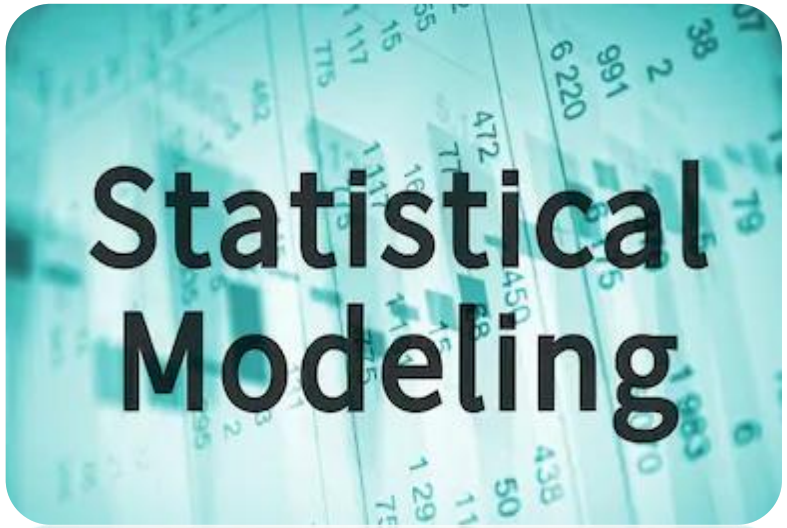
Subject : Introduction to Actuarial Models

Topic : Introduction to Modelling

What is statistical modeling?

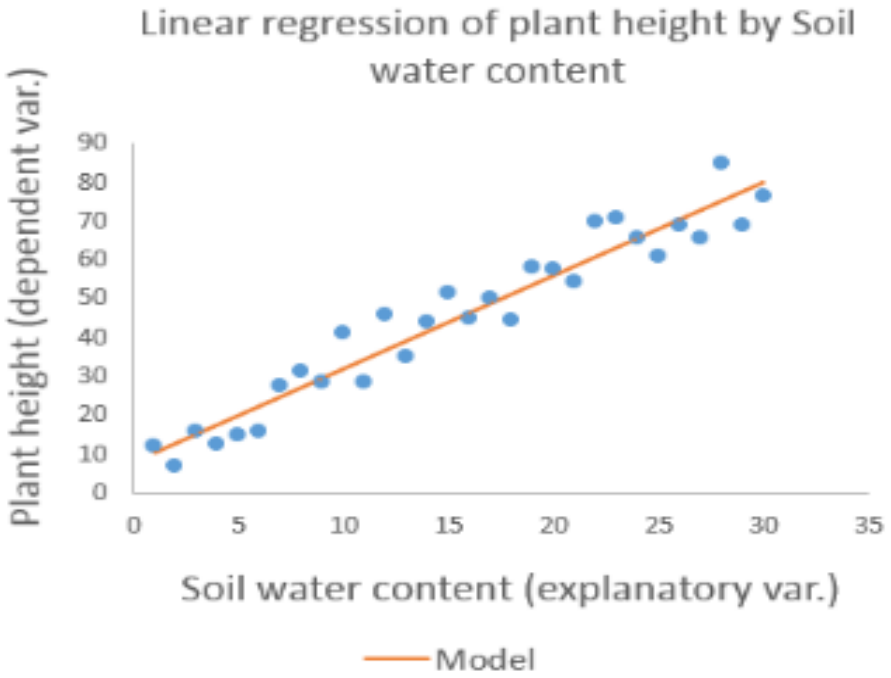


- In simple terms, **statistical modeling** is a simplified, mathematically-formalized way to approximate reality (i.e. what generates your data) and optionally to make predictions from this approximation.

- The statistical model is the mathematical equation that is used. Statistical models contain variables that can be used to explain relationships between other variables.
- A statistical model will have sampling, probability spaces, assumptions and diagnostics etc, to make inferences.

