Course material

Course:

ENG606 / PHYS 442

Video:

DOE_lesson5_part2_Anova

Concepts (extracted from automatically generated subtitles):

Sum of square of the individual vectors. Toxicity of products. Least square feet algorithm. Cocktail effect. Type of investigation. Products p. Typical model. Different coefficients. First product. Sum of square. Classical way. Main effect. Value of the concentration. Chemical model. Order of magnitude.

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

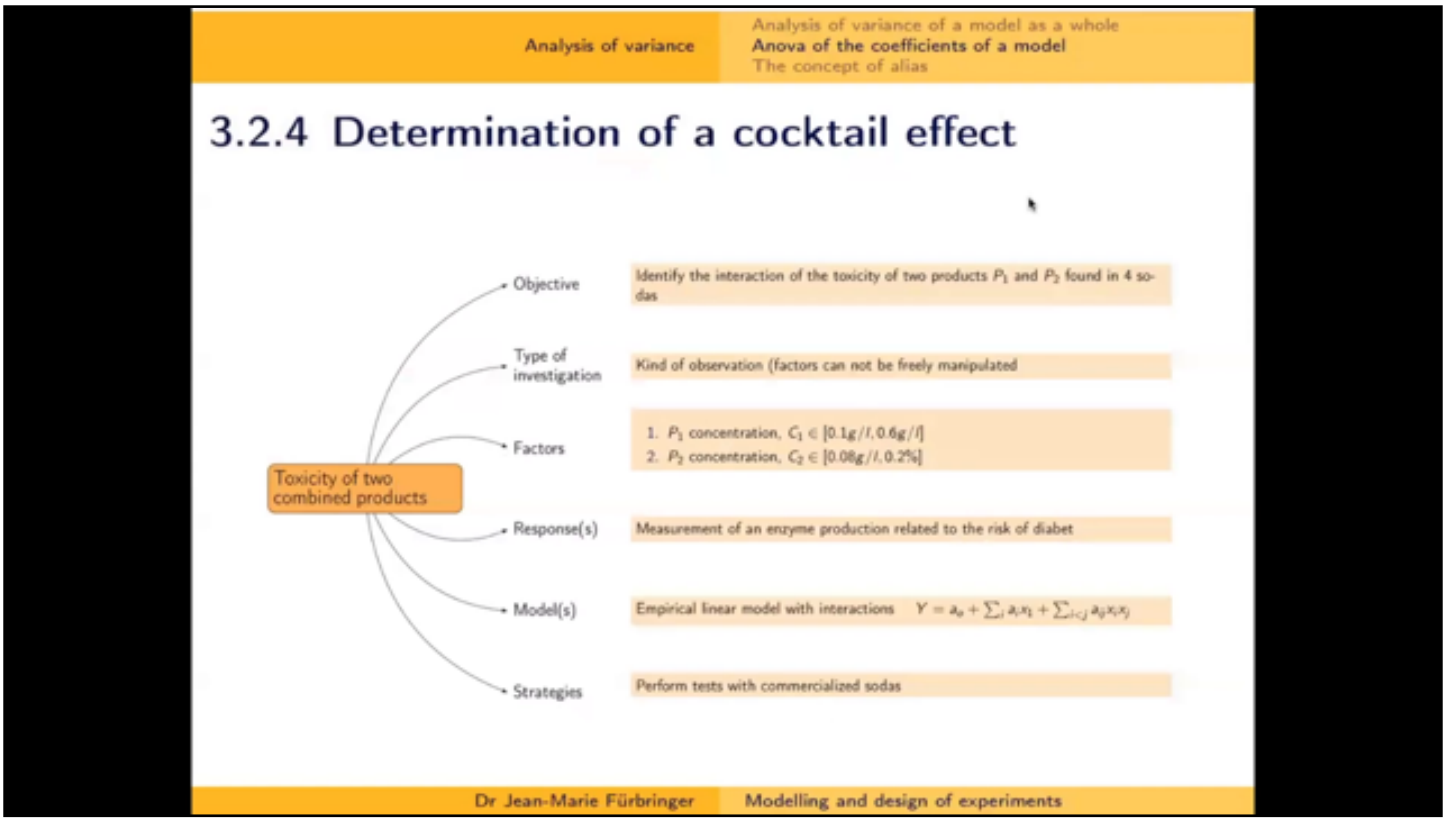
to video

...

notes

summary

0m 0s

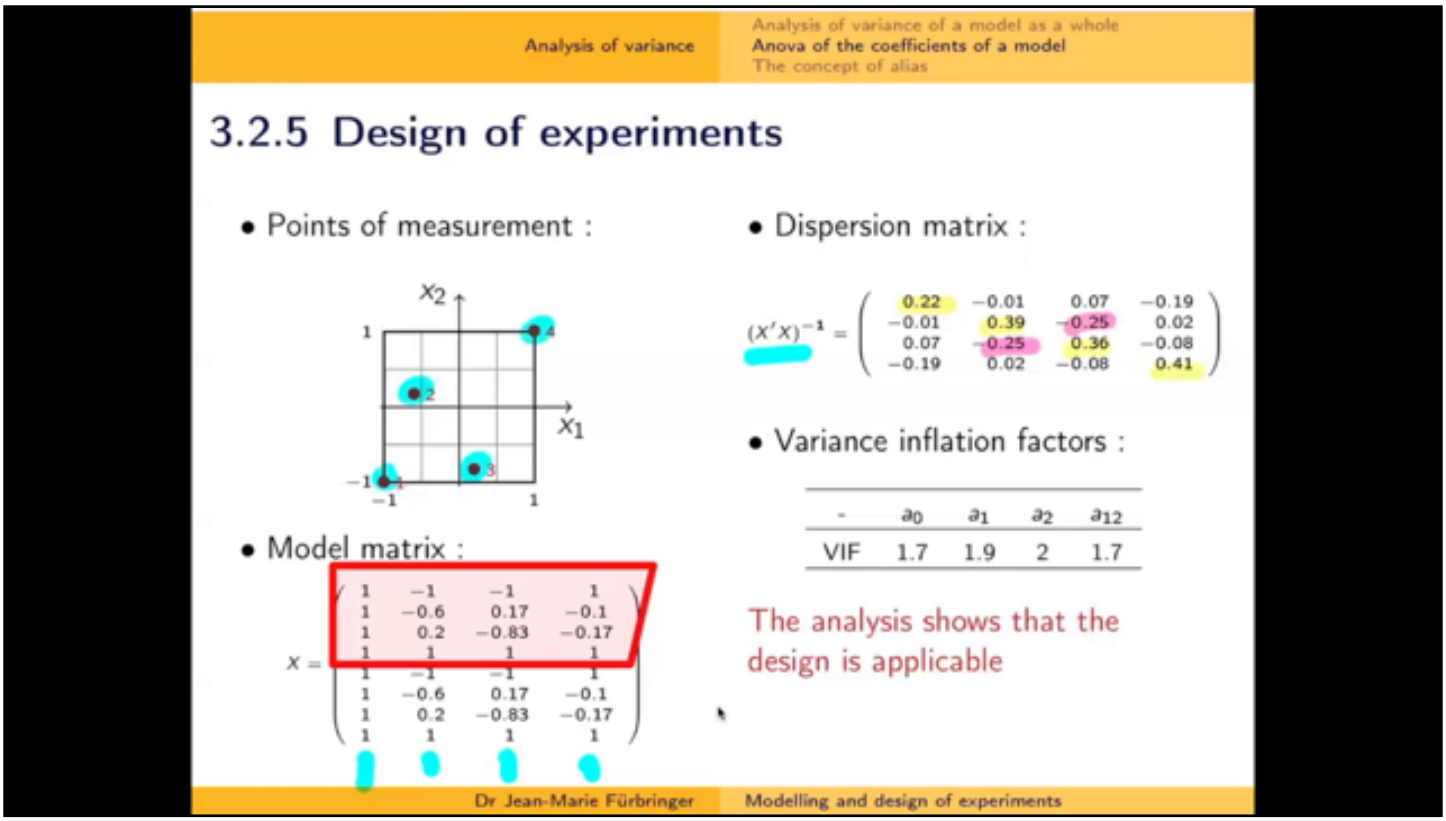


These subtitles have been generated automatically Okay, so this is an example, a cocktail effect, you know that in toxicity, this is a very

