LOCAL DEVELOPMENT II
URBAN PROJECT

Submitted to

USAID / CAIRO

Submitted by

WILBUR SMITH ASSOCIATES

in association with

PUBLIC ADMINISTRATION SERVICE

DELOITTE HASKINS AND SELLS

DEVELOPMENT CONSULTING OFFICE

ENGINEERING AND GEOLOGICAL CONSULTING OFFICE

Local Development II Urban Project

Programming & Budgetting

Technical Paper

FIVE YEAR FORECASTING MODEL USER'S GUIDE

December, 1988

LDII-URBAN
FIVE YEAR FORECASTING MODEL
A USERS' GUIDE

I. Introduction

The introductory forecasting system you have just received will help you to generate forecasts based upon historic revenue data. The system uses two models, a linear regression and a curvilinear regression model to perform time-trend analysis of the available final account revenue data. The two basic tools of this model are the /Data Regression command in Lotus and a standard statistical table, the Student's t distribution. This guide explains the way in which these models were created in Lotus 123 and demonstrates step-by-step the way in which governorate personnel can update, replicate or modify the original forecasting models.

II. Forecasting Methodology -- An Overview

Let's begin by outlining the basic steps of a forecasting methodology. Table II.1 summarizes these steps.

A. Identify the factors which you feel should influence the variable which you wish to forecast.

Throughout this discussion we will refer to these factors as revenue "Drivers". In the time-trend model with which we are currently working, we are de facto defining the Driver to be Time or Fiscal Years. For a more complete discussion of the potential revenue Drivers for each line item please see Volume I - Revenue Sources.

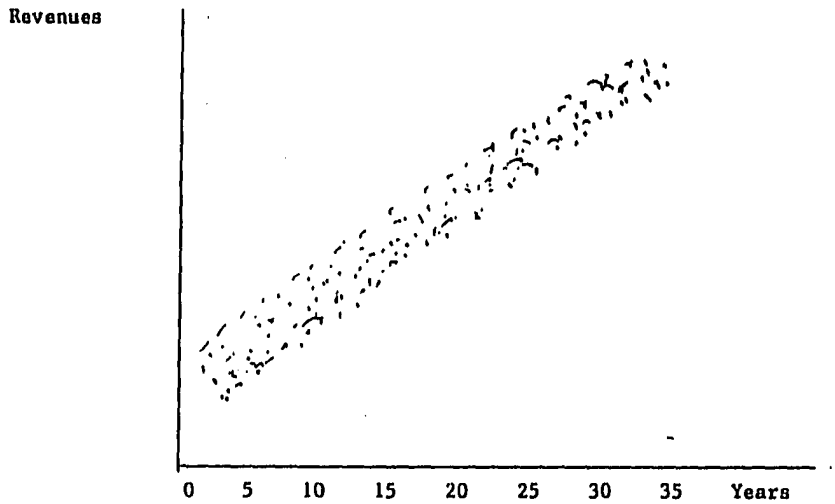
B. Establish the conceptual relationship between the Dependent Variable, Revenue, and the Driver, Time (Years)

The most basic method to describe the relationship betweenRevenue and Years is a linear equation. A straight line, or linear relationship between the two variables Revenue and Years implies that a change in the Driver, results in a proportional change in the left-hand, or dependent variable, Revenue. Such an assumption may be not truly represent the configuration of your data however. Think about your revenue collections in recent years. If you were to plot actual revenue collections over time, how would they look? Would the diagram most resemble a straight line? Figures II.1. and II.2. demonstrate this point. The scatter diagram in Figure II.1. illustrates an example of a dependent variable which moves in proportion with the movement of the Driver, or independent variable. Notice that the plot of the data in Figure II.1. closely approximates a straight line. Figure II.2., on the other hand is a scatter diagram demonstrating a curvilinear relationship between the two variables. This second pattern suggests that the analyst should experiment with a regression model that uses non-linear forms of the data.