Example

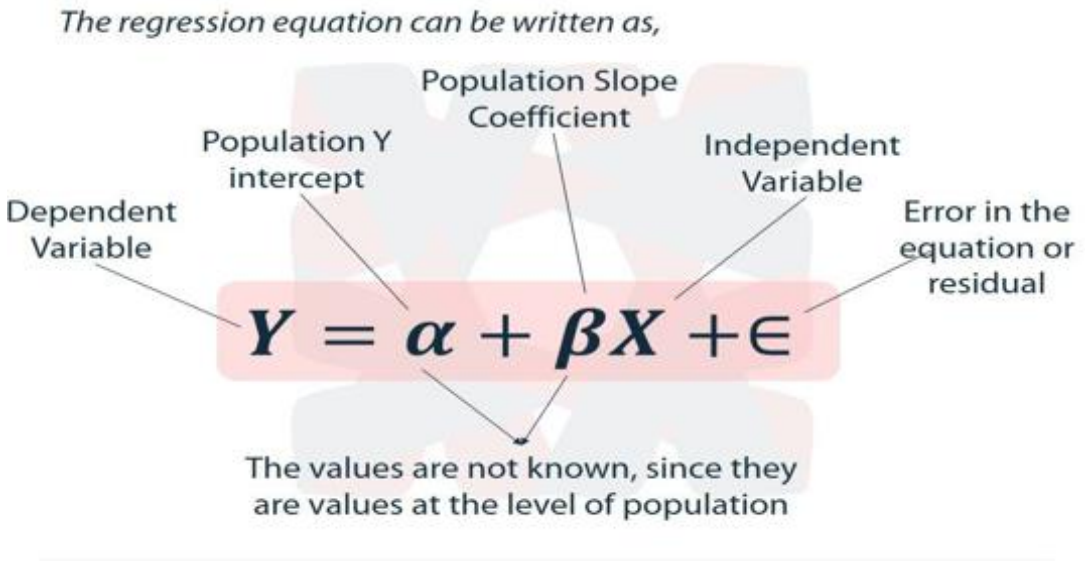
An example is attempting to represent the height of plants according to soil water content by a straight line characterized by a slope and an intercept drawn after an experiment on a sample of plants submitted to a growing soil humidity. This particular model is called simple linear regression. This is a very basic example, statistical models can be used to solve complicated problems with a great deal of understanding.



Terminology that Statisticians use.

- The general basic form of a model is;

$$y_i = \alpha + \beta x_i + \varepsilon_i .$$



- The response variable is the one we want to describe, to explain, to predict.
- Explanatory variables, also referred to as independent variables, are the ones we use to explain, to describe or to predict the dependent variable(s).
- The mathematical function consists of two parts. These parts are the predictor variables x1, x2,… and the parameters, $\beta_1 \beta_2$,…
- The random errors are unknown. They are simply the difference between the data and the mathematical function. They are assumed to follow a particular probability distribution, however, which is used to describe their aggregate behaviour

What are the models used for?

- Models are used for four main purposes:

1. Estimation, 2. Prediction, 3. Calibration, and 4. Optimization.

- Lets look at each one of them:

1] **Estimation=** The goal of estimation is to determine the value of the regression function (i.e., the average value of the response variable), for a particular combination of the values of the predictor variables. A critical part of estimation is an assessment of how much am estimated value will fluctuate due to the noise in the data

2] **Prediction=** The goal of prediction is to determine either
a). the value of a new observation of the response variable, or
b). the values of a specified proportion of all future observations of the response variable, for a particular combination of the values of the predictor variables.

Continued..

3] **Calibration=** The goal of calibration is to quantitatively relate measurements made using one measurement system to those of another measurement system. This is done so that measurements can be compared in common units or to tie results from a relative measurement method to absolute units

4] **Optimization=** Optimization is performed to determine the values of process inputs that should be used to obtain the desired process output. Typical optimization goals might be to maximize the yield of a process, to minimize the processing time required to fabricate a product, or to hit a target product specification with minimum variation in order to maintain specified tolerances.

What are some of the different statistical methods for model building?

- ***Selecting an Appropriate Stat Method for Modeling***

In order to build a statistical model, we need to be very careful in selecting the method. There is often more than one statistical tool that can be effectively applied to a given modeling application.