notes

summary

0m 1s

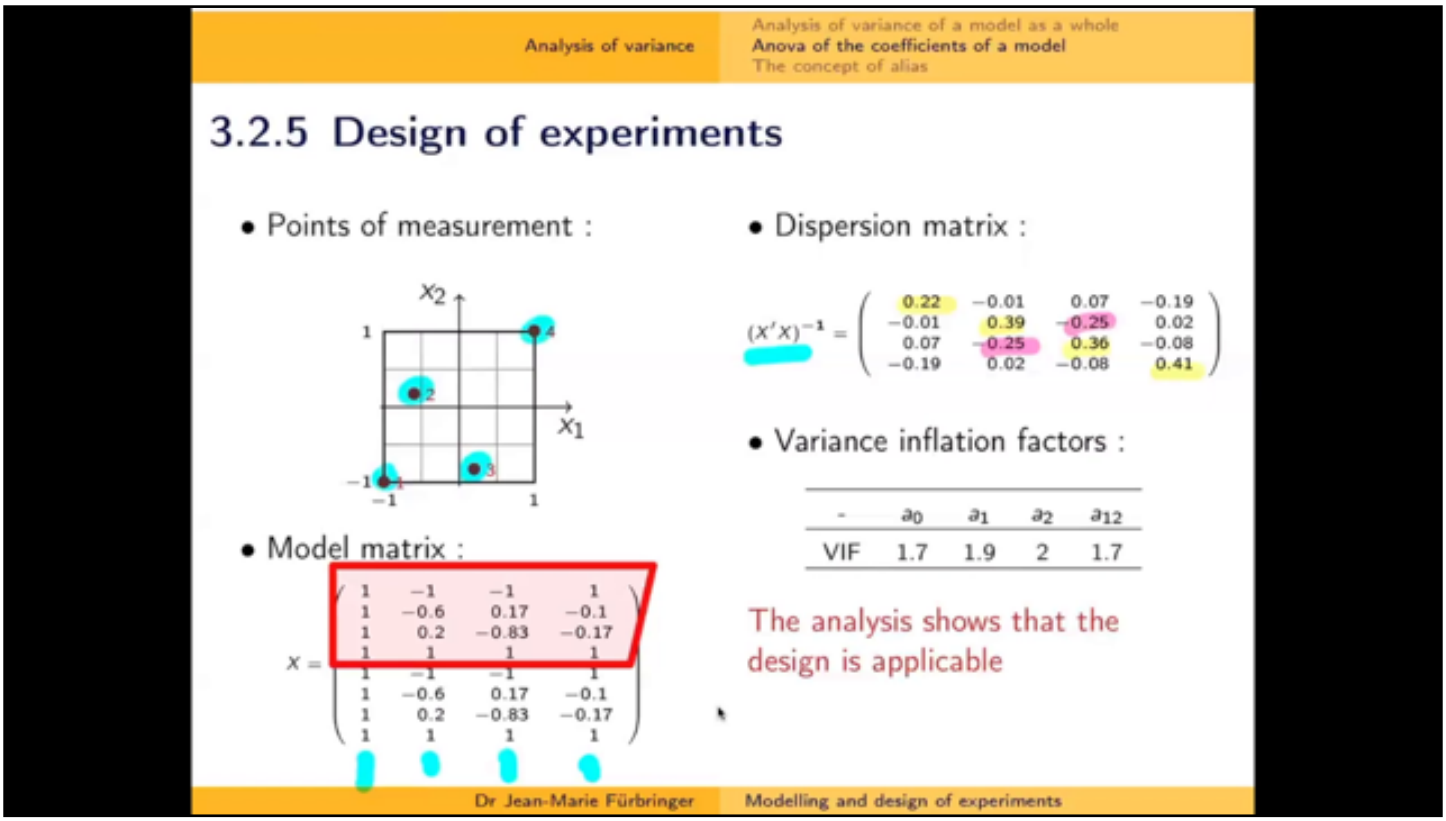


important thing. Usually toxicity of products are tested product by product and the idea is to make a small experiment to check if there are what is called cocktail effects or an interaction between the toxicity of two products. So again, it's absolutely fake data. So imagine that you are interested to checking if the product put in the sodas for replacing sugar are bad for the health. And the idea is to see if two products we call P1 and P2 have interactions that make them worse than each product alone. So the objective is to identify the interaction of toxicity of two products P1 and P2 found in four sodas. The type of investigation is observation. Imagine that in fact, I just starting your boss is not really convinced about what you want to do. So you do not have access to those products, but you have to use the products that exist in the markets. You can have concentration in those sodas before the P1 between 10th of a gram or a 6th 10th of a gram for the first product for the second product is between is a lot smaller bit eight hundreds of a gram per liter or the other sorry, there are an error is not a person. It's a 0.2 grams per liter. And the response in this case, imagine that you can make the test with cells. You can put the products or the sodas on cells. And after a few days, you are able to analyze the cells and to see the cells have a problem. And we can imagine that the analysis of the cell will tell you if eventually this product can cause problem to the to the house. The model, you have no typical model, chemical model, biological model. So an empirical model would work well. And as you are looking for a cocktail effect, it's evident that

notes

summary

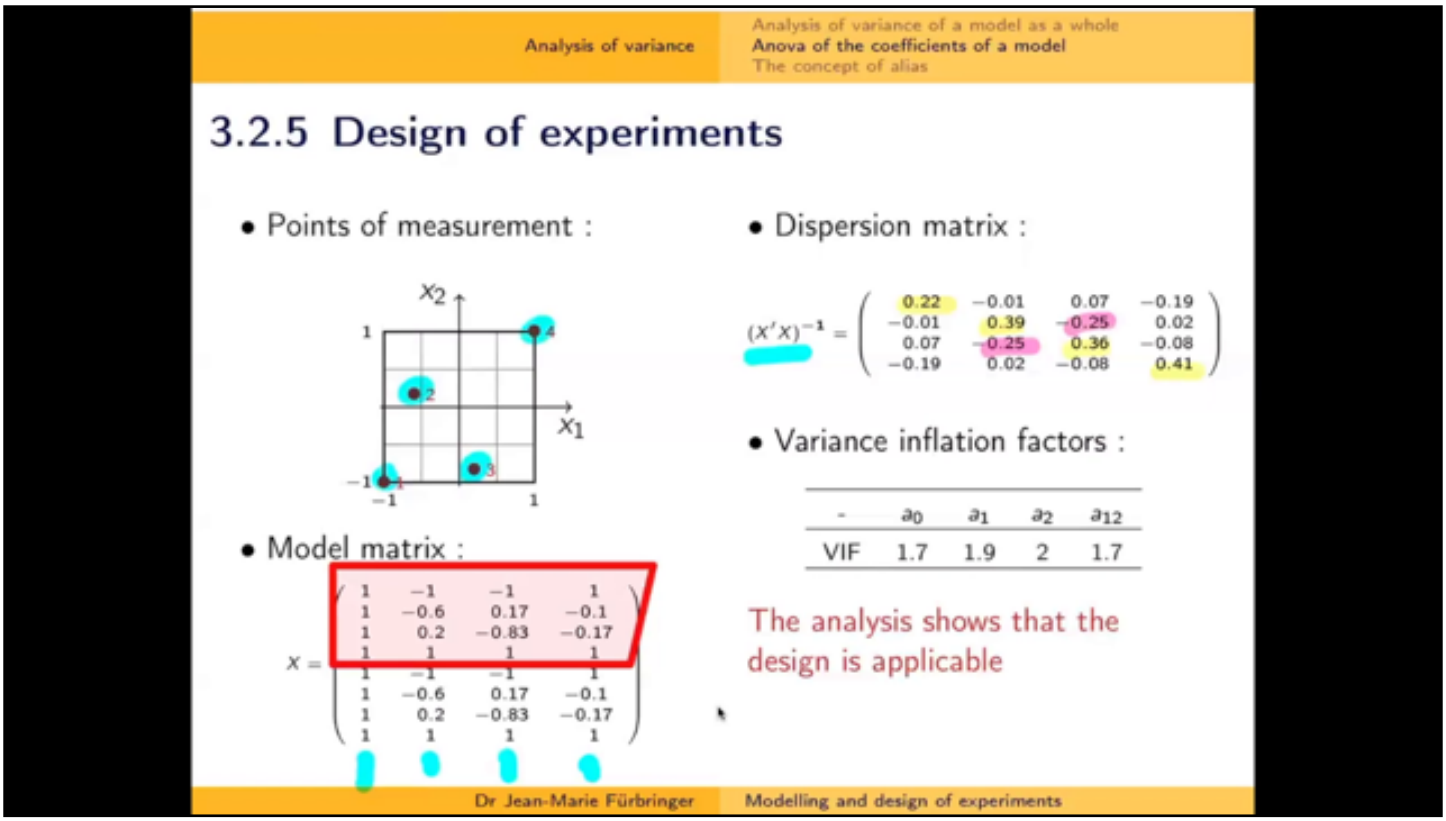
0m 5s



you are interested by having the constant, the main effect and the interaction between the product one and the product two is exactly the objective of your analysis. And your strategy, as I said, you do not have a lot of money for doing it. So you don't have access directly to those products as a given purity for making the experiment. So you use sodas. We can also say that you want don't want to test the product pure. You want to test it are they are used with other products. So you go in the market and you find poor sodas with the different concentration. And you see they are distributed in this two dimension plane, not orthogonally, because you have no choice. You are obliged to use a soda that exists. It's this happened many times in the experimental situation. So you have here the model matrix corresponding of making the measurement. I have four points, but I'm doing two times. So I have a model matrix with eight row. I'm looking at this time. I have analyzed the eight rows in a while. So you have one column for the constant. I have one column for the product P1, one column for the product P2. And I have one column in my model matrix for the interaction, which is the product of the column P1 and the column P2. And you can see that if you look at them, we have two times the same matrix. They have repeats. So we have two replicates. As I mentioned in a comment to some question outside of the video, that when do we use for analyzing, we usually use with only one genuine replicate, only one replicate to see if the design is OK. When I'm making the analysis of my data, I make with my replicates. So now I am able to calculate the dispersion matrix. And

