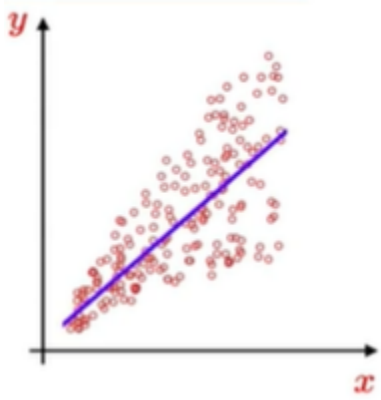
Introduction to Supervised Learning

Prediction through Regression

We're going to do prediction using regression methods. Prediction is the most important problem - it shows up all over in data science and machine learning. We want to predict quantities that are not known ahead of time.

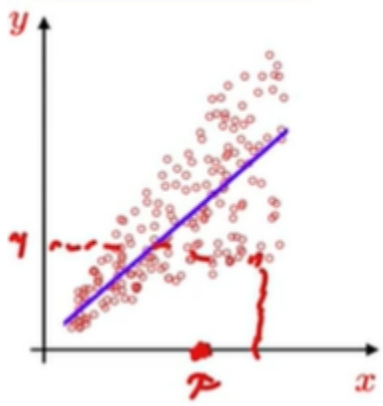
For example:
We have lots of individuals and each individual has a certain attribute or related variable x and has a variable y of interest, which we want to predict.



So the idea is that the typical person is going to show up - they will have a value of x and we will make a prediction of the corresponding value of y.

How can we do that?

We do that using lots of examples. We have a dataset or the red points in the below diagram.



So a typical person in the dataset has a value of x and the corresponding value of y. Our job now is to predict the y's for new people that are coming in. For the new people, we only have their x's and we need to predict y.

This file is meant for personal use by collinsbenda360@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

In this module, we will be dealing with the cases where y ranges over a continuous set of variables and will make those predictions using a line. We call these linear predictors.
For example: Whenever a new person comes in we look at the value that line suggests and we make our prediction.

Classification is a different type of prediction problem in which we're trying to predict the type of an individual. For example, is any individual sick or healthy and so on. Classification problems can be addressed using related methods but are not quite the same as the linear regression methods.

What is Supervised Learning?
It's called supervised learning because we're given several examples through which we can supervise the process of learning the mapping from attributes of a datapoint to its label. Each one of those points in our data set is an example. It's a fully labeled example in the sense that we know the value of x but also we know the corresponding value of y. So think of y as a label for an individual with characteristics x.

Let's understand the big picture. Suppose that you are a doctor and you're examining patients.

$\mathbf{X}$: symptoms, test results, etc.
Y: state of health

A typical patient comes with a data record and it's a vector x. That vector tells you the symptoms of that patient, fever, blood pressure, test results, and all that. These are the attributes of a typical individual. We're interested in the label of that patient y, which is the state of health of that patient. Whenever a patient comes in, we see their symptoms, and we want to say something about the state of their health. **How do we do that?** We get trained.
The training happens by having seen lots of patients in the past, and for a typical patient, we see their attributes, and we also know the state of their health. These are labeled past patients, and we have seen loads of such past patients. We've seen their data records and we know whether they weren't sick or not and then a new patient comes in. Now, the question is to make a prediction of the state of health of that patient based on the attributes of that particular patient.

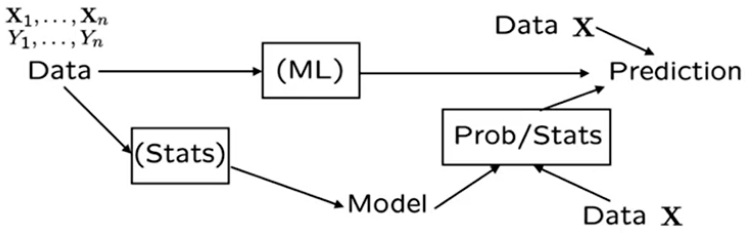
This file is meant for personal use by collinsbenda360@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

These predictions could be of two types,

1. Classification - It could be just to decide whether the patient is sick or not and that would be a binary prediction problem.
2. Regression - One example could be to try to estimate the life expectancy of a person in which case y takes a continuous range. It's a more quantitative type of prediction that we're trying to make.

How do we do regression?

In machine learning, one starts with a dataset. $(X_1, X_2, \ldots X_n)$, $(Y_1, Y_2, \ldots\ldots Y_n)$ are the labeled examples that we have on the basis of which we are to train a predictor. Then a machine learning algorithm takes over and does some preprocessing, gets trained, and builds that way of making predictions. Then at a later time, a new person comes in, and based on the training that we have done and the data, the attributes of the new person. We use these to make a prediction. So that's sort of a straight pipeline for coming up with predictions.



But sometimes we're not just interested in making a prediction. We're also interested in building some understanding of what is going on. That is besides just making predictions, we want to create a theory and understand the mechanism of how the x's cause the y's. In such cases, one wants to deploy statistical methods to create a full probabilistic model that relates the x to the y's. Once we have a model that ties x's to y's when a new x comes in, we can use that model to make a prediction for the y. So this is an indirect approach for addressing the same problem. It turns out that in the context of linear regression, whether you follow the straight path or whether you go the indirect way, the formulas are actually the same and you get essentially the same results. On the other hand, the interpretation is quite different. In one case, we're just looking to make predictions. In the other case, we're trying to model a situation and understand the underlying mechanism.

