Regression Forecasting

PREREQUISITE TOOLS

None.

USAGE

PURPOSE

Regression forecasting obtains the relationships between two (or more) variables based on pairs (or sets) of past data values.

USES

Regression forecasting is used to:

1) Obtain economic forecasts
2) Forecast demand for services and products.
3) Forecast any variable where past behavior is assumed to continue.

KEY DEFINITIONS

1) An *independent variable* is the non-random variable which is used for forecasting other variables using the regression. M is the independent variable in:

$$B = a + (b \times M) \qquad [1]$$

where

B = number of births
M = number of marriages registered
a,b= constants

2) A *dependent variable* in regression forecasting is the variable being forecast. It is written in the regression equation as being dependent on the independent variable. For example, in [1] the dependent variable is "number of births."

3) A variable is *regressed* on another when the former is dependent on the latter. In [1], the "number of births" is regressed on the "number of marriages registered."

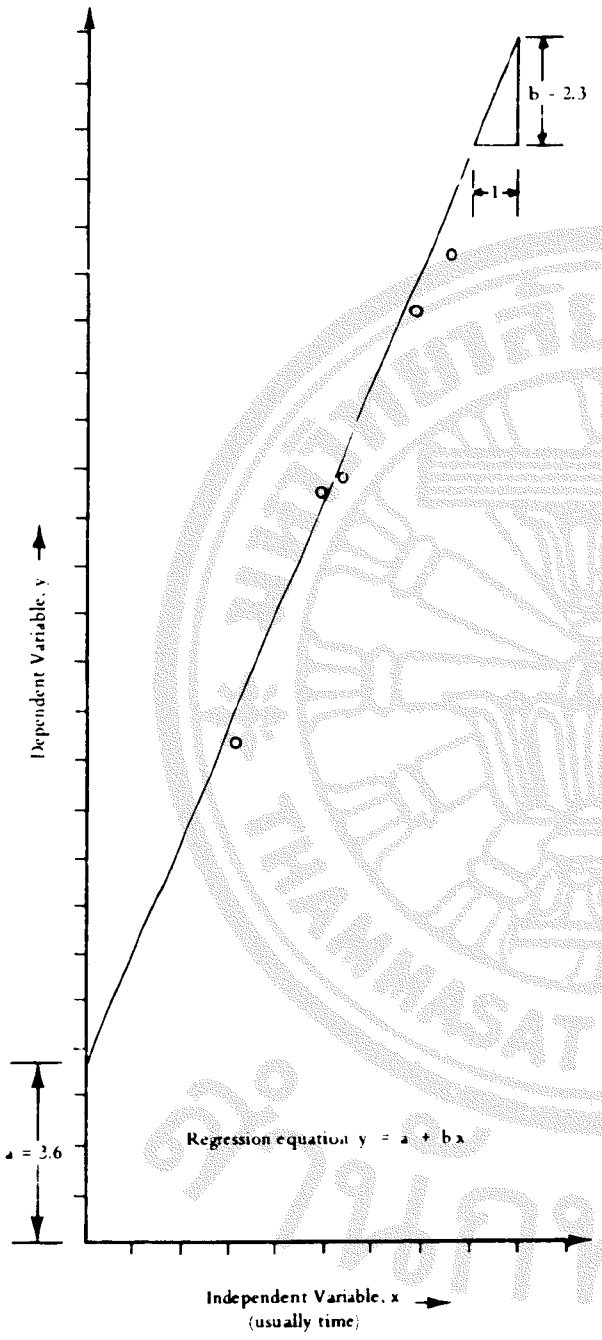
4) *Correlative behavior* is an assumed relationship between two or more variables in which the changes in one variable may be associated with predictable changes in the others. The change, however, is not necessarily cause-effect.

SHORT DESCRIPTION

Regression relates a *dependent variable* with an *independent variable* in the form of a mathematical equation. The independent variable is usually time, and regression extrapolates the past into the future. The equation is obtained from past data gathered in pairs (a value of the dependent variable corresponding to a value of the independent variable).