notes

summary



you see that it's a non-orthogonal situation as my dispersion matrix is not a diagonal matrix. I have standardized. I didn't mention it, but the model of matrix was standardized. The value of the concentration was a little bit different, but I have standardized normalized. Sorry. I don't remember if somebody make me. I believe you make the comment. So I check and you write. We have to use normalized for what I'm doing. And standardized is usually when you are going to the average and the standardization, which for me it's disturbing because normalized don't use normal distribution, but it's OK. From now I will follow what is the classical way of calling those things in the statistical test. Now that is a dispersion matrix. You see values of 22%, 40%, and for the different coefficients. So this is the amount of the experimental variance that will be transferred to my coefficients. But you see that we have values outside of the diagonal. They are not quite high in comparison to the diagonal. We usually accept when they are order of magnitude smaller is what we appreciate 10 times or 20 times smaller than the diagonal. It's not the case. You see that here they are quite one important correlation, but it's OK.

notes

summary

3.2.6 Inference of the coefficients

After the experiments

Experimental data :

Expériences	1	2	3	4
Y(set 1)	80.4	70.8	67.1	270.0
Y(set 2)	89.7	58.9	53.7	275.3

Model coefficients :

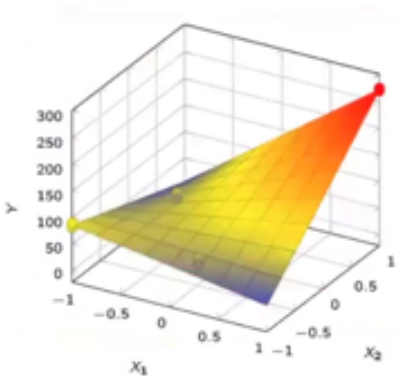
-	a_0	a_1	a_2	a_{12}
α_i	97.7	52.6	41.2	81.2
α_i/α_o	-	54%	42%	83%

Estimator :

$\hat{\alpha} = (X'X)^{-1} X'Y$

Model :

$Y = 97.7 + 52.6x_1 + 41.2x_2 + 81.2x_1x_2$

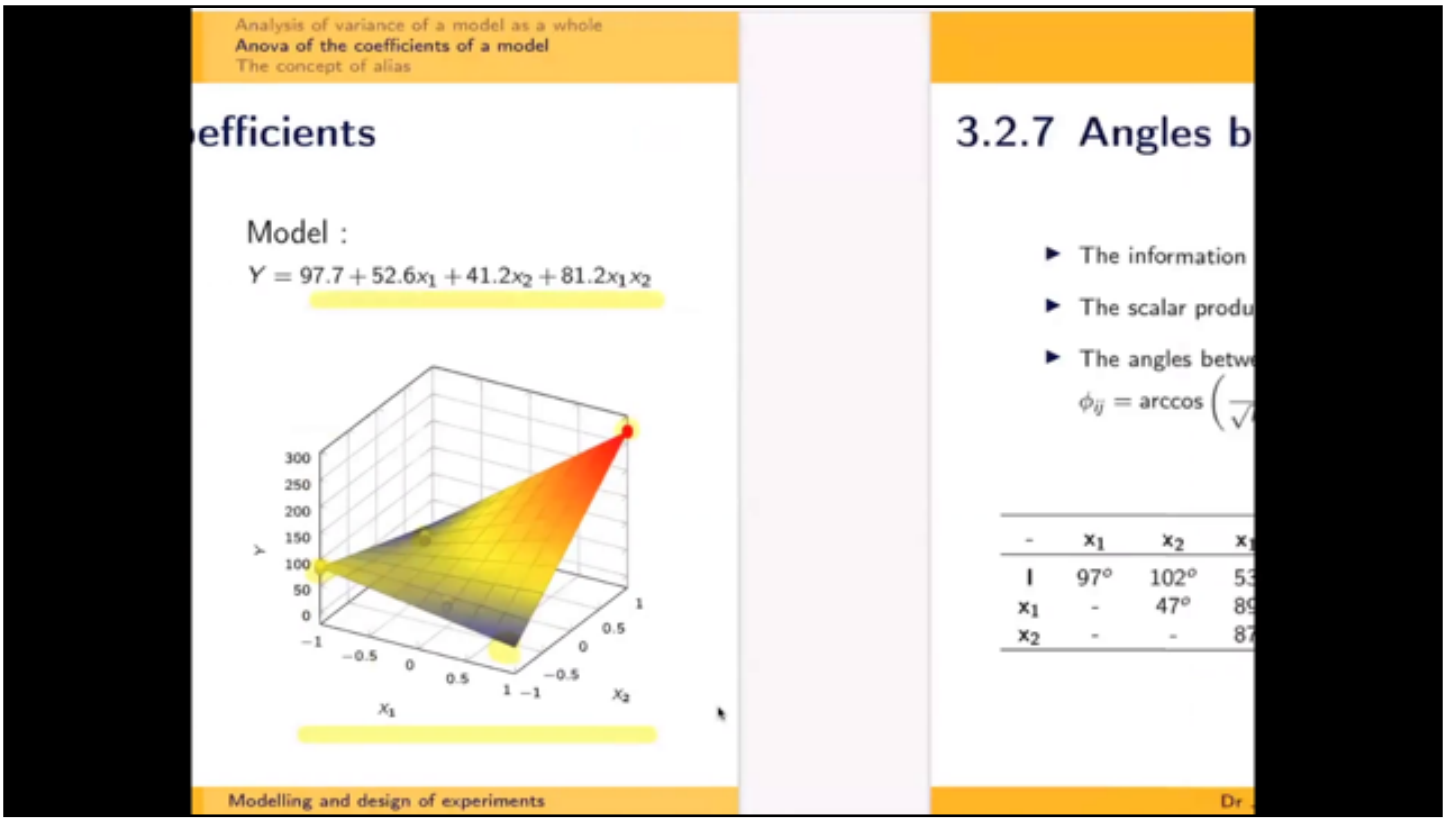


We can make the calculation for calculating the model is not in itself a problem. It could be a problem with the ANOVA, but it's not a problem for calculating. So I'm analyzing using the least square feet algorithm. And so I'm able to produce some coefficient. So the variance inflation factor, you see that even as I have extra diagonal elements, the variance inflation factor are not so bad.