This file is meant for personal use by collinsbenda360@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

So models are interesting because sometimes we want to understand the mechanism. We're interested in the model. Models can also be interpreted just as a means toward an end. Maybe they facilitate the task of making predictions. Now, models are always incorrect. Any interesting real phenomenon will always have a true model which is too complicated to deal with. So models are always approximate. In some sense, models are always wrong. But there's a famous saying that **all models are wrong, but some of them are useful.** It's part of the art of doing statistics to figure out how to make good use of models, even when you know that your modeling assumptions are not quite right, or perhaps they're approximate.

Let's take a few minutes to understand some of the notational conventions,

1. Vectors by boldface symbols. For example,

$$\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

2. One-dimensional scalar real numbers using normal fonts.
3. When we deal with multiple data records, we use boldface symbols to indicate a typical record and the subscript will be telling you the position of the data record. For example, X_2 is the second person or the second data record.
4. For any vector, vectors will always be column vectors. We can take the transpose of that vector. The transpose vector is denoted by,

$$X^T = [X_1 X_2 X_3]$$

5. Let's say we have two vectors X and Y, and we take the transpose of the X, which makes the row vector and multiply it with a column vector Y then we will get the inner product from those two vectors, which is given by the sort of familiar form,

$$X^T Y = X_1 Y_1 + X_2 Y_2 + X_3 Y_3$$

6. n denotes the total number of data records.
7. 'star' is used to indicate true quantities. So if there's some unknown parameter theta, theta star would stand for the true one (θ^*).
8. 'hat ' indicates that it's an estimate. So hats always indicate estimates($\widehat{\theta}$).
9. Real numbers or constants indicated will be denoted by lowercase words.
10. Random variables will be indicated by uppercase words. For example, lowercase y which stands for the realized value of some random variable capital Y.

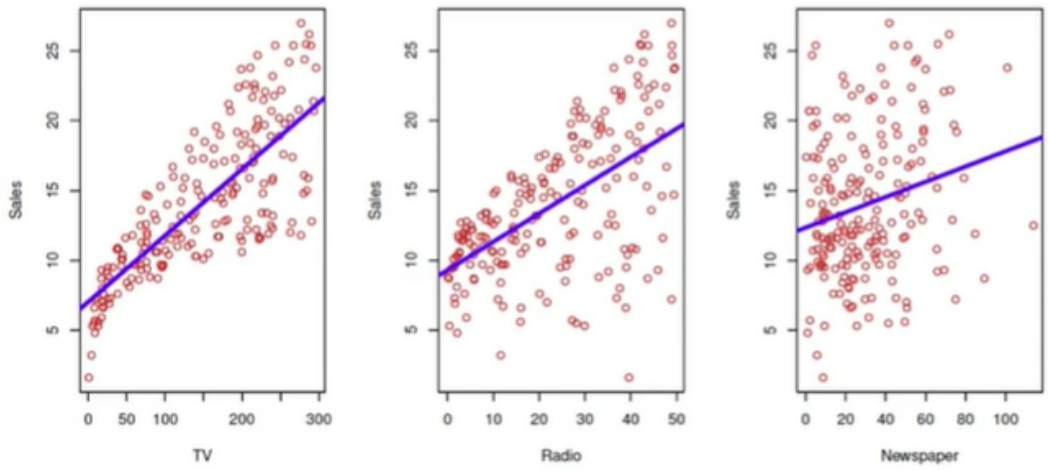
This file is meant for personal use by collinsbenda360@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Let's start with the fundamentals of linear regression,

We will understand this using an example. The example comes from advertising and sales. There is a company that operates in two hundred different markets and in each one of those markets it has advertisements. It has a budget and spends money advertising through different channels, such as TV, radio, and newspapers. In each one of those markets, they also observe what kind of sales they get. The data record has three variables X_1, X_2, and X_3 corresponding to TV, radio, and newspapers respectively and we're interested in the outcome of our target variable which is the sales (Y). So for each one of the markets, we have a data record of this type, which gives us the X and the Y. We have 200 such records. We have collected this data and these are our training data. Once we have them, we might pose some questions,

1. Is there a systematic relation between how much you advertise in the different channels and the sales?
2. Could it be that advertising in certain channels results in higher sales or maybe not? We want to investigate whether there is a relation of this kind and if there is a relation, can we quantify it?
3. If I tell you the x two budgets in a new market, can you use that to make a prediction of the sales? It can be useful to the marketing department so that they can decide how to adjust their advertising levels in the new markets or perhaps even adjust them in existing markets.

Before we do any math the useful first step is to always plot the data.



Now, in this case, the x's are three-dimensional and there's one more dimension for the y's. So a complete plot will be in 4 dimensions which, unfortunately, we cannot create. We can plot two dimensions at a time. So in this first plot, on the horizontal axis is how much we spent on TV. and the vertical axis is how much sales we have and similar plots for the other channels.

This file is meant for personal use by collinsbenda360@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Each one of the red points corresponds to one of the 200 markets that we're operating. We eyeball this data and we see that there is a kind of trend. For example, in the first plot when X1 is larger, Y also tends to be larger, and we can capture that relation by just putting it down as a line through the data that captures the strength. There are similar trends in the other two diagrams as well, although they're not as clear and strong. So we eyeball the data and the data tells us there seems to be some trend. Can we now capture those trends a little more mathematically? More precisely, can we quantify them? And can we draw some conclusions from them?

This is what linear regression will now try to do systematically. Instead of doing it by looking at one explanatory variable at a time, we will try to do that in higher dimensions by taking into account all of the axes simultaneously.

This file is meant for personal use by collinsbenda360@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.