed, the Box-Jenkins ARIMA models have come to be quite highly regarded and results from them carry a greater degree of acceptability than do models based on either exponential smoothing or curve estimation procedures.

The SPSS Trends routine provides the researcher with the tools to specify and evaluate the ARIMA model. Based on the results provided, the researcher can choose the model with the "best fit" and use it to develop projections of the modeled time-series data into the future.

ARIMA Parameters

ARIMA stands for AutoRegressive Integrated Moving Average based on the model's three components. The general model (not considering seasonality) is written as ARIMA (p, d, q), where p is the order of autoregression, d is the degree of differencing, and q is the order of moving average involved. Researchers specify levels for each of these parameters according to the guidelines established in the ARIMA module of SPSS Trends. The following paragraphs discuss in turn each of the parameters and their specification process.

The p parameter is the order of autoregression. In any autoregressive process, each value is a linear function of the preceding value or values. In a first-order autoregressive time-series model, only the single preceding value is used in model building; in a second-order process the two preceding values are used in building a model; and so on. The coefficient for the autoregressive parameter usually is greater than a -1 and less than a +1, indicating that the influence of earlier observations dies out exponentially.

In a first-order autoregressive process, the current value is a function of the preceding value, which is in turn a function of its preceding value. Thus, "each shock or disturbance to the system has a diminishing effect on all subsequent time periods."

The d parameter is the differencing parameter, providing adjustments needed to make the time-series data stationary. Time-series data are stationary when two consecutive values in the series depend only on the time interval between them and not on time itself. A time-series with a constant mean value over time is consistent with this notion. However, "real-world time-series are most often nonstationary; that is, the mean value of the time-series changes over time, usually because there is some trend in the series so that the mean value is either rising or falling over time." The nonstationary properties need to be removed from time-series data prior to an attempt at specifying the model.

Indeed, time-series data often reflect the cumulative effect of some process. "The process is responsible for changes in the observed level of the series, but is not responsible for the level itself. Inventory levels, for example, are not determined by receipts and sales in a single period. Those activities cause changes in inventory levels. The levels themselves are the cumulative sum of the changes in each period. A series that measures the cumulative effect of something is called integrated series. You study an integrated series by looking at the changes or differences from one observation to the next. The differences of even a wandering series often remain fairly constant."

As noted in the discussion of the cross-correlation between a dependent variable and a leading indicator variable, differencing is a common approach to bringing about stationarity to a data series. The researcher can specify a differencing parameter of either 1, for first-differences, or 2, for second-differences.

The third ARIMA parameter is q, the order of the moving average. In a moving average process, each value is determined by the average of the current disturbance (i.e., error term) and one or more previous disturbances. The order of the moving average process specifies how many previous disturbances are averaged into the new value.

It is important to differentiate between the autoregressive parameter and the moving average one. "Each value in a moving-average series is a weighted average of the most recent random disturbances (i.e., error terms), while each value in an autoregression is a weighted average of the recent values of the series. Since these values in turn are weighted averages of the previous ones, the effect of a given disturbance in an autoregressive process dwindles as time passes. In a moving-average process, a disturbance affects the system for a finite number of periods (the order of the moving average) and then abruptly ceases to affect it."

Steps in Using ARIMA

ARIMA modeling involves three distinct phases: identification of the underlying processes of the time-series data through specification of the three parameters; model estimation based on the specified parameters; and model diagnosis. The researcher can use the diagnosis to re-specify the parameters and re-estimate the model until the model is satisfactory. The ARIMA process is iterative and highly flexible.

The identification of the values of the three parameters involves a systematic procedure. Since the identification process for both the autoregression and the moving average parameters requires stationarity, a researcher must transform the data series, if necessary, in order to obtain a stationary series. The most frequent method of obtaining a stationary series for time-series data is differencing. The selection of a first or second-order differencing results in the determination of the d parameter in the ARIMA identification process. This parameter is most frequently either a zero or a one. It should be noted that while differencing is the most common method of data transformation, the ARIMA routine in SPSS Trends provides for logarithmic and square-root transformations – useful in the situation in which there is more short-term variation where the actual values are large than where they are small.

Once the differencing parameter is identified, the researcher must select the autoregressive and moving average parameters. The ARIMA package provides the researcher with autocorrelation functions between the time-series variable of interest and its lagged value at 1, 2, 3,...lags. In addition, the researcher is provided with the partial autocorrelation function, controlling for autocorrelations at intervening lags. Based on theses functions and their plots, researcher are guided in their selection of both the AR and the MA parameters.