notes

summary

6m 25s

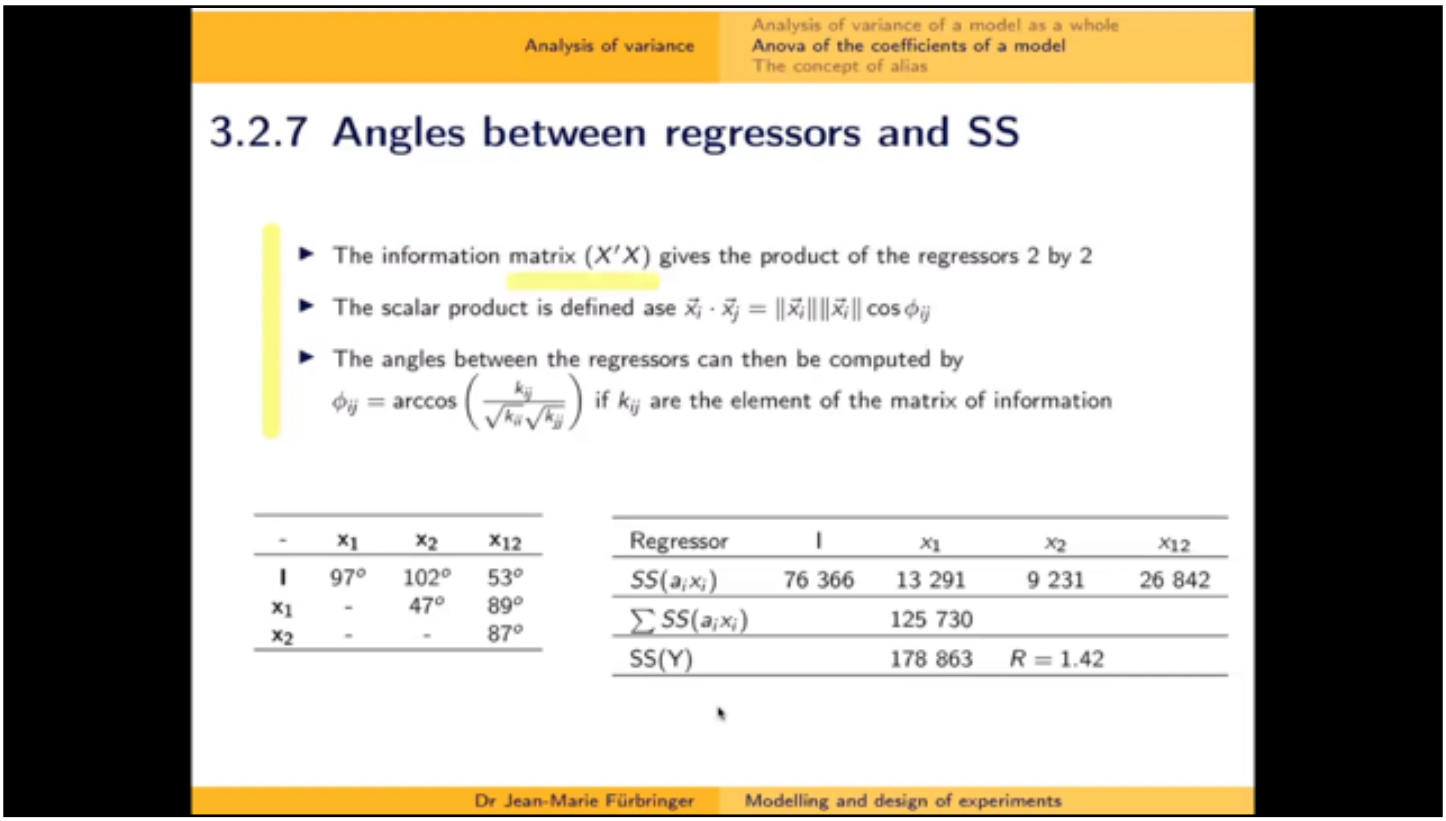


data and then I'm able to calculate the coefficient of my model. I have a constant of quite around 197.7. I have an effect for the factor one for the product 152. I have an effect of the product 241. I have an effect of interaction, which is even higher than the effect. So the unit forgets the unit. It's just fake data. Just consider that it's the importance of the damage that you can do to your health with those products. I'm calculating here something which is quite interesting is the relative coefficients. I'm dividing the coefficient by the constant. So it's giving me an angle to understand if my effects are important or not in comparison to the constant. So you see that with the product one, we can change the toxicity of 50%, 54%. We can change the toxicity of 42% with the product two. And in the instruction, we can change it of 83%. So it's a way of understanding if you have high effect, low effect in comparison. This is the estimator I've used. This is the model that I have used using the coefficients. And I even have made a model, a graphic of the model. I have two dimensions. So it's possible. And I even have reproduced the measurements. You see that the replicates are quite OK and quite close. I have a model that could be quite OK.

notes

summary

7m 9s

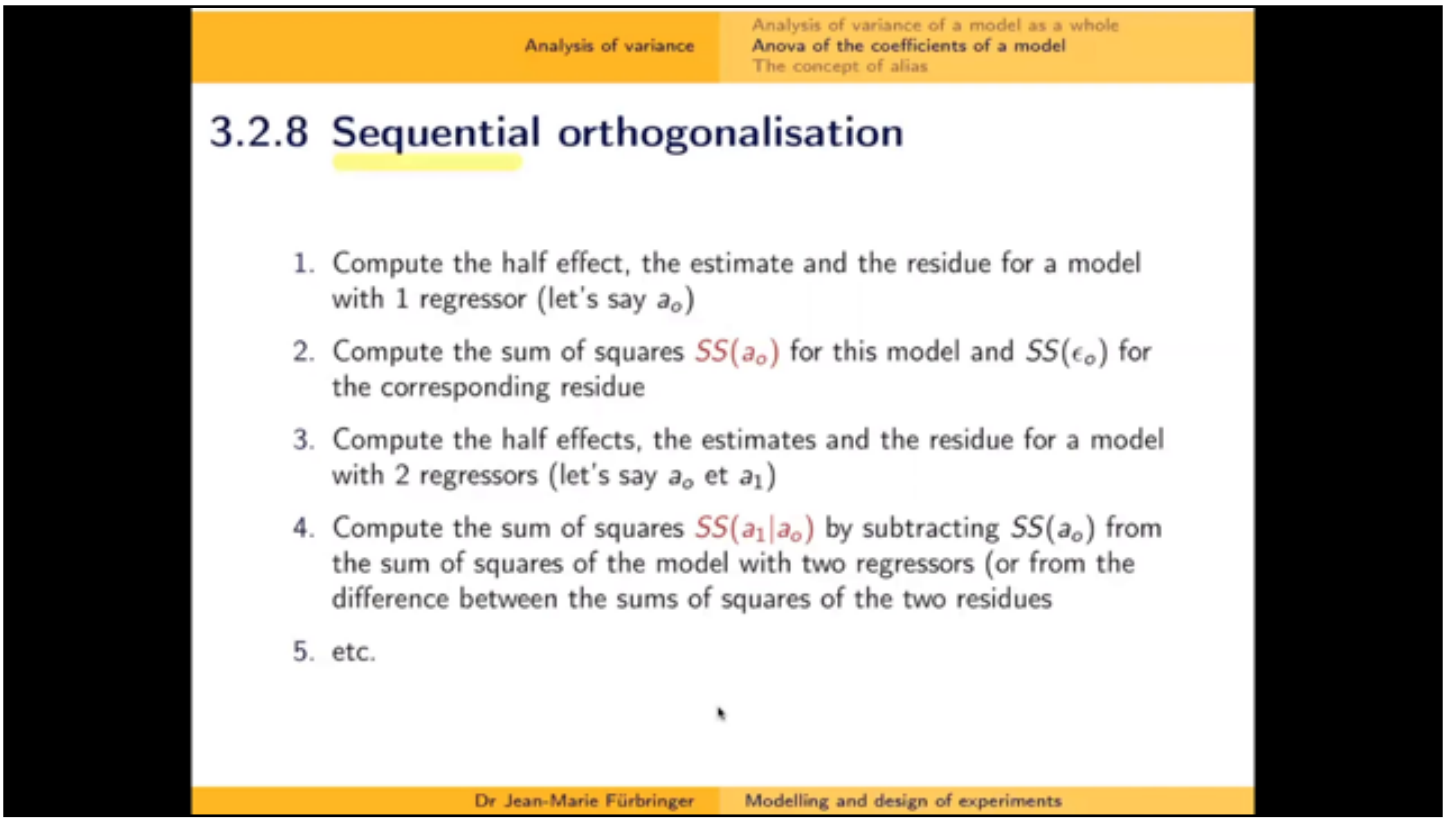


So I do that for trying to understand how we do the ANOVA of this situation. I'm able to calculate the angle between my vectors. The vectors I'm talking are the vectors on which I'm projecting my vector of results. And I'm trying to see if they are orthogonal or not. I know they are not orthogonal because my dispersion matrix was not diagonal. But what is the angle between them? So here is a small rationale.

