Course material

Course:

ENG606 / PHYS 442

Video:

DOE_lesson10_part3_CompositeDesign

Concepts (extracted from automatically generated subtitles):

Data points. Higher degree model. Result of standard routines of linear fit. Sum of square. Balanced design. First approach. Linear models interactions. Second degree model. Good p value. Estimate of the variance of the sample variance. Main things. Degree of freedom. First thing. Sum of squares. Last hour.

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

to video

These subtitles have been generated automatically So, let's start again. So I show you a first approach with the fingers, understanding why we could need to go to a higher degree model. After I present you a very simple case with a test for explaining you the statistical perspective in this lack of fit. After I present you the result of standard routines of linear fit of ANOVA that normally you have to look in the ELP if you are in Python and other, but normally they have this option of calculating the lack of fit. But the main things to remember from last hour is that for calculating lack of fit, you need to have repetitions for having an estimate of the variance of the sample variance and you need some data points that are more than the minimum scheme for evaluating your model. So if you are evaluating linear models interactions, you need absolutely some data points that are outside of the scheme. And there is another example I will try to present you comprehensively. So we have two variables and we are, no we have one variable X and one answer Y and you can see here the model matrix corresponding to the estimation of the coefficient of this model. So you have one column of one, one, one corresponding to a zero. You have one column corresponding to X and will correspond to the coefficient A1 of your model and we have one column for the quadratic terms. So it corresponds to X by X and it corresponds to the coefficient A1. So the tricky question is do I prefer the linear model, the linear response model or do I prefer the model with the quadratic? You can observe in this case that because those numbers are the fits of my data and what you can observe which is common in the situation that the linear

notes

summary

0m 0s

coefficients stay the same. What usually change if you have a balanced design equilibrated around your zero is the constant and the quadratic term. Obviously the quadratic term are aliased with the constant. So the fact that you calculated yes or not your quadratic term will change the constant will not change the linear coefficients and interactions if you have a balanced design. So let's make the calculation for this case. So I have repetition in this case. You can observe that each time I have made two measurements it's why normally a part one point where the two points are one above the other for each measurement. So you have measurement at minus one, at minus one half at zero, at one half and at one I have five level of measurement and I have made the measurement two times. I have my two conditions. I have repetitions in this case is not the repetition of one point is the repetition of all the points but only one time. But it will let me calculate sample variance and I have more than the minimum points because for calculating a linear model I just need two points. I have five and for calculating a second degree model I need three points at the minimum and I have more than three minimum points. It would work with four points it will also work. So I have the sufficient condition for calculating a lack of it. So let's calculate by hand a lack of it. So you have here your points of measurement one one two three three four four. So this was my answers. Nine point five eleven eleven point seven seven point nine etc. There are my different responses for my data my data points. I'm calculating here below the sum of square of them because I think you already have understood that all the discussion is with sum of

notes

summary

squares because everything is within the sum of squares. And I get for this measurement four hundred thirty six point eight as sum of square of my data. That means I have now to explain these sum of squares and part it sort it in different sum of square that must be if I can also go on one to the other. So first thing that I can do I can calculate average for each measurement point for each position each x value. So the average of nine point five and eleven point seven make ten point six and after I have another average seven point seven three point five four point six and three point six I have made each time the average of the data points that correspond to the same value of my factor. And after I'm calculating the difference to the average. So the first measurement point has a difference of minus one point eight and I just have two points the other is the opposite. And each time I'm for each of these two column I can calculate the lack of fit. So I see that I have a lack of a sum of square and I have a sum of square of four hundred thirty three for the average and I have a sum of square of three point eight for the difference between the average and the real data points. And you can see that they are orthogonal the two vectors you can consider those columns as vectors. You understand that you have y bar which is orthogonal to y minus y bar. As you can see that the sum of squares correspond to the total sum of square. It's a decomposition between two vectors that are orthogonal. And I need that because I would like to be able after to make some rational between the sum of square and if they are

