LINEAR REGRESSION ANALYSIS

MODULE – I

Lecture - 1

Introduction

Dr. Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Linear models play a central part in modern statistical methods. On the one hand, these models are able to approximate a large amount of metric data structures in their entire range of definition or at least piecewise.

Linear models and regression analysis

Suppose the outcome of any process is denoted by a random variable *y*, called as dependent (or study) variable, depends on *k* independent (or explanatory) variables denoted by $X_1, X_2, \ldots, X_k$. Suppose the behaviour of *y* can be explained by a relationship given by

$$y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ..., \beta_k) + \varepsilon$$

where *f* is some well defined function and $\beta_1, \beta_2, ..., \beta_k$ are the parameters which characterize the role and contribution of $X_1, X_2, \ldots, X_k$, respectively. The term ε reflects the stochastic nature of the relationship between *y* and $X_1, X_2, \ldots, X_k$ and indicates that such a relationship is not exact in nature. When $\varepsilon = 0$, then the relationship is called the mathematical model otherwise the statistical model. The term "**model**" is broadly used to represent any phenomenon in a mathematical frame work.

A model or relationship is termed as linear if it is linear in parameters and nonlinear, if it is not linear in parameters. In other words, if all the partial derivatives of *y* with respect to each of the parameters $\beta_1, \beta_2, ..., \beta_k$ are independent of the parameters, then the model is called as a **linear model**. If any of the partial derivatives of *y* with respect to any of the $\beta_1, \beta_2, ..., \beta_k$ is not independent of the parameters, the model is called as nonlinear. Note that the linearity or non-linearity of the model is not described by the linearity or nonlinearity of explanatory variables in the model.

For example

$$y = \beta_1 X_1^2 + \beta_2 \sqrt{X_2} + \beta_3 \log X_3 + \varepsilon$$

is a linear model because $\partial y / \partial \beta_i$, $(i = 1,2,3)$ are independent of the parameters β_i, $(i = 1,2,3)$. On the other hand,

$$y = \beta_1^2 X_1 + \beta_2 X_2 + \beta_3 \log X + \varepsilon$$

is a nonlinear model because $\partial y / \partial \beta_1 = 2\beta_1 X_1$ depends on β_1 although $\partial y / \partial \beta_2$ and $\partial y / \partial \beta_3$ are independent of any of the β_1, β_2 or β_3 .

When the function f is linear in parameters, then $y = f(X_1, X_2, ..., X_k, \beta_1, \beta_2, ..., \beta_k) + \varepsilon$ is called a linear model and when the function f is nonlinear in parameters, then it is called a nonlinear model. In general, the function f is chosen as

$$f(X_1, X_2, ..., X_k, \beta_1, \beta_2 ..., \beta_k) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k$$

to describe a linear model. Since $X_1, X_2, \ldots, X_k$ are pre-determined variables and y is the outcome, so both are known. Thus the knowledge of the model depends on the knowledge of the parameters $\beta_1, \beta_2, ..., \beta_k$.

The statistical linear modeling essentially consists of developing approaches and tools to determine $\beta_1, \beta_2, ..., \beta_k$ in the linear model

$$y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

given the observations on y and $X_1, X_2, \ldots, X_k$.

Different statistical estimation procedures, e.g., method of maximum likelihood, principle of least squares, method of moments etc. can be employed to estimate the parameters of the model. The method of maximum likelihood needs further knowledge of the distribution of y whereas the method of moments and the principle of least squares do not need any knowledge about the distribution of y.

The regression analysis is a tool to determine the values of the parameters given the data on y and $X_1, X_2, \ldots, X_k$. The literal meaning of regression is "to move in the backward direction". Before discussing and understanding the meaning of "backward direction", let us find which of the following statements is correct:

S1: model generates data or

S2: data generates model.

Obviously, S1 is correct. It can be broadly thought that the model exists in nature but is unknown to the experimenter. When some values to the explanatory variables are provided, then the values for the output or study variable are generated accordingly, depending on the form of the function f and the nature of phenomenon. So ideally, the pre-existing model gives rise to the data. Our objective is to determine the functional form of this model. Now we move in the backward direction. We propose to first collect the data on study and explanatory variables. Then we employ some statistical techniques and use this data to know the form of function f. Equivalently, the data from the model is recorded first and then used to determine the parameters of the model. The regression analysis is a technique which helps in determining the statistical model by using the data on study and explanatory variables. The classification of linear and nonlinear regression analysis is based on the determination of linear and nonlinear models, respectively.

Consider a simple example to understand the meaning of "regression". Suppose the yield of crop (y) depends linearly on two explanatory variables, viz., the quantity of a fertilizer (X_1) and level of irrigation (X_2) as

$$y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

There exist the true values of β_1 and β_2 in nature but are unknown to the experimenter. Some values on y are recorded by providing different values to X_1 and X_2. There exists some relationship between y and X_1, X_2 which gives rise to a systematically behaved data on y, X_1 and X_2. Such relationship is unknown to the experimenter. To determine the model, we move in the backward direction in the sense that the collected data is used to determine the unknown parameters β_1 and β_2 of the model. In this sense such an approach is termed as regression analysis.