notes

summary

8m 59s

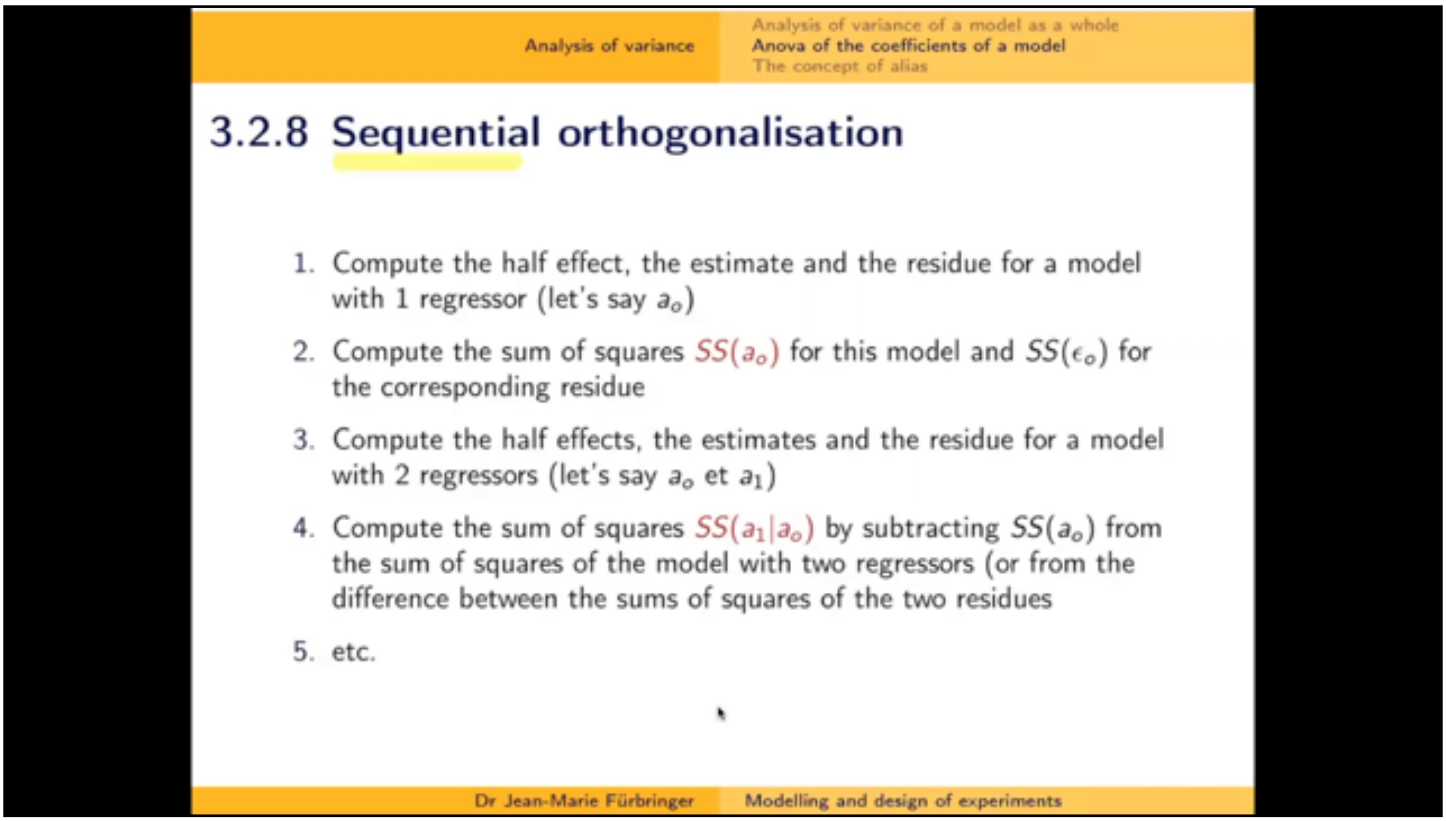


So I use what we call the matrix of information, which is the product of all the vectors 2 by 2. So this is related to the cosine of the angle between the vectors. So I'm able to evaluate my angle. I just do that for the course. We don't do that usually. Just for illustrating my course, just showing you that they are not orthogonal. But usually we don't do this type of calculation. But I'm able to calculate the angle between two of the vectors by calculating the arcos value of the element of my matrix of information. I call the element of my matrix of information k. And so I'm dividing the k, i.g. by the roots. It looks like the matrix of covariance. Now it's very close to the matrix of covariance. But the matrix of covariance is the same calculation with the inverse of this matrix, the inverse of the matrix of information. So it's a sort of inverse of correlation between my coefficients. And you see here the angle, 97% between the constant e for identity. I call the columns of identity sometimes e because it's 1, 1, 1, 1, 1, 1. So I call it e. And between the vector for x1, the 97 degree, 100 degree between x2 and 53 degree, etc. Just for showing you that I have very different angle in my situation and it's really not an orthogonal situation. And for showing you that the problem could be dramatic, look the sum of square of the individual vectors. So I have 76,000, 13,000, 9,000, 26,000 as the sum of square for each of those vectors. And if I make the sum of square of them, I have 125,000. But when I'm calculating the sum of square of my data, I have 180,000. You have a difference of 40% of the length of my vectors. So it's really not orthogonal.

notes

summary

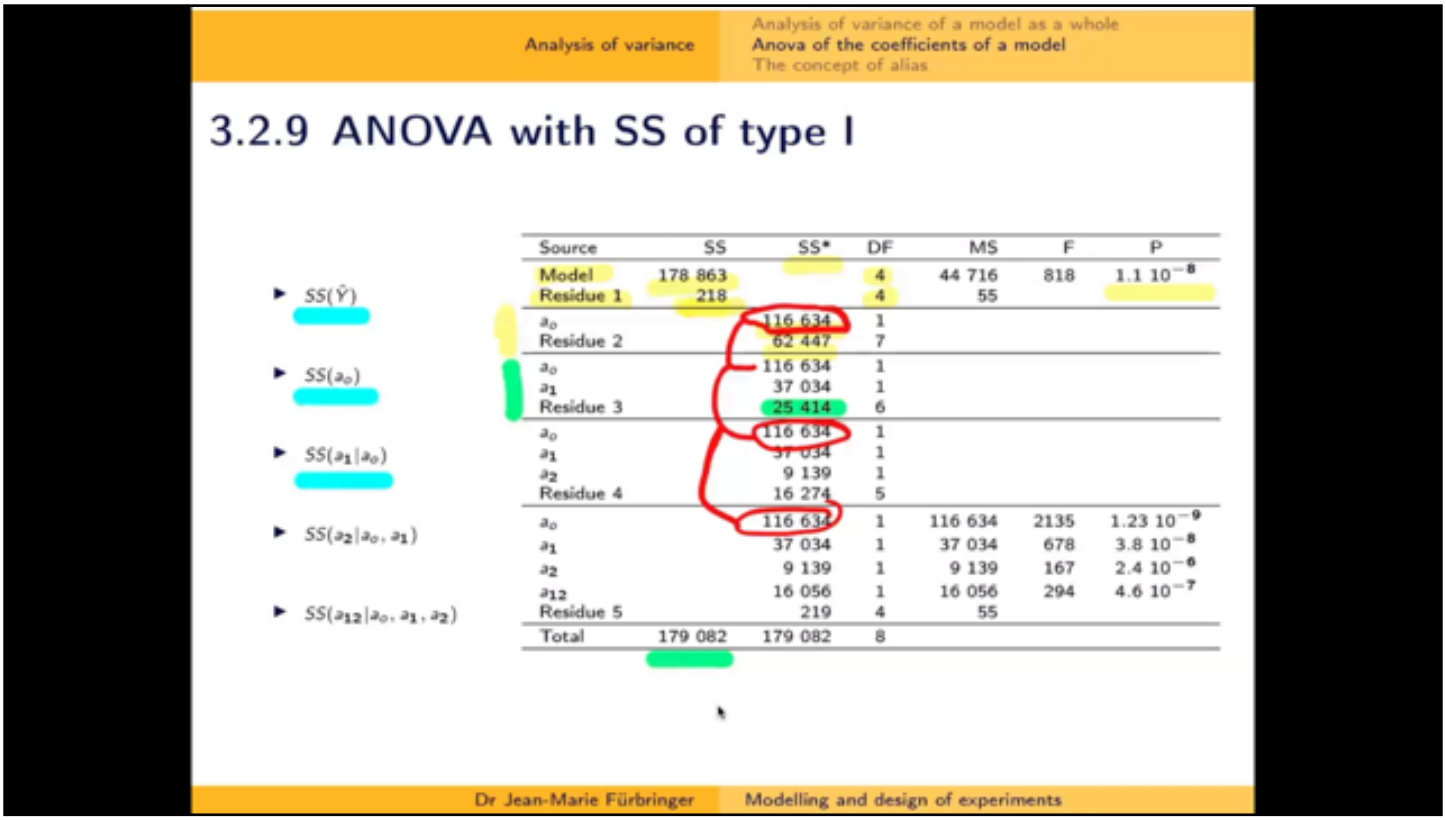
9m 37s



And if I'm making my calculation, my p-value would represent nothing because you don't know what you are corresponding. And you have some confidence which is shared between different coefficients. So this is just for showing you the problem is severe. If you don't want to enter in this calculation, don't enter in the calculations just for showing you the problem is severe. If I want to calculate a sequential orthogonalization, saying that the first one will take the common sum of square, the second one will have only the orthogonal part of the sum of square, which comes from the first one, etc, etc. It's a sequential way of doing things. And in this, the problem is that the order in which you consider your factor is important.

