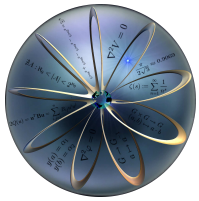

planetmath.org

Math for the people, by the people.

regression model

Canonical name	RegressionModel
Date of creation	2013-03-22 14:30:31
Last modified on	2013-03-22 14:30:31
Owner	CWoo (3771)
Last modified by	CWoo (3771)
Numerical id	10
Author	CWoo (3771)
Entry type	Definition
Classification	msc 62J02
Classification	msc 62J05
Synonym	univariate regression model
Related topic	LinearLeastSquaresFit
Defines	regression function
Defines	regression coefficient
Defines	simple regression model
Defines	multiple regression model
Defines	linear regression model
Defines	polynomial regression model
Defines	non-linear regression model

In statistical modeling of N data observations ($N < \infty$), two types of variables are usually defined. One is the response variable or variate, usually denoted by Y, and the other is the explanatory variable or covariate X. While there is only one response variable, there may be one or more than one explanatory variables. The response variable is considered random, where as the explanatory variable(s) may or may not be random.

Based on the above setup, a *univariate regression model*, or simply *regression model*, is a statistical model with the following assumptions:

1. all of the variables, random or not, are *continuous* in nature (as opposed to categorical in nature)

2. the response variable Y can be expressed as the sum of a function $f(\mathbf{X})$, called the *regression function*, where $\mathbf{X}$ represents the row vector of explanatory variables, and an error term ε_i:

$$Y = f(\mathbf{X}) + \varepsilon = f(X_1, \ldots, X_p) + \varepsilon$$

   where p is the number of explanatory variables. $f(\mathbf{X})$ is called the systematic component, and ε is the random error component.

3. the error component and the systematic component are independent

4. random error variables ε_i for the N observations are iid normal with mean 0 and variance σ^2

Any unknown variables appearing in the regression function f, other than the covariates, are called the *regression coefficients*.

Remarks

- The conditional distribution of Y, given $\mathbf{X}$ is normal, or Gaussian, with mean $\mu = \mathrm{E}\left[Y \mid \mathbf{X} = \boldsymbol{x}\right] = \mathrm{E}\left[Y \mid X_1 = x_1, \ldots, X_p = x_p\right]$ and variance σ^2. In addition, the random variables Y_i corresponding to the reponses are independent.

- Sometimes, Condition 4 above is skipped to encompass a wider class of regression models. Those models that observe Condition 4 is generally called a normal, or Gaussian regression model. Otherwise, they are classified under the non-linear regression model discussed below. Some well known non-normal regression models are the logistic regression for binary data and the Poisson regression for count data.

- A regression model can be classified by the number or explanatory variables. If there is only one explanatory variable, it is called a *simple regression model*. Otherwise, it is a *multiple regression model*.

- A regression model can also be classified by the form of the regression function f. If f can be expressed as a linear combination of the regression coefficients:

$$f(\mathbf{X}) = \beta_0 z_0(\mathbf{X}) + \cdots + \beta_k z_k(\mathbf{X}),$$

where the functions $z_i(\mathbf{X})$ do not contain any regression coefficients, then the model is called a *linear regression model*. Two examples of linear regression models are:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

and

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_k X^k + \varepsilon$$

The last one is called a *polynomial regression model*. Linear regression models belong to a more general class of statistical models called the general linear model, where explanatory variables are no longer restricted to be continuous ones only. When f can not be expressed linearly in terms of the regression coefficients, the model is known as a *non-linear regression model*. An example of a non-linear regression model is

$$Y = \beta_0 + \frac{1}{\beta_1 + \beta_2 X} + \varepsilon$$

- The univariate regression model can be generalized to what is known as the *multivariate regression model*, where at least two response variables are considered.