Course material

Course:

ENG606 / PHYS 442

Video:

DOE_lesson5_part1_Anova

Concepts (extracted from automatically generated subtitles):

R square. Sum of squares. Analysis of variance of a model. Anova analysis of variance. Part of the model. Second degree. Vector of results. Anova table. Values of the coefficient. Variance of the model. Length of y hat. Invention of mr. fisher. Mean square of the model. Only way. P coefficients.

to video sequence search
(within ENG606 / PHYS 442.)

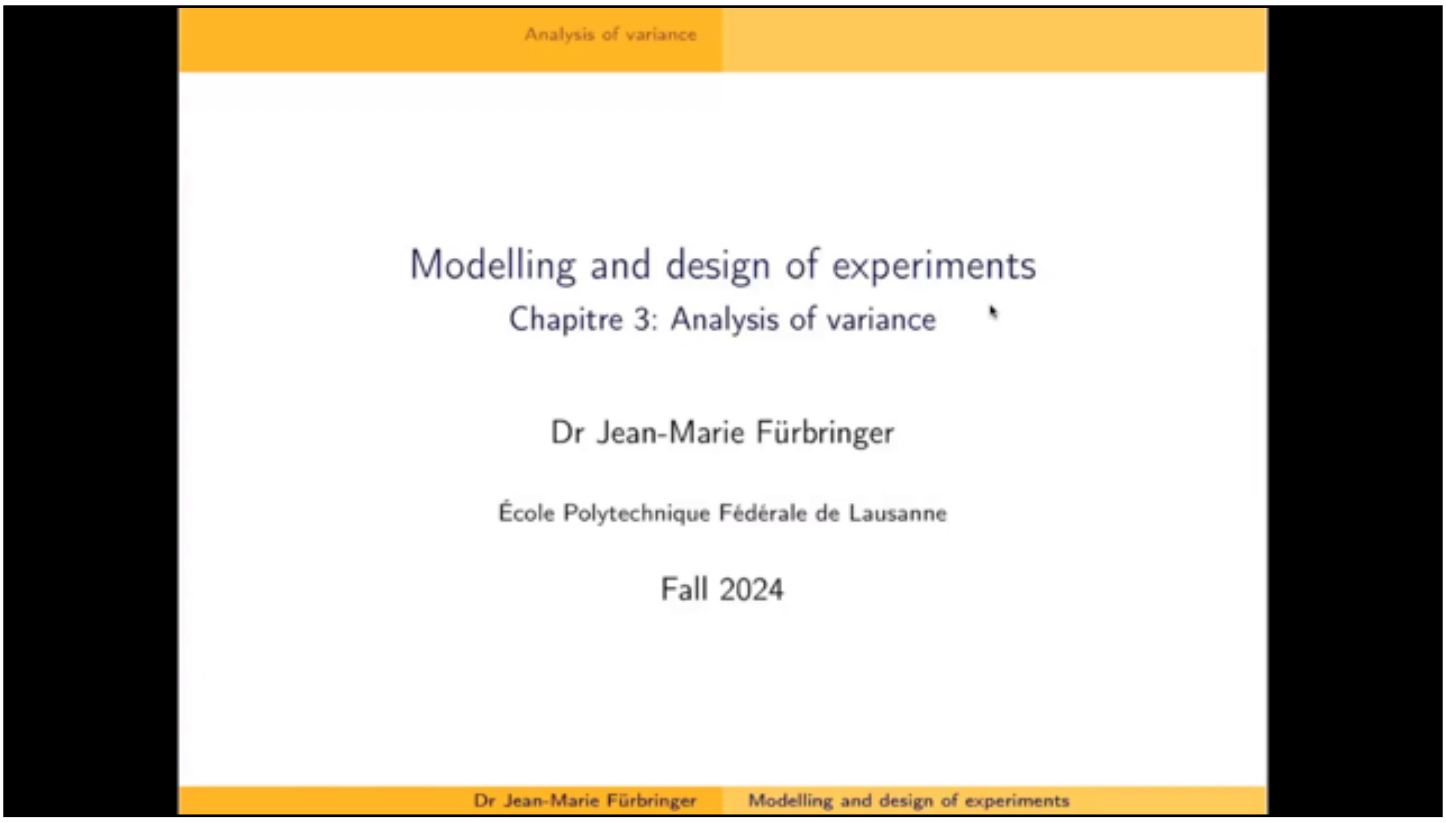
to video

...

notes

summary

0m 0s

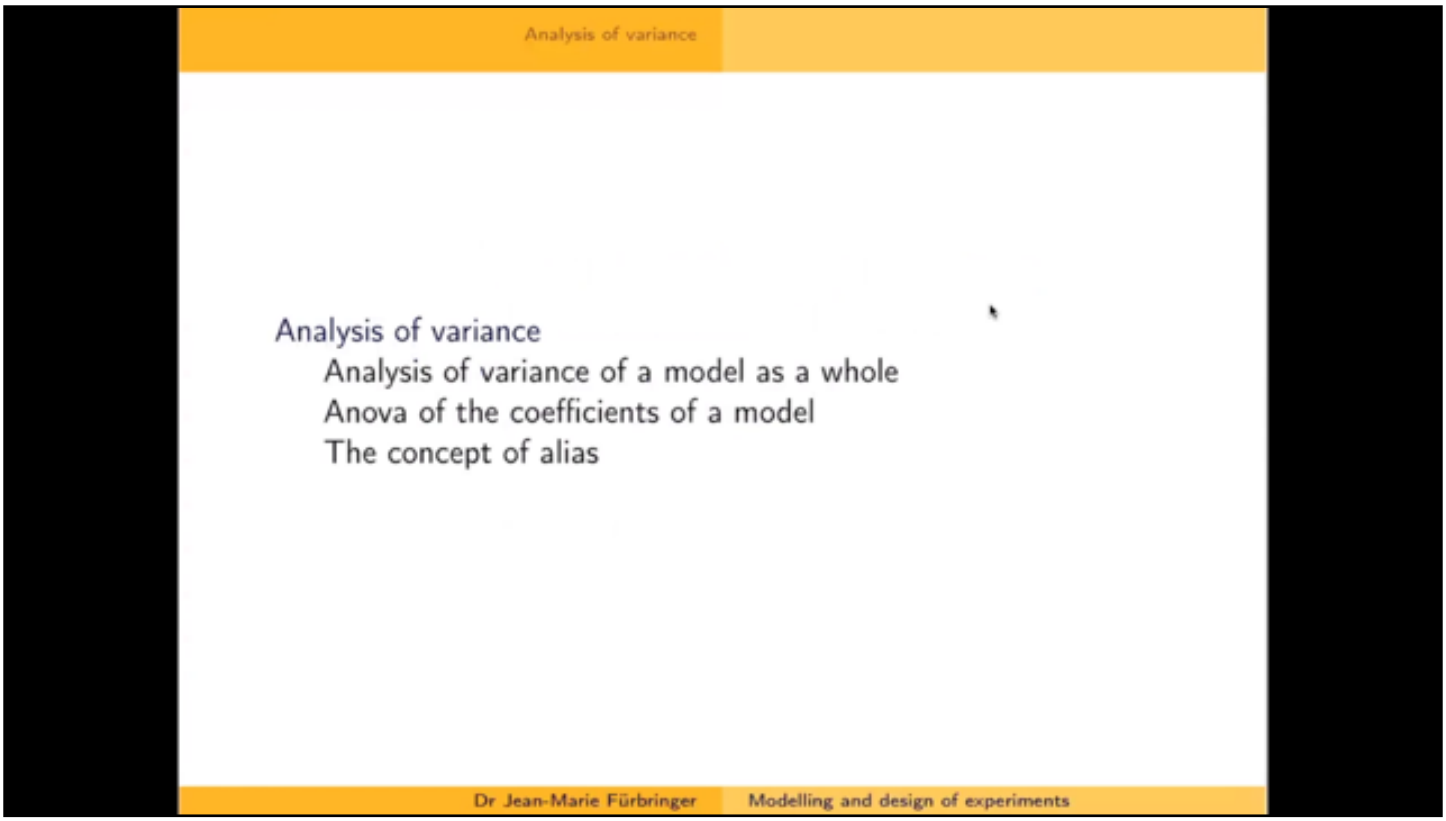


These subtitles have been generated automatically So, be welcome.

notes

summary

0m 1s

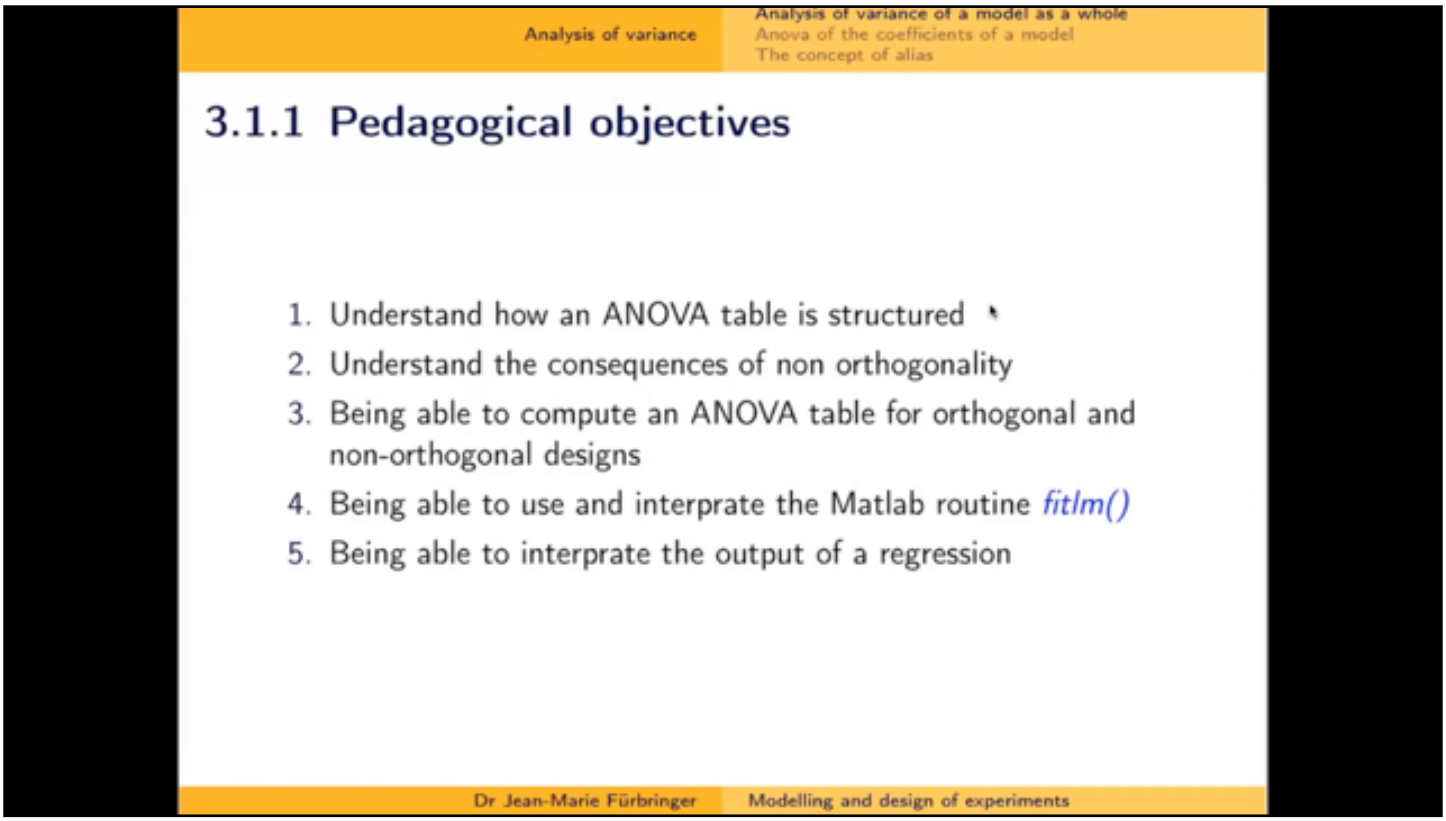


Today, we will see a very important concept. So, it's a concept which has its importance after the experiment. The last lesson, we do a lot of things for understanding how to organize the experiments and how to qualify the experiment before the experiment. Now, we would like to see how we can see after the experiment if the data is of confidence. The data is OK if it's effective. So, we call that ANOVA analysis of variance. So, the variance is the consequence of your experiment because you vary your factors and the variance is also the consequence of random errors. So, what we want is to separate variance between what was the noise, what was caused by the random variation in your measurement and what was the result of your factor in a whole for the model. Also, factor by factor terms per term because you have factors but you also have interactions after we could have a second degree, third degree element. So, it's what we will learn. It's also an invention of Mr. Fisher, Ronald Fisher. And we can say that it's a scientific proof. When we would like to prove something, it's not the only way to do it and it's a statistical perspective on the proof, but in fact, it's the proof that something is concrete.

notes

summary

0m 5s

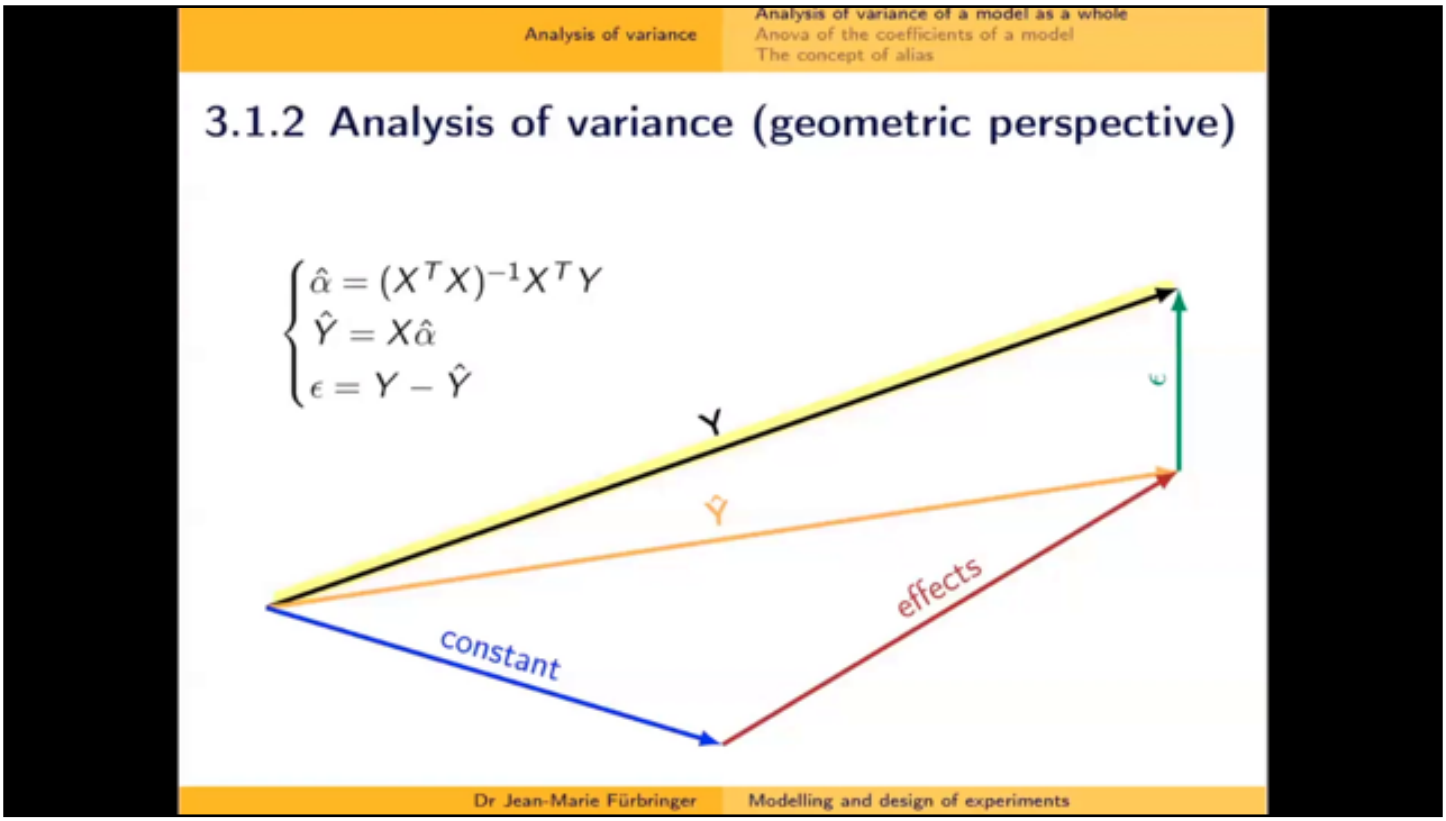


We have in this chapter 3, sub-chapter, I will show you the analysis of variance of a model as a whole. So, separating the variance which comes from your experiment and the variance coming from the random variation of your data, of your measurement. After, I will show you how we can separate this variance of the model in the variance for each of the terms or part of the model. And after, we will see a very interesting concept, alias, that to understand what is the relation between coefficients and when you have non-orthogonal situation. When you have orthogonal situation, you have no alias, you have a diagonal matrix of dispersion and everything is perfect. But in most of the situation, we are not working with diagonal matrices, so we have to understand what is the influence of the covariance between the regresses.

