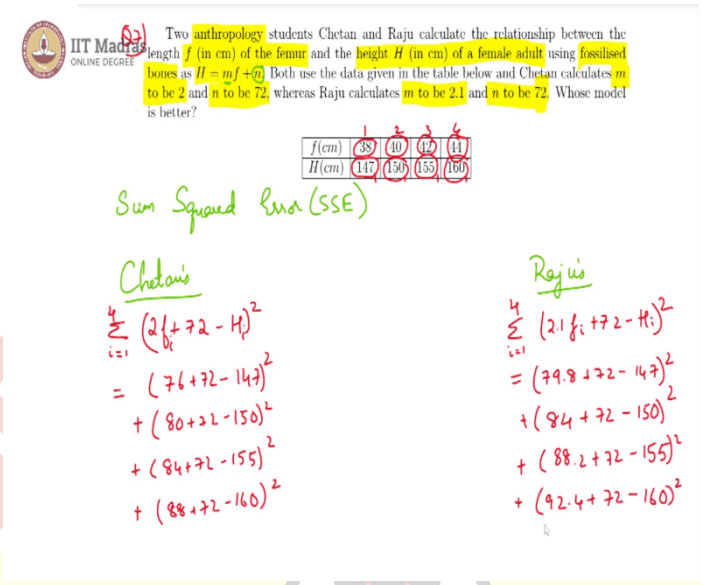
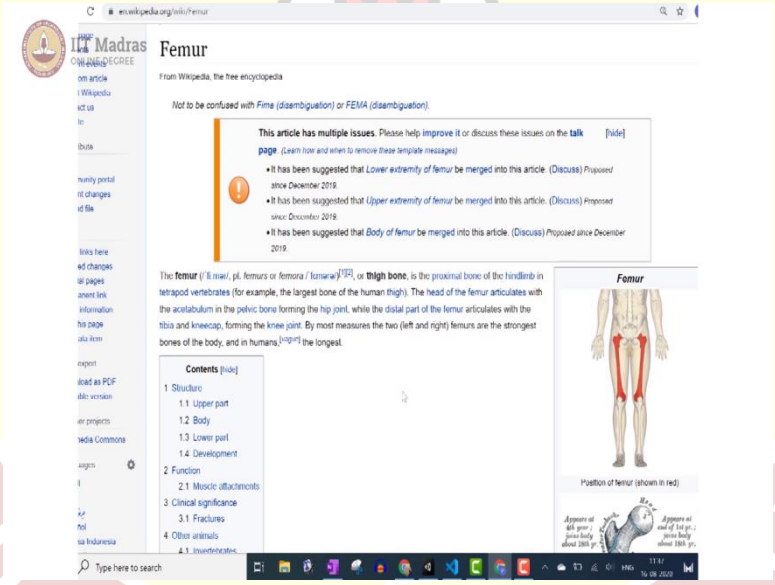
IIT Madras

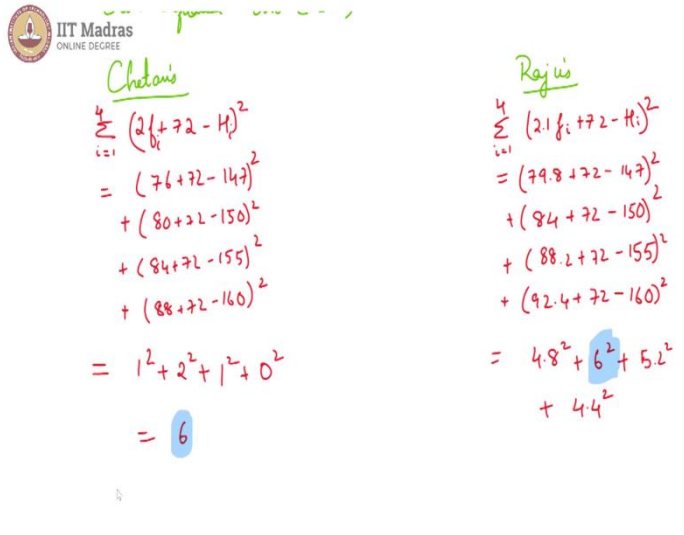
ONLINE DEGREE

Mathematics for Data Science 1
Week 03 – Tutorial 07

(Refer Slide Time: 0:14)







In our seventh question, we have this interesting thing, where there are two anthropology students and they are calculating the relationship between the length f of the femur and the height H of a female adult using fossilised bones. So, what is exactly happening here? What is femur?