In the process of developing the model we will often come across situations where we build a first basic model and then run the model, perform calculations and try to improvise the model.

Here, we discuss some of the most popular and well-established statistical techniques that are useful for different model building situations;

- ***Modelling methods that we will discuss***

1. Linear Least Squares Regression

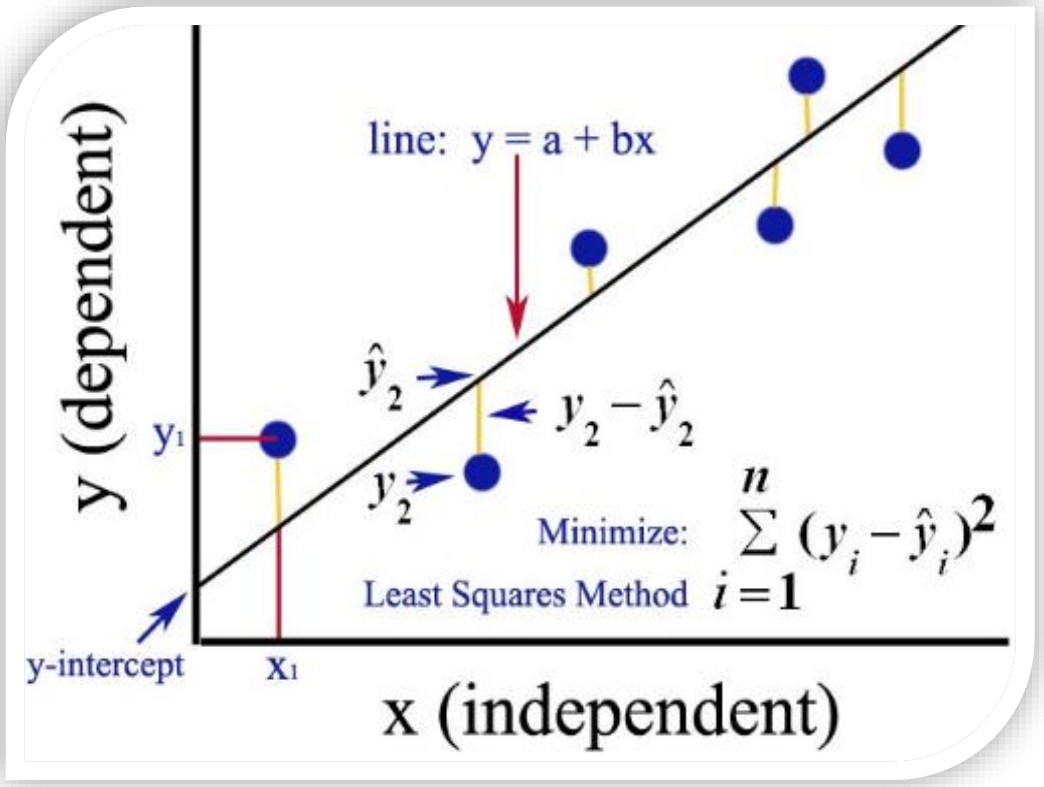
2. Nonlinear Least Squares Regression

3. Weighted Least Squares Regression

Linear Least Squares Regression

Least squares regression method is by far the most widely used modelling method.

The least squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve.