notes

summary

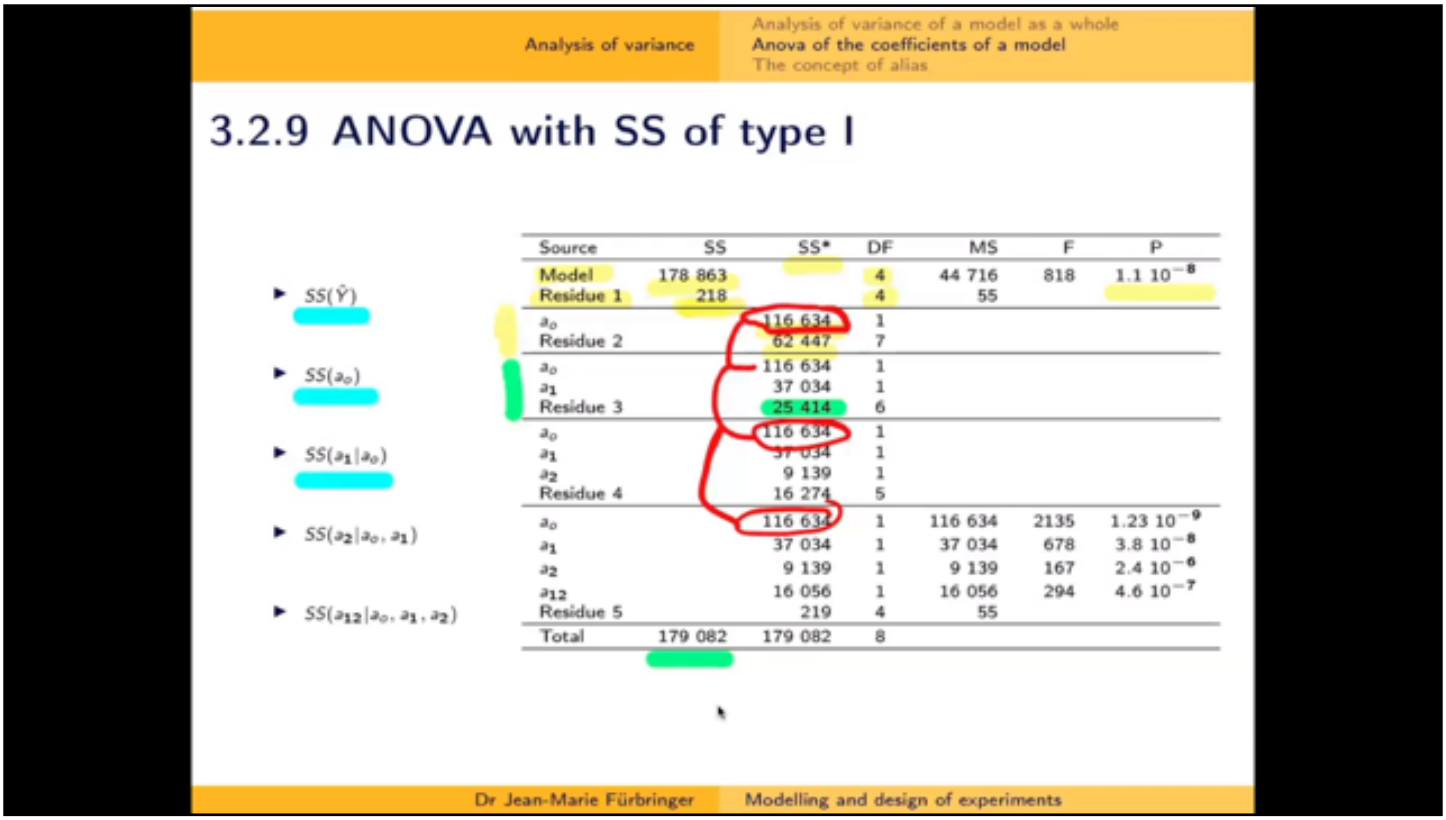


If one comes before another, it's taking the sum of square, which is common. So how do we do that? We compute the half effect. We have done it. We estimate, we have the estimate of the effect, the estimate of the value, what we call estimate. You recalculate the value, the theoretical value that your measurement should give you for identifying the residue. And we do that for the model with only one regressor. We forget the other. We calculate one only. And after we calculate the sum of square of this coefficient calculated alone. And we see the notation will became important now. SSA0 is the sum of square of A0 alone. And for this model, we can calculate a sum of square of the residue when I'm only calculating the first coefficient. Only one regressor. I forget I have the other ones. And after I will calculate the half effect for the model with two regresses. Remember, I have in fact four regresses, constant, first effect, second effect, interaction. So I'm calculating first the constant alone, forgetting the other. After I'm calculating the constant and the first main effect. And after we'll calculate the three first and after the fourth. And I will keep my calculations. And each time, so for the second one, I'm calculating as I'm able to consider the change of my residue will change. And each time my residue will be diminishing. When I'm integrating more coefficient in my model, I will have my residue diminishing. And I will consider that the diminishing in my residue is also diminishing in the sum of square of my residue. This is what is attributed to the new coefficient. It's what I call it sequential. Buckle up because this slide is probably is considered by the students the most complicated slide in my course. I find it easy because I made it, but I understand from

