Statistical Process Simulation - A Linear Regression Approach

An Online Continuing Education Course for Engineers

Course Number: I-3003

Credit: 3 Hours / 3 PDH / 3 CPD

Statistical Process Simulation - A Linear Regression Approach

Manuel Calzada, P.E.

Course Objectives

- Understand the purpose of process simulation
- Understand different ways of simulating processes
- Basics of a statistical process simulation and procedure to solve complex problems
- Linear regression techniques for single and multiple variable cases
- Become familiar with Microsoft Excel statistical tools
- Develop linear regression formulas and analyze results
- Optimize results using Microsoft Excel tools

This course is designed to present different simulation processes and concentrate in a statistical simulation method by developing an example from beginning to an optimization phase. Further studies in this area can lead to potential and simple artificial intelligence techniques for a process or equipment. This will not be discussed in this course, but the student may find these techniques interesting as they may lead into different areas of simulation.

What is and why use simulation processes?
Simulation processes are techniques which are used to estimate, predict and forecast behaviors. These tools and techniques can be extremely helpful and can be used as tools to design, reduce cost and time, improve success and / or reduce risks.

Typical techniques to simulate processes
This course is designed to illustrate the use of different techniques that can be used to simulate processes. Some processes are well known and there are many accurate mathematical equations (some of which can be difficult to handle or very complex) that can help to predict the performance of a process with a fairly high degree of accuracy. We will call these **Theoretical**

Methods.
Unlike the theoretical methods, other processes are not as clear or don't have mathematical equations (or extremely complicated to work with) that can help predict the behavior of some variables in the process. In these cases we have to go to experiences or experimentation to develop such equations so we can simulate their behavior. This second method will be called

Statistical Method.
The author has written courses where mathematical simulations ("Finite Difference in Heat Transfer") are used. This course will concentrate only on the second method **Statistical Method**.

Theoretical Method
In this first category the processes can include some physical behaviors such as process simulators like flight simulators that most people are familiar with. In this case, even though the process can be rather complex, it is possible to develop some mathematical equations to understand the drag, lift, temperature dependencies, or any other variables that may affect the process.

In general, mathematical processes can be simulated rather accurately under many different circumstances once all the proper equations are taken into account. This type of simulation can sometimes be very difficult and complex, but very accurate and can provide great information on the output of a given process.

Statistical Method
In the second category we can include processes for which there are not mathematical equations found in the typical literature that would be consistent and can be applied to every case. Example of these types of processes would be to determine how temperature and heater location may affect the reaction of a certain type of yeast on food and its taste. Much of this knowledge is determined through experimentation. Proper design of experiments is critical and the results are based on statistical measurements. Most results from these experiments can be analyzed using statistical tools and the conclusions expressed in units of probability. These types of simulations have some advantages and disadvantages. The main advantage of this method is that we don't really need to "fully" understand the process itself. Some knowledge of the process is required to properly select the right variables and design of experiments. As a disadvantage, it is important to know that if we need to make small changes in some of the variables (beyond our test limits), we may need to repeat the process and analyze the data using another experiment. In other words, this method is not as flexible as the theoretical one, but extremely valuable when there is no mathematical information available.

When to use one method or another?
If you are an expert in a field and have a very good understanding and knowledge of mathematical equations that are used in the process, it probably would be advisable to use the **Theoretical method.** On the other hand, if you have a good experience in a field, BUT there are no specific mathematical equations for the variables that you need to consider, or you are not very familiar with those mathematical equations, then it will probably be recommended to use the **Statistical method.**

In this course we will cover the Statistical method. The course titled "**Heat Transfer Process Simulation Using Finite Difference and Controls**" would be a great example of a theoretical method that the student can always refer to it.

Statistical Method Examples
The following example is a simple case that will help the student to become familiar with some of the basic statistical tools that will be used in the more general COURSE EXAMPLE. The discussion below will cover single and multiple variable and linear and non-linear regressions.

Single Variable Linear Regression

One tool that we will select in this course is **Linear Regression**. Sometimes the data may suggest that would be better to use **Non-Linear Regression (quadratic, exponential, polynomial, etc.)**, but this is not included in this course. Most students would be familiar with simple and single linear regression. In a single variable linear regression, we have an independent variable and a dependent variable. The relationship between these two variables is a linear relationship and this can be simply represented in a plane.

An example of this type of relationship may be the height of children in relationship to their age. We can clearly see that as children get older their height also increases, not necessarily all of them at the same rate (some faster than others), but in general we will all be in agreement that as a child gets older, he or she will grow taller.

In the previous illustration, we know that there is no mathematical equation that will give us the height of a child if we know his/her age. This relationship may depend on what they eat, exercise, genes, race, etc.

This example (age versus height) will ignore any other influence on the child. This model will be good only for a population of interest. This may not be a very good model for a universe where many other populations are included. As a first step, this will help us to get closer to a general concept (age vs height). Therefore, we need to be careful about drawing conclusion out of the population sample.

If we were to take this example a step further, we can also consider the type of diet that the children eat. Now we have two independent variables to consider (age and diet) and one dependent variable (height). Graphically this is represented as a plane in a three dimensional coordinate space. It makes sense that this new model may be slightly better than the first, but still there are a lot of other variables that are not taken into account and may skew the results significantly. Now stretch your imagination and use another independent variable (genes). Now the model may start to get better accuracy as we move forward. However, the newer model would be harder to be represented graphically. We would all probably start to agree that the model starts getting better as a tool to predict the height of the children.