The theory and fundamentals of linear models lay the foundation for developing the tools for regression analysis that are based on valid statistical theory and concepts.

Steps in regression analysis

Regression analysis includes the following steps:

- Statement of the problem under consideration
- Choice of relevant variables
- Collection of data on relevant variables
- Specification of model
- Choice of method for fitting the data
- Fitting of model
- Model validation and criticism
- Using the chosen model(s) for the solution of the posed problem and forecasting.

These steps are examined below.

1. Statement of the problem under consideration

The first important step in conducting any regression analysis is to specify the problem and the objectives to be addressed by the regression analysis. The wrong formulation or the wrong understanding of the problem will give the wrong statistical inferences. The choice of variables depends upon the objectives of study and understanding of the problem. For example, height and weight of children are related. Now there can be two issues to be addressed.

(i) Determination of height for given weight, or

(ii) determination of weight for given height.

In the case (i), the height is response variable whereas weight is response variable is case (ii). The role of explanatory variables are also interchanged in the cases (i) and (ii).

2. Choice of potentially relevant variables

Once the problem is carefully formulated and objectives have been decided, the next question is to choose the relevant variables. It has to kept in mind that the correct choice of variables will determine the statistical inferences correctly. For example, in any agricultural experiment, the yield depends on explanatory variables like quantity of fertilizer, rainfall, irrigation, temperature etc. These variables are denoted by $X_1, X_2, \ldots, X_k$ as a set of k explanatory variables.

3. Collection of data on relevant variables

Once the objective of study is clearly stated and the variables are chosen, the next question arises is to collect data on such relevant variables. The data is essentially the measurement on these variables. For example, suppose we want to collect the data on age. For this, it is important to know how to record the data on age. Then either the date of birth can be recorded which will provide the exact age on any specific date or the age in terms of completed years as on specific date can be recorded. Moreover, it is also important to decide that whether the data has to be collected on variables as quantitative variables or qualitative variables. For example, if the ages (in years) are 15,17,19,21,23, then these are quantitative values. If the ages are defined by a variable that takes value 1 if ages are less than 18 years and 0 if the ages are more than 18 years, then the earlier recorded data is converted to 1,1,0,0,0. Note that there is a loss of information in converting the quantitative data into qualitative data. The methods and approaches for qualitative and quantitative data are also different. If the study variable is binary, then **logistic** and **probit regressions** etc. are used. If all explanatory variables are qualitative, then **analysis of variance** technique is used. If some explanatory variables are qualitative and others are quantitative, then **analysis of covariance** technique is used. The techniques of analysis of variance and analysis of covariance are the special cases of regression analysis .

Generally, the data is collected on n subjects, then y denotes the response or study variable and $y_1, y_2, ..., y_n$ are the n values. If there are k explanatory variables $X_1, X_2, .., X_k$ then x_{ij} denotes the i^{th} value of j^{th} variable, $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, k$. The observation can be presented in the following table:

Notations for the data used in regression analysis

Observation Number	**Response** y	**Explanatory variables**			
		X_1	X_2	$\cdots$	X_k
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1k}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2k}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{nk}

4. Specification of model

The experimenter or the person working in the subject usually helps in determining the form of the model. Only the form of the tentative model can be ascertained and it will depend on some unknown parameters . For example, a general form will be like

$$y = f(X_1, X_2, ..., X_k; \beta_1, \beta_2, ..., \beta_k) + \varepsilon$$

where ε is the random error reflecting mainly the difference in the observed value of y and the value of y obtained through the model. The form of $f(X_1, X_2, ..., X_k; \beta_1, \beta_2, ..., \beta_k)$ can be linear as well as nonlinear depending on the form of parameters $\beta_1, \beta_2, ..., \beta_k$. A model is said to be linear if it is linear in parameters. For example,

$$y = \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \varepsilon$$
$$y = \beta_1 + \beta_2 \ln X_2 + \varepsilon$$

are linear models whereas

$$y = \beta_1 X_1 + \beta_2^2 X_2 + \beta_3 X_2 + \varepsilon$$
$$y = (\ln \beta_1) X_1 + \beta_2 X_2 + \varepsilon$$

are non-linear models. Many times, the nonlinear models can be converted into linear models through some transformations. So the class of linear models is wider than what it appears initially.

If a model contains only one explanatory variable, then it is called as **simple regression model.** When there are more than one independent variables, **then it** is called as **multiple regression model.** When there is only one study variable, the regression is termed as **univariate regression.** When there are more than one study variables, the regression is termed as **multivariate regression.** Note that the simple and multiple regressions are not same as univariate and multivariate regressions. The simple and multiple regression are determined by the number of explanatory variables whereas univariate and multivariate regressions are determined by the number of study variables.

