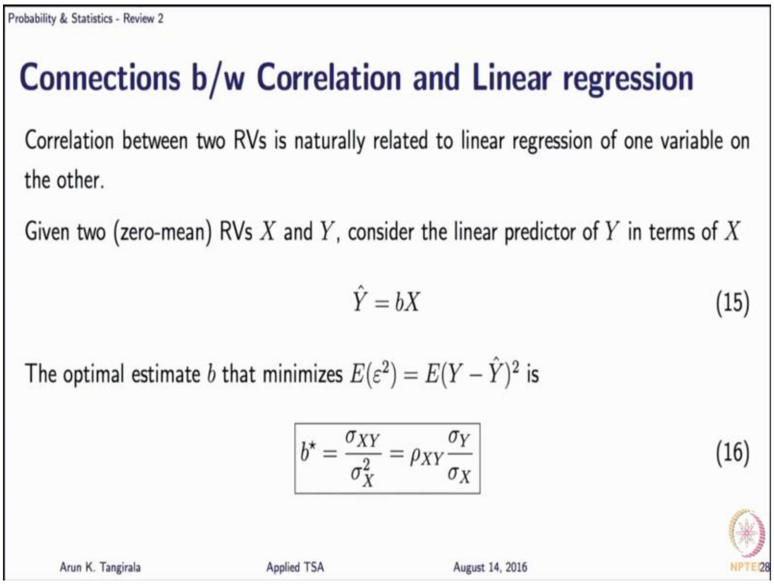
Applied Time-Series Analysis
Prof. Arun K. Tangirala
Department of Chemical Engineering
Indian Institute of Technology, Madras

Lecture - 16
Lecture 07C - Probability and Statistics Review (Part 2)-6

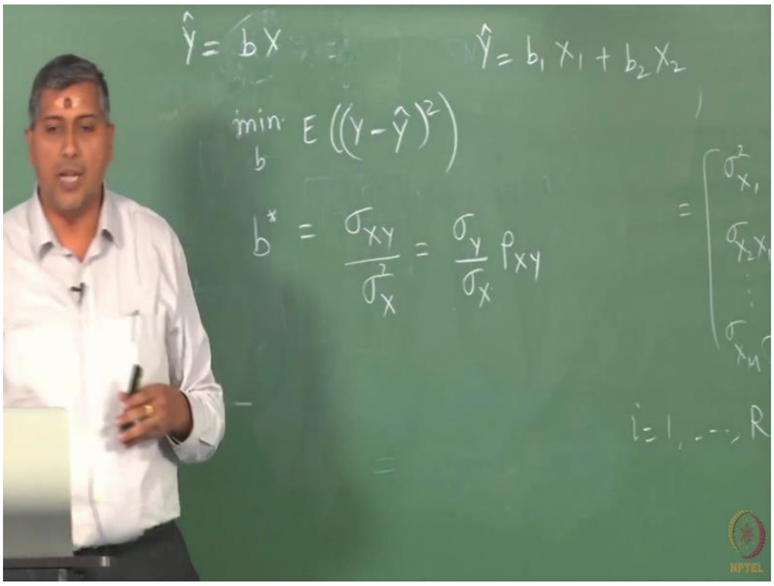
Let us now look at this aspect of connections between correlation linear regressions.

(Refer Slide Time: 00:17)



A bit more closely and then we will of course continue this discussion even tomorrow. Remember I said as far as linear models are concerned it is sufficient to know the first and second order moments. And here is an example, very simple example; in fact just now we talked about it I said if you are given data for y and x how do you estimate alpha. In fact, now that is what you see on the slide, which then establishes the connection between the correlation and linear regression.

(Refer Slide Time: 01:06)



So, it turns out that when I want to fit a linear model; so let us say I assume perfect linear relationship y hat equals b x, Remember now I have not written y equals b x, I am saying that I am going to predict y using a linear function of x; this is the linear regression problem. When I do this then let us say I obtain estimate of b by minimizing expectation of y minus y hat square. We have done this before, but in a slightly different context.

Now when you try to do this, that is when you try to estimate b by minimizing the some square error or mean square error; that not some square error, the resulting estimate is what you see on the slide but I will also write this on the board; b star you can actually do this in fact it is a small homework you should do that: b star is sigma x y by sigma square x. In fact, I could also have b x plus a, but I did not just for the sake of convenience include that.

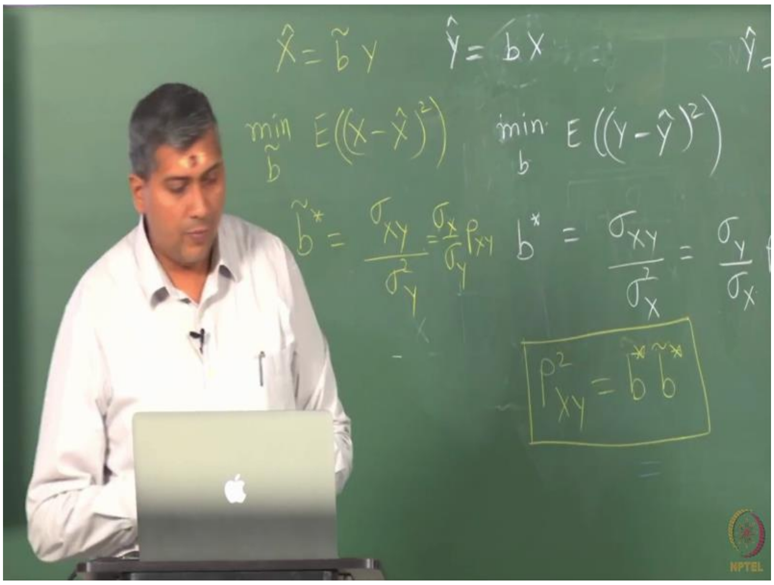
So, you can see straight away that as far as estimation of a linear model is concerned or as far as linear regression is concerned I only need to know sigma x y and sigma square x. What are they? They are the sigma x y is the second moment of the joint pdf and sigma square x is the second moment of the marginal pdf. That is all I need, I do not need to know anything more as far as the linear regression is concerned, but if you want to fit a non-linear model then you will need the knowledge of higher order moments.