notes

summary

1m 53s

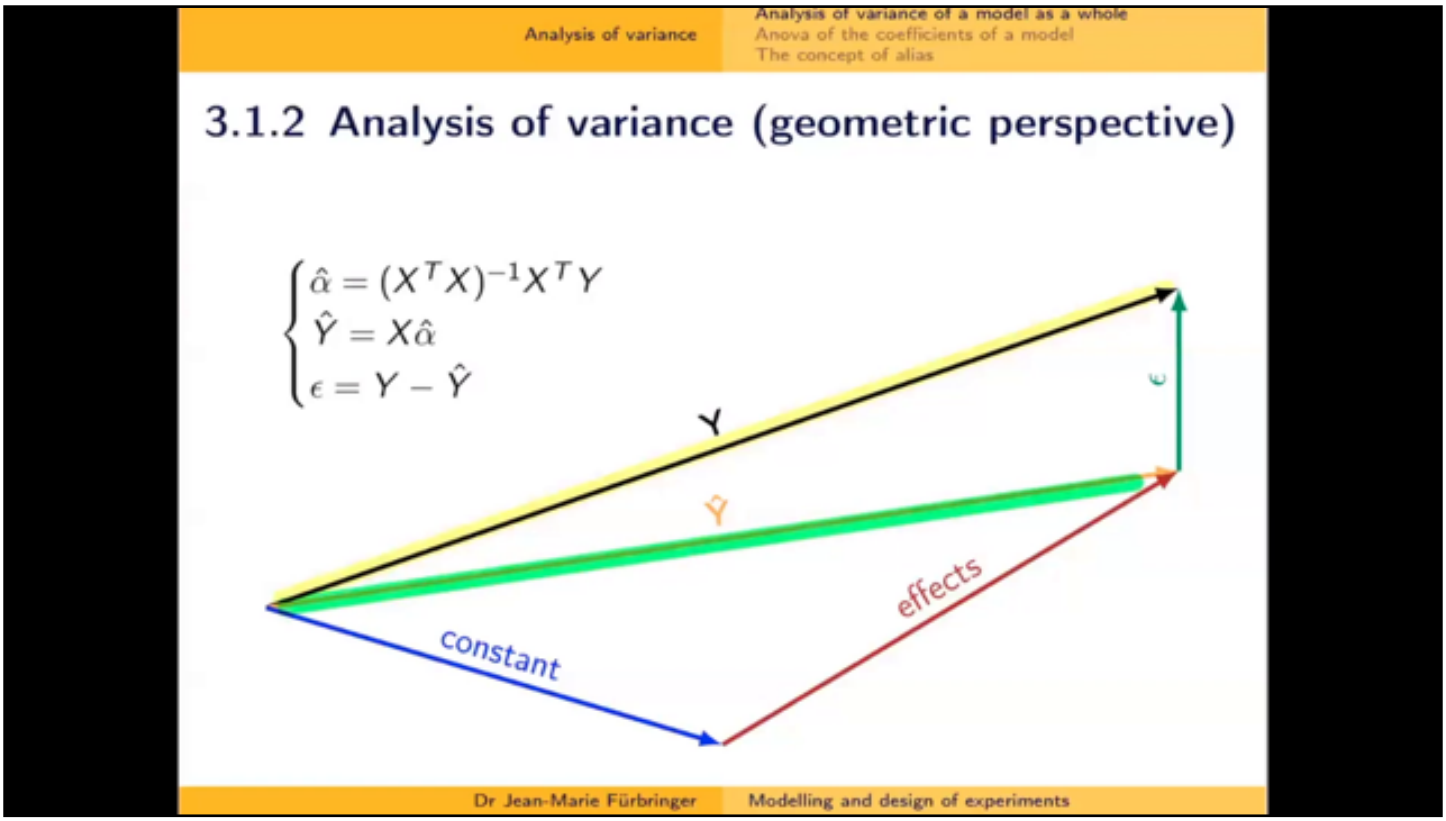


So, the objective for today, I will see if I'm finishing the course, but is to understand what is ANOVA, how ANOVA table is structured, understand the consequence of the non-orthogonality for the coefficient and then, we will see if we can automate the coefficient. After, you have also to learn how to do it, so being able to compute the ANOVA table, it's evident that we do it through algorithms, so you have MATLAB, Python, R algorithms for making it, but it's nice to know how to do it. So, I have one routine, which is fit LM for fit linear model. I will present you the structure in MATLAB of the linear model and I would like that you are able to analyze the output. And finally, I will make some slides about how to interpret the output of regression, generally not only within MATLAB, but generally speaking, what are the things that you have to check after you have made. So, I'm starting with the same element that I already have presented. It's the geometrical interpretation of what is a fit. A fit is a projection and in this, we would like to make a difference between the data, which is a vector of results, typically in your experiment with the Jung modulus, it was a 27 length vector, and we will make a model projecting this vector of results on a subdomain.

notes

summary

2m 59s

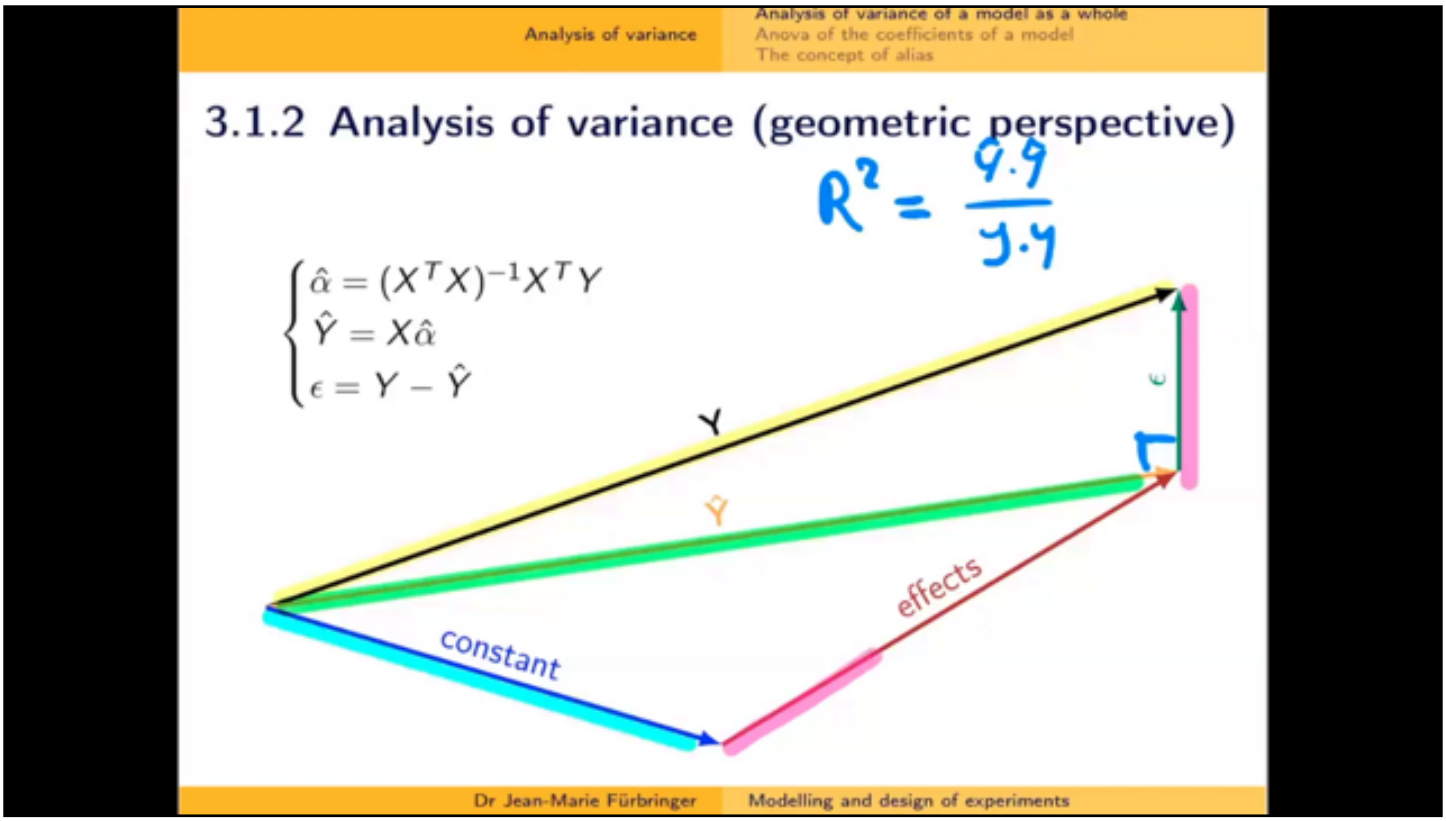


We call the model and we put a hat on it for indicating that it's a model.

notes

summary

4m 49s

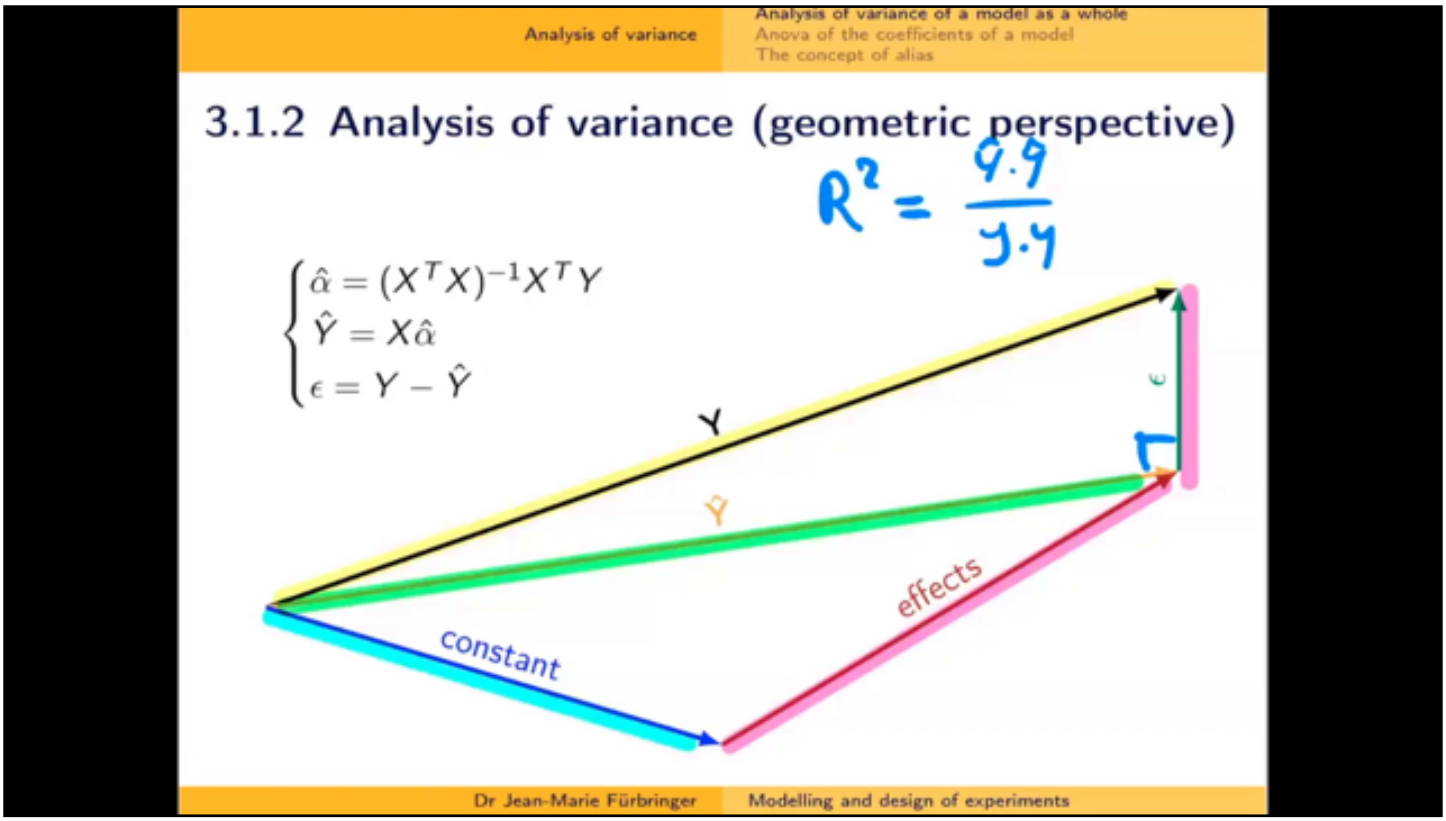


And what we do now, we would like to evaluate if the model is representing correctly the data. That's what we have to check. There are a standard way of checking the R square. So, the R square is in fact the ratio between the length of Y hat, my model, dividing by the length of Y. And so, I did that with sum of squares. So, you can write it like that. This is the R square. This is what does the R square. And usually, if, as you have a right angle here, if they are very close, if the distance is the residue is small, the distance between the two vectors are small, and then the lengths of the two vectors are the same. We usually have R square around 90%, 95%, 99%. It's okay. But it's not a so good statistic, in fact, because in it, you have a big importance of the constant. And I mentioned it already. In fact, the constant represents in your data, the constant is here. It's represent what you didn't consider in your experiment, what you didn't change during your experiment. It could be very huge. Sometimes we are checking effects that are 1% or 1 ppm of the result. So, the constant have a very huge effect in this. So, it's not correct to evaluate the quality of your model, but integrating the constant. And you will see, we will take it out. So, now what we will do, and what we do when we have ANOVA, we compare the lengths of the residue with the length of the effect.