Choosing the wrong type of regression can result in inaccurate forecasts. Experimenting with other models can, if nothing else, demonstrate whether a simple straight line is indeed the best way to describe the relationship between Revenue and its Driver(s). Some examples of the non-linear forms of data which are frequently used in forecasting are:

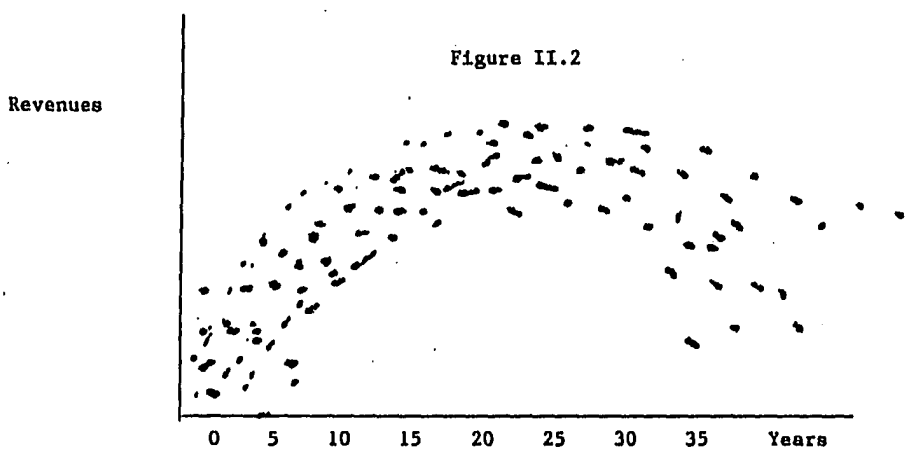
Logarithms
Square Roots
Power Polynomials
(quadratic, cubic and so on)

Figure II.1



Scatter Diagram Demonstrations
Linear Pattern of Revenue Growth
Over Time

Figure II.2



Scatter Diagram Demonstrations
Curvilinear Pattern of Revenue Growth
Overtime

Trigonometric Functions

In the model which you have received, we experiment with the linear model and with one form of Power Polynomials called a quadratic model. In the linear model, revenue will be described as a function of a single Driver, Years. In the quadratic model, the dependent variable Revenue, is related to two Drivers, "Years" and "Years" raised to the second power, "Years^2". You will use Years (and Years^2) as the Driver from which you will predict future revenue collections.

The spreadsheet also includes standard statistical indicators and tests which we will use to help you to determine whether the more fully described quadratic model results in a significant improvement to your forecast.

C. Quantify the relationship between the Dependent Variable and the Driver(s)

By following the instructions in this User's Guide you will learn to perform regression analysis in Lotus 123. The output of the regression will be a quantifiable estimate of the relationship between the dependent variable, Revenue and the Driver(s), Years or Years + Years^2.(2)

Using the regression output, you will be able to construct an equation which follows the general format below:

$$Y = Con + B_1 (X_1) + B_2 (X_2) + \ldots B_n (X_n)$$

Where:

Y = the Dependent Variable (in our case Revenue)

Con = some constant number (and may be zero in some cases) representing the Y intercept. The Y intercept is the value of the Dependent Variable when the Driver is equal to Zero.

B = the X Coefficient for the independent variable or Driver (Time or Time^2 in our case)

X(n) = the nth X Variable, or Driver, which you use to describe your equation

We will examine these elements in greater detail in section IIID.

D. Forecast future revenues using the equation you develop in Step C and your knowledge or estimates of the Driver data in future years.

Having quantified the estimated relationship between Revenues and the Drivers you must next supply data which represent your knowledge or best estimates of what the Drivers will be in the future. Solving the equation which you produced in step C using these future Driver data will yield your forecasts for Revenue. In our example, the Driver "Years" is simple to calculate; we need only increase the count of years by one for each year into the future that we wish to forecast.(3) This step is explained in detail in the tutorial in Section IIIA.

E. Determine the validity of both models and choose the model which is most appropriate for each line item.

Once you have created your two forecasts you will want to determine which result appears most reasonable. By judging the statistical performance of each model you can decide the amount

of confidence which you can place in the products of those models--your forecasts. Just as each revenue line item represents a different source of funds and follows a unique growth patterns, each model will have varying degrees of success estimating the future revenues across line items. In this section, you will learn graphing and statistical techniques what will help you to determine on a line item per line item basis, whether one, both, or neither of the models presented generate reliable forecasts.