notes

summary

not orthogonal it could be tricky so I prefer that are orthogonal. And usually using the average they are by definition of the after I have a first model. So I have here the calculation for the linear model and after I have the calculation for the quadratic model. So I've fitted you see in the previous slide I have fitted the both models. So I have my coefficients I have what I need for calculating the y hat. You see a lot of small detail in my notation. Y bar. Why hat the estimate of the answer. So y hat one is the estimate for the linear model. So they correspond to the points in my model corresponding to my measurement points my predictions. So they are the same for the points that have been repeated because the value of x is the same so the prediction must be the same. So 9.4 for the first one 7.7 for next one etc. There are my different estimates and now I'm able to calculate a residue. Epsilon one is the residue between the model one and the measurement point. Point 12, 2.29 etc. There are my residues and now I'm able to calculate the difference between my estimate and my average point. So it's another type of residue. And now you can check that those different vectors because each columns can be considered as a vector in my data point space. I have that my model is representing 415 to which I have to add 21.7. And this will correspond to 436.8 with some problem of rounding numbers. But I'm also able to calculate the difference between the estimate point and the average point. And now I get another value which is 18. I do the same for my quadratic model. So I have my estimate of my quadratic model. So they correspond to they will correspond to the point

notes

summary

on my quadratic model. And then I can calculate the residue with this quadratic model. And again I'm able to calculate the difference between my estimate and my average. And now I obtain that my model is a little bit better than before. I'm closer to quite 10 unity different from the sum of square of my data. I'm explaining more of my data. I have sum of square of residue which is now smaller. It was 21, quite 22 for the linear model. Now is quite 10 for the quadratic model. And I also see the difference, the sum of square of the difference between my average point. Okay, those are different statistics. And here I just calculate sum of square of what PE? The lack of, pure error. PE means pure error. And L, O, F equal lack of feet. So we are now able to calculate and see where are some of the values. So we have, so this was for the linear model. And here is for the quadratic. So the first part is clear. We already see it. Is the ANOVA, the standard ANOVA that we have made usually. So we have for the linear model, we have for the constant, a sum of square of 356.8. And we have for the linear part 50, 58. And we have a residue, which is 21, sum of square 21. We have one degree of freedom for the constant, one degree of freedom for the linear coefficient. Then we can calculate the mean square. And we are able to check a P value for the linear part, which is quite okay, quite less than 5%. So it was an adequate model, perhaps the lack, so you understand that the P value tell you if it's sufficient and the lack of feet will tell you if you need more. So now lack of feet and pure error. And

notes

summary

you see those values, 18 and 3.8, they are coming from this. And the difference between this and this, because this is the 21.8 is this value. So now 3.8 is 21 minus 18. And it's a pure error. And now we have to check the degrees of freedom. We have eight degrees of freedom for the residue, because we have 10 measurements and we are estimating a constant and a linear coefficient. First, accounting eight degrees of freedom for my residue. And now after the next step is calculating the degree of freedom of the pure error. And so I have five measurement points repeated each one two times, five times the possibility to calculate a difference. So I have five degrees of freedom in my variance. So the three is what stay? So the three is eight minus five. Because the sum of the degree of freedom in my analysis of lack of feet, my two columns lack of feet and pure error must be the same as my residue. I'm just decomposing my error in two errors. And they must be orthogonal if I would like to separate them. And they are because it's calculated with means. So now I'm able to calculate mean square. So 18 divided by three make six. And three point eight divided by five made zero point 75. And I'm able to calculate able to calculate a Fisher ratio, which is the ratio is six divided by point 75. It's make quite eight. And now I'm able to calculate a p value, which is the Fisher. So now it's a Fisher test. The first slide I present you was a student test, but you have understood that we can play there are a lot of things we can play with students, we can play with Fisher. So with quite eight for three degree of freedom for the numerator, five degrees of freedom