notes

summary

4m 53s

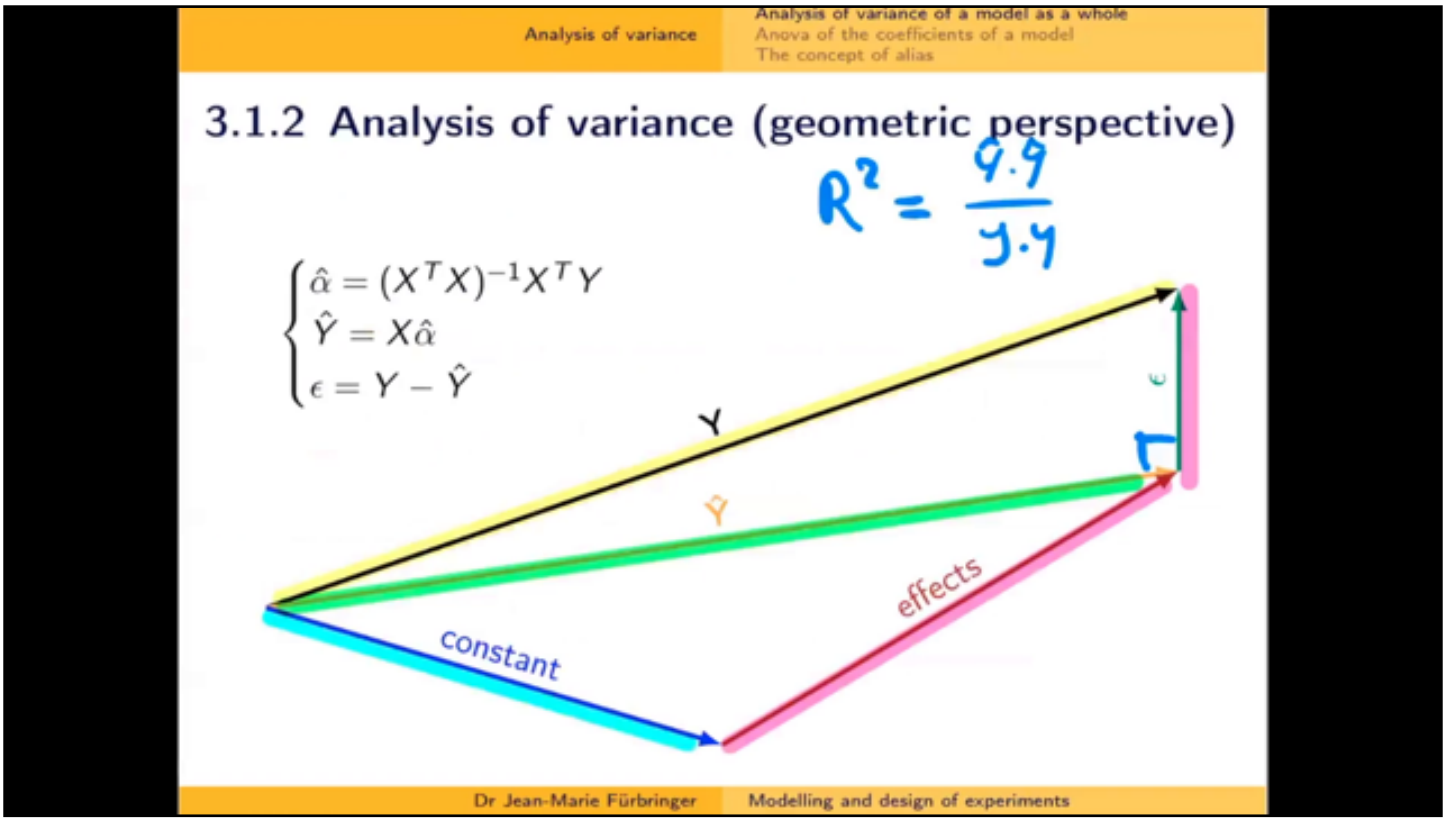


So, here you just have one vector for different effect, but in fact, it could be several dimension. We can have one dimension per effect. And we can consider, because each time you have an alpha 1, x1, alpha 2, x2, so each time have one degree of freedom per coefficients. And so, we would like to confront those values of the coefficient with the residue, with the idea that usually it's better having an effect which is bigger than the residue. Because if you have an effect which is of the same size of the noise, it doesn't represent nothing. So, the rule of Samp, you would like to have effects

notes

summary

6m 59s

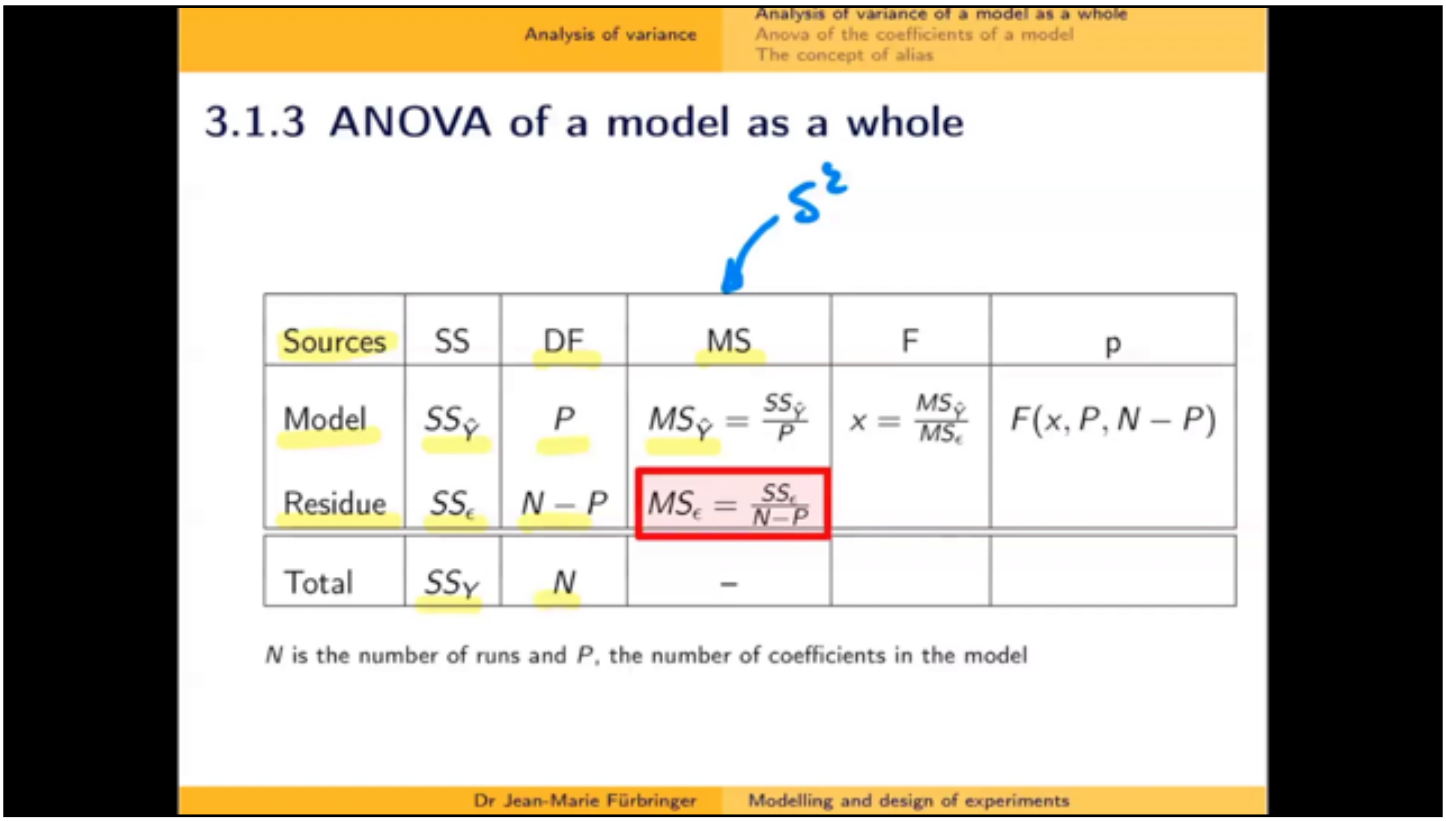


that are at least three times higher than the noise. So, it's what we can calculate in an ANOVA table. So, ANOVA for analysis of variance. So, we have columns and we have rows. So, the first column is the source. So, from where is coming the variations, the variance. So, mainly you have what we say the model, what you would like to model, the subspace and the coefficient of your model is a first origin of the variation and the other is the residue.

notes

summary

7m 47s

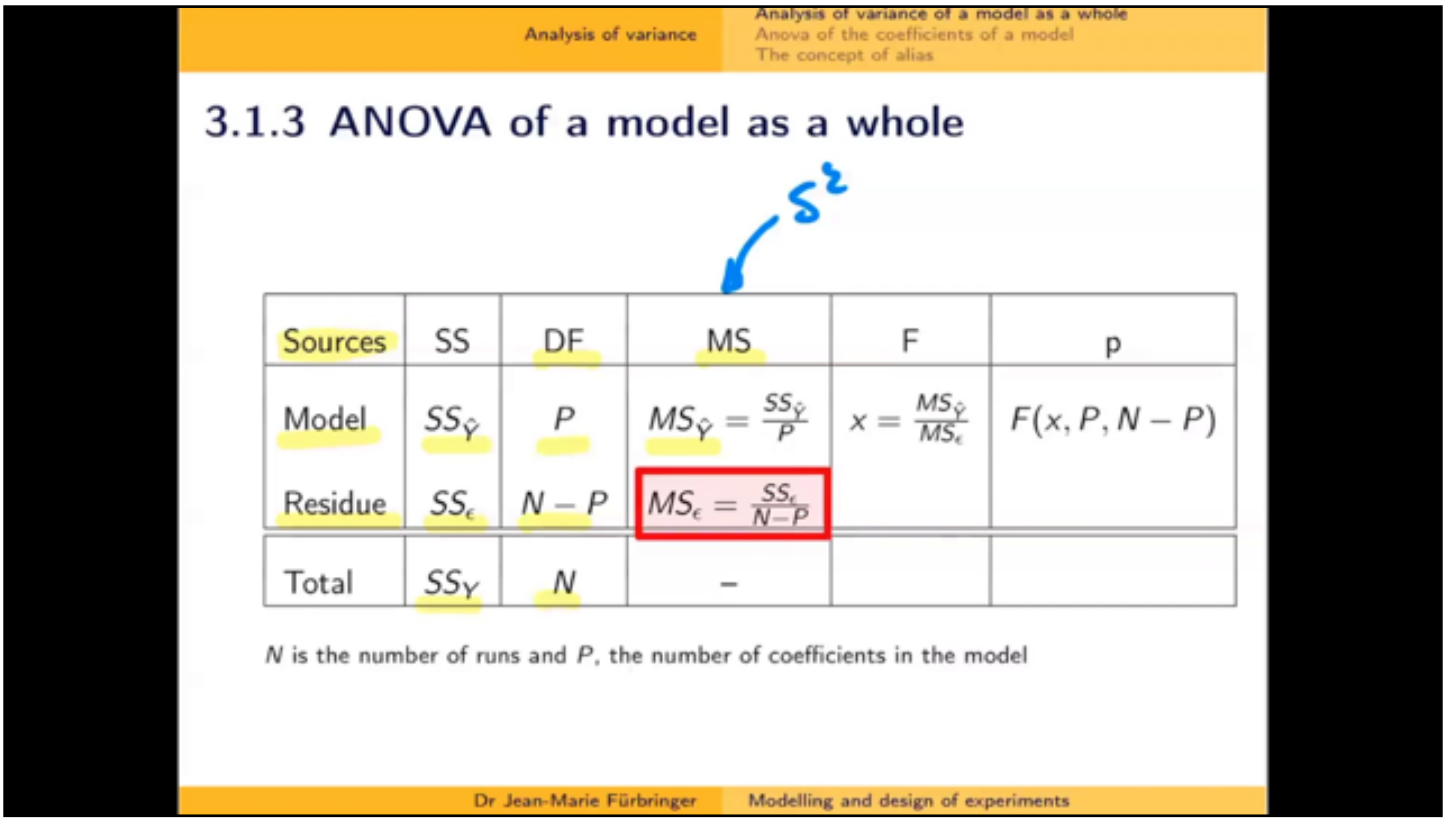


And as I mentioned, there is a right angle between them. So, it's nice because when you calculate sum of square, thanks to Pythagore, we can say that the total sum of square is equal to the sum of square of your model plus the sum of square of your residue. Thanks to this right angle. So, it's what we write here. We have the sum of square of my model, ssy hat, so sum of square of the model, and we have the sum of square of the residue. So, the square of the length of the vector model and the square of the length of the vector residue. And we can calculate the total sum of square is the sum of square of your data. And where is the honesty in measurements, you have to explain all your sum of square. So, you are not allowed to hide sum of square. You are not allowed to create artificially sum of squares. The sum of square is coming of your experiment, you have variation, and you have to interpret it. And the part, it's okay, say it's my random variation and you can analyze it like that. And the other part is from your model. That's, we have subspace of different dimension. If you have, like we have in our exercise three, we have, for example, 27 dimension in my vector y, because I have 27 measurements. And my model has only, if I remember, five coefficients. If I remember the constant three factor interactions of five coefficients. So, the vector of the model y hat can move more. And the vector of the residue will have the size between 27 and 5. So, 22 dimension. So, a vector then can vary in more spaces. Randomly can have, can be longer. So, there are justice in statistics. Not in the world, but there are justice in the statistic. And

notes

summary

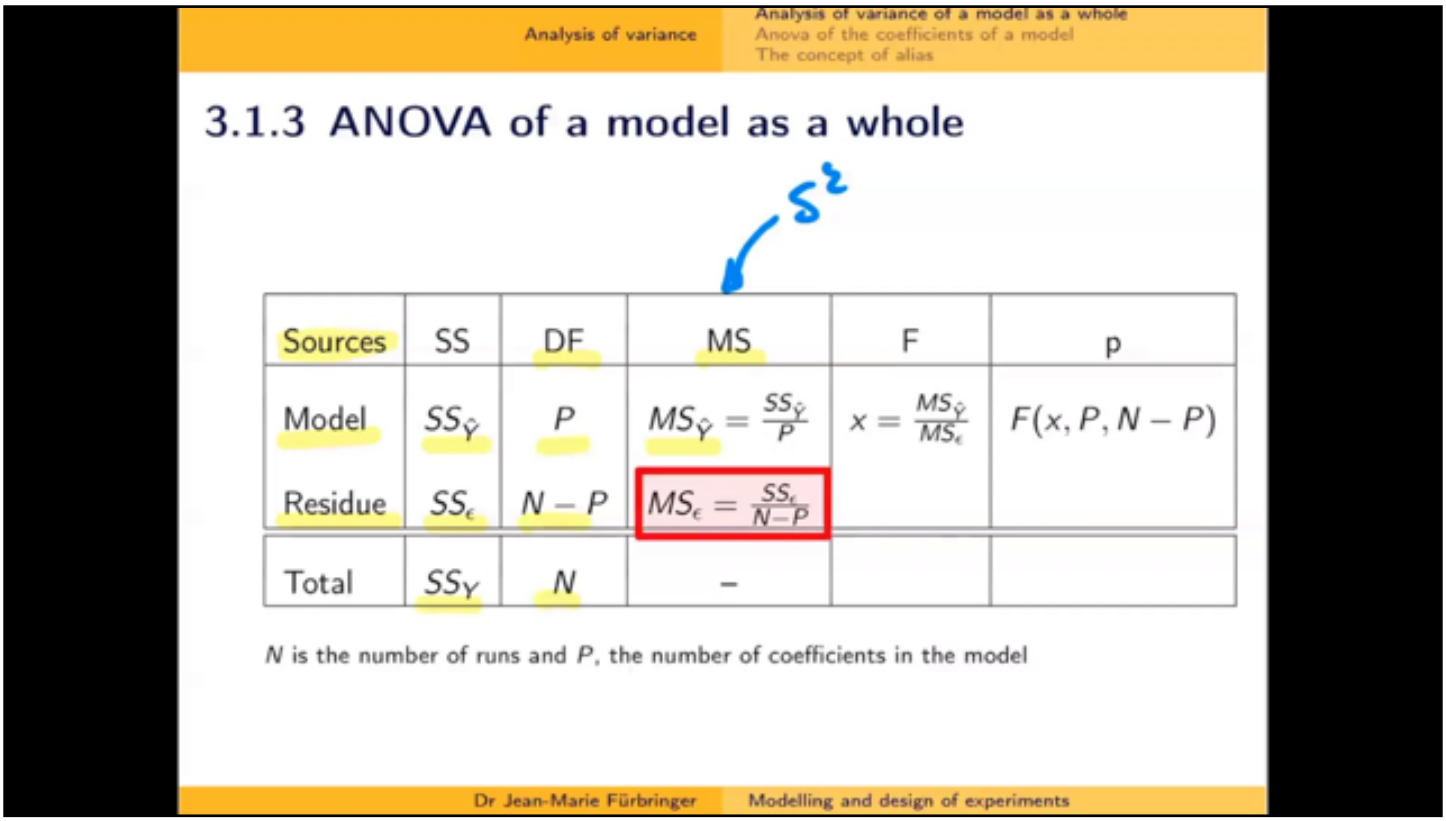
8m 32s



so, we calculate what we call the degree of freedom. Df is a degree of freedom. So, we calculate what is the dimension of the subspace in which the vectors are moving. So, the total which represents the measurement is moving in, and is the numbers of experiment 27 in the exercise three. The model have p coefficients. That means that is moving in a subspace of p dimension. And as we are not allowed to create dimension, the dimension of the residue is the difference between the dimension of the measurement and the dimension of the model. So, I say that what is correct is, it's like David trial. We cut the baby in part. So, we say, okay, we divide the sum of square by the dimension and we can compare it. It's not perfect, but it's work. The mean square of the model, ms, y hat, is just the sum of square of the model divided by the numbers of parameter. And the mean square of the residue is the sum of square of the residue, divided by the dimension of the residue. The degree of freedom of the residue is the numbers of measurement minus the numbers of factors in your model. And you see that after the total, it doesn't work anymore. As we have divided the sum of mean square is not meaningful. So, this column corresponds to a variance, what we usually call a variance. But usually when you calculate the variance of a series, you divide it by n minus one, because you consider that you have an average of the series that is using one degrees of freedom. So, it's why you make n minus one. Now, we are using n minus p for the variance, the experimental variance. So, this data is what we use to call r square, the experimental, the sample variance. We make an approximation because they