F. Update the system to perform forecasts in future years.

To use this system in coming years all that you will need to supply is additional years of historical data revenue and a few minutes of time to format the worksheet to accomodate more data.

III. Tutorial: Creating Forecasts in Your Lotus 123 Model

A. The Worksheet Setup

This section provides step-by-step instructions for using the Lotus 123 model to generate your own forecasts. Let's begin by entering Lotus and looking at the current model. The Lotus model and its forecasts are based upon the data which were made available to the OMED Group.

> NOTE: In general, a minimum of ten to fifteen years of historical data are necessary to increase the effectiveness of the forecasts. Data requirements are discussed further in Section III.D.1

When you have gone through a file and are ready to begin the tutorial enter /Worksheet Erase in order to reconstruct a similar sample worksheet.

> NOTE: Throughout this discussion, text which are enclosed in brackets <> should NOT be interpreted as commands but as words which are to be typed as text; do not type the brackets.

Let's use Cairo Bab I, Line Item 4, Automobile Tax as an example. In an empty worsheet, label the following columns accordingly:

Worksheet Address:	Label:
A1	<Governorate: Cairo>
A2	<Bab I Revenue Forecast>
A4	<Fiscal>
A5	<Year>
B4	<Actual>
B5	<Collection>
C4	<Year>
D4	<Year^2>
E4	<Forecast>
E5	<Linear>
F4	<Forecast>
F5	<Quadratic>
G5	<AC = Con + Year>

4

L5	<AC = Con + Year + Year^2>
Q5	<Student's T Test Analysis>
P6	<Regression>
P7	<Linear>
P8	<Quadratic>
Q6	<R-squared>
R6	<Increase>
S6	<Student's t>
T6	<DF>
U6	<Signif.>

Notice that you have left at least four columns reserved to receive the output of each regression analysis.

You may want to insert columns of one or two spaces in width for ease of format. You may do so with one important exception.

IMPORTANT:

****Do not insert a column between columns, X and X^2, labeled Year and Year^2 or Lotus will not be able to perform the regression****

Don't forget to center the labels using /Range Label Center.
Select /Range Format , 0 to format ranges A1..F16. Select/Range Format, 4 and highlight J9..O9. Select/Range Format, 4 and highlight Q7..S8. Starting in A7 enter <82/83>, the label for the first fiscal year for which you have data. Continue to enter the fiscal years up through 92/93 in column A. Next, enter the historical data for final accounts starting at cell B7.

9,000,000
9,350,000
12,828,000
15,000,000
18,348,476
19,054,615
19,224,376
20,130,899

Beginning with column C, row 7, enter numbers from 1-11. These years represent the six years of historical data plus the five years in the future which we are trying to forecast. In column D, enter the value of each entry in C raised to the second power. For example, if the value in C9 is 3 then D9 will be (3)12 or 9.

B. Performing the Regression

Now let's instruct Lotus to compute the linear regression. Select /Data Regression X-range and highlight range C7..C12, the range which contains the Years for which we have final account data. To specify the Y-range, highlight B7...B17. Finally, select Output-Range, enter cell G6 and press Go. Your data analysis is presented in G7..J12. We will use this information later when we assess the reliability of the models and their forecasts.

Now let's calculate the quadratic regression. Choose /Data Regression X-Range. The program will default to the X-range which you highlighted for the linear model. Expand the X-range to C7..D12. Next, change the Output-Range from G6 to L6 and press Go. The new regression results are written to the L7 range. Notice that since you increased the X range to include both Year and Year^2, the regression now generates two X coefficients.

C. Calculate the Forecasted Revenues

We will now use the results of the linear regression to calculate the forecasted revenues. If you have structured the worksheet according to Section III A, you may simply copy the following formula into cell E7, and the copy cell E7 to cover the range E7..E17:

+J$7+(C7*I$13)

This formula adds the regression constant (J7) to the product of each X value (in row C) and the X coefficient (cell I13). If you have inserted blank columns between the rows of data, you must adjust this formula accordingly.

