CH5230: System Identification

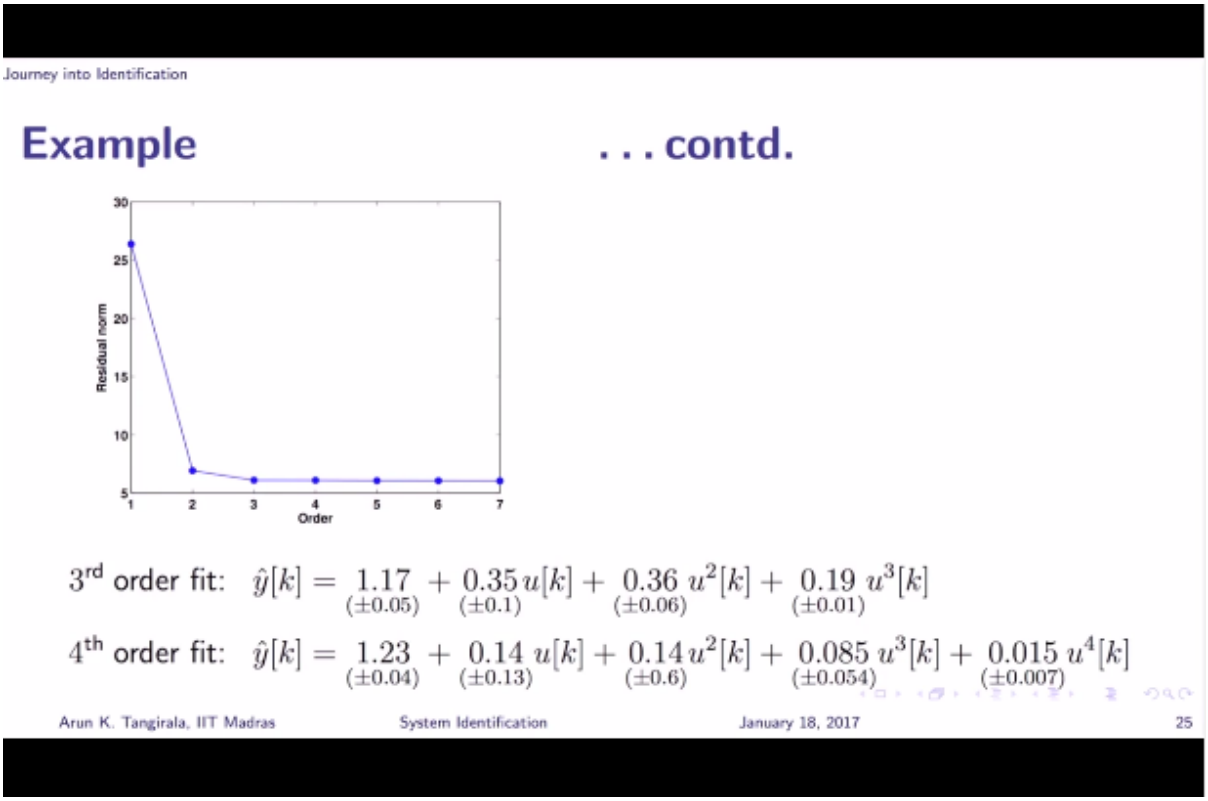
Journey into Identification

(Case Studies) 7

And that is what happens in any over fitting, when the global variations are actually confused, the global trends are confused with the local variations. And therefore one has to be careful. Now how

could I have avoided this over fitting. One is of course having a fresh data, right. Because how do I know in practice what is a true order. I will never know. I have fit third, fourth and fifth and I have to choose among these. If I were to only look at this plot here, the residual norm verses order that plot alone will not tell me that a third or whether a third or fourth or fifth is better.