notes

summary

for the denominator, I have a value of 2%. And now be attentive. So 2%, which is smaller than 5%. That means that my output, my hypothesis of a lack of fits is acceptable. Like you like the the coefficients a zero and the linear coefficient was acceptable. So I have a lack of fit. So that means that my model is not sufficient with the data that I have. I have in my data sufficient information for telling me that the model is not optimal is is okay because I have a good p value, but is not optimal. I can't do better. I have a lack of fit. I'm not fitting sufficiently. Because the question is why do we go one step further when we see that our model is adequate. So my answer is that is engineering must must manage. So if I'm okay with the linear model, I can do what I want with the normal. It's okay. I know I have a lack of fit. I know I have a curvature, but okay, I'm happy with what I have. But I know I could do further. I have sufficient data to go further. Is it necessary or not? Statistical elements say yes. Practical elements could say okay, be happy with what you have. So when I go, I can go a little bit more rapidly now. When I go with my quadratic model, I have one degree of freedom for my coefficient, for my constant, two degrees for the part of my model because I have one linear coefficient, one quadratic coefficient. And then now I have seven degrees of freedom for my residue. When I'm checking the sum of squares so you can see where they are coming from. So you have this 9.57 and the 3.8. If you go here, you have the yeah, the 5. So the 3.8 is the difference between

notes

summary

these two and the degree of freedom, the 9.75. It's a problem of numbers of significant number, but the data is coming from here. Now I'm able to calculate the mean square. So I have mean square 100 for the quadratic part, or the quadratic unlinear in this case, so the quadratic model, this quadratic plus linear. And the residue. So I have a model which is even better than before. The previous one was okay, but this one is even better. The fact that I still have the digit 17 is just a lack. And after the lack of fit, I have 5.8 and 3.8 for my different lack of fit and pure error. This is the 5.8 and the 3.8 is the difference between them. And so I'm calculating the mean square 2.9 and 0.75. You see that the pure error have not changed. It's normal because it's the same data points. What I've changed is the lack of fit of the model that have diminished now. And now I'm making all the calculations, I get the p-value of 10%, which is bigger than 5%. Indicating me that my lack of fit is not acceptable. So I don't have what I was talking about, double negative. I'm not accepting a lack of fit. So I'm happy with what I have. I'm not saying that, definitively, my model is second degree, because I don't know, it's a problem I can go in the outside of my range of measurement, but I do not have data telling me that my model is not sufficient for explaining. And my scheme was sufficient large for being confident. So it doesn't tell you, lack of fit doesn't tell you everything. You have to check with the scheme of your measurement. Did you make sufficient measurement to be sure what could be? If I have, I don't know, I could have a real model.

notes

summary

My real model could be something, trying to invent something. So, okay, I'm not checking that my model is not this, okay? It could be because between two points, something could happen. So I'm not checking that, definitively, is the only model possible. I'm just telling that, worst data point I have is the best model that I can, but best I can perhaps check also not the type of curvature, eventually, I don't know. Low-garry is with exponential, perhaps it could be even better, but at least it's sufficient. And if I'm working with porcimony, I'm happy. I'll let you ask your question. Okay, so now I've finished with this first part. What I can check when I have a model to see if I need to go further and is all the time, remember, sorry to insist, the two things you have to check, you need to have repetition for a variance, pure error, you need pure error, and you need a scheme of measurement, which is larger than the minimum for detecting your model, for checking if it could be another model. Okay, so now let's see. We still have 10 minutes. So I will present some of the very classical designs and we will finish next time with the rest. Okay, so the first design I would like to present is not the best one, like that we will go increasing the quality of the design. It's quite an expensive one. It's the first one we have in mind, as when you have a straight line for having a quadratic model, you need three points. So when you have more dimensions, you multiply that for each dimension. So it would be a three power k design. So if we already were complaining that the two power k was exponential, what to say to the three power k, it's increasing also a lot. If you are playing