HELPFUL HINT: Because of the large number of revenue line items I suggest the following advice to facilitate replication of the model for each of the line items to be forecast. After copying the equation in E7 to the E7.. E17 range, return to E7 and remove the dollar signs ($) which act as address anchors. Use the commands /Name Range Create to name the formula in E7. Choose a name that will be easy to remember like "Lin" for "linear model." Later, when you replicate the model for other line items you can copy the forecast formula to the new cell simply by copying from the cell address entitled "LIN." Replace the address anchors (the $ signs) and copy the formula to the new range.

The forecast formula for the quadratic equation follows the same logic but is expanded to include both X variables and both X coefficients. If you have structured the worksheet according to section III A, you may simply copy the following formula into cell F7, and then copy cell F7 to cover the range F7..F17:

+O$7+(C7*N$13)+(D7*O$13)

You now have both a linear and a quadratic forecast based upon time-trend regression analysis. The second large task is to determine the validity or relative helpfulness of each model. We will assess the appropriateness of each model using the statistical information supplied by the regression and by plotting the actual collections against the forecast data. Both methods can give us information about the "Goodness of Fit" or the extent to which the forecast data conforms to actual data.

C. Graphing the Forecasts

To create a graph of the Linear Forecast data, execute the following Lotus commands. Remember that brackets <> indicate words which are to be typed as text; do not type the brackets.

1) Select /Graph Type XY

2) Select X and enter range C7..C17. Notice that this range accounts for years of historical data plus the five forecast years

3) Press A and highlight B7..B12

4) Press B and highlight E7..E17

5) Choose Options Format A Symbols; Select B Lines Quit

6

6) Select Scale X-Scale Manual Lower and enter a value of <0>; select Upper, enter <15> and then Quit.

7) To create titles, select Titles First, and enter <Cairo: Automobile Revenues>. Next select Titles Second and type <Linear Model>. Select Titles Y-Axis and type <Egyptian Pounds>. Select Titles X-Axis and type <Years, Year 1 = 1980/1>.

8) Choose Legend A and type <Actual Collection>; choose Legend B and type <Forecast>

9) Select Quit and View, to see the graph you have just created.

10) Select /Save and enter a name which you wish to call the graph. Once saved, you can send the graph to the printer at a future.

11) Choose /Name Create and enter a name by which you wish to identify this graph for future viewing and editing.

**********IMPORTANT**********

Unless you name your graph you will NOT be able to modify it in the future, even if you have saved the .PIC file onto your disk. It is always a good idea to name and save your graphs.

****

To create the graph depicting the quadratic model forecast simply requires two alterations to the linear graph that you just created.
First, choose Option Titles Second and type <Quadratic Model>. Finally, select Reset B Quit B and highlight F7..F17 and continue, performing steps 9-11 as outlined above.

Now compare the two graphs. Notice that the general pattern of the historical data takes the shape of a slight curve. The first graph illustrates that given this pattern of historical data, a linear model forecast fits the data adquately but under-estimates revenues in Years 3 through 6 and overestimates revenues in the remaining years. The second graph highlights the better fit of the quadratic regression curve. This better fit is reflected in the corresponding increase in the R-squared value associated with the quadratic equation. We will discuss the R-squared statistic and its implications in the next section.

D. Statistical Interpretation of Regression Results

1. Review the Regression Output

While this user's guide is not intended to be a statistics course, the following section provides a brief overview of the statistical output which the /Data Regression command generates.

a) The constant is the Y-intercept, or the value of the dependent variable when the Driver is zero.

b) The standard error of the Y-estimate is an indicator of the probable error in the Y-intercept. Ideally, the standard error of the Y-estimate should be small relative to the size of the constant. In general, one should distrust any data regression with a standard error that is more than half the value of the constant.

