

1.1: Prelude to Linear Regression

Data mining is a phrase that has been popularly used to suggest the process of finding useful information from within a large collection of data. I like to think of data mining as encompassing a broad range of statistical techniques and tools that can be used to extract different types of information from your data. Which particular technique or tool to use depends on your specific goals.

One of the most fundamental of the broad range of data mining techniques that have been developed is regression modeling. Regression modeling is simply generating a mathematical model from measured data. This model is said to explain an output value given a new set of input values. Linear regression modeling is a specific form of regression modeling that assumes that the output can be explained using a linear combination of the input values.

A common goal for developing a regression model is to predict what the output value of a system should be for a new set of input values, given that you have a collection of data about similar systems. For example, as you gain experience driving a car, you begun to develop an intuitive sense of how long it might take you to drive somewhere if you know the type of car, the weather, an estimate of the traffic, the distance, the condition of the roads, and so on. What you really have done to make this estimate of driving time is constructed a multi-factor regression model in your mind. The inputs to your model are the type of car, the weather, etc. The output is how long it will take you to drive from one point to another. When you change any of the inputs, such as a sudden increase in traffic, you automatically re-estimate how long it will take you to reach the destination.

This type of model building and estimating is precisely what we are going to learn to do more formally in this tutorial. As a concrete example, we will use real performance data obtained from thousands of measurements of computer systems to develop a regression model using the R statistical software package. You will learn how to develop the model and how to evaluate how well it fits the data. You also will learn how to use it to predict the performance of other computer systems.

As you go through this tutorial, remember that what you are developing is just a model. It will hopefully be useful in understanding the system and in predicting future results. However, do not confuse a model with the real system. The real system will always produce the correct results, regardless of what the model may say the results should be.

This page titled [1.1: Prelude to Linear Regression](#) is shared under a [CC BY-NC 4.0](#) license and was authored, remixed, and/or curated by [David Lilja \(University of Minnesota Libraries Publishing\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.