Let us now introduce the linear least mean squares formulation.

The setting is the usual one-- we have an unknown random variable and another random variable, which is our observation.

We're given enough information so that we can, for example, calculate the joint distribution of these two random variables.

What we would like to do in the least squares methodology is to come up with an estimator, such that the mean squared error of this estimator is as small as possible.

And we have seen the general solution to this problem.

If we consider arbitrary estimators, it turns out that the best possible estimator, the best possible function g, is this particular function of the observations.

Our estimator is a conditional expectation of Theta, given X.

Now, let us look at an example that we considered earlier.

Suppose that X and Theta have a joint PDF, which is uniform over this particular region.

We did consider this example and we found that the optimal estimator was a function that had this particular shape.

So this blue curve here corresponds to the function, which is the conditional expectation of Theta, given the value of the observation that we have obtained.

We notice that this function is nonlinear, but it is only mildly nonlinear.

The fact that it is nonlinear is a little bit of a nuisance.

It makes it somewhat of a complicated function.

Wouldn't it be nicer if our estimator had turned out to be a linear function of the data, such as this one?

It would have been nicer, but, unfortunately, that's not the case.

By what if we impose it as a constraint, that we will only look at estimators that are linear functions of the data.

What does that mean?

Mathematically speaking, it means that we will only consider estimators that depend linearly on the data X.

Now, a and b here are some parameters that are for us to choose.

If I choose a and b differently, I'm going to get a different red curve here.

Which one is the best red curve?

Well, we need a criterion.

But let us stick to our mean squared error criterion.

And in that case, we're led to the following formulation.

We want to find choices for a and b.

That is, we want to choose a particular red line, so that the resulting estimation error, the resulting mean squared estimation error, is as small as possible.

So what we have here is a random variable.

Here is the value that's going to be given to us by our estimator.

And we look at the associated error, square it, and take the expectation.

So this is the linear least mean squares formulation.

We're looking for an estimator, which is a linear function of the data.

And we want to choose the best possible linear function.

How does it compare to the earlier problem of picking the best estimator?

Here we were considering an arbitrary function g and we were trying to find the best possible function of the data, which would be our estimator.

So this was really an optimization over all possible functions.

Here we only have an optimization with respect to two numbers.

So at least mathematically, this should be a simpler problem.

And we will see that it has a simple solution.

Before going on to the solution, however, let me make one comment that in some cases, the linear least squares estimation problem is relatively easy to solve.

And these are the cases where the conditional expectation turns out to be linear in the data.

This is the best possible estimator.

If it happens to be linear, it's at least as good as any other linear estimator, so it's also going to be the optimal linear estimator.

That is, if the optimal solution turns out to be already linear, by imposing the extra constraint of sticking to linear estimators is not going to make any difference.

But for the general case, in general, this is not going to be the case.

The conditional expectation may well turn out to be a nonlinear function of the data, as in this example.

And in those cases, the linear least mean squares estimator is going to turn out to be different.