(Refer Slide Time: 01:00)

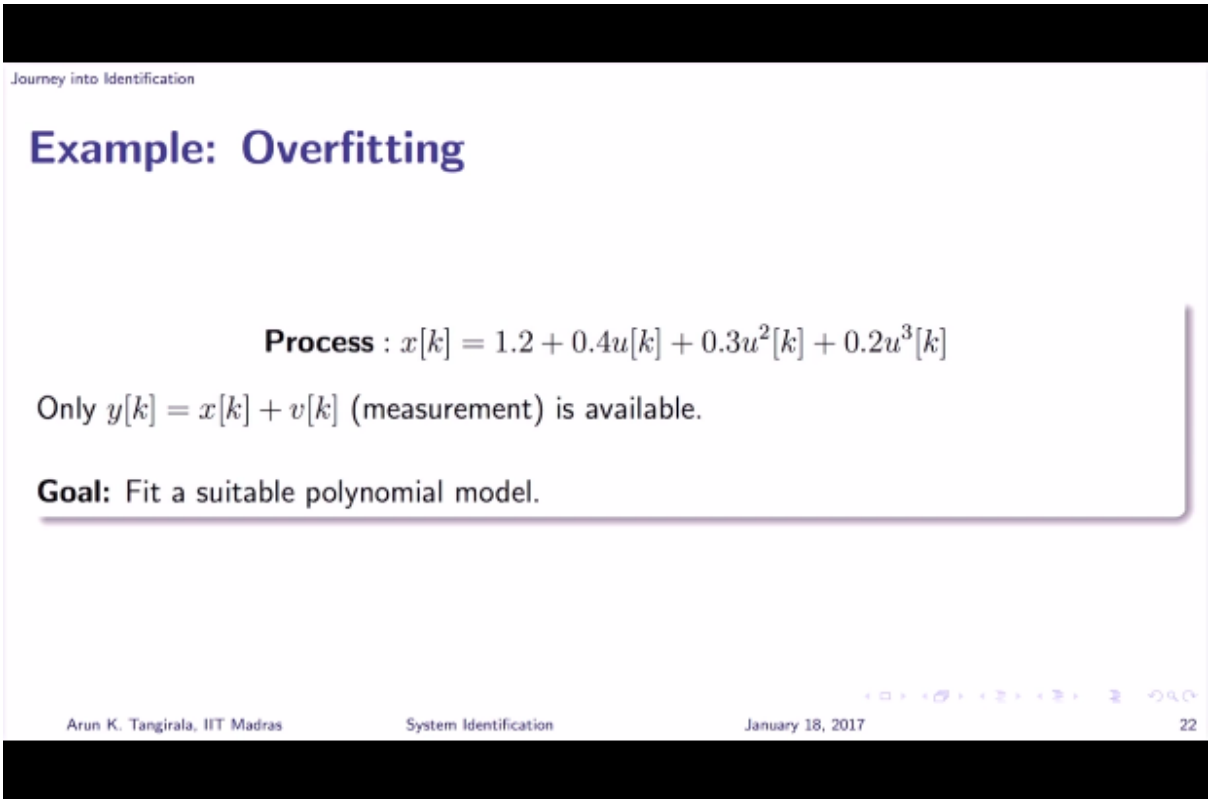


In fact, this plot tells me, fifth order is the best, correct? Yes. As far as the fit is concerned fifth is best but as far as the performance of the model of fresh data is concerned third is the best. So what has happened here is, the third order model is the best compromise between a fit on the training data set and the performance on a fresh data. And that is a tradeoff that we are looking for. That is one way of figuring out whether we have over fit, always having fresh data and seeing the performance of the model and the fresh data. The other symptom of over fitting is as someone pointed out, large errors in parameter estimates. So I'm showing you here the estimates of the third and fourth order polynomials, as you see here, the errors in parameter estimates, I don't know how well you can see from the back but let me zoom that for you. Let's look at the third order polynomial model. You see that all the errors, the standard errors are reported in the brackets underneath and what you see is that parameter estimates and the standard errors they're quite far apart. That means the errors are not as large as a parameters estimates, these are qualitative statements but there is a statistical way of determining if a parameter should have been included in the model or not. Or if a parameter estimate is significant or not.

What I mean by significant is, now recall that the original process that we have used for generation is a third order polynomial. If you recall that slide, let me take you back to that slide so that you can also take a look at the true values. So the true values are one-point-two, point-four, point-three and point-

two that is in the increasing order of the powers of u. And it's the third order polynomial. If I look up on this as a fourth order polynomial the last coefficient would be zero. Right?