The Trends ARIMA procedure then estimates the model and its coefficients based on the parameters specified. The researcher supplies the three parameters p, d, and q from the analysis of autocorrelation and partial autocorrelation functions, while ARIMA performs the iterative calculations needed to determine the maximum-likelihood coefficients associated with each of the parameters. The ARIMA software also adds new series to the data file representing the fitted or predicted values, the error (residual), and the confidence limits for the fit.

The diagnosis of the ARIMA results requires an investigation of whether the model's residuals are correlated and/or whether the residuals show a pattern. If either the residuals are correlated or they show some time of pattern, then the researcher needs to return to the identification process and re-evaluate the parameters entered into the model.

The model provides the researcher with the ability to calculate the autocorrelation and partial autocorrelation function among the error terms or

residuals. If the first or second-order correlations are large, the researcher has probably misspecified the model.

The residuals should be without pattern. That is, they should be white noise. The ARIMA package provides a test for whether the residuals have a pattern. The test is called the Box-Ljung Q Statistic, also called the modified Box-Pierce statistic.

The focus on determining the appropriateness of a Box-Jenkins ARIMA model is on the error terms – to insure no autocorrelation and no residual pattern. It is not on whether each of the model's coefficients is statistically significant.

Once the researcher is satisfied with the model's coefficients, its results can be used to predict future values of the time-series variable of interest. It is expected that projections resulting from the ARIMA method will benefit from its enhanced features and its simultaneous treatment of the order of autoregression, the degree of differencing, and the order of the moving average.

■ E.5 Intervention Analysis

In the transportation field, events will frequently occur that result in major changes in an established time-series pattern. For example, major deregulation legislation, passed in the late 1970s and 1980 significantly altered the competitive relationship among transportation modes. In addition, major changes in fuel prices during the 1970s and 1980s resulted in major disruptions and shifts in modal patterns. Transportation planners are frequently called upon to estimate the impacts of major events on, for example, levels of truck traffic or, alternatively, levels of rail traffic. The Box-Jenkins ARIMA models can be adapted to include a specific assessment of the impact of an intervention (e.g., passage of a major piece of transportation legislation or major fuel price increase) on a time-series data. The following pages explain the process through a technique called intervention analysis. Again, the SPSS Trends program provides an option to incorporate intervention analysis in the ARIMA model.

Researchers initially estimate an ARIMA model for the data without regard to the intervention event or its corresponding impact. Thus, the researcher follows the procedure detailed in the preceding section to specify the three required parameters of the ARIMA model – i.e., the autoregressive parameter (p); the difference parameter (d); and the moving average parameter (q). As noted, initial specification of the parameters is based on determination of whether the time-series data is stationarity; the transformation of data through differencing; and an

examination of the autocorrelation and partial autocorrelation plots of the time-series data at various lags.

Assessment of the impact of the intervention on the time-series requires that an intervention variable be added to the analysis. The coefficient of this intervention variable will represent its impact on the change in the time-series variable of interest at a particular time controlling for the impact of the other three parameters in the model.

The intervention variable is what econometricians label as a "dummy" variable, taking on a value of "1" from the time of the intervention on to the present time and a value of "0" prior to the intervention. Thus, if a transportation planner had a time-series data of motor carrier market share and wanted to assess the impact of the Motor Carrier Act of 1980 on that traffic, the planner would create a new variable to include in the ARIMA model. This intervention dummy variable would equal zero for all time-series data points prior to 1980 and a value of one from 1980 to the present.

The SPSS Trends ARIMA program gives the researcher the ability, once the p, d, and q parameters have been specified, to specify one or more predictor variables (also called regressors) for the time-series data being modeled. The ARIMA program treats these predictors much like predictor variables in regression analysis. It estimates coefficients for them that best fit the data. The coefficients, indeed, are interpreted just like regression coefficients. Positive signs indicate that the intervention event adds positively to the change in the modeled variable, while negative signs indicate the opposite.

The specified ARIMA model with regressors must be diagnosed in a fashion similar to the ARIMA model without regressors. The autocorrelation of the residuals must be evaluated as well as their pattern. If autocorrelation is found or a distinct pattern emerges, then the researcher must return to the model identification phase and reevaluate the situation.

■ E.6 Seasonal Decomposition and Weighted Least Squares Regression

Frequently, planners work with transportation data having distinct seasonal trends. For example, small package shipments peak during the holiday season; auto traffic peaks during the summer months; truck traffic slows during the winter months; household goods shipments peak in the spring and summer and fall off rapidly in the winter.