notes

summary

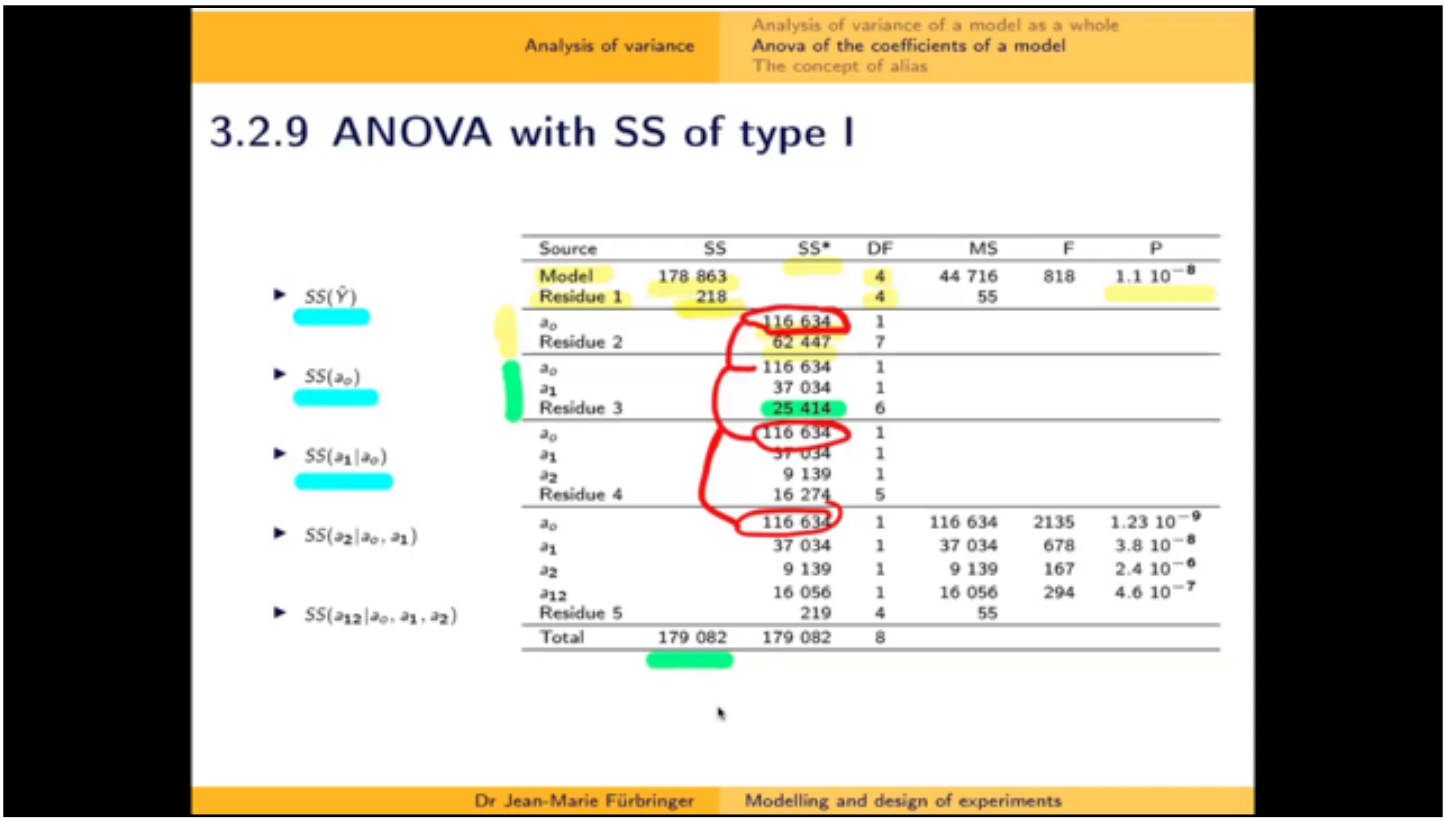
13m 1s



outside it could be complicated. So what I just explained you, the sequential, I have done it step by step in this situation. So first part here, I have my model and my residue. OK, I consider my model with my four coefficients, my model, my final model, the one I want, constant, two main effects, one interaction. Four degree of freedom, eight measurements, so four residue. And I'm calculating the sum of square of my model and the sum of square of my residue. And I'm able to calculate a p value for my model, 10 power minus 8. So in general speaking, my model is good. Now I want to dig within my model and see what is good and what is less good in my model. Typically, somebody would like to criticize my cocktail effect. My coefficient of interaction is valid, not only my model as a whole, because it's clear as you increase the numbers of coefficient each time your model is better. You also have a confidence interval of your coefficient as an argument, but I would like to use the argument of the sum of square in this ratio. So as I said, first part, I consider the constant alone. So if I'm considering only a constant, my A0 will be the average of my measurement and it gives me a sum of square of 116. And see, I create a new column in my ANOVA table with a star, corrected sum of square. You will find this notation in book of statistics, the star meaning corrected sum of square is not my pure sum of square. If I'm analyzing all my coefficient together, it's corrected sum of square. I have a residue and you can see that the residue here, 62,000, is a lot bigger than my residue when I'm considering my four coefficients. So now what I will do, and this is

notes

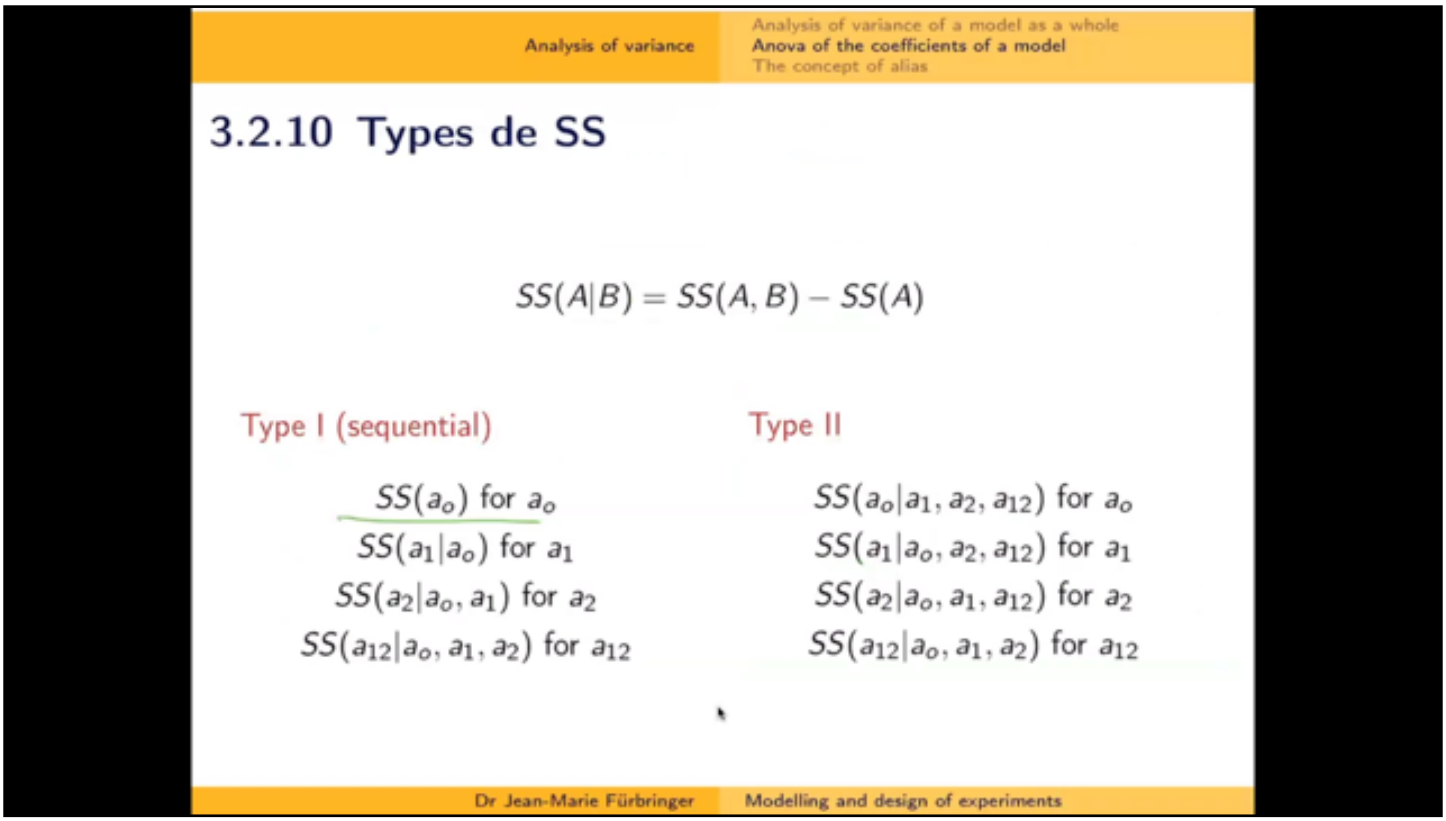
summary



the trick, the way of doing it, I will consider that this is the corrected sum of square for my constants and I will keep it. You see, each time I put, I will go step by step, but each time I will consider now my sum of square for the first coefficient for each step as if it was alone. You have to take the sum of square for him, and now the rest of sum of square will only be the orthogonal part of that. So when I'm going to the second step, now I have a model with two coefficients, a constant and the first main effect. So I'm calculating the model, I'm able to calculate a new residue. So now the residue is now 62,000 is 25,000. And the difference was before I attribute it to the first effect because this is a part because I know that 25,000 is orthogonal to my model. So I want that 25,000 plus 116,000 plus what I don't know yet is the sum of square of my measurements. So Anna, I'm just doing that step by step. So for here, I'm calculating the sum of square of my model, my full model, my model with the four coefficients. Here I've calculated the sum of square of a zero alone.