If the relationship is assumed to be linear, the regression of the dependent variable on the independent variable is a straight line when plotted on a graph (see figure 1). The simple linear regression equation is used to obtain forecasts of the dependent variable for a given value of the independent variable. A dependent variable may be *regressed* on two or more independent variables, but this is

FIGURE 1
Graph of Regression Line



not easily visualized on a graph. The forecasting method is similar, however, to simple linear regression.

ADVANTAGES

1) Regression is a simple and straightforward process.
2) Regression can be used in a wide variety of situations and is often the only recourse for forecasting. For example, in many social and economic contexts, causal or predictive models based on theoretical grounds are difficult to construct. Regression gives an empirical model which can be used for forecasting.

LIMITATIONS

1) Regression models estimate the correlation between variables. This *correlative behavior* is often mistaken as meaning "a change in the independent variable causes a change in the dependent variable." This may lead to false assumptions about causal relationships (see Oval Diagramming, OVD, page 81).

2) Regression forecasting extrapolates the data in order to obtain the forecasts. The relationship obtained from the data does not necessarily hold outside the range of available data, and erroneous forecasts are obtained. For example, the relationship between two variables may be linear only in the region examined and non-linear in other regions.

REQUIRED RESOURCES

LEVEL OF EFFORT

The effort required is minimal if the data for the regression model are available. However, a considerable amount of effort may be required if data collection is necessary. For example, surveys (SVY, page 36) may be needed to obtain the data.

SKILL LEVEL

Some statistical knowledge is needed to fully understand and use regression.

TIME REQUIRED

The time required to gather the data depends on the nature of the variable and the amount of data needed. Adequate regression models can be obtained using 20 to 50 pairs of data points. Once the data are obtained, the calculations require only a few hours. Regression on more than one variable takes more time depending on the number of variables being considered.

SPECIAL REQUIREMENTS

A calculator or a slide rule is useful in making the calculations.

DESCRIPTION OF TOOL

SUPPLEMENTAL DEFINITIONS

1) The *regression coefficient* is the coefficient of the independent variable in a regression equation. In [1], "b" is the regression coefficient.

2) Symbols:
x is the independent variable
y is the dependent variable
a and b are regression coefficients.

REQUIRED INPUTS

Knowledge about the variables defined in the regression equation is needed. Between 20 and 50 sets of data points are needed to obtain the regression equation. The higher the number of data sets, the higher the reliability of the regression equation.

TOOL OUTPUT

The output is the regression equation model relating the variables. The model may then be used to forecast values of the dependent variable for given values of the independent variable.

IMPORTANT ASSUMPTIONS

Regression models assume that the independent variable is deterministic (non-random) and can be measured with an accuracy that is much higher than that involved in measuring the dependent variable. Often both the variables are random, and the variable which can be measured with less error is chosen as the independent variable. However, most often observations of one variable are made at intervals of time. Time becomes the independent variable in the regression equation, and the assumption about the independent variable is then valid.

METHOD OF USE

GENERAL PROCEDURE

Linear regression fits a linear equation between the variables (see figure 1). The *regression coefficients* of the equation are selected so that the data values have minimum deviation from the line.

The following procedure is recommended to develop a regression equation.

1. **Obtain the data.**
Once the independent and dependent variables are determined, the data values are obtained in pairs, i.e., a datum point for the dependent variable corresponds to each value for the independent variable. The data should be recent and should be representative of the trend. Consider a situation where the total industrial output in a region for the next year is to be forecast. The industrial output is known to be correlated to the annual steel production. The industrial output will be regressed on the steel production. Data for past five years are used to obtain the equation (see figure 2).