notes

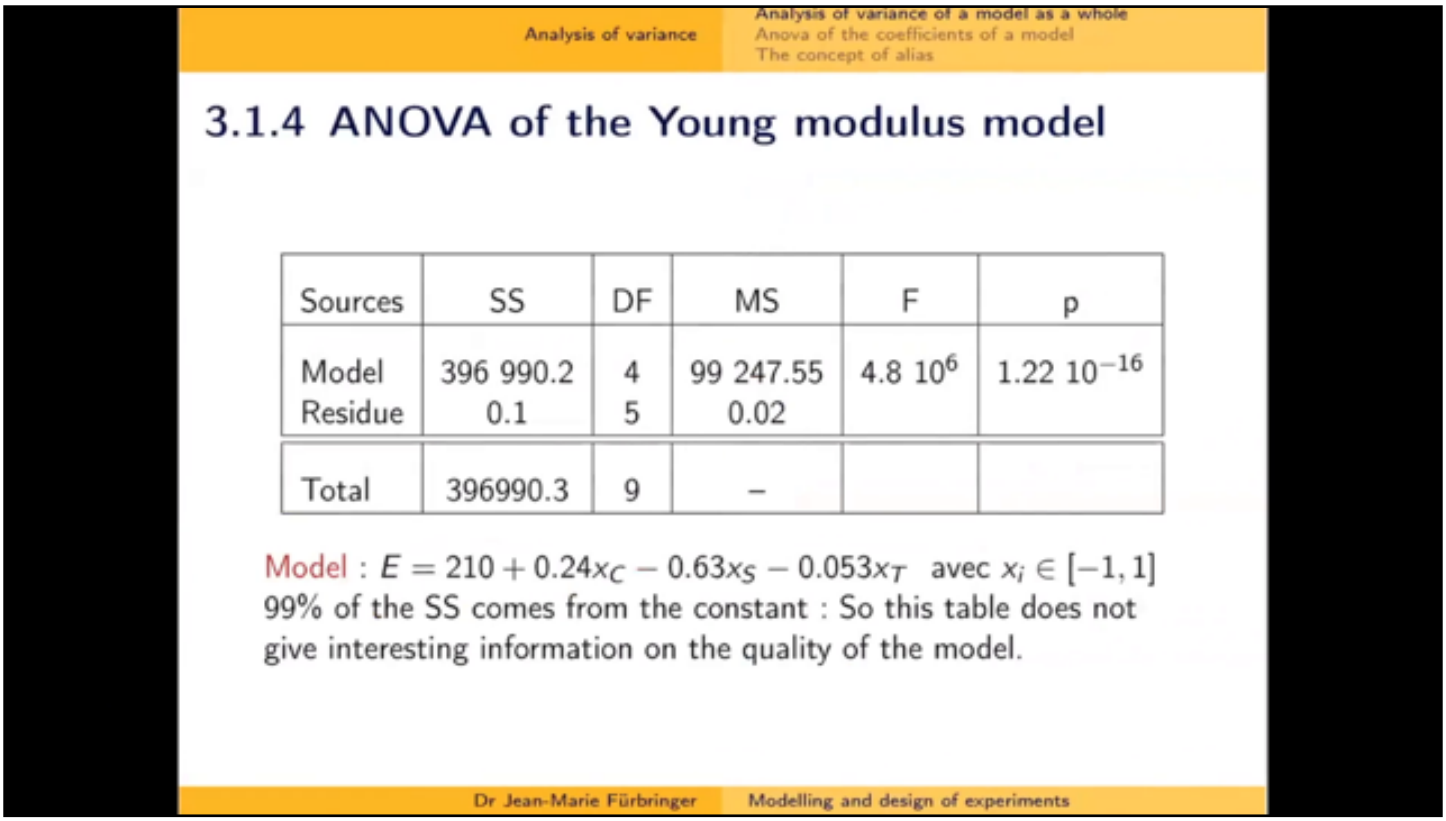
summary



are in the residue, you have two things. You have the fact that your data is randomly varying when you are making a measurement, because your instruments are not perfect. But eventually also you have what we call the lack of fits, the fact that you choose a model which is not correct. We will separate that in three, four lessons. We will look how to separate this sum of square in two parts. The pure error, which really is considered as your experimental error and the lack of fits, what's come from the fact that eventually you are not following the correct model. Your reality is not following the model you would like, it's following.

notes

summary

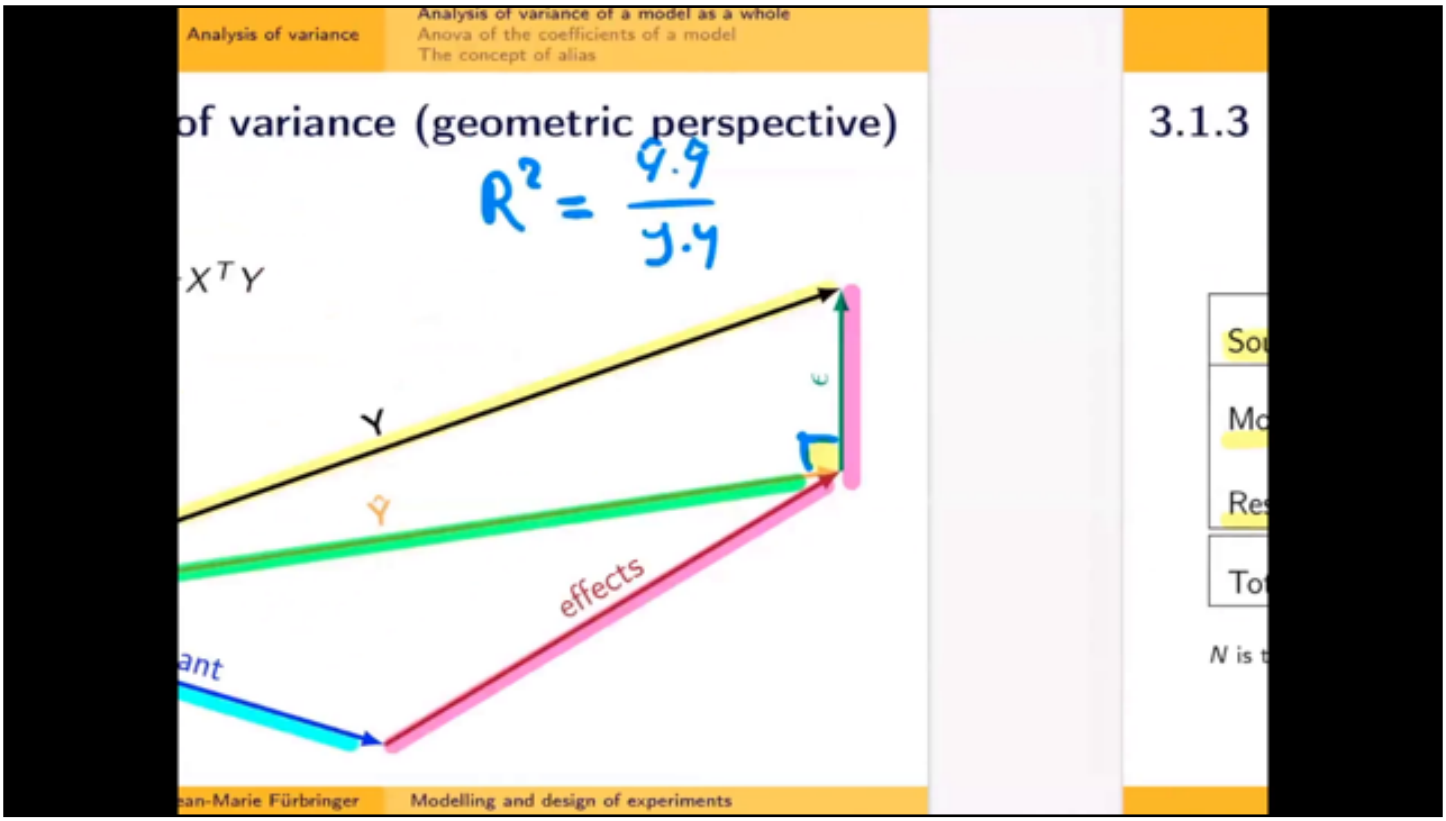


After we have the Fisher ratio, here's the ratio between the mean square of something that interests you, divided by the mean square of the residue. So, when you just have one line, it's easy, it's the first one divided by the second one. The Fisher ratio, the question was what is the name of this column, which has an F letter, and it's the Fisher ratio because all that have been invented by Fisher. And after you have the p value, the probability. So, this column is the probability of getting the Fisher ratio by chance. The chance is okay when you are gambling, but when you are making science, you don't like having things by chance. So, you would like this p value small. In fact, it's the probability to be wrong when you pretend that you have an effect. This p value, be careful, the F letter is used two times. So, I have an F letter, which is classical for the column of the Fisher ratio. And after I have the Fisher distribution to find in a table that you calculate with MATLAB, Excel, Python, or R. And it's telling me it's the Fisher value. For the value of X, you see X here. So, for a given Fisher ratio, and a number, a degree of freedom, p is the numbers of parameters, so the degree of freedom of your numerator in your ratio and the degree of freedom of your denominator. And so, you have the p value, you have the... how you can be confident that your effect is real.

notes

summary

14m 37s

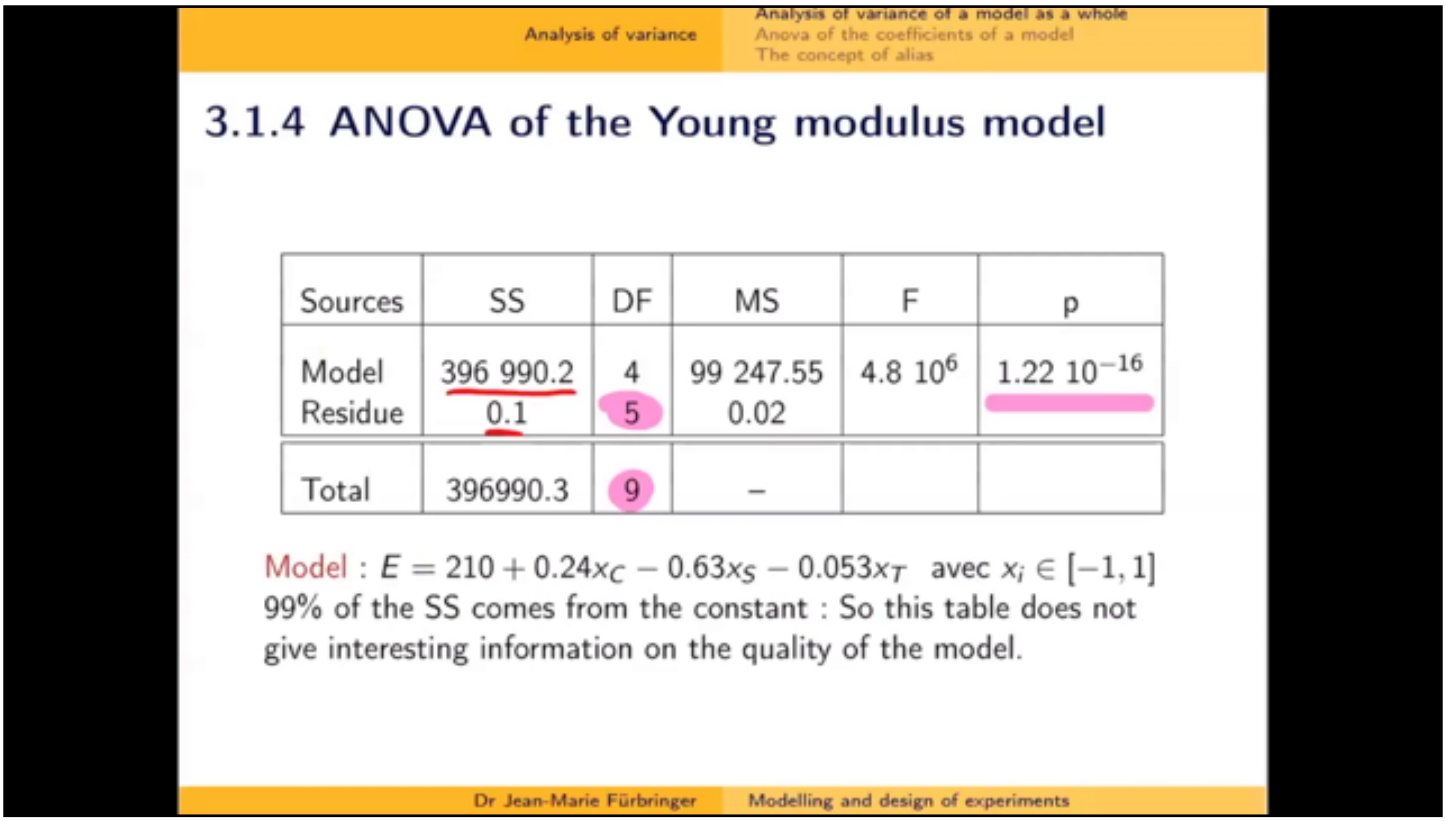


So, this is the ANOVA table of exercise 3 for the Jung modulus experiment. So, perhaps it's not the data of exercise 3, perhaps more the data of my slides, which are not exactly the same. So, you see, for the model, I have calculated the sum of squares. So, I have calculated the estimate, what my model estimate as a response, and I take the square of each result and sum them. And I get a big number, usually it's a very big number, quite 400,000. And I have calculated my residues, so I make the difference between my measurement and my estimates. And I make a sum of square and I get a very small number. My model seems quite very good. I calculate the degree of freedom. So, in this case, the degree of freedom of the model was 4, a constant and one degree of freedom for the carbon, one for the sulfur, one for the temperature. I'm not calculating as you did in your exercise, the interaction between carbon and sulfur. As I have nine experiments, my degree of freedom for my residue is 9 minus 4, so it's 5. So, I'm able to calculate the mean square. I get something as 100,000 for the mean square for my model. And I have a mean square for the residue of 200. So, I'm very happy already. I understand that my error is very small in comparison to my model, my vectors.