notes

summary

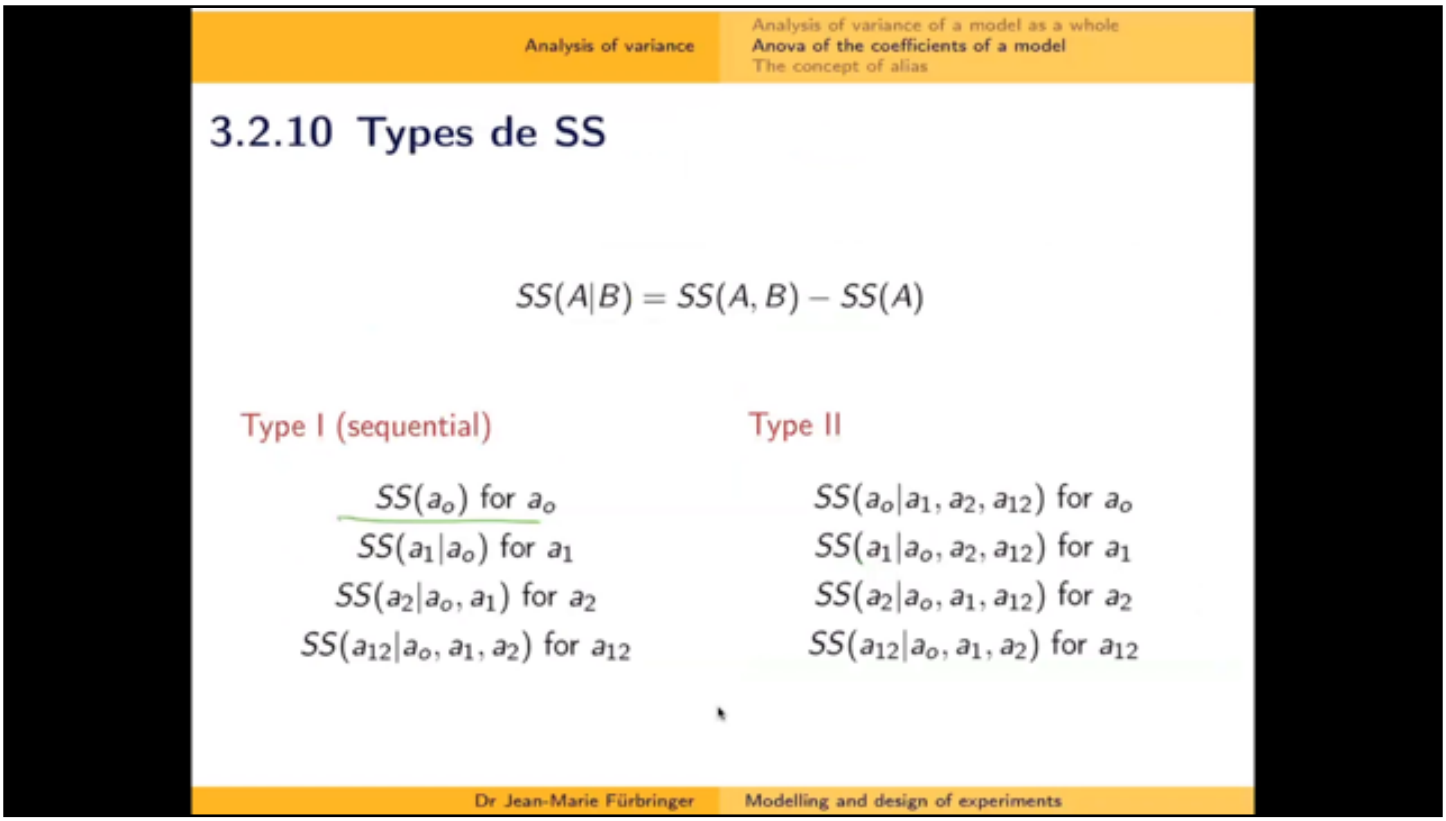


And now I'm calculating the sum of square of the first main effect, a one, knowing the bar a zero. So the notation means that that I'm calculating the sum of square after I've already subtract the common sum of square coming from a zero. And so we have used this notation of the probability, dependent probability of the vertical bar, meaning knowing. And after the next step, I will calculate the sum of square of a two, knowing a zero and a two. And at the end, I will calculate the sum of square for a one, two, the instruction, knowing a zero, knowing a one, knowing a two. And I just do like that each time, calculating my sum of square. And if you see here, I have calculated the sum of square for a one. And after I keep it 30,000, I keep it in all the other steps. So you understand that the last one is the last of the family. You don't have to eat too much because all the common sum of square have been attributed to the other. So it's not really fair, or at least you understand that the order in which the coefficient have been presented, if you made a model constant, instruction, and after main effect, the result, the p-value would have been different. So this is a sequential way of making a sum of square. It's also called type one. Remember in type one, the order in which you put your coefficient as an influence on the p-value. If they are all OK, it's perfect. The problem is one is not OK, and you say, OK, I don't consider this instruction. That's depend to your job. Statistician will say, OK, I prefer consider main effect and after instruction. But perhaps my objective was to prove that I have a cocktail effect. So I'm interested to have all my coefficient

notes

summary

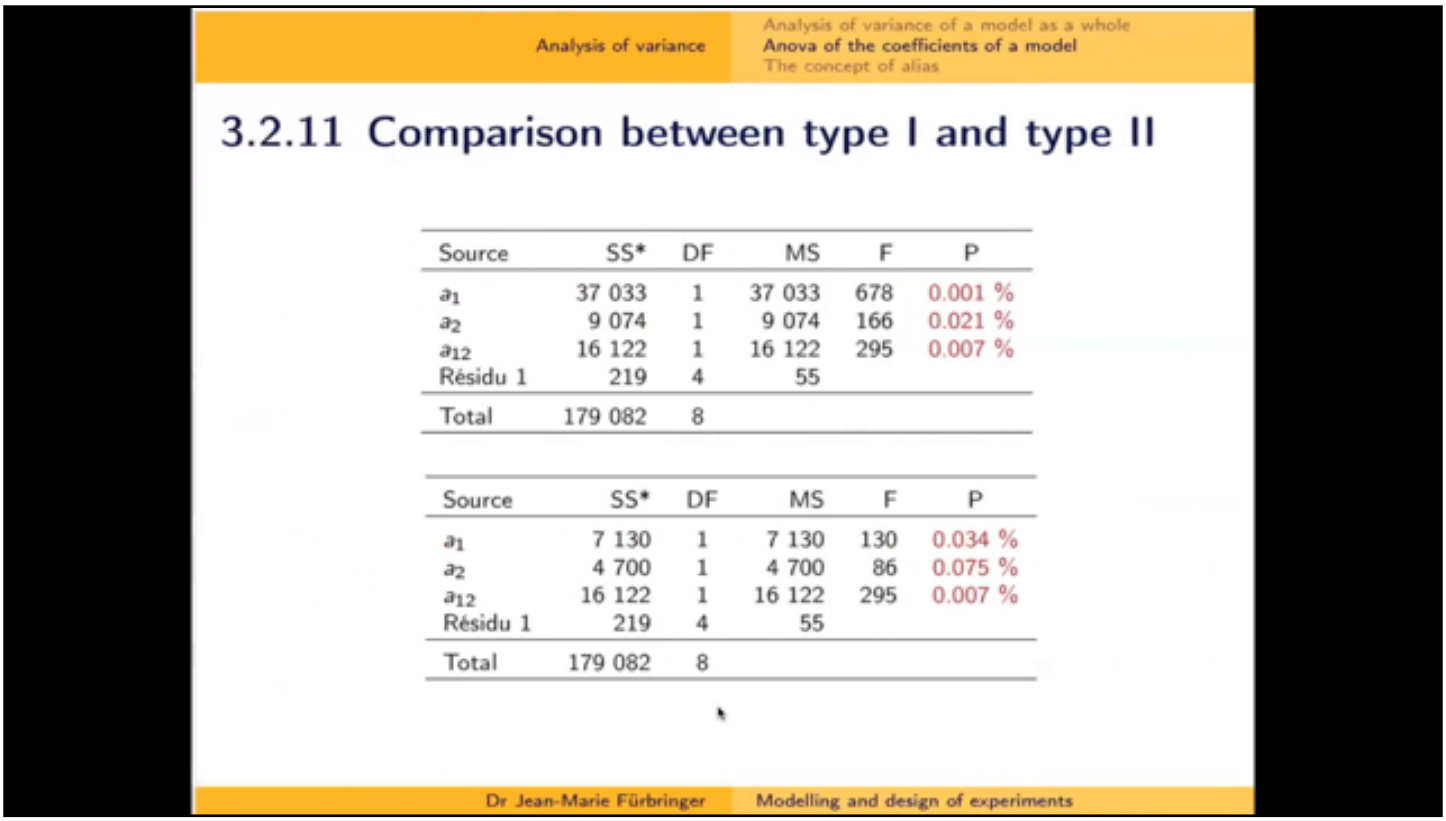
19m 25s



and also the instruction valid. I could play on the order, which will not be fair. People will criticize me, so I will present you the type two. Here you have the final p-value I can calculate it at the end. You see that all coefficients are OK, and a lot of people publish that like that and don't make them. This is distinction. But in your algorithm, you will see for ANOVA, you have to say if you want to type one and type two. So I don't know if you knew it, but now you can make the decision knowing type one, the order in which you put your coefficients as an importance. And the first one is advantage in comparison to the last ones. The sequential, we have a sum of square to consider. First, the sum of square alone.

notes

summary



The second sum of square knowing the first one, knowing the first one is the second one, knowing the three first. So a fair way of comparing things if you have really different factors. So when you have instruction in the main effect, we can we can argue because statistician will say that instructions are less important than main effect. The category of models, the hierarchy of models, they will stick to that. But in other situation, if you have a lot of different factors, be careful. The last one for most of people will probably, especially if it's not a subgeneral experiment, which usually is the case, most of the experiment are not are not not so good. No, it's better to consider for each one as if it was the last one. So we take each one in the way. So that means that the total sum of square will not fit. Because now the common part of the sum of square is is forget. It's forgotten. We only consider the sum of square, which is pure for this coefficient. Only on this orthogonal to the rest. This is the two way of orthogonalizing. And as I said, sorry, this lessons a lot of explanation. You don't use it. Usually you just decide type one type two, but you have to remember why you have type one type two and what is the difference between type one and type two. And you also have this the explanation. You can go from one to the other. So the sum of square of a knowing B is the sum of square of a and B. Minus the sum of square of B.

notes

summary

22m 37s

So here and it will be my last my last slide for today. So now we can compare the two situations in the first situation. I have put the corrected sum of square and you see the P values are quite very good for the first one, not so good for the second one. But it's still very good, but less good for the second one. P value of 0.021 is a lot smaller than 5%. So it's OK. It's good people. It's fake data so I can do what I want. But the data seems to be to be OK. And the last one seems in between in the quality. In reality, if you treat them all the same way, then the interaction is a lot more certain than the other. They are all certain. There are no no no question about the confidence. But in fact, the second one is still the less precise. The first interaction is the most precise and the first one is coming after. When you have an ANOVA table that you read, you are referee for a paper. Just check if it's type one, type two. So this is all for today. OK, so thank you very much for your attention. I'm staying here for the exercise, so happy to help you answer questions.

notes

summary

24m 32s