2. **Determine the equation coefficients.**
If the relationship between the variables is assumed to be linear (see figure 1), the regression equation used is:

$$y = a + bx. \qquad [2]$$

The regression coefficients are calculated using:

$$b = \sum^{n} (x - x')(y - y') / \sum^{n} (x - x')^2 \qquad [3]$$

where
x', y' = averages of n data points for x and y
$\sum^{n}$ = summation of all terms in parentheses computed from data points

The calculations for [3] are easily done using a table (see figure 2b). The data points for x and y are first filled in and the averages x' and y' calculated. Using these averages, the rest of the table is filled in. The totals for columns $(x\text{-}x')$ and $(y\text{-}y')$ should be zero. This can be used as a check for calculations. The ratio of the totals in columns 5 and 6 then gives the value of b:

$$b = 26.86/11.70 = 2.29.$$

The coefficient a is calculated by

$$a = y' - bx'. \qquad [4]$$

In the example,

$$a = 16.2 - (2.29 \times 5.5) = 3.6,$$

the regression equation is

$$y = 3.6 + 2.29 \times x. \qquad [5]$$

3. **Forecast using the equation.**
The forecast of a new value of the dependent variable is made by fitting the corresponding new value for the independent variable into the equation. For example, steel production is known to be six million tons. The industrial output is estimated by substituting in [5], so that:

$$(y) = 3.6 + 2.29 \times 6 = 17.34 \qquad [6]$$

The industrial output (y) is estimated at \$17.34 million using the linear regression equation.

EXAMPLE

Many examples of regression analysis may be found in the literature. Fredericks' (1976) analysis of cooperative

FIGURE 2
Regression Computation

a) Data for Forecasting Industrial Output

Year	1970	1971	1972	1973	1974
Steel Production/yr (million of tons)	7.5	6.8	3.1	5.2	4.9
Industrial Output/yr (millions of $)	20.3	19.2	10.3	15.8	15.6

b) Table for Computing Regression Coefficients

	(1) x	(2) y	(3) $x - x'$	(4) $y - y'$	(5) $(x - x')(y - y')$	(6) $(x - x')^2$
	7.5	20.3	2.2	4.1	8.2	4.0
	6.8	19.2	1.3	3.0	3.9	1.69
	3.1	10.3	-2.4	-5.9	14.16	5.76
	5.2	15.8	-0.3	-0.4	.12	.09
	4.9	15.4	-0.6	-0.8	.48	.36
TOTALS	27.5	81.0	0.0	0.0	26.86	11.70
AVERAGES	$x' = 5.5$	$y' = 16.2$	0.0	0.0	$b = \frac{26.86}{11.70} = 2.29$	

movements in West Malaysia is illustrative and instructive. Twelve structural variables were included in the analysis of structural development.

THEORY

Regression equation models are widely treated in statistics texts (Fryer 1966, or Wetherill 1972). The theory is based on summing the squares of the deviation of each data point from the corresponding value of the model, and then selecting coefficients of the model which minimize this sum. If the model equation is a straight line, the coefficients fit a linear regression model. Non-linear regression models are used to fit equation coefficients to data which do not appear to fall on a straight line.

Multiple regression models use the same basic principle to fit the observed data to two or more independent variables. The forecaster is referred to specialized texts (Draper, 1966) for details.

Bedworth (1973) has an excellent presentation of regression forecasting when the independent variable is time.

BIBLIOGRAPHY

Bedworth, D. D. *Industrial Systems.* New York: Ronald Press Company, 1973, pp. 84-113.

Draper, N. R., and Smith, H. *Applied Regression Analysis.* New York: John Wiley and Sons, 1966.

Fredericks, L.J. "A Trend Analysis of the Structure of the Cooperative Movement in West Malaysia, 1922-1967." *Cooperative Information,* ILO, January 1976.

Fryer, H. C. *Concepts and Methods of Experimental Statistics.* Boston: Allyn and Bacon, 1966.

Wetherill, G. Barne. *Elementary Statistical Methods.* London: Chapman and Hall, 1972.