5. Choice of method for fitting the data

After the model has been defined and the data have been collected, the next task is to estimate the parameters of the model based on the collected data. This is also referred to as **parameter estimation** or **model fitting**. The most commonly used method of estimation is the least squares method. Under certain assumptions, the least squares method produces estimators with desirable properties. The other estimation methods are the maximum likelihood method, ridge method, principal components method etc.

6. Fitting of model

The estimation of unknown parameters using appropriate method provides the values of the parameters. Substituting these values in the equation gives us a usable model. This is termed as model fitting. The estimates of parameters $\beta_1,...,\beta_k$ in the model

$$y = f(X_1, X_2,..., X_k, \beta_1, \beta_2,..., \beta_k) + \varepsilon$$

are denoted as $\hat{\beta}_0, \hat{\beta}_1,..., \hat{\beta}_k$. which gives the fitted model as

$$y = f(X_1, X_2,..., X_k, \hat{\beta}_1, \hat{\beta}_2,..., \hat{\beta}_k).$$

When the value of y is obtained for the given values of $X_1, X_2,..., X_k$, it is denoted as $\hat{y}$ and called as fitted value.

The fitted equation is used for prediction. In this case, $\hat{y}$ is termed as **predicted value.** Note that the fitted value is where the values used for explanatory variables correspond to one of the n observations in the data whereas predicted value is the one obtained for any set of values of explanatory variables. It is not generally recommended to predict the y - values for the set of those values of explanatory variables which lie outside the range of data. When the values of explanatory variables are the future values of explanatory variables, the predicted values are called forecasted values.

There are different methodologies based on regression analysis. They are described in the following table:

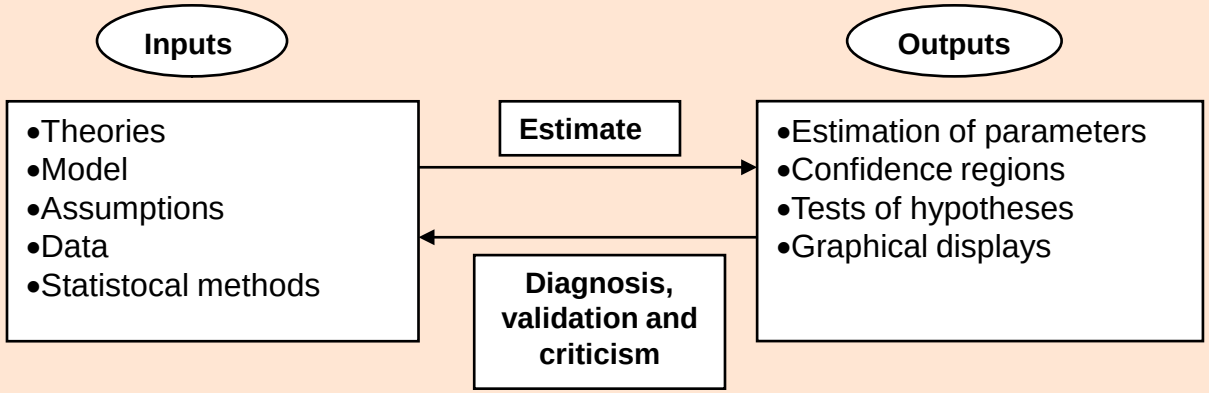
Various classification of regression analysis

Type of Regression	Conditions
Univariate	Only one quantitative response variable
Multivariate	Two or more quantitative response variables
Simple	Only one explanatory variable
Multiple	Two or more explanatory variables
Linear	All parameters enter the equation linearly, possibly after transformation of the data
Nonlinear	The relationship between the response and some of the explanatory variables is nonlinear or some of the parameters appear nonlinearly, but no transformation is possible to make the parameters appear linearly
Analysis of variance	All explanatory variables are qualitative variables
Analysis of Covariance	Some explanatory variables are quantitative variables and others are qualitative variables
Logistic	The response variable is qualitative

7. Model criticism and selection

The validity of statistical method to be used for regression analysis depends on various assumptions. These assumptions are essentially the assumptions for the model and the data. The quality of statistical inferences heavily depends on whether these assumptions are satisfied or not. For making these assumptions to be valid and to be satisfied, care is needed from beginning of the experiment. One has to be careful in choosing the required assumptions and to examine whether the assumptions are valid for the given experimental conditions or not. It is also important to decide the situations in which the assumptions may not meet.

The validation of the assumptions must be made before drawing any statistical conclusion. Any departure from validity of assumptions will be reflected in the statistical inferences. In fact, the regression analysis is an iterative process where the outputs are used to diagnose, validate, criticize and modify the inputs. The iterative process is illustrated in the following figure.



8. Objectives of regression analysis

The determination of explicit form of regression equation is the ultimate objective of regression analysis. It is finally a good and valid relationship between study variable and explanatory variables. The regression equation helps in understanding the interrelationships among the variables. Such regression equation can be used for several purposes. For example, to determine the role of any explanatory variable in the joint relationship in any policy formulation, to forecast the values of response variable for given set of values of explanatory variables.