notes

summary

16m 39s

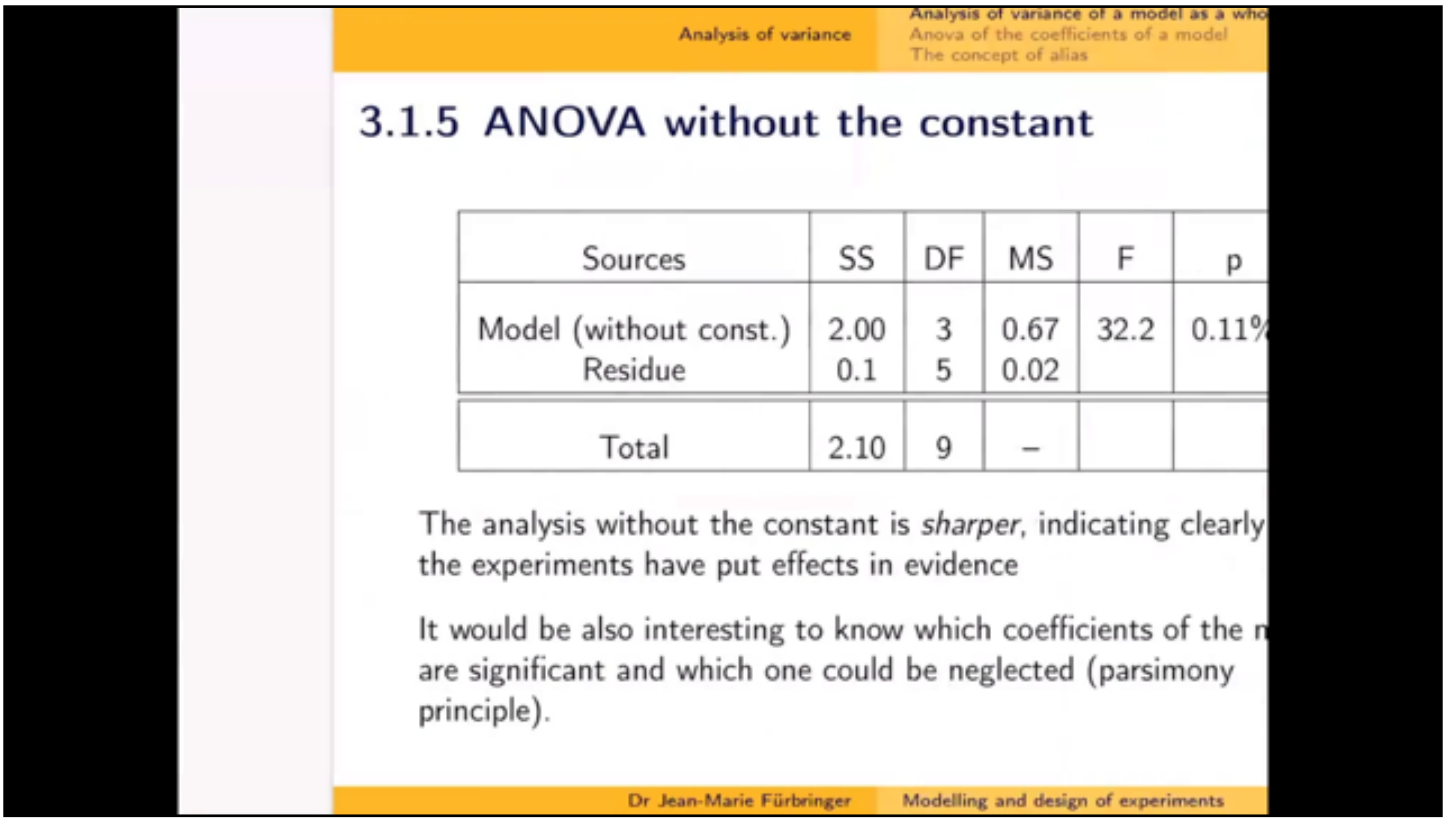


So, I'm comparing the size of my vectors, and I understand that my green vector is something as, I don't know, 10,000 times bigger than the size of my residue. So, the model is very close to my measurement. And when I make the ratio, you see I obtain something as 5 ppm part per million, 10 power 6 for the ratio. My model is something as 5 million times bigger than the ratio. And I'm calculating the probability to get that by chance. And my probability to get that by chance is 1 at the power 10 minus 16. So, very, very small. So, my probability to be wrong when I'm pretending that I have a linear relation between my measurement of my young modulus

notes

summary

18m 30s

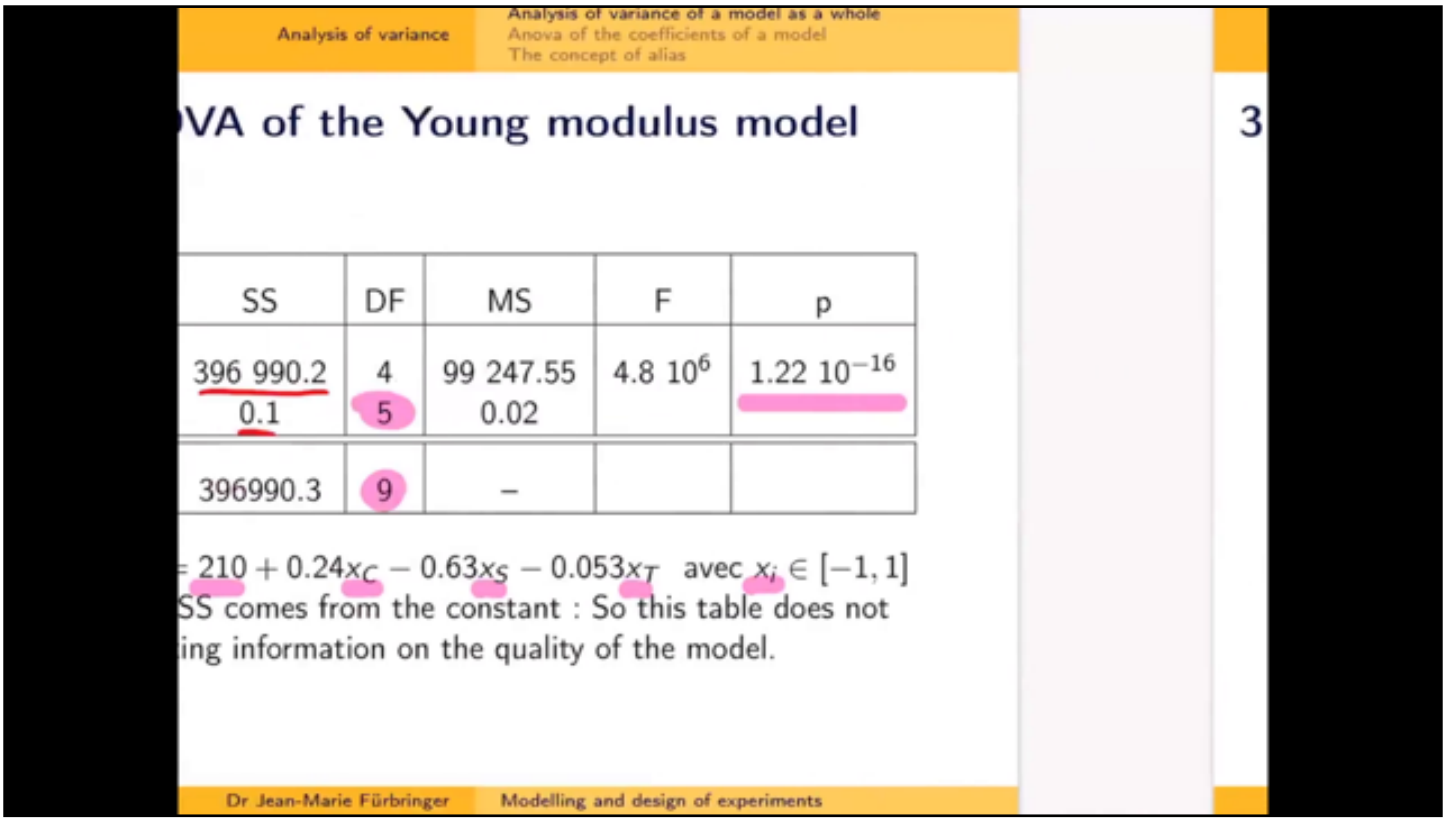


and the change in my carbon, and the temperature is very small. I'm confident in my model. I'm also able to calculate the confidence interval of the coefficient, but that is also a result. It's a result for the model in a whole. It doesn't tell me if I'm confident more in the carbon, more in the sulfur, perhaps one of our big values, another small values, to remember the temperature coefficient was very small. This is a result in a whole. So, I'm able to say that my young modulus, my probability was something as 200, 210,000 kPa, so mega-pascals, and I have the different coefficients for the carbon, sulfur, and temperature, 0.24 for the carbon, minus 0.6 for the sulfur, and minus 0.05 for the temperature, when I have standardized variable, and I have very, very good model. I'm very confident in this model. But what could come from carbon? What could come from sulfur? Are my models okay in all its coefficients? This I don't know at that level. What is an acceptable p-value? Very good question. 5%. We consider, usually, that we could accept a model experimentally with 5% risk to be wrong, which is still a lot. It's big. Usually, companies about very important results would like to have PPM. So, it's why you have companies that are called 6-Sigma. That means they would like to have a confidence interval corresponding to 6-Sigma. So, values of PPM for the p-value around part of a million. And even sometimes you accept something with risk. You know that you have risk, but you accept something which is high.

notes

summary

19m 37s

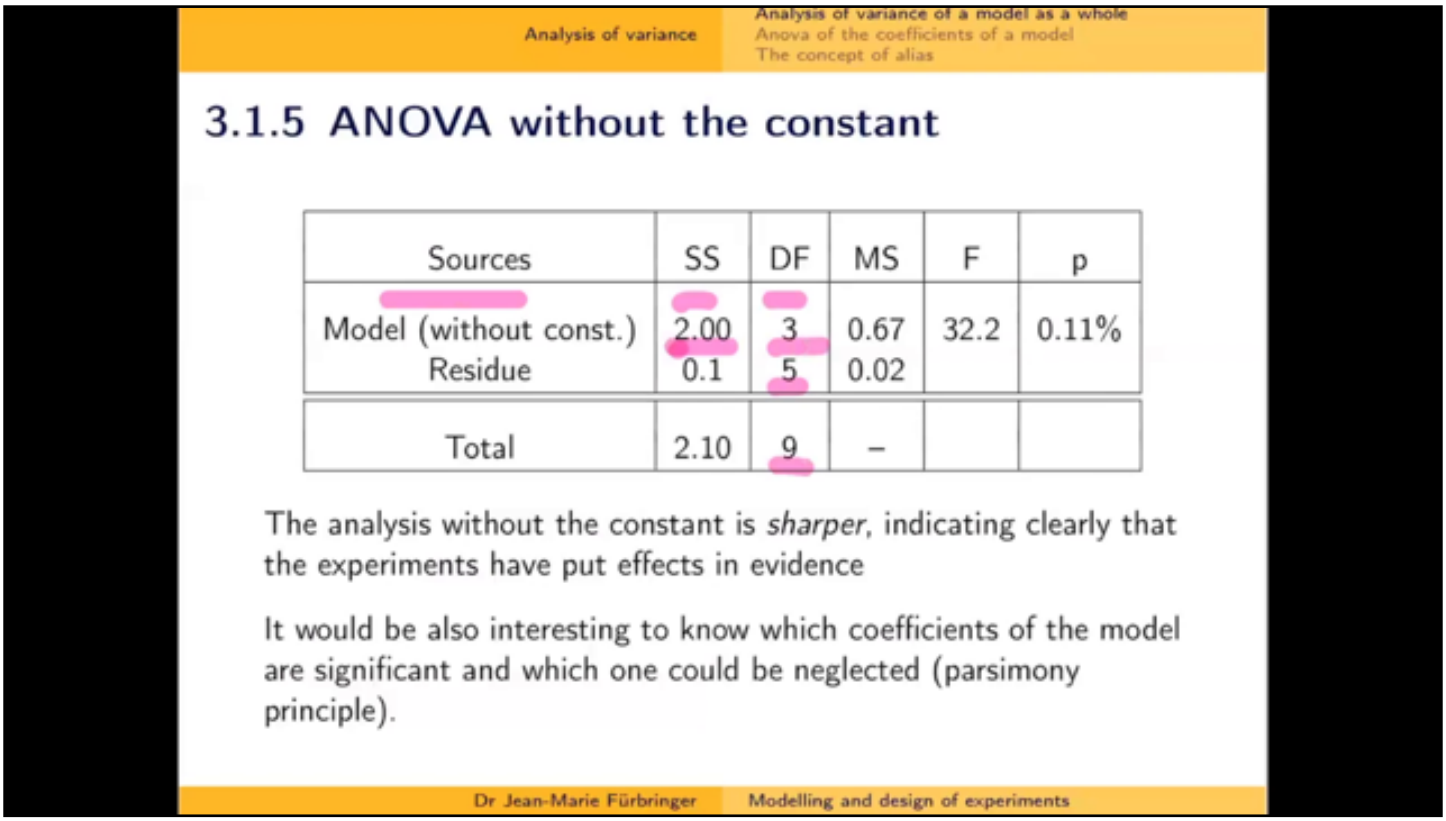


The threshold, the official threshold is 5%. But as I tell you, it was not very interesting to compare the model with the constant. In this case, you have C.

notes

summary

22m 13s

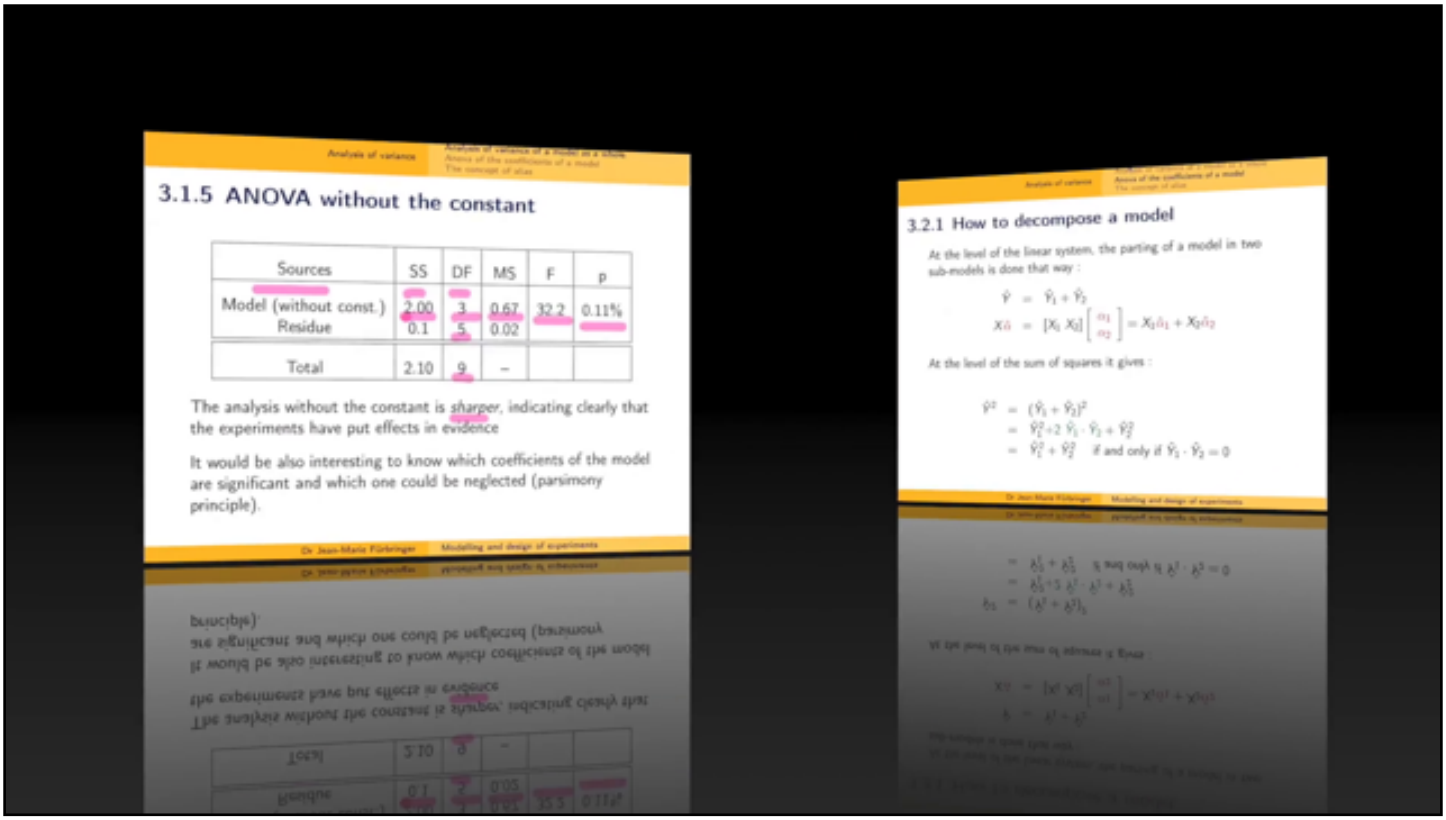


It's why I show the values of my model. The constant is something as 1,000 times bigger than my coefficients. So, most of the things, it's not your work when you make the measurement. It's the nature making it. So, usually, we make another table without the constant. Like that, you are really comparing in your model the part which depends on your coefficient, not the part that is not depending on your measurement. So, if you do that, you see that the sum of square, just because of your measurement and the change, just because of, without the ground mean, is not before it was 400,000. In fact, your variation was very low because it's not easy to change the value of the young modelers just changing the carbon concentration after you don't have any more steel or you have a metal which is not correct. So, in fact, the variation was quite small. So, the sum of square of your variation only was also very small. So, now you compare not 400,000, but 2 when you have taken out the sum of square of your constant, you have no sum of square for your model without the constant of 2. And the residue is the same. So, if you do not change the model, we are just changing what we consider for the ANOVA. So, we take out the constant of the ANOVA. And now you see the value. So, the degrees of freedom for the model is not 4, it is 3. And if you see there are one degree of freedom where it appears, the degree of freedom of the constant. So, eventually I could put the constant here and I have some sum of square here and the degree of freedom one here, but usually we don't do it. When you look at the tables prepared by MATLAB,