c) R-squared indicates the "goodness of fit" or how well the forecast values match the historical values. R-squared will always be between zero and one. A zero signifies that there is no relationship between the two data sets while an R-squared of one means that the two data sets are perfectly matched. An alternative explanation is that the R-squared represents the percentage amount which the Driver(s) can account for the value of the dependent variable. We will use this statistic in the calculation of the Student's t statistic.

d) Number of observations is simply a count of the data points in the X and Y ranges. The fewer the number of observations, the weaker the predictive powers of the model and the fewer the options available to manipulate the model. Standard statistical laws require at least 15 observations. The validity of and thus the confidence which you can place in your forecasts will be enhanced as you add historical data to the data base.

e) The degrees of freedom is the number of independent observations. Another way to derive the degrees is freedom is to think of it as the total number of Y data points MINUS the constant and the number of X variables. Notice the difference in the degrees of freedom between the linear and the quadratic models. In our linear example, we have six (6) years of data on actual collections, one constant and one Driver or X variable, and six (6) degrees of freedom accordingly. In our quadratic model, we still have six (6) years of actual collection data, but they are are restricted by one constant and two X variables for a total of four (8 - 3 = 5) degrees of freedom. The concept of degrees of freedom reiterates the importance of having sufficient numbers of historical observations. For example, it is impossible to run a quadratic regression with fewer than three degrees of freedom.

We will use the degrees of freedom to calculate and to analyze the Student's t test statistic. Interpretation of the Student's t test and, therefore the resulting level of confidence which you can accord a given model, will be very sensitive to the number of degrees of freedom available.

f) Standard error of the coefficient, similar to the standard error of the Y-estimate, measures the probable error in each X coefficient and ideally should be as small as possible. In general look for the standard error of the coefficient to be no greater than one third the X coefficient.

2. The Student's T Test (1)

In this section, we will use the information generated by the regression to determine which model, linear or quadratic provides the best fit of the data and a sufficient base for making forecasts.

Each time that you add a Driver to the regression equation, your R-squared or measure of goodness of fit, will increase. Reliance upon R-squared alone is insufficient to determine the validity of a model. You must expand your analysis to determine whether the increase in R squared resulting from the additional term (Years^2) is statistically significant. By incorporating the statistical information we have generated into a formula for the Student's t Test, we will be able to expand our analysis beyond Goodness of Fit to include estimates of the significance of each of the models.

First, modify your spreadsheet to perform the t test,

In cell address:	Enter:
Q7	+J9 (R-squared Statistic linear model)
Q8	+O9 (R-squared Statistic quadratic model)
R7	+Q7
R8	+Q8-Q7
T7	+J11 (Degrees of Freedom, linear model)
T8	+O11 (Degrees of Freedom,quadratic model)

In cell S7 we will enter the formula for the t statistic. The t statistic can be calculated in various ways. For our purposes the easiest method is:

t statistic = [(increase in R-squared)*(degrees of freedom)/(1-R-squared)]^.5

which translates in the language of Lotus 123 to:

@SQRT((R7*T7)/(1-Q7))

Next, copy the formula in S7 to S8 to calculate the T statistic for the quadratic regression. You will need the standard statistical table called the Student's t-distribution table in order to determine whether the t statistics in column S are significant. A copy of this table is presented in Exhibit III.1 of this guide.

In our example, Cairo automobile tax revenues, the Student's t for the linear equation is 8.64. Even with only 6 degrees of freedom, a Student's t score of 8.64 is greater than the value which corresponds to the highest level of significance. From these results you can determine that the linear regression produced a highly signficant fit of the forecast and historical data. The quadratic equation does not perform quite as well as the linear equation. At five degrees of freedom the quadratic equation's t score of 2.23 is signficant only at the 90% level. A signficance (or confidence) level of 95% is generally the lowest acceptable level of confidence. In this case, the limited number of available historical data suggest that the linear model more closely approximates the relationship between auto tax revenues and time than does a quadratic model.

If, however, you are willing to accept the 90% level of statistical signficance you would choose to use the quadratic model. Given a set of equations which produce statistically significant fits of the data, the equation with the highest R-squared should be your best choice. When you review the accompanying preliminary forecasts you will see that the performance of both models will vary across line items. You should watch for changes in the R-squared and therefore the Student's t test as you continue to add actual collection data to the model in the future. When you feel comfortable with the model you may also wish to experiment with other mathematical transformations (5) of the Revenue and Driver data.