This is typical I mean all they have just taken the bivariate case you can extend this to even the trivariate or multivariate case. In other words if I were to predict y using x 1 and x 2; for example, if I were to write y hat as b 1 x 1 plus b 2 x 2, then even when you sit down to optimally estimate b 1 and b 2 you will see that only covariances and the variances will appear apart from that you will not run into any higher order moments.

So of course, now you can rewrite this as sigma y over sigma x times rho x y and there is a reason why we write that. You can see straight away first of all that when covariance is 0, what is optimal estimate of b 0 sigma square x cannot be 0; so when correlation is 0 there is no hope for any linear model you cannot come to the conclusion that y and x are not linearly related and so on; that is got in practice in practice because you will be working with estimates, but theoretically yes when correlation is 0; that means there is no linear relationship between y and x whatsoever your linear model will not do be of any help to you.

Now, we will close this today's lecture with a very interesting observation which is; suppose I interchange the roles of y and x that is I would like to predict x using y.

(Refer Slide Time: 04:45)



I say x hat let us say is being predicted with using y, I can do that. But now let us let me denote again I am using a linear predictor, let me denote the coefficient as b tilde; and again going through the same thing that is fine b tilde such that the mean square a prediction error is minimized. So, you can think of this as some kind of a forward model if you like it and this as a reverse or you can think of this as a forward, this as a reverse it does not matter because at this moment we have not said whether x is the cause and y is the effect I do not really worry about that, I am given one I want to predict the other.

So, here I am given x I am trying to predict y, here I am given y I am trying to predict x; what would be b tilde star?

Student: (Refer Time: 05:49).

Sigma.

Student: (Refer Time: 05:55).

I hear the numerator synchrony or symphony, but denominator.

Student: Sigma squared y.

Sigma squared y; so the denominators are also in single. So, sigma I mean even think of sigma y x, but it does not matter the order of subscripts does not matter here sigma square y; which again I can rewrite as sigma x by sigma y times rho x y.

Now, you should straightaway observe that the squared correlation is b star times b tilde star. What is the importance of the result? I mean it is it is adding colour to the board and so on, but what is the importance of this result do you see any importance of the result or it is at another relationship for you. Can you infer anything nice from this result useful?

Student: (Refer Time: 07:15).

Sorry.

Student: (Refer Time: 07:18).

What is less than one?

Student: (Refer Time: 07:20).

Why, rho xy is always less than or equal to 1 in magnitude but why you say from this relation you infer that, because you think the product have to be some kind of less than 1; interesting but not something that we are looking for. Anything else, Earlier we said something about correlation that it does not know any direction, it has effects of both directions. You can see this that the squared correlation captures the effects in both directions; that means if I give you squared correlation or correlation alone you will not be able to figure out how much is b till and how much is b tilde.

That is how much is a contribution in one direction versus the other, you will not be able to say that it is a factor. So, this result actually clearly tells you that correlation is I mean why also reinforces the fact that correlation is symmetric apart from that it is says that it captures the effects in both directions. Of course, it is another way of computing your correlation, you can fit a linear regression of a linear model in both directions, it gives you a parametric way of estimating correlation. Suppose you want to estimate correlation, you can fit a linear model between x and y with x as a regressor and y as a predictor and vice versa and compute the coefficients the product of that is the squared correlation.

Now, another use of this result here appears later on in fact tomorrow when we talk of partial correlations. Today we have talked about correlation extensively about covariance and correlation and established that correlation is a measure of linear dependence; that there is a strong connection between linear regression and correlation; that it only suffices to know the first and second order moments when it comes to fitting linear models.

And so many other useful things about covariance and correlation, but there are some limitations of correlations apart from the fact that it does not have directionality embedded in it, does not know the physics there is yet another limitation of just working with correlation which is that of confounding. We will take up this concept tomorrow in detail, that is when I find x and y correlated; sorry it does not necessarily mean that they are directly related. And there are some very classic examples doing that and one of the; sorry, examples is let us say I look at the number of fire engines being sent out from a fire station you know and the damages actually caused by fire in a in the nearby locality. If I had this data number of trucks going out of a fire station and the damage caused by the fire, do you think they would be correlated or not, what do you think? Yes or no?

Student: Yes.

No, then why would you send out trucks?

Student: (Refer Time: 11:05).

What do you mean by that? Fire is a fireman.

Student: (Refer Time: 11:12).

Does not matter, but the more trucks that you send is it a inverse correlation or direct; in the sense positive sign or negative sign the correlation.

Student: Positive.

Positive, so the more trucks that you send more damage you will see.

Student: We know damage (Refer Time: 11:37).

Now, first let us ask what is the sign of the correlation; first of all do we agree there is a correlation between the number of trucks sent out from a fire station and the damage cause, assuming that trucks are going there they are not going some Pokémon or anything like that. But ok assuming they are going to this locality do we do see a correlation. Now, do we expect to see a negative valued correlation or a positive valued correlation?

Student: Negative.

Negative; your answer is in the negative. It is right: so the more tracks you sent the lesser the damage should be. But tomorrow we will take up some data where we will see if this is true. We will see if this is true, and ask why is it that we do not see what we expect to see. We will meet tomorrow.