## MITx Video | 6.041x | Spring 2014 | Video transcript | L16-1: LMS estimation without any observations

In this segment, we introduce the subject of least mean squares estimation. But as a warm-up, we will start with a very simple special case. Suppose that we have some random variable that we wish to estimate and that we have the probability distribution of this random variable-- a probability mass function if it's discrete or a probability density function if it's continuous.

As a concrete instance, suppose that our random variable is uniformly distributed over a certain interval. We would like to estimate this random variable. We're interested in a point estimate. However, we look at the very special case where there are no observations available. All that we have is this probability distribution.

How do we estimate this random variable? Well, we can use the rules that we have already developed, for example, the maximum a posteriori probability rule, what would it do? In this case, since there are no observations, the posterior distribution of Theta is the same as the prior. There are no observations that could change the prior. So we need to find a point at which this distribution is highest.

Well, because this distribution is flat, the MAP rule does not give us a unique answer. Any value of theta inside the interval from 4 to 10 would be an acceptable answer. So any particular estimate in this interval would be fine.

So this rule is not particularly helpful. How about a different estimator? We have seen the conditional expectation estimator. Once more, in our case, since we do not have any observations, the conditional expectation is the same as the expectation. And for this particular example, it would give us the midpoint of the distribution, namely an estimate equal to 7.

Now, this rule was inconclusive. This rule gave us a number. How can we choose and decide that one of these is the right estimate?

We can do that if we introduce a specific performance criterion. What is it that we wish from our estimators? And the particular criterion, the one that will we be focusing on, is the mean squared error.

If you come up with a certain estimate, you look at how far is your estimate from the true value that you're trying to estimate, take the square of that, and average it. And this leads us to a formulation whereby we will try to find an estimate theta hat that minimizes this mean squared error over all possible estimates. Let us now look at this formulation and see how we can solve it.

This is a function of a single variable theta hat. And we can try to minimize it using conventional methods. To carry out this minimization, let us first expand this expectation into a sum of terms. We have the expected value of the square of the random variable, then a cross term minus 2 the expected value of Theta times theta hat.

However, theta hat is a number that we're trying to choose. It's not random. Therefore, we can pull it outside the expectation. And similarly, the last term, the expected value of theta hat squared is just theta hat squared itself.

This is what we want to minimize. How do we minimize it? We take the derivative with respect to theta hat and set it to 0. And this gives us minus 2 the expected value of Theta plus twice theta hat equal to 0.

And when we solve this, we find is that theta hat, the optimal estimate is equal to the expected value of Theta. So this is the answer to our optimization problem. The optimal estimate, according to the least squares criterion, is the expected value of the random variable.

Now, it's interesting to derive this result, also, in a second way. Since it is a quite important and fundamental result, let us see whether there is a different way of establishing it that stays closer to the probabilistic world rather than the calculus world. So let us look at this criterion that we have here.

The expected value of the square of a random variable is always equal to the variance of that random variable plus the expected value of that random variable squared. This is what we're trying to minimize. Now, theta hat is a constant.

The variance of a random variable plus or minus a constant is the same as the variance of that random variable. And there is nothing that we can do about this term. So what we're trying to do is, essentially, just try to minimize this term with respect to theta hat.

Now, here we have the square of something. The way to minimize this is to try to make this quantity as small as possible, make it 0 if we can. Well, we can make it 0 if we set theta hat equal to the expected value of Theta.

So this term here is minimized. When theta hat is equal to the expected value of Theta, it is minimized because this is a choice that will make this term equal to 0. So this was a second derivation of why this is the optimal estimate of Theta. Once we adopt this particular estimate, the mean squared error is going to be equal to this, because this is our theta hat. And, of course, we recognize that this is the variance.

So the variance of Theta is the least possible value of the mean squared error that we can obtain using any particular estimate. And this is our final conclusion. And in the next segment, we will exploit the conclusions that we found here and apply them to a more general situation.