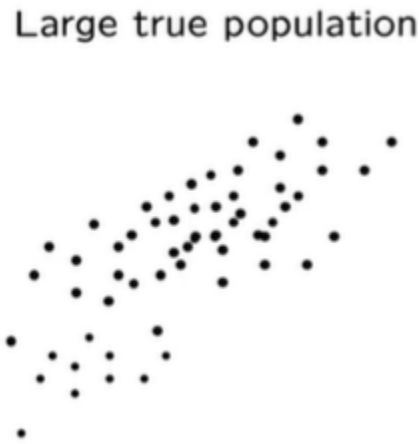
Interpretation and Justification

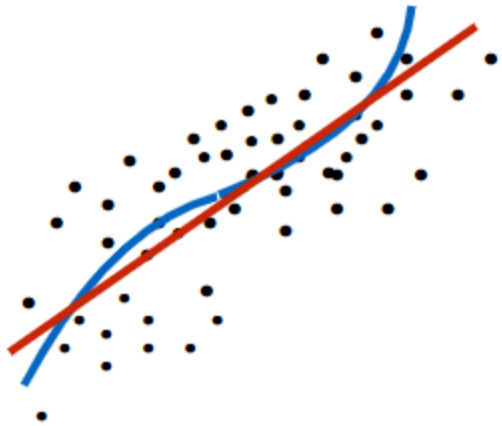
What is it that we're really doing in this linear regression?

There are a few different ways of interpreting it mathematically. So, let's walk through two different interpretations and compare them.

Here's one interpretation. There is a large true population of individuals.

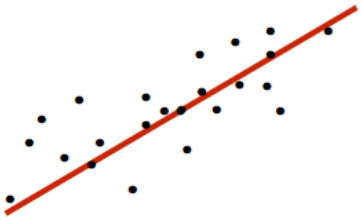


The horizontal axis is the X and the vertical axis is the Y. The phenomenon that we're observing is captured by some complex nonlinear relation. However, we want to create predictions using a linear predictor.



So, we're looking into something, like the red line, for making such predictions, and we're interested in constructing the best possible linear predictor, ideally for the overall population, but we do not know the overall population. The only thing that we have available is a finite data set. It's a finite sample that has been drawn from a large population and what linear

regression does is it tries to create a line by finding the best possible linear predictor based on this finite population.



- Finite sample: find best linear fit

$n \to \infty$: recover "population best"
(as long as samples are drawn representatively)

Because the finite population is not the same as the true population, the red line that we get by doing this linear fit is not going to be exactly the same as the best line for the true population. On the other hand, if we have enough samples, then it will be the case that the red line constructed on the basis of the small population or sample, is going to be approximately the same as the best line for the true population. We're trying to approximate the best linear predictor. The best red line for the true population is by using the limited data that we have. This is one interpretation. The only assumption here is that there is some underlying population from which we are drawing samples. There's a caveat here that for this to work, the samples need to be drawn in a way that's representative of the large population, that is we're sampling individuals at random.

Maximum likelihood Estimation (second interpretation):

In this interpretation, we consider the case where X is one-dimensional. We make a much stronger assumption that the world is described by a linear model. In the previous interpretation maybe the true relation is nonlinear and complicated. Here, we're making the stronger assumption that the true relation is actually linear.

Illustrate for $m = 1$

- assume structural model: $Y_i = \theta_0^* + \theta_1^* X_i + W_i$

Under that assumption it is linear, but there is also some idiosyncratic noise. W in the equation stands for the noise. It's a noise term. it's not a perfect linear relation. it's a linear relation plus some noise. This now looks more like a concrete probabilistic model, and within such a

probabilistic model, one way of estimating parameters is the maximum likelihood methodology.

What maximum likelihood does ?

It looks at the following expression,

$$\max_{\theta} \ \mathbb{P}(\mathbf{Y} \mid \mathbb{X}; \theta)$$

Given a data record X. All the data records together and for a particular choice of parameters, there's a certain probability of seeing the Y's that we actually saw. Then we try to find the parameter value theta for which this probability is as large as possible. So, we're asking which is the theta under which what we saw was most likely to be seen. There is one more assumption inorder to use the maximum likelihood methodology. The assumption is about the distribution of the noise terms, that is all those noise terms are independently drawn. They all have the same distribution, and their distribution is actually normal. They have zero mean, there is no systematic term and they have a certain variance sigma squared. Once we make that assumption, the probability of observing certain Y's is given by these formulas,

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(Y_i - \theta_0^* - \theta_1^* X_i)^2}{2\sigma^2}\right\}$$

This is the formula for the density of a normal random variable. So Y_i is normal with $\theta^*_0 + \theta^*_1 X_i$ mean under a particular model. The variance comes in this denominator. It's a messy formula, but this is the quantity that needs to be maximized with respect to theta in order to find the maximum likelihood parameters. Maximizing a product is the same as maximizing the logarithms of that product, and when we take logarithms, the logarithm of a product becomes a sum. So, differently said the product of the exponential is the exponential of the sum of the different terms. So it's enough to maximize just the sum of these expressions. There's a minus sign after we take it out, it boils down to minimization and it's a minimization of the sum of those terms here. The denominator is a concern, so we can get rid of it and so we end up with the sum of these quadratic expressions. So, the maximum likelihood methodology amounts to minimizing the sum of those quadratic expressions with respect to theta but this is just the same as the sum of squares minimization that we have in linear regression.

So, to summarize, the math basically tells us that the maximum likelihood methodology, when these assumptions are satisfied, gives us the usual linear least-squares minimization problem. And so it's an alternative interpretation.

Why is that interpretation useful?

Because there's a rich theory behind maximum likelihood estimation and it gives us theoretical guarantees, for example, it tells us that if we have a large enough dataset, we're going to recover exactly the true parameters. So, what's happening here is that maximum likelihood doesn't just give rise to a predictor, it actually learns the true values of an underlying model.

To summarize what we have said, to compare the two interpretations we have two different settings and two different ways of thinking about linear regression. One is that we have a population. We drew samples from that population and we learned the best linear predictor for that particular population. So, the task is just to build a predictor and we try to build the best one. An alternative way of thinking about what linear regression does is that we make a strong assumption that the world is linear. We know the structure of the world, and in that case, we try to learn the coefficients of the model that describes the world. So, this is more of a modeling task as opposed to just a pure predictive task. If we're interested in understanding the mechanics, what is really going on in a particular phenomenon? If we're interested in building a model, then it's the second interpretation that we're working with. But that second interpretation has stronger assumptions, that is we believe, or we pretend to believe that the world is approximately linear. The formulas are exactly the same, in the two interpretations. On the other hand, the interpretations and the underlying assumptions are significantly different, and it's important in any particular application to be cognizant to know which of the two are we trying to do. Are we just trying to build a good predictor, or are we also trying to learn something about the physics or the mechanics of the world?"