IV. Performing Additional Forecasts/System Modification

Your forecasts will improve as you enter additional years' data to the database. To update the system with additional revenue data simply use the /Worksheet Insert Row command to make room for the incoming data. Go to the last row containing your formulas in Columns E and F ("Forecast Linear" and "Forecast Quadratic"). Remove the dollar sign ($) from each formula and copy the remaining formula into the newly created rows reserved for the additional forecast years. Once you have entered the new revenue data you will be able to recreate your regression. Simply repeat the steps in Section III to create and to analyze your new forecasts. The new regression output will overwrite the previous results. (6) Lotus will use the formulas in columns E and F to

Table 3 Percentiles of the *t* distribution

	Pr							
df	.80	.60	.40	.20	.10	.05	.02	.01
1	0.325	0.727	1.376	3.078	6.314	12.706	31.821	63.657
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660
120	0.254	0.526	0.845	1.289	1.658	1.980	2.358	2.617
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576

†Pr represents the probability that the *t* value will exceed each number in the table in absolute value. This is appropriate for two-tailed tests. For one-tailed tests simply divide each probability in half. For example, .325 in row 1, column 1 tells us that the probability of *t* being less than −.325 *or* greater than .325 is .8.

SOURCE: Obtained from Table III of Fisher and Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, with the permission of the authors and publishers (Edinburgh: Oliver & Boyd, Ltd.)

calculate new forecasts from the new regression output automatically.

You may also wish to experiment with other forms of the data such as those outlined in Step 2 of the forecasting methodology.

> **NOTE:** The easiest and safest approach is to save your worksheet under a different name before attempting any transformations of the data.

In this manner you can preserve the original database and forecasts for comparison with future system modifications. Next, you can use the mathematical functions in Lotus to transform the Actual Collection and/or the Year data as you choose. Once you finish manipulating the data you simply repeat the Steps in Section III to create and analyze your new forecasts.

Finally, as your forecasting efforts expand you may alter the revenue system by adding or substituting different sets of Drivers for which you have adequate historical data. Please refer to Volume 1-Revenue Sources for a discussion of the revenue definitions and Drivers formulated to date.

11

NOTES

(1) Acknowledgment: Adapting the T-statistic to Lotus format as presented by Dr. Lee-Ann Hayek, Dr. Alan Cheetham, "Curvilinear Regression in Release 2", Lotus Magazine, May 1987, pp 78-83. Traditionally, the Student's t score is used to guage the statistical significance of an estimated X-coefficient.

(2) Of course, you can modify the model to accept alternative Drivers which affect Revenues. More sophisticated specifications of the model introduce statistical issues which must be taken into consideration when choosing the appropropriate mix of Drivers. These issues, which include but are not limited to: multicollinearity, serial correlation, heteroskedacticity, and simulatancous equations, are beyond the scope of this User's Guide. Should the Governorate wish to expand its system to include alternative Drivers, it should also expand its statistical analysis to test and to connect for statistical basis which may effect the proposed regression equations.

(3) Future assessments of the Driver, Years^2 will also be simple to generate. Just raise to the second power each additional year which you forecast into the future.

(4) Lotus Ttips, Tricks ans Traps pg. 101

(5) See Section II B, page 2 for sample transformations.

(6) If you wish to compare the results you can move the the contents of the Regression Output cells to a free area of the worksheet before creating the new regression.

TABLE II.1

SUMMARY FORECASTING METHODOLOGY

A. Identify the factors which you feel should influence the variable which you wish to forecast (the dependent variable).

B. Establish the conceptual relationship between the dependent variable and its Driver(s).

C. Quantify the relationship between the dependent variable and the Driver(s)

D. Forecast future revenues using the equation you develop in Step 3 and your knowledge or estimates of the Driver data in future years.

E. Determine the validity of each model and choose the most appropriate model for each line item.

F. Update the system to perform forecasts in the future.