notes

summary

with infinity, it doesn't make any difference, but if you are playing with numbers of k around 10, below 10, it makes a big difference. So if you have three factors, you need 27 experiments and now you have three level per factor. So it's the network three by three. It's too expensive. Look at this one, more interesting. You make the factorial points, you make experiment at the center. Eventually you repeat it. Like in my first example, you make a lack of fit. The lack of fit tells you you can go to a second degree and you perform, let's change the color, you perform some measurement at the center of your faces. And now for three factors, only 15 measurements, you can have a very good design for estimating the quadratic model for three factors. So before it was 27, now you have made 12 experiments, so quite half less, not exactly half less, but a little bit. So you win one week of holidays or one month of holidays and you have good quality. This design is called composite. Why is composite? Because we start with the red points that I call the factorial scheme and after we add the blue scheme, which is a one factor at a time type scheme. So the one I was criticizing a lot at the start. So this one factor at a time is bad to start with, but afterwards it could be interesting to perform it. So it's composite because we have added two designs. I present to you the case with three factors and it's a factorial plus a star design. Factorial plus star design, but it could be a fractional factorial. You are not obliged to calculate all the factorial points. Three dimension is not very interesting, but at more dimension it became interesting. And there are one parameter also in this is this distance that

notes

summary

I'm calling alpha. So in this case, I have put alpha equal one. So my point is exactly at the surface. So it's a cubic centric, so a centric cubic if you are familiar with the crystallography. Fifteen measurements, very interesting, very robust, not cheap, not too expensive. It was the favorite of Barks. You publish a lot of papers with this design. I really appreciate it. You can play another game changing the value of alpha. So this distance here now is 20% 22% higher and they are an optimum. What you would like is eventually having something which is closed to a sphere. So it's interesting. So what is the interest of that? That in your phenomena, eventually you have some direction which is better than other. And the risk that you show the direction and it's not the right one after when you have your data. So if you have what we said covariance per rotation, the fact that your variance is not changing if you are moving your design around, you don't make a choice of direction. So after your data could be in any direction, your phenomena could have any direction, favorite direction, your data is still good. We will talk about this value. We can improve it for different objectives. In the algorithm, if you go in MATLAB, you have MATLAB, call it CC design. You have one routine of MATLAB. He's asking you what you want to favor it. It could favor the numbers of points to go at the center. It could favor this covariance per rotation, is variance per rotation and things like that. So quite the same situation as before except the dimension. Your question. So you understand it with this type of strategy, we are not limiting minus one plus one. We are going a little bit far away, but it's correct and it's even very good because usually

notes

summary

when you start making your fractional factor design, you don't want to be really at the limit of your phenomena. Usually you try to be a little bit in a smaller portion because you don't want having too much nonlinear reactions, nonlinear phenomena. So you better have smallest reasonable domain for understanding what's what's matter. And after when you want to go and want to see the nonlinearity, it's interesting to go further. So it's also a very good strategy managing your domain of variation of the range. So after depending, look, this value was 22%. There are a relation between the number that you do at the center and what is interesting are far you are interested to go outside of your domain. So now we are at something as 33% outside of your domain because we are making more. So these points are in fact one above the other. It's just for showing you that I have three that I have not put them exactly at the same position. They are at zero zero, the three points at the center, they are exactly the same value. So you have and will be my last comments for today and perhaps I will start just present you, but I will start again with that. There are a way of optimizing the iso variance per rotation and also the orthogonality. So we quite have the same discussion that I could have with the rest shaft near design and the three quarter design. Sometimes you want to give profit to one side. Okay, we have two balance. They are not perfect, but depending on the number of factors as you have, what type of design you have chosen for the factorial step? Do you have a fractional factorial or do you have a full factorial? An how many measurements you are doing at the center? If you are interested to make a

notes

summary

lack of fit or you are not so much interested by that, there are different values. So you go between the distance of being on the face or the extreme. Here you see when you have six, seven, eight factors, you are quite the double size of your range. So we can manage. So today we have seen two of those designs, the expensive one, three by three, and a more interesting one, the composite one, with some special recipes, depending what you would like to do. I'm finished for today and next week we will continue discovering a few other designs interesting. So two designs have one characteristic, my last word, that you can start with the factorial because you can do the same with this. You can start with the factorial points and after add some other points. So it's a step by step approach. First you manage the linear aspects and after you go for the curvature. I will show you also all the strategies. Okay. Now there are problems, a new series. I know you are not in Advent official series, but I will just load. I forget to load it tomorrow, so I will load a new series. So I'm finished for today for the course, still the exercise or questions on your project. If you have, I will be available.

notes

summary