

Now that we have found the solution to the linear least mean squares estimation problem, it is time to offer a few comments, make some observations, and provide some insights.

A first important observation is the following.

In order to implement this estimator, you do not really need to know everything about the distribution of X and Θ .

The only thing that you need to know is the mean of the two random variables that are involved, the variance of X , and the covariance of Θ with X .

So it's only a few pieces of information that we need, and that means that we do not need to be so careful about modeling in a particular problem, as long as we know what the means, variances, and covariances are.

This is a very desirable property, because it tells us that this could be simpler to implement in the real world.

Now let us start looking at the form of the solution, and let's try to give some interpretation.

Suppose that the correlation coefficient is positive.

Then what does this estimator do?

It starts with a baseline estimate, which is the expected value of Θ .

And then provides us with a correction term that's based on the data.

In particular, if it happens that we see an observation that's larger than expected, in that case, our estimate is going to be above the expected value of Θ .

So we started with our baseline estimate, but if we get a big observation, then our estimate will also be large.

And conversely, if X happened to be on the lower side, below the expected value, then our estimate would also be below the expected value of Θ .

Of course, there's an analogous story.

If ρ was negative, then a same argument would apply, except that it would work in the opposite direction.

When ρ is negative, if we see a large X , then we will come up with a low estimate for Θ .

Let us look at another special case now.

Suppose that the rho is equal to 0, X and Theta are uncorrelated.

Then this last term here disappears, and in this case, our estimate is going to be just the expected value of Theta.

In other words, our estimate doesn't make any use of the data.

Essentially what's happening is that a linear estimator exploits the correlation between the two random variables to come up with an estimate.

But if the two random variables are uncorrelated, then [there is] nothing that it can do, and it does not give us anything useful.

It just reports back the expected value of Theta.

Let us now look at the mean square error that's obtained when we implement this linear estimator.

Let us write down what this expression is.

And to keep things simple, let us assume that we have 0 means and 0 variances.

So just for the purposes of this derivation and to simplify the algebra, we will work with the 0 mean case.

So in this case, what we have-- let me first write Theta, and then put here the estimator.

The estimator is rho times sigma Theta, sigma X times X.

Basically, I took this formula, but I put 0s for the expected values.

Now let us expand this quadratic.

We obtain the expected value of Theta squared.

That's the variance of Theta, since we assume 0 means.

And the variance is the square of the standard deviation.

Then we have a cross term.

Is going to be twice the expectation of the product of Theta with this.

Now we can take out this constant outside of the expectation, so it's 2 rho sigma Theta, sigma X. And then we are going to have the expected value of Theta times X.

Because we assume 0 means, the expected value of Theta times X is the covariance of Theta and X.

And the covariance is rho sigma Theta, sigma X.

And finally, we have a last term, which is rho squared, sigma Theta squared, divided by sigma X squared times X squared.

We have an expected value, so that's the expected value of X squared, which is just the variance of X or sigma X squared.

OK, this looks like a big mess, but in fact, we get nice cancellations.

This term cancels with that term.

This term cancels with that term.

We notice that we have sigma Theta squared in each one of those terms, so we can factor them out.

And then here we have minus 2 rho squared plus rho squared.

When we subtract those two terms we're left just with a minus rho squared term.

So after you do that algebra carefully, you find that the answer takes a very simple form.

It's 1 minus rho squared times the variance of Theta.

I should add here that this formula remains valid, even without the assumption that X and Theta have 0 means.

What's the interpretation of this?

The variance of Theta describes the initial uncertainty we have about Theta.

But after we carry out the estimation, the uncertainty gets reduced, and it gets reduced by a certain factor.

What is this factor?

If rho is equal to 0, then this coefficient is 1 and we do not have any variance reduction.

After all, when rho is equal to 0, this estimator is not very useful.

It doesn't help us estimate Theta better.

So the expected value of the squared error is the same as the variance of Theta.

On the other hand, when ρ is large, then this term here becomes small, and this means that the mean squared error is small.

In fact, there is an extreme case that's interesting, namely if ρ is equal to 1 in absolute value.

So if the random variables are perfectly correlated, then this term here becomes 0, and the mean squared estimation error is 0.

Which essentially means that our estimate is going to be equal to the unknown value.

So the special case of a unit correlation, in absolute value, corresponds to a case where we can perfectly estimate Θ , using a linear estimator.

So to summarize, the correlation coefficient plays a crucial role in linear least squares estimation.

It determines the form of the estimator, and also, it determines how much the uncertainty in Θ will be reduced through the process of estimation.