The description of this model is quite logical and we see that as we add more independent variables to the model, we will be fine-tuning the output and hopefully improve the results. The more independent variables we consider in the model, the more accuracy, but also more complication and longer experimentation is required. This is a trade off that the student would have to decide if he/she has the budget, time and resources to conduct the simulation model. The illustration on Fig. 1 shows a graph of the example above. In this case only one independent variable is used (Age). It seems clear that as the children get older, their height increases. However, notice that not all the children grow at the same rate.

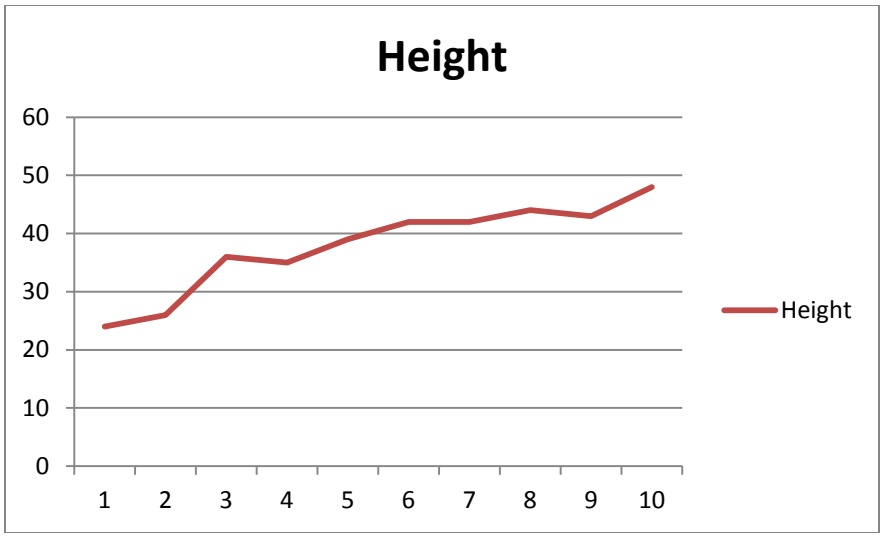

Fig. 1

The next graph on Fig. 2 illustrates the also the regression line (best-fit line for the data). Looking at this line we can make some estimates as to what the height of a child at age 7 may be. The height would be approximately 42" (based on this data alone).

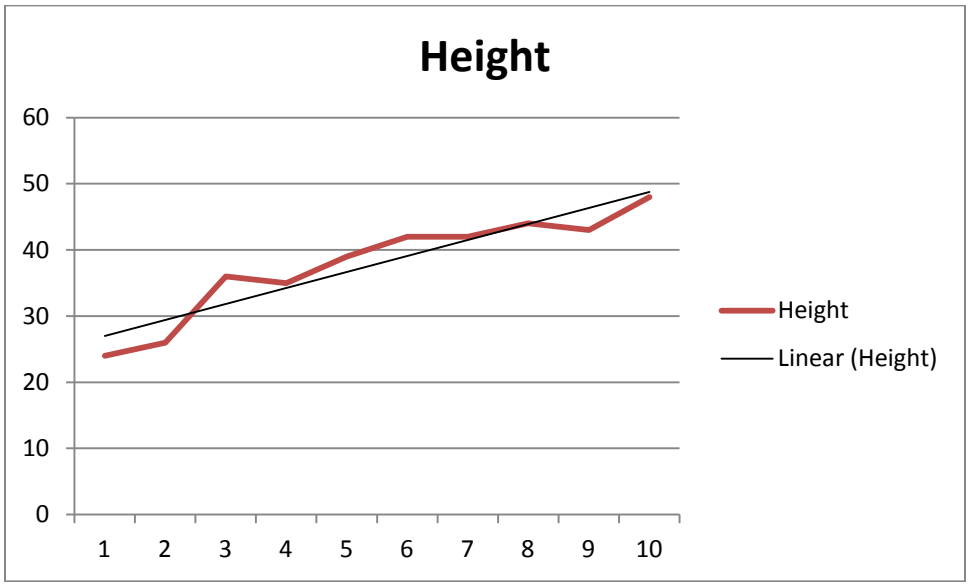

Fig. 2

In this graph we have used an equation that is linear with only one independent variable. The equation of this line can be expressed as:

$$Y = a + b\,X$$

Where

Y = is the dependent variable (height),
a = is the **intercept** of the line (inches)
b = is the **Coefficient** slope of the line (inches / year)
X = is the independent variable (years)

In this particular case the equation is:

y = 2.4182 (X) + 24.6
a = 24.6
b = 2.4182

The value of the co h is very good and indicates that the tw

Example:
Using the example at is 8 years old.

Solution:
Use the regression equ
Y = (2. er

Notice that this is only population sample

Multi-Variable Linear
The previous illustratio s we all know, most cases in real life are not e behavior of several variables and not just on

If we were to have an equ more independent variables, then the equation would be somewhat similar to what we have previously seen, but we need to include the other variables:

$$Y = a + b1\,X_1 + b2\,X_2 + \ldots\ldots + bn\,X_n$$

Where

Y	is the dependent variable
a	is the intercept (inches)
$b_1, \ldots b_n$	are the coefficients of each variable (slopes for each variable)
$X_1, ..X_n$	are the dependent variable

In the example that we will study in this course we will be using this type of equation since we will considering several variables in our model.