We have presented the complete solution to the liner least mean squares estimation problem, when we want to estimate a certain unknown random variable on the basis of a different random variable X that we get to observe.

But what if we have multiple observations?

What would be the analogous formulation of the problem?

Here's the idea.

Once more, we restrict ourselves to estimators that are linear functions of the data, linear functions of the observations that we have.

And then we pose the problem of finding the best choices of these coefficients a1 up to a n and b.

What does it mean to find the best choices?

It means that if we fix certain choices, we obtain an estimator, we look at the difference between the estimator and the quantity we're trying to estimate, take the square, and then take the expectation.

So once more, we're looking at the mean squared error of our estimator and we try to make it as small as possible.

So this is a well-defined optimization problem.

We have a quantity, which is a function of certain parameters.

And we wish to find the choices for those parameters, or those coefficients, that will make this quantity as small as possible.

One first comment is similar to the case where we had a single measurement [and] is the following.

If it turns out that the conditional expectation of Theta given all of the data that we have is linear in X, if it is of this form, then what happens?

We know that this is the best possible estimator.

If it is also linear, then it is the best estimator within the class of linear estimators as well and, therefore, the linear least mean squares estimator is the same as the general least mean squares estimator.

So if for some problems it turns out that this is linear, then we automatically also have the optimal linear estimator.

And this is going to be the case, once more, for certain normal problems with a linear structure of the type that we

have studied earlier.

Now, let us look into what it takes to carry out this optimization.

If we had a single observation, then we have seen a closed form formula, a fairly simple formula, that tells us what the coefficients should be.

For the more general case, formulas would not be as simple, but we can make the following observations.

If you take this expression and expand it, it's going to have a bunch of terms.

For example, it's going to have a term of the form a1 squared times the expected value of X1 squared.

It's going to have a term such as twice a1, a2 times the expected value of X1, X2.

And then there's going to be many more terms to some of them will also involve products of Theta with this.

So we might see that we have a term of the form a1 expected value of X1 Theta.

And then, there's going to be many, many more terms.

What's the important thing to notice?

That this expression as a function of the coefficient involves terms either of this kind or of this kind, or of that kind, first-order or second-order terms.

To minimize this expression, we're going to take the derivative of this and set it equal to 0.

When you take the derivative of a function that involves only quadratic and linear terms, you get something that's linear in the coefficients.

The conclusion out of all this discussion is that when you actually go and carry out this minimization by setting derivatives to zero, what you will end up doing is solving a system of linear equations in the coefficients that you're trying to determine.

And why is this interesting?

Well, it is because if you actually want to carry out this minimization, all you need to do is to solve a linear system, which is easily done on a computer.

The next observation is that this expression only involves expectations of various terms that are second order in

the random variables involved.

So it involves the expected value of X1 squared, it involves this term, which has something to do with the covariance of X1 and X2.

This term that has something to do with the covariance of X1 with Theta.

But these are the only terms out of the distribution of the X's and of Theta that will matter.

So similar to the case where we had a single observation, in order to solve this problem, we do not need to know the complete distribution of the X's and of Theta.

It is enough to know all of the means, variances, and covariances of the random variables that are involved.

And once more, this makes this approach to estimation a practical one, because we do not need to model in complete detail the distribution of the different random variables.

Finally, if we do not have just one unknown random variable, but we have multiple random variables that we want to estimate, what should we do?

Well, this is pretty simple.

You just apply this estimation methodology to each one of the unknown random variables separately.

To conclude, this linear estimation methodology applies also to the case where you have multiple observations.

You need to solve a certain computational problem in order to find the structure of the best linear estimator, but it is not a very difficult computational problem, because all that it involves is to minimize a quadratic function of the coefficients that you are trying to determine.

And this leads us to having to solve a system of linear equations.

For all these reasons, linear estimation, or estimation using linear estimators, is quite practical.