From Wikipedia, we can see that the femur is the thigh bone, which is this particular bone. So, what is happening is, in our question, there are fossilised bones and these anthropology students, anthropologists try to study the nature of humans and their societies as they were evolving.

So, here we have fossilised bones and suppose we have the femur of what we know to be a female adult, then we are estimating the height of that female adult from the length of the femur bone, from the thigh bone. So, it is given that this relationship is linear, we have $H = mf + n$. Both use the data given below, so this is the data that is available. We have the femur length and the height of the adult female.

So, from this we are trying to develop this model and Chetan has found $m = 2, n = 72$, whereas Raju has calculated m to be 2.1 and n to be 72. So, both of them agree on n, this parameter is already fixed. It is the m that we are trying to see, whose m is better. So, in terms of linear equation, m is basically the slope of the line. So, how do we do this? We want to use the concept of Sum Squared Error, which we call SSE.

So, in both cases, we are going to look at what is being predicted in terms of height and what is the actual data. So, let s look at case one, let us look at Chetan's case here and Raju's case

here. In terms of Chetan's case, we would have the $H = mf + n$, where m is 2, so we have $\sum_{i=0}^{4}(2f_i + 72 - H_i)^2$ and we sum it over how many items 1, 2, 3, 4.

So, let us call this f_i, H_i and i goes from 1 to 4 and in case of Raju's measurements, this error would be again, i goes from 1 to 4 and we have $\sum_{i=1}^{4}(2.1f_i + 72 - H_i)^2$. So, I think we just need to do the calculations now. So, let us look at this here, so this is case 1. So, this is case 2, this is case 3, this is case 4, f_1 is 38 and H_1 is 147.

So, when we put in 38 here we get 2 times 38 is 76 + 72 - 147 the whole square and then in case 2, we have 40 and 150 as f_2 and H_2. So, we will get 80 + 72 - 150 the whole square and then we have, f_3 is 42 and H_3 is 155. We have (84 + 72 – 155) ^{2} and lastly, we have f_4 is 44 and H_4 is 160. So, we have (88 + 72 – 160)^2.

$$\sum_{i=0}^{4}(2f_i + 72 - H_i)^2 = (76 + 72 - 147)^2 + (80 + 72 - 147)^2 + (88 + 72 - 160)^2$$

So, this is the total sum squared error for Chetan. Whereas in case of Raju, we would get 2.1 times the same thing. So, 2.1 times 38 is 79.8 + 72 - 147 the whole square and in case 2 we get 84 + 72 - 150 the whole square + in 3 we get 88.2 + 72 - 155 the whole square and lastly, in case 4, we have 92.4 + 72 - 160 the whole square. We calculate these values then we get 1^2 + 2 ^{2} + 1 ^{2} + 0 ^{2}.

$$\sum_{i=1}^{4}(2.1f_i + 72 - H_i)^2 = (79.8 + 72 - 147)^2 + (84 + 72 - 150)^2 + (92.4 + 72 - 160)^2$$

So, this is 6. So, sum square error for Chetan is 6. Whereas in Raju's case we would have 4.8 ^{2} + 6 ^{2} + 5.2 ^{2} + 4.4 ^{2}. Now clearly, in this error, there is a 6 ^{2} which has to be greater than 6, which means Raju's error is much more than Chetan's error. Therefore, Chetan's line fit is better.