Advanced Statistics : Regression Analysis

retnosubekti@uny.ac.id

Introduction

The term *regression* is known from the work of Sir Francis Galton (1822-1911), a famous geneticist, who studied the size of seeds and their offspring and the heights of father s and their sons. In both cases, he found that the offspring of parents of larger than average size tended to be smaller than their parents and that offspring of parents og parents of smaller than average size tended to be larger than their parents. Galton called this phenomenon as *regression toward mediocrity*.

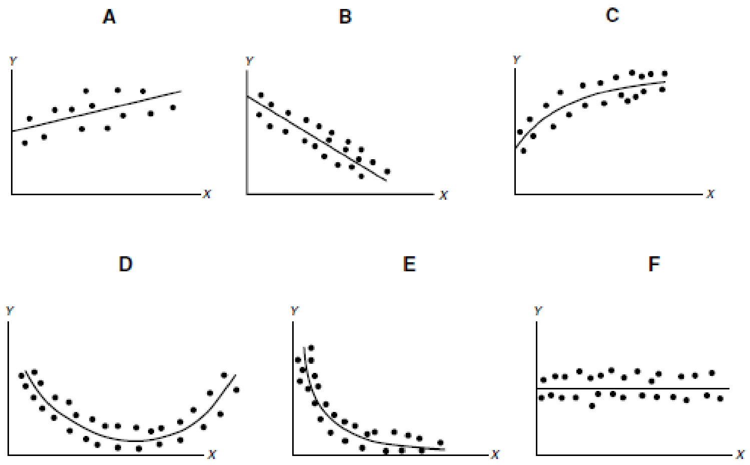
In regression analysis there are two kinds of variable, Independent variable and dependent variable. Independent variable is usually denoted by X-variable, it was also called as predictor variable or explanatory variable. Dependent variable is usually denoted by Y-variable, it can be said as a response variable.

For example, a manager of mall is interesting to learn the relation between profit and numbers of visitor and its' advertorial. What did you think about that case? Which one as an independent variable and which one as a dependent variable?

Simple Linear Regression

CONCEPT A statistical technique that uses a straight-line relationship to predict a numerical dependent variable *Y* from a *single* numerical independent variable *X*.

INTERPRETATION Simple *linear* regression attempts to discover whether the values of the dependent *Y* (such as store sales) and the independent *X* variable (such as the size of the store), when graphed on a scatter plot would suggest a straight-line relationship of the values. The figure below shows the different types of patterns that you could discover when plotting the values of the *X* and *Y* variables.



The patterns shown above can be described as follows:

- Panel *A*, positive straight-line or linear relationship between *X* and *Y*.
- Panel *B*, negative straight-line or linear relationship between *X* and *Y*.
- Panel *C*, a positive curvilinear relationship between *X* and *Y*. The values of *Y* are increasing as *X* increases, but this increase tapers off beyond certain values of *X*.
- Panel *D*, a U-shaped relationship between *X* and *Y*. As *X* increases, at first *Y* decreases. However, as *X* continues to increase, *Y* not only stops decreasing but actually increases above its minimum value.
- Panel *E*, an exponential relationship between *X* and *Y*. In this case, *Y* decreases very rapidly as *X* first increases, but then decreases much less rapidly as *X* increases further.
- Panel *F*, values that have very little or no relationship between *X* and *Y*. High and low values of *Y* appear at each value of *X*.

Scatter plots only informally help you identify the relationship between the dependent variable *Y* and the independent variable *X* in a simple regression. To specify the numeric relationship between the variables, you need to develop an equation that best represents the relationship.

Determining the Simple Linear Regression Equation

After you determine that a straight-line relationship exists between a dependent variable *Y* and the independent variable *X*, you need to determine which straight line to use to represent the relationship. Two values define any straight line: the *Y intercept* and the *slope*.

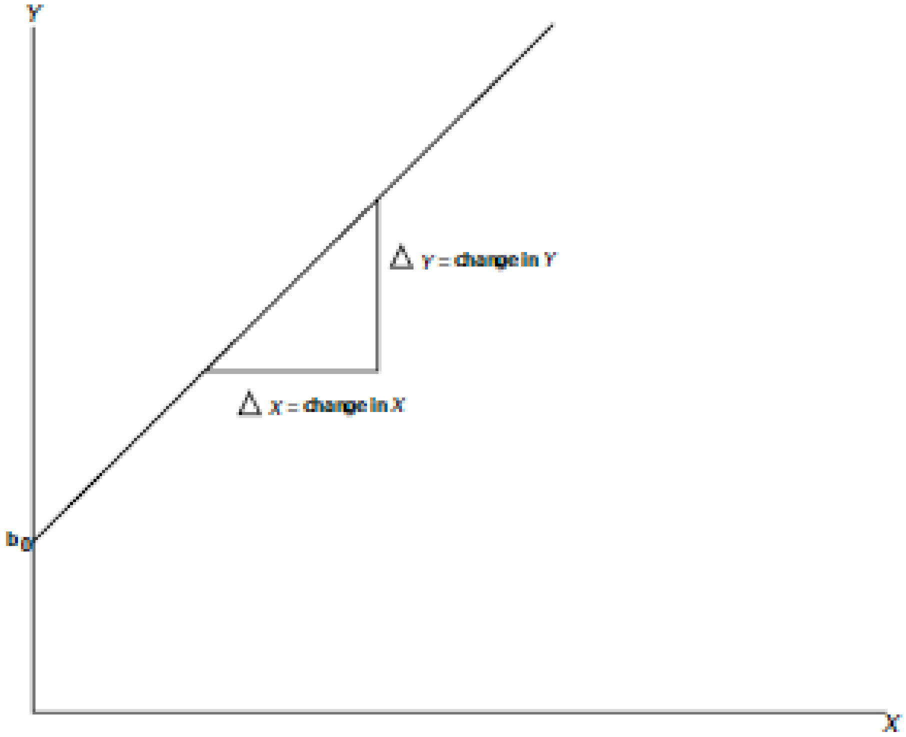
Y intercept

CONCEPT The value of *Y* when *X* = 0, represented by the symbol *b0*.

Slope

CONCEPT The change in Y per unit change in X represented by the symbol $b1$. Positive slope means Y increases as X increases. Negative slope means Y decreases as X increases.

INTERPRETATION The Y intercept and the slope are known as the **regression coefficients**. The symbol $b0$ is used for the Y intercept, and the symbol $b1$ is used for the slope. Multiplying a specific X value by the slope and then adding the Y intercept generates the corresponding Y value. The equation $Y = b0 + b1X$ is used to express this relationship for the entire line. (Some sources use the symbol a for the Y intercept and b for the slope to form the equation $Y = a + b\ X$.)



Least-Squares Method

CONCEPT The simple linear regression method that seeks to minimize the sum of the squared differences between the actual values of the dependent variable Y and the predicted values of Y.

INTERPRETATION For plotted sets of X and Y values, there are many possible straight lines, each with its own values of $b0$ and $b1$, that might seem to fit the data. The least-squares method finds the values for the Y intercept and the slope that makes the sum of the squared differences between the actual values of the dependent variable Y and the predicted values of Y as small as possible.

Calculating the Y intercept and the slope using the least-squares method is tedious and can be subject to rounding errors if you use a simple four-function calculator. You will get more accurate results faster if you use regression software routines to perform the calculations.

Other topics : Check out http://besmart.uny.ac.id/course/view/1076