The SPSS Trends package includes a Seasonal Decomposition routine to "decompose" or break down a time-series variable into the following

components: a long-term trend component, a seasonal adjustment factor, a cyclical component, and a random or irregular component. Indeed, the Seasonal Decomposition routine takes the original time-series data and adds the following information: (a) a seasonal adjustment factor for each season; (b) a seasonally adjusted data series (i.e., the original data with the seasonal component removed); (c) a deseasoned trend and cycle component; and (d) an error component.

In the Seasonal Decomposition routine, the time-series dependent variable is treated as a linear function of the following independent components: trend, seasonal, cyclical, and irregular or random. This multiplicative model is appropriate when seasonal variation is greater at higher levels of the series. If seasonality does not increase with the level of the series, an alternative additive model is available. Each of the model's components are estimated separately by the methods discussed below. The components are re-assembled and used to generate forecasts of the time-series variable from either the multiplicative or additive models.

Estimation of Seasonal, Trend, Cyclical, and Error Components

The Seasonal Decomposition routine initially removes the seasonality effect, i.e., it deseasonalizes the data, and calculates a seasonal adjustment factor for each season (e.g., each quarter). By removing the seasonal variations in the data, the long-term trend and cyclical components can be more easily identified.

The Seasonal Decomposition routine removes the seasonal variance by calculating moving averages whose number of terms equals the periodicity of the time-series (four quarters in our example). This removes the seasonality by averaging the high and low points of each quarter for every period in the time-series. A ratio is established between each quarter's value of the time-series data and the average value for the four quarters in the period (that quarter and the subsequent three quarters). If this ratio is greater than one, the quarter has a positive seasonal impact on the value of the series. The specific seasonality index for each quarter is based on the average of this ratio for each quarter throughout the entire time-series. This seasonal index is the first component of the four needed to develop a time-series decomposition forecast.

The second component needed for the decomposition forecast is the trend component. The trend component is developed from a regression between the seasonally-adjusted time-series and a time variable that increments one unit for each quarter or time period in the database. A positive coefficient for the trend variable would indicate growth in the series over time, while a negative coefficient would suggest decline over time. The trend coefficient in the regression is used to estimate a moving average trend for each quarter in the time-series. This moving average

trend value is the second component of four required to generate a forecast by the decomposition method.

The cyclical factor, the third component, needed for a decomposition forecast is the ratio of the seasonally-adjusted moving average and the moving-average trend. If this ratio is greater than one, there is an indication that the deseasonalized value for that period is above the long-term trend in the data. If the cyclic factor is less than one, the reverse is true.

By combining the trend, seasonal, cyclic, and error terms together, the Seasonal Decomposition routine can be used to predict values of the time-series data for both the historical and evaluation periods as well as for forecasting in the future. As shown, here, however, the Seasonal Decomposition routine requires separate estimates be developed for each component of the equation. After developing each component's estimates, they can be re-assembled to develop estimates of the time-series data for forecasting purposes.

Seasonal Adjustments with Dummy Variables

If, however, the seasonal factors are treated as dummy variables in a larger regression model, the seasonal effects and the trend can be evaluated simultaneously. The simultaneous evaluation of the trend and seasonal factors simultaneously make the use of the Seasonal Decomposition routine somewhat less cumbersome. Positive coefficients for a dummy seasonal variable would be indicative of a positive seasonal impact, while a negative coefficient would suggest the opposite.

Use of Weighted-Least Squares to Adjust for Heteroscedasticity

When the seasonal effects are estimated simultaneously with other independent factors, such as the trend component, researchers must be aware of and make adjustments for heteroscedasticity – violations of the assumption that regression residuals have constant variance. It is often the case that there are differences in variance of a time-series variable depending upon the specific time period. For example, while truck traffic has a seasonal component (with declines in the winter months), the variance of truck traffic in the winter will be depend greatly on the severity of the winter. Since there are fluctuations in winter's severity, a researcher should expect greater variation in truck traffic in the winter months.

Thus, when using seasonal dummy variables in a regression analysis, the transportation planner needs to evaluate a scatterplot of residuals against the values predicted from the regression. If this scatterplot indicates greater dispersion in residuals depending on the predicted value of the time-series variable, then heteroscedasticity adjustments should be made.