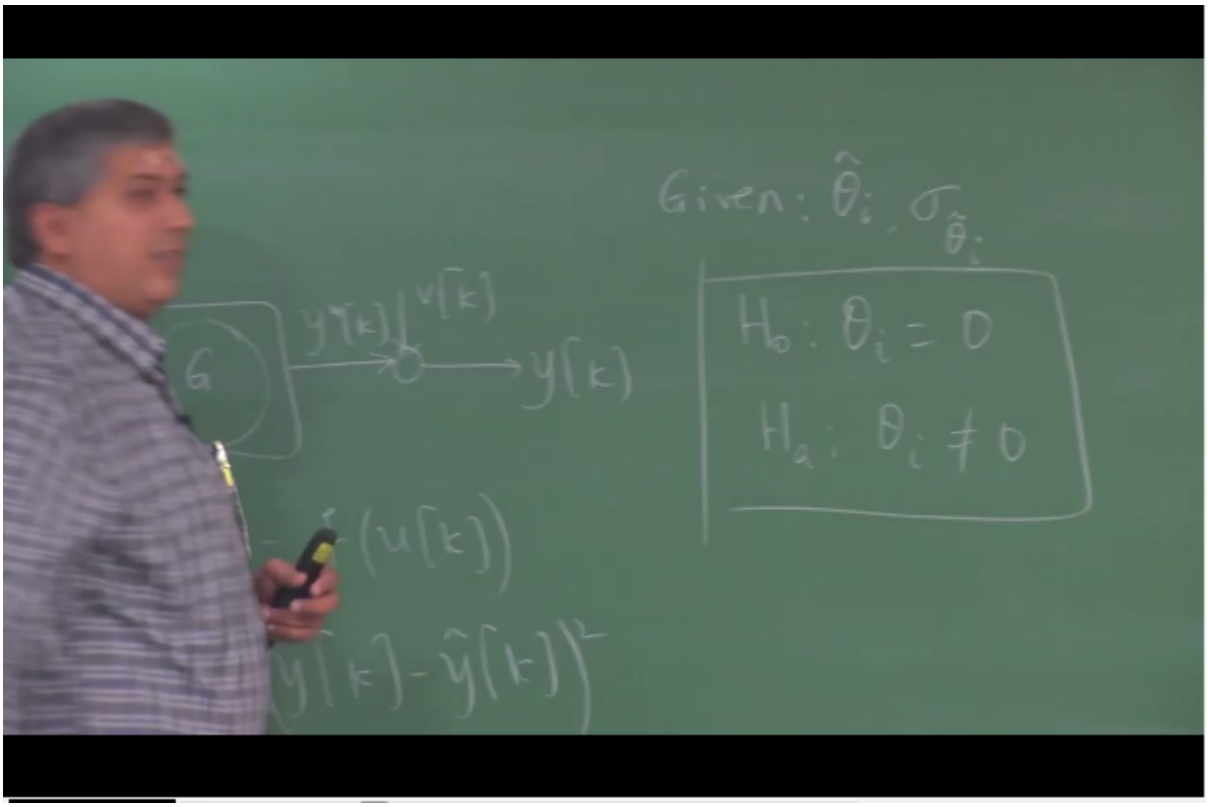
(Refer Slide Time: 03:05)



And let's go to the estimate here. What has happened to the fourth order polynomial here. So you have one-point-two-three, point-one-four, point-one-four, point-zero-eight-five and point-zero-one-five. Assume that momentarily the errors in the estimates are not given to you. Then when you look at the point estimates of the fourth order polynomial, you'll find some to be small value. For example the last one, however one cannot say by looking at the point estimate whether that should be thrown out of the model. Typical is our point-zero-zero-one, I think it should be thrown out. It's very small. But that is with reference to some number that we have, maybe truly that number is actually that parameter is small value. So we cannot look at the point estimate and conclude that that particular parameter estimate is insignificant