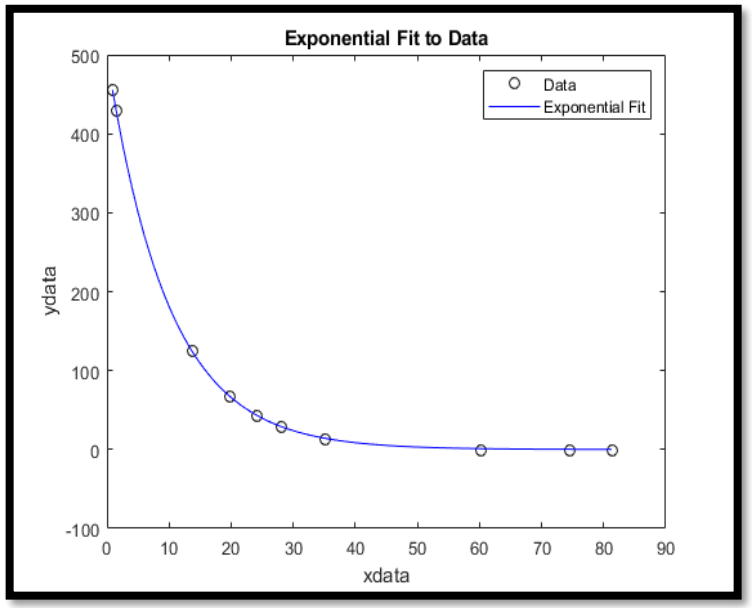


Linear least squares regression also gets its name from the way the estimates of the unknown parameters are computed – using the "method of least squares". In this method the unknown parameters are estimated by minimizing the sum of the squared deviations between the data and the model.

Nonlinear Least Squares Regression

- Nonlinear least squares regression extends linear least squares regression for use with a much larger and more general class of functions
- Nonlinear regression is a regression in which the dependent variables are modelled as a non-linear function of model parameters and one or more independent variables. The reason that these models are called nonlinear regression is because the relationships between the dependent and independent parameters are not linear.
- As the name suggests, a nonlinear model is a model of the basic form,

$$Y = f(\vec{x};\vec{\beta}) + \varepsilon,$$



Eg. of a Nonlinear curve

in which

i) the functional part of the model is *not linear* with respect to the unknown parameters, β0,β1,..., and

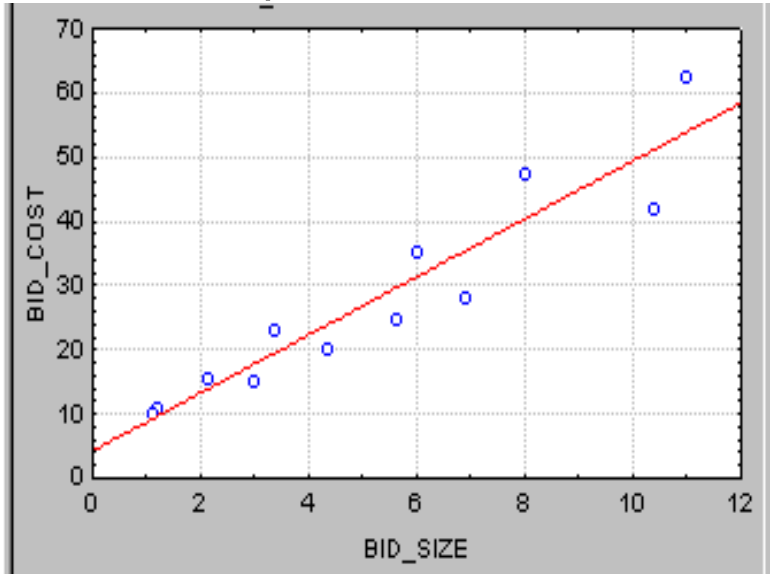
ii) the method of least squares is used to estimate the values of the unknown parameters

Weighted Least Squares Regression

- One of the common assumptions underlying most process modelling methods, is that each data point provides equally precise information about the deterministic part of the total process variation. Weighted least squares method assists us to overcome this assumption.

- Weighted Least Squares is an extension of simple regression. Non-negative constants (weights) are attached to data points. Weighted least squares reflects the behavior of the random errors in the model; and it can be used with functions that are either linear or nonlinear in the parameters.

For example, suppose a construction company wants to estimate the relationship between the size of a bid and the cost of preparing the bid. It makes sense to assume that the larger the project, the greater will be the residual variability of the cost about the estimated regression line.



Note that the size of the bid (on the horizontal axis) is scaled in terms of millions of dollars, and the preparation cost is in terms of thousand of dollars. As you can see, the variability of the residuals about the regression line tends to be larger for larger bids. It is apt to use weighted least squares in this case.