The SPSS Trends routine provides for the use of weighted least squares as an adjustment for heteroscedasticity. One approach would be to weight each time-series observation by the standard deviation of its residual. However, the package evaluates a number of different weighting approaches and selects the best "weighting" factor and, then, uses that factor in re-estimating the regression equation with heteroscedasticity removed.

Advanced Methods for Seasonal Adjustments

While the discussion in the previous section focused on the use of the Seasonal Decomposition routine for handling time-series data with seasonal patterns, the SPSS Trends package includes other methods for handling seasonal adjustments as well. In fact, the procedures for making seasonal adjustments in the Seasonal Decomposition routine are based on procedures developed by the U.S. Bureau of the Census in the 1950s for seasonally adjusting census data. New methods have been developed that constitute refinements over the originally approaches. The SPSS Trends package includes, for example, a seasonal adjustment method, labeled the X-11 ARIMA approach, adopted by researchers at Statistics Canada. These researchers noted that when new data were added to a time-series, the seasonal adjustment factors estimated with the Seasonal Decomposition method often were different. Forecasts resulting from the method changed every time new data became available. While some changes in the seasonal adjustment factors are inevitable with the addition of new data, researchers felt that the level of change was too great in factors with the Seasonal Decomposition method.

The X-11 ARIMA method attempts to reduce the size of changes in seasonal forecasting when new data is added to the series. The approach adds forecasts and backcasts (obtained through ARIMA modeling) to the ends of the original time-series data and then calculates seasonal adjustment factors on the extended series with ARIMA modeling.

As discussed, the ARIMA procedure requires the researcher to specify three parameters in modeling a time-series with no seasonal pattern. The process of specifying the p, d, and q parameters was presented above. When a seasonal pattern exists in the data, the ARIMA model requires the researcher to specify three additional parameters for the p, d, and q parameters to reflect the seasonal factor. The SPSS Trends package fully supports the specification of an ARIMA model with a seasonal component. Thus, the ARIMA model can be used to develop backcasts and forecasts for the original time-series data under the assumption of seasonality. These values are subsequently added to the original time-series and the X-11 ARIMA procedure is used to develop a new model with seasonal adjustment factors and better forecasts.

This section will not go into detail regarding the modifications in the ARIMA procedure needed to incorporate the seasonality factor. Suffice it to say that the SPSS package fully supports this process and provides the researcher the flexibility to evaluate each specified model and to re-estimate the model based on intermediate results. Like the Seasonal Decomposition procedure, X-11 ARIMA produces four new series and adds them to the original time-series file. These new series are the seasonally adjusted series, the seasonal factors, the trend-cycle component, and the error component.

■ E.7 Other Software Packages

The previous discussion has shown the SPSS Trends software to be very extensive and supportive of the entire range of time-series techniques available. Its use requires user knowledge and interface with the program. Frequently, the researcher needs to examine output, determine, for example, whether error terms are correlated or whether they show some distinct pattern. Based on this examination, the researcher must modify parameters and re-estimate models. This required interaction and feedback has many desirable characteristics. It gives the researcher maximum control over the process and allows for modifications based on the unique characteristics of the time-series. To many, this type of control and input is a necessary condition for an effective tool.

However, there are in the market place some time-series packages that provide an "expert" system component for selecting the "best" model from among the range of alternatives – i.e., exponential smoothing, curve estimation, ARIMA, etc. These programs make decisions about the parameters that have to be specified – e.g., the p, d, and q parameters in the ARIMA model and make decisions about what adjustments need to be made in those parameters based on an analysis of the initial results. For the regression with leading indicators, these programs will examine up to fifty leading indicator variables in order to determine which, if any, are appropriate indicators of the time-series data. Furthermore, the techniques determine the appropriate time lag for any selected variable. In short, these "expert system" software packages automate many of the decisions that the SPSS Trends routine requires researchers to make on their own.

The advantage of the "expert system" software packages is that the planner has the benefit of "expert" statistical advice on the most appropriate method for establishing a time-series data and using that estimation for forecasting purposes. Certainly, the planner without any detailed background and training would be in a position to produce better forecasts than would be possible in the absence of the software. On the other hand, planners with some knowledge would prefer the control over the process

that is afforded by the SPSS Trends routine. Indeed, these planners would object to the "cookbook" aspect of the "expert system" software.

Certainly, planners would benefit from a combination of approaches. That is, they should analyze the time-series data to the best of their ability with the SPSS Trends package and then compare their projections with projections generated from an expert system package. Indeed, the expert system software packages have features that allow the planner to override the "expert's" choice and to substitute their own evaluation in place of that of the computer.