notes

summary

22m 25s

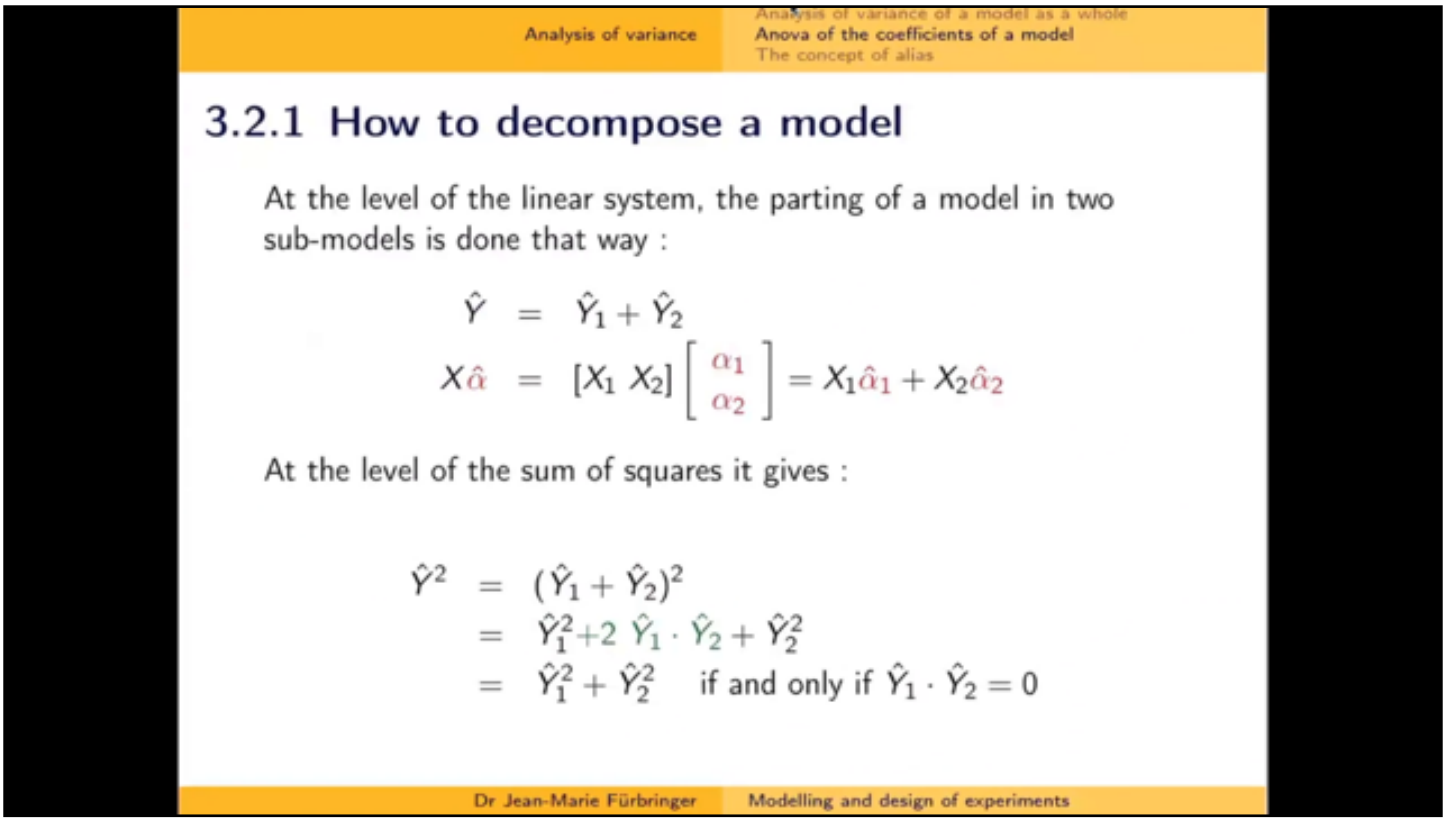


by Python routines, by AR, they are just forgetting the constant. It's why you are missing a degree of freedom between your model. It considers the degree of freedom of the model P minus 1. And now the mean square is a lot smaller. 60, 67.2. It makes now a Fisher ratio, which is smaller again, but it's still bigger than one. So, I'm very happy because if you have one, you have 50% probability to be wrong. So, you don't appreciate that as 50, 50. And now I have a probability value, which is probably more real to qualify my model in my situation, which is a tenth of a person. So, we say here that the analysis without the constant is sharper. So, now we would like to understand if all the coefficient of our model are okay. Because we would like to work with the per se money principle. So, eventually if one coefficient doesn't bring a lot of information, it's not necessary in my model. So, we would like to dig further in my model to see what is the probability for each terms to be correct or not correct.

notes

summary

24m 37s

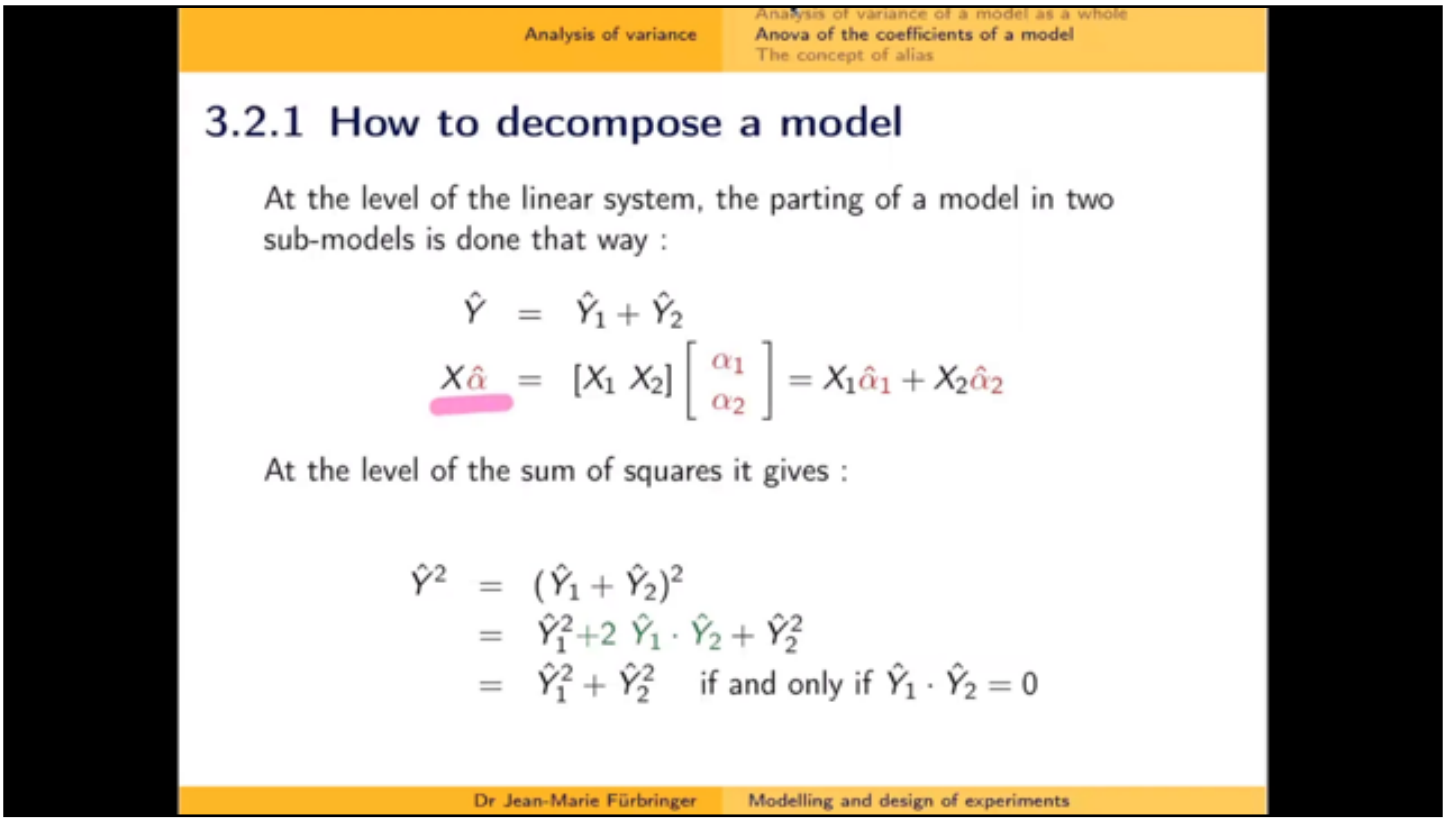


So, now I would like to separate my model in different layers.

notes

summary

26m 1s

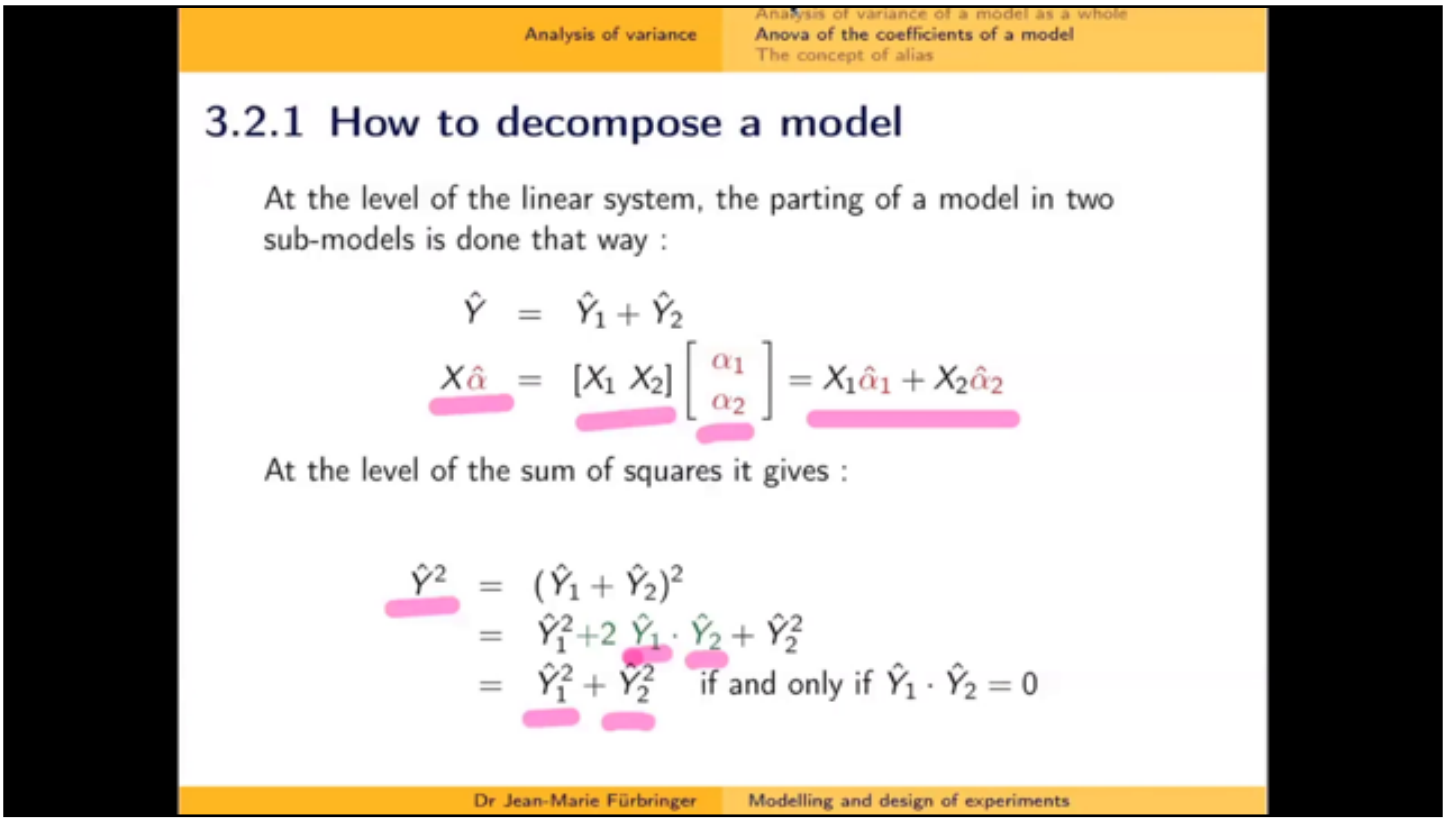


So, let's separate first in two layers. And when we separate as we are working with variance, we are working with sum of squares. The orthogonality is important. So, when we have a model, why hat and we want to separate in two parts, let's call why hat one and why hat two. Matricially is very easy to do it.

notes

summary

26m 2s

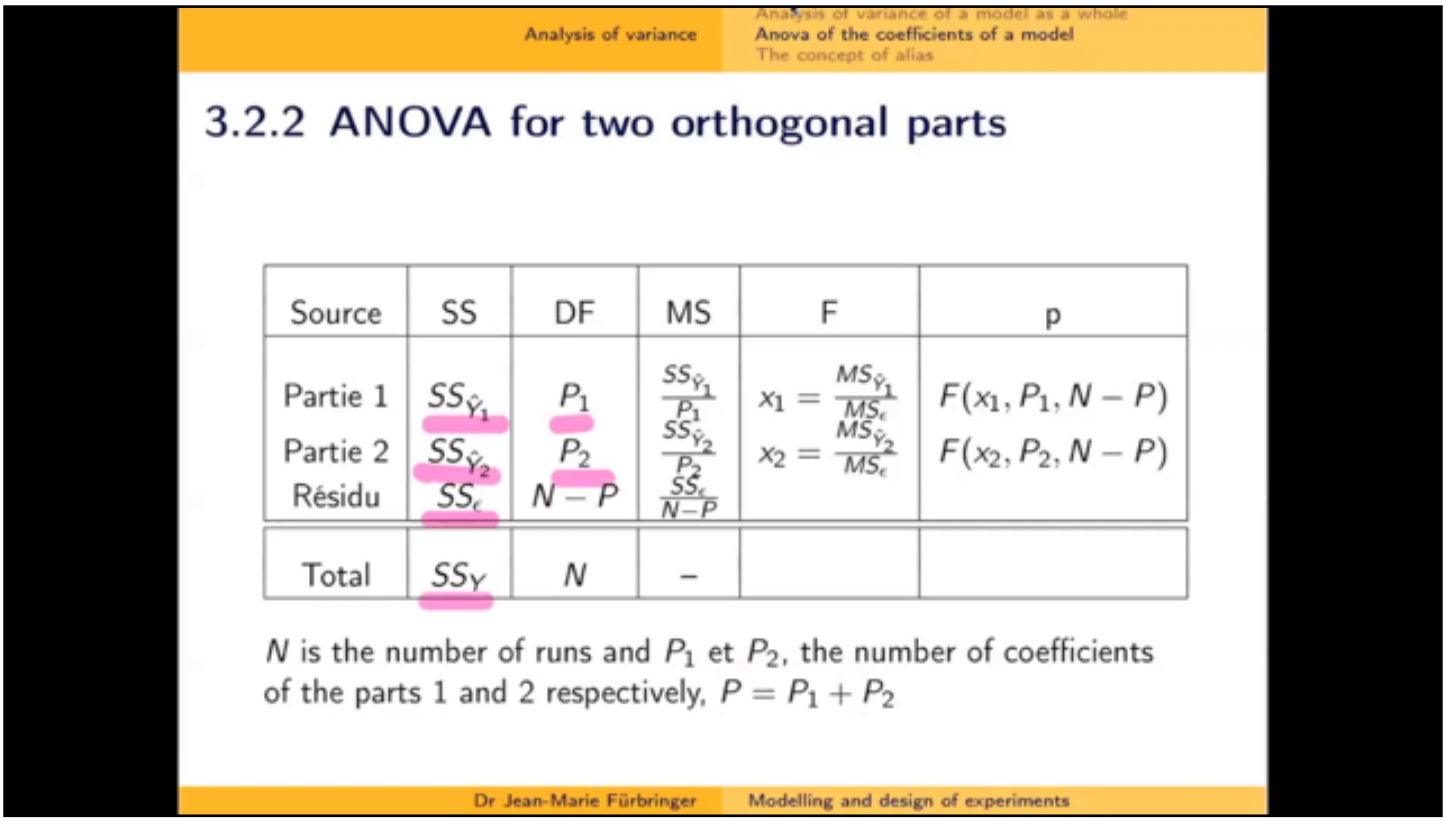


So, we have our model X multiplying the coefficient alpha, alpha being a vector of coefficient. And we can do that by concatenating, subdividing some part of the model matrix. X with separated columns per column, what's belong to the first part of the model, we call it X1. And what's belong to the second part of the model, X2. It's easy to do it. It's a linear model. So, we can do that column per column. And our vector of coefficients is separate in two parts. No, they are separated vertically. I call the coefficient of the first part alpha one and the coefficient of the second part alpha two. And then I'm able to calculate my model as two parts, being X1 multiplying alpha hat one plus X2 multiplying alpha hat two. So, this line is just for showing you that algebraically, it's work well. Now, what's happened with the sum of squares? So, if you have the sum of square or Y hat, we call it Y hat square, I'm calculating the square of it. And you know from high school that it's equal to the square of Y1 plus the square of Y2 plus two times the multiplication of Y1 and Y2. So, if Y1 and Y2 are orthogonal, it's easy to subspace that are orthogonal, have the products that are zero. So, if they are orthogonal, the square of Y hat is the sum of square of Y1 and Y2.

notes

summary

26m 38s

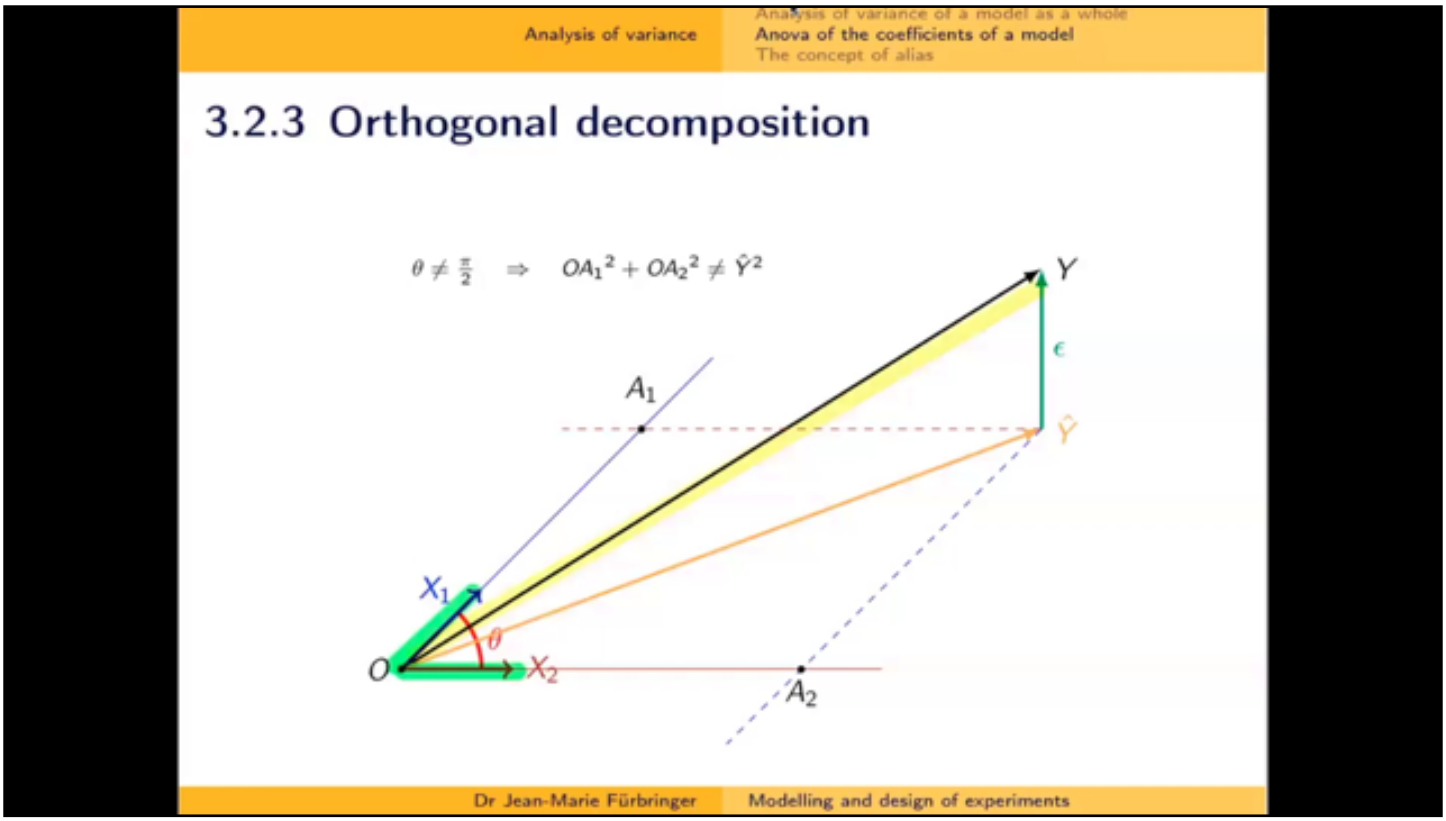


But this only if Y1 multiplied by one two equals zero. So, in this case, it's okay. I can have two lines for my model and the sum of square will work. So, I will have the sum of square of Y1, the sum of square of Y2, my residue when I have my two models, my two parts of the model together, and it should correspond to the sum of square of my measurements. Now, I have degrees of freedom that depend how I have part my model,

notes

summary

28m 47s

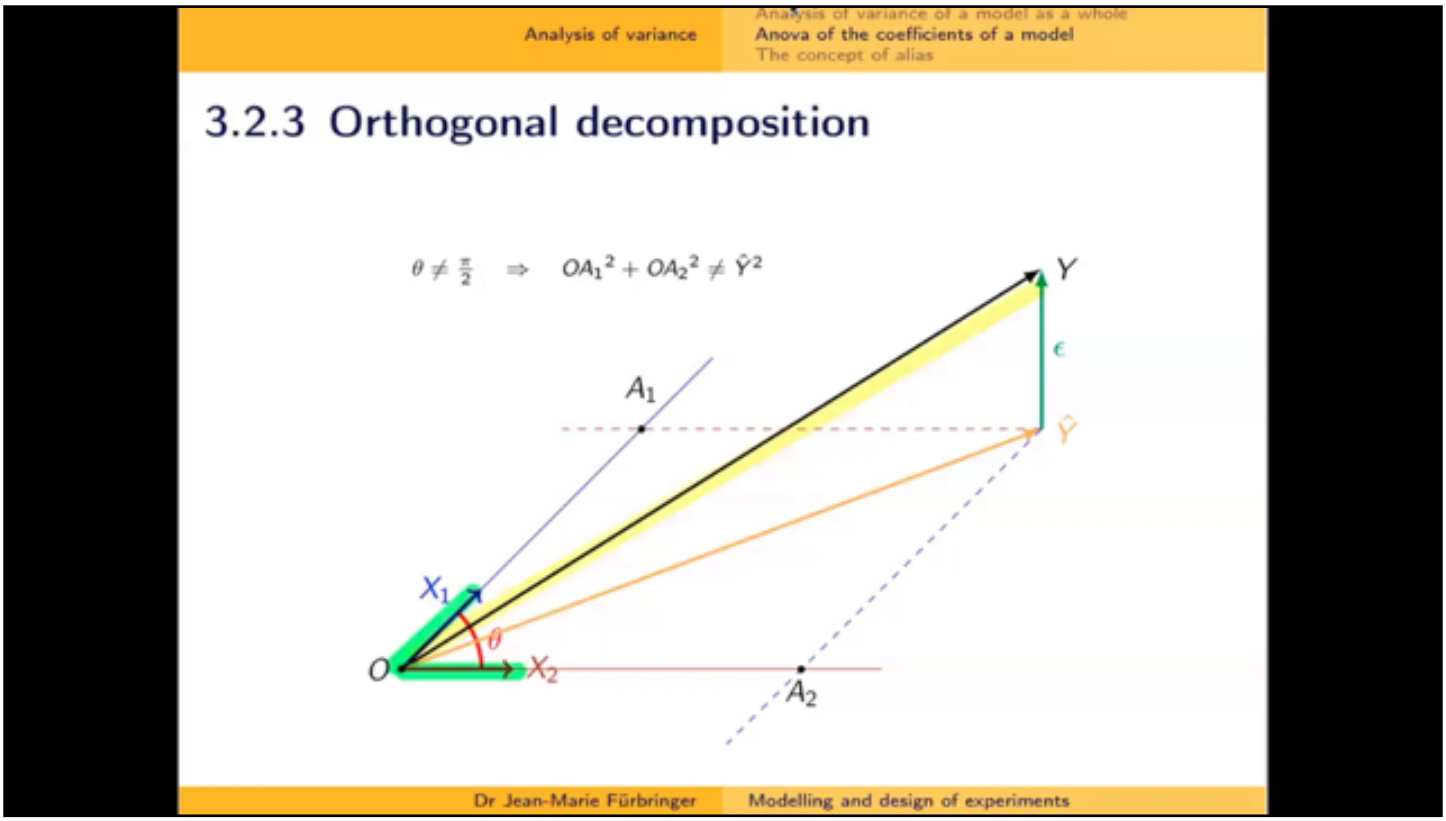


P1 for the first part, P2, but the P1 plus P2 should correspond to P, the numbers of coefficient in my model. And my degree of freedom for my residue is N minus P. And my total degree of freedom is my numbers of experiments. So, I'm calculating after that the mean square. So, the mean square is the sum of square divided by the degree of freedom corresponding to each part, the first one by P1, the second one by P2. And finally, the sum of square of the residue for epsilon in the slides and divided by N minus P. And I'm able to calculate a Fisher-Ratio value, X1 and X2, for the different layers. So, the different layer, what you want, it could be the carbon and on the other aspect, the sulfur on the temperature. It could be the carbon and the sulfur together against the temperature. It could be when you have exercise three, the linear part of your model and the interaction part of your model. But in this case, as I'm writing this table, what I'm obliged is to have an orthogonality between the layer Y1 and the layer Y2. If it's not, I'm not allowed to calculate that because the sum of square will not fit if it's not orthogonal. And then I'm able to calculate P values for each of the layers of my model and this will be, I hope, smaller than 5%. And if it's smaller than 5%, each one is smaller than 5%, I will consider that each layer is valid. If I have a layer which is, I don't know, 50% P value, I say, oh, this part of the model is not true. And by persemini principle, I will not consider it. So, like that, another table lets you select what part of your model and after each part of the model will be each coefficient.

notes

summary

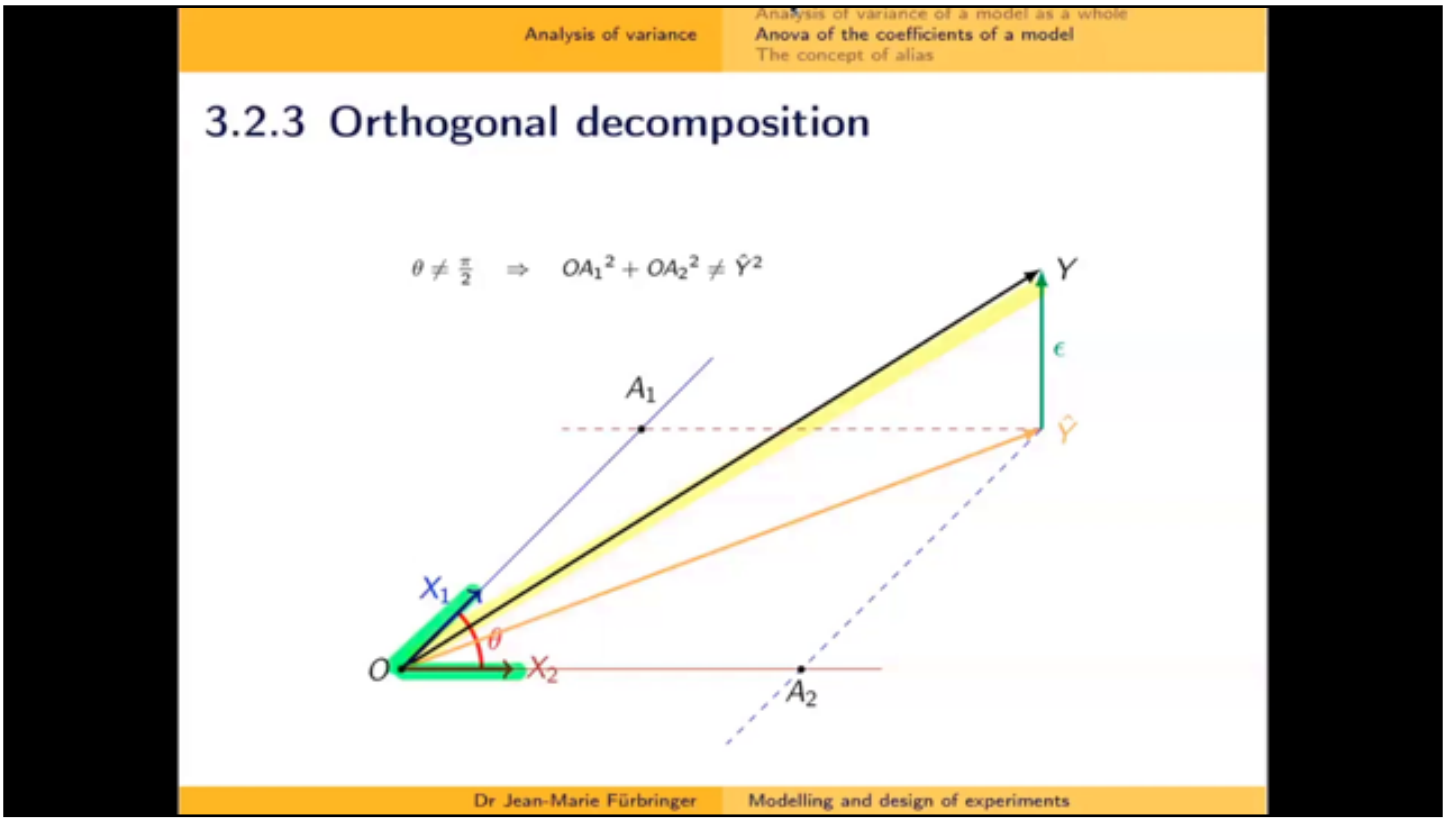
29m 25s



You can make an ANOVA table with each coefficient. I just presented dividing into, separating into, but in fact, you can separate with the same principle. You can separate in how many layers that you want. So it's why ANOVA is used when you have a lot of factors for selecting the factors that are effective, factor effective. It's the same word which is behind. You would like to know if a factor is really a factor, or it's not a factor. So it's ANOVA, it's a way of testing, we say regressors. We make the difference between factors and regressors because factor, in this case, will be, for example, carbon, sulfur, and temperature. And the regressor would include also eventurized instructions that I'm considering. The second degree or the third degree I would eventurize. So it's why it's very important the ANOVA because it lets you select at the end of your experiment what finally is effective and what finally is not effective. And the thresholds usually is 5%. We are used to reject part of your model coefficients whose p-value is higher than 5%. We can have special decision in different situations. What happens when your coefficients are not orthogonal? That means that the vectors on which you are projecting your experiment are not. That means that the column of your model matrix are not orthogonal. You can consider the columns of your model matrix as vectors. And you can check if they are orthogonal or not. When they are not orthogonal, you have to take that into account. So here I'm explaining what's happened. How do I orthogonalize? I make a not-so-gonal decomposition of something which is originally not orthogonal. So you have, if you see my graphic, you have here my measurement, the black vector. And I have projected my model into axes. X1 in blue, X2 in red. That's the angle. You don't see it

notes

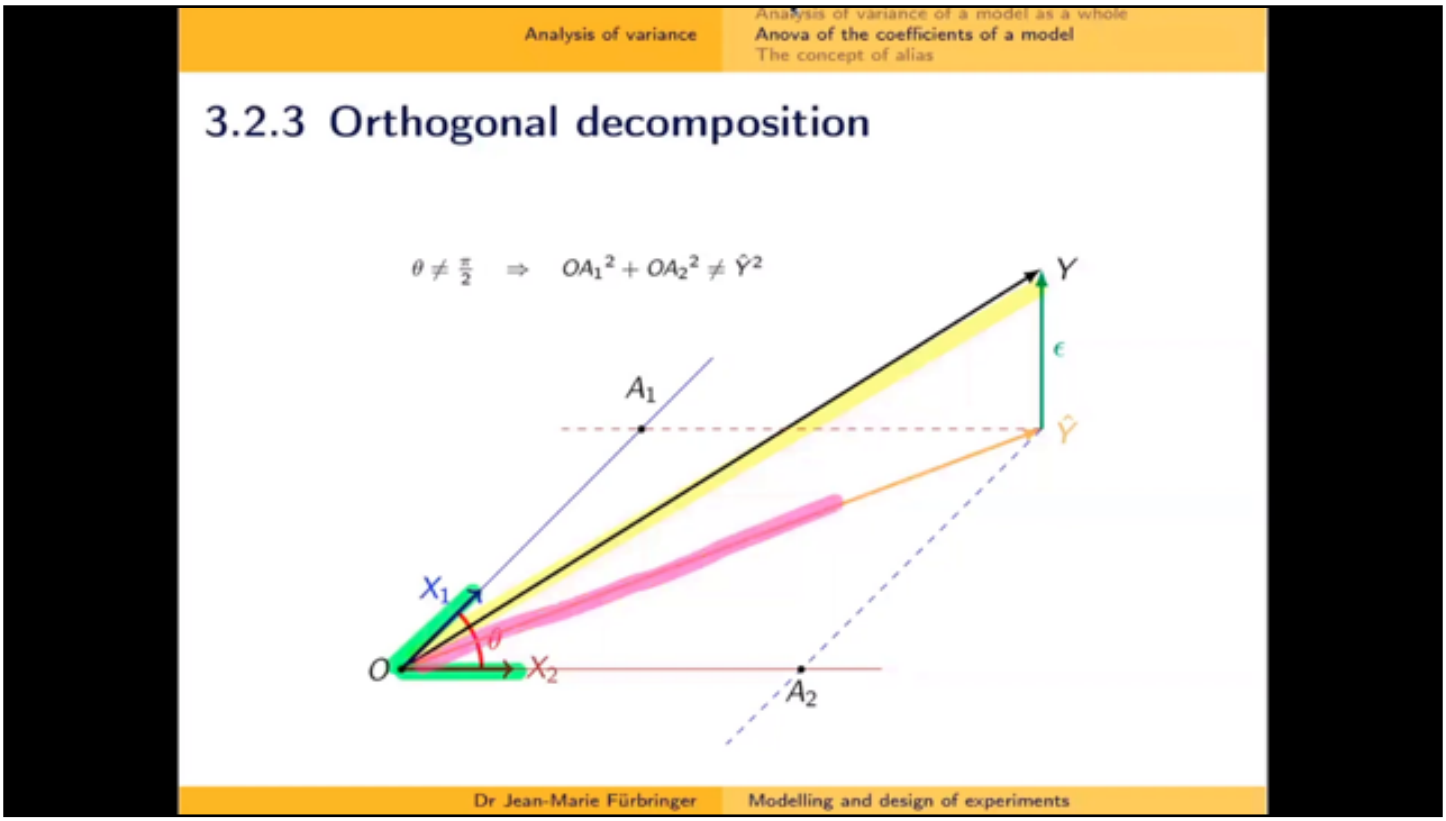
summary



because it's a projection so you cannot really see the angle. But I tell you that this angle is not the right angle. It's an angle theta, which is, I don't know, 70 degrees,

notes

summary

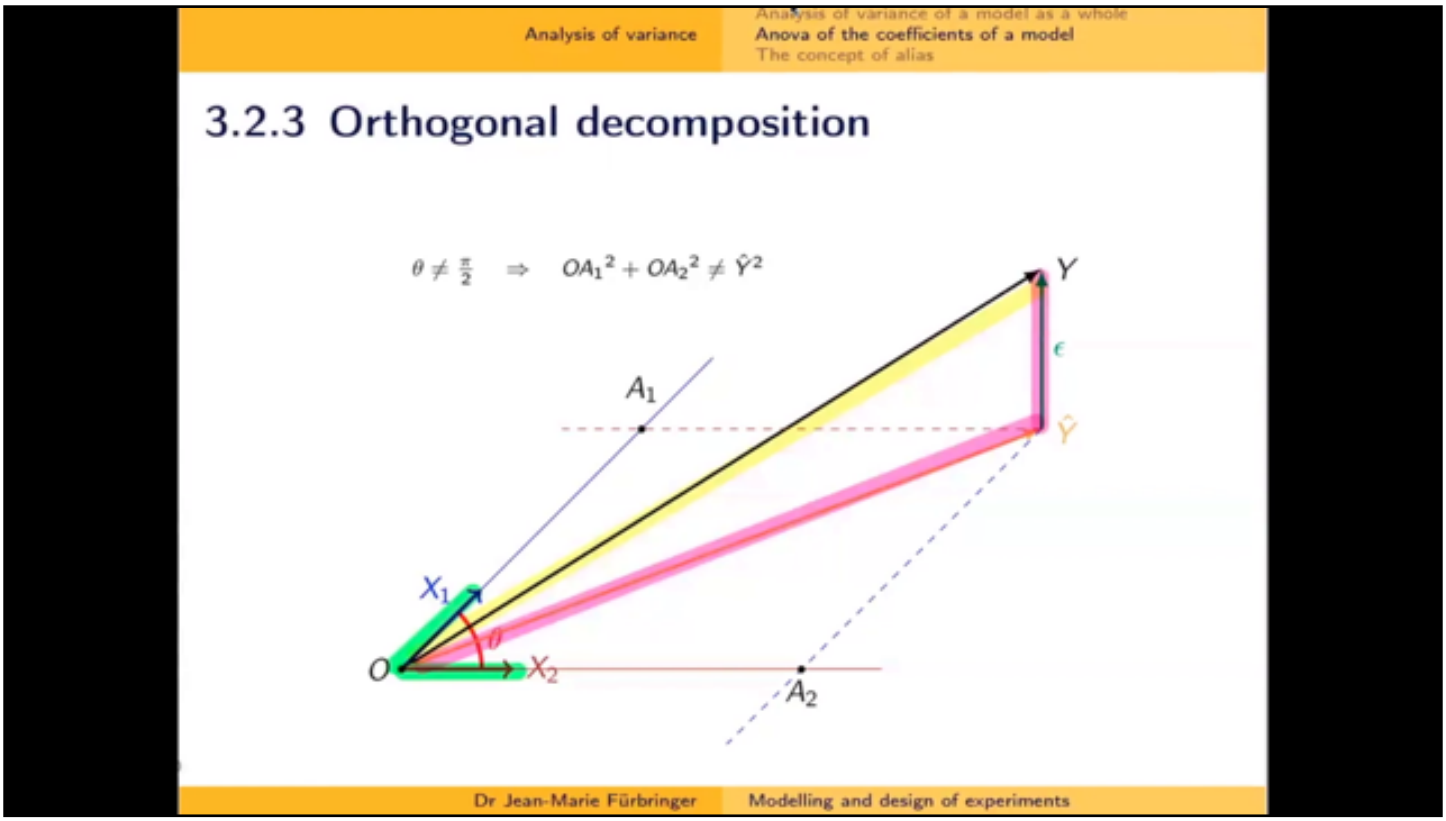


60 degrees, et cetera. Why? Because you were not free to make the measurement you wanted. You have some steel samples. This has decided the angle. And the angle is not right between two of the vectors that interest you. Let's say the constant and perhaps the carbon or between the carbon and the surface. That means that it's OK to consider a sum of square of your model, this part.

notes

summary

34m 25s

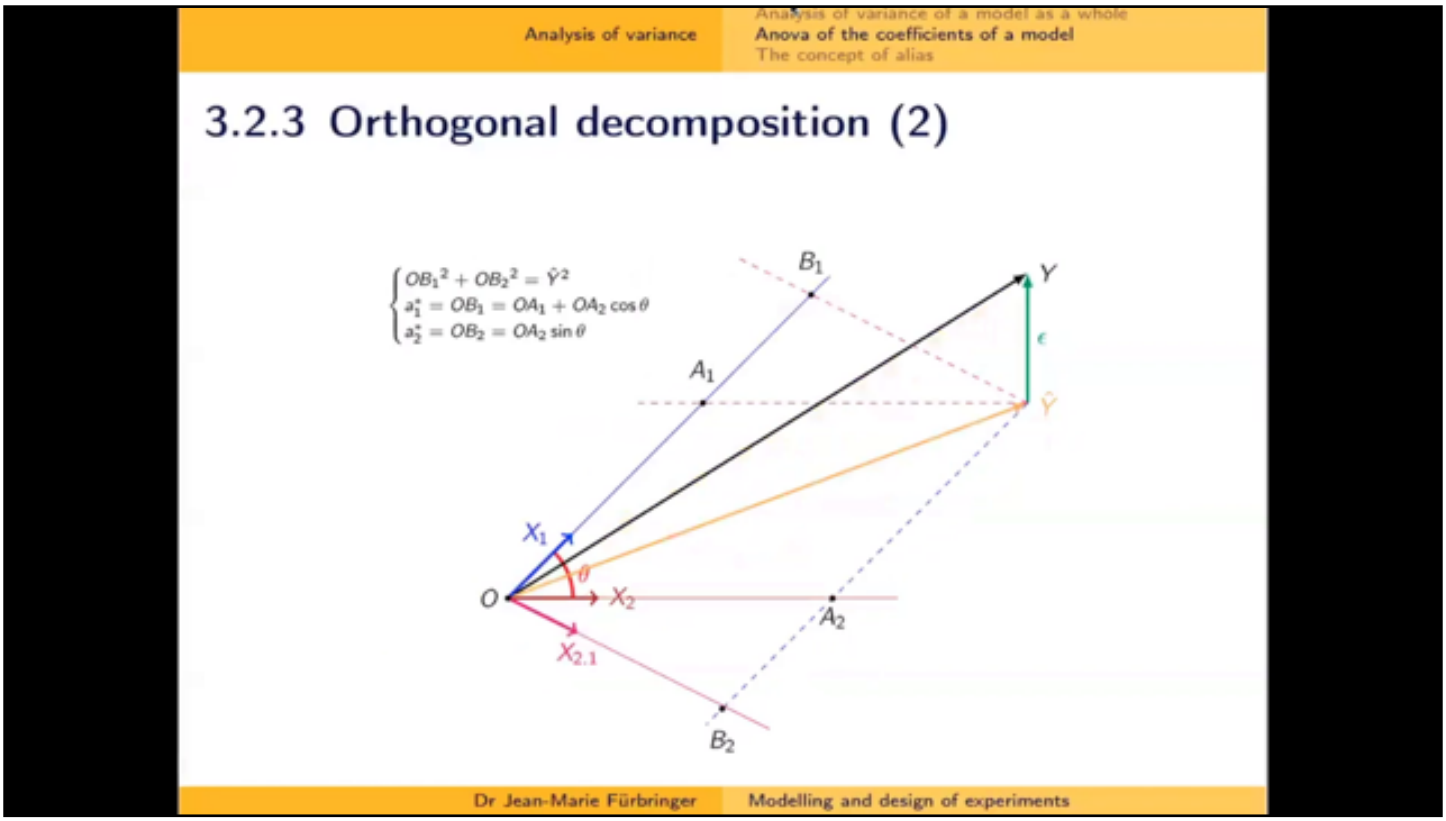


You can say that the sum of the model plus the sum of your residue is the sum of square of your data.

notes

summary

35m 2s

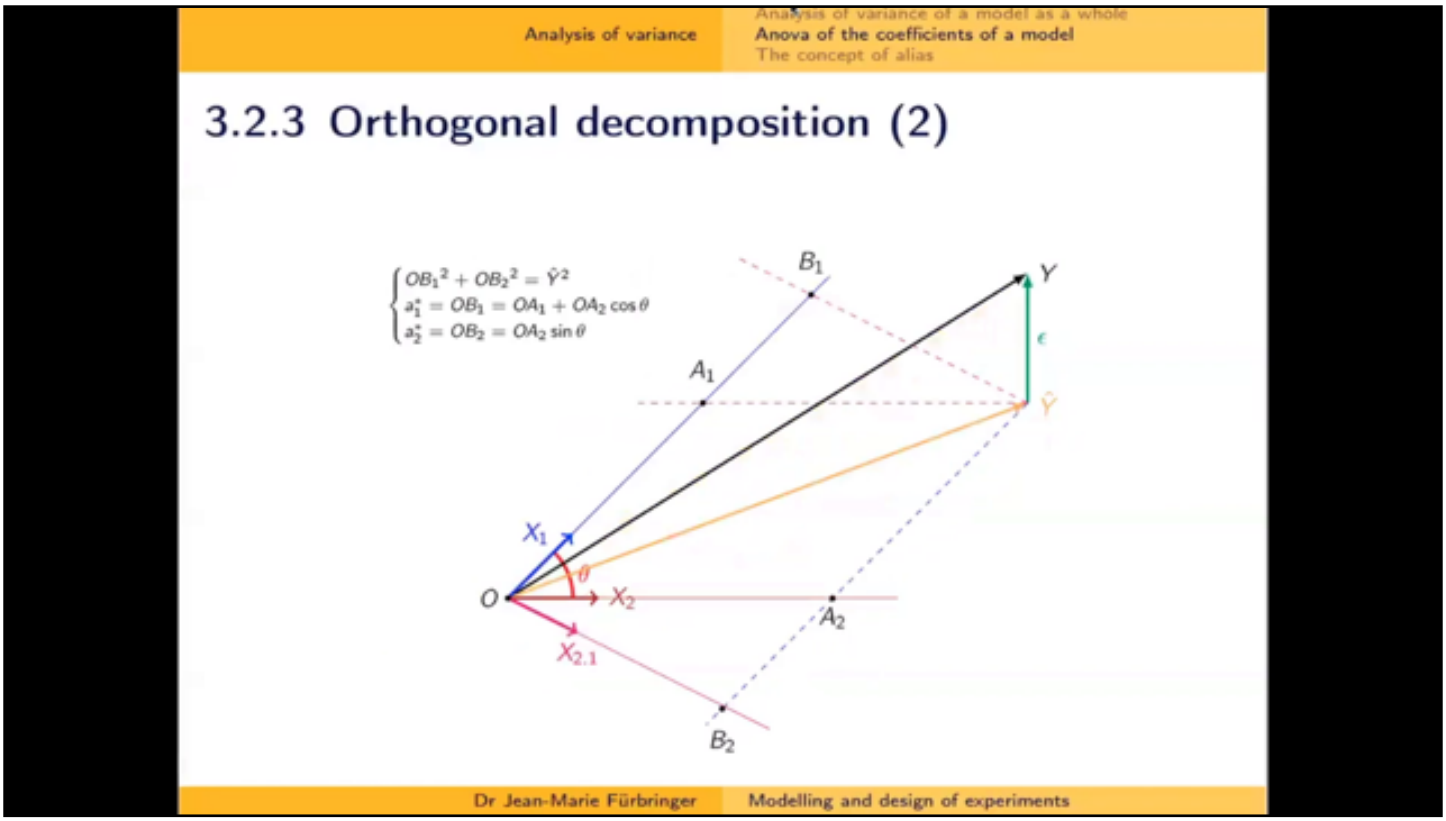


This is OK because here you have a right angle. But here you don't have a right angle. So when you want to calculate what's happened on your coefficient you have to make some correction for having corrected sum of square. You remember your geometry. You do that probably in geometry, analysis one, perhaps in physics one. You have to check forces and make sum of forces and things like that. So that means that you have to orthogonalize. So you have one vector x and you have to build a vector which is orthogonal to it. I call it x2.1 is the notation of geometry, the dots indicating the fact that you are orthogonal. So this new vector x2.1 is orthogonal to x1. And now I have to project my model on it. So that means that the value, in fact, when you correct your angle, the value for the first coefficient that was the value between O and A1 now is here. It's a new value that you have to consider. And the same thing before you have to make the measurement between O and A2 and now you have to consider that. If you want to decompose your vector in orthogonal components, so it's changing the value of your coefficient. So you don't change your model. It's just for making the anovar that you make this correction. Because in fact, when you have no non-orthogonality, you have a part of the sum of square that you are sharing between the two factors because you have correlation and you have to do something. So in this way, I will present you a new way after, but in this way, the first one gains the common parts. So the second one have only the part of the sum of square which is orthogonal to the first one. It cannot take profit of something which have been already count in

notes

summary

35m 14s



the first factor. We will not do that geometrically, just for explaining what we are doing. We will do it algebraically with algorithm. But this is what we have to do. It's why we have to correct. So we talk about corrected sum of squares. And this correction is related to the angle between my regressors, between my vectors. So this is the sum of square type one. I will present the first one, take all the second one, or the third one, or the fourth one is only what the other have not used already. They are common sum of square and the sum of squares is attributed to the first. So the order in which you analyze have an influence. You don't change your model. Again, the model is working well because the model is making projection. And for the projection, you can project on your vector without straight angle. But now you will change the sum of square. You don't change your coefficients. You consider A1 and A2 as your coefficient. You don't change your model. It's just for the ANOVA. Because we want a sum of squares at work. As I tell you, you are responsible of the sum of square. You have made measurement. Now you have to explain to the world, what is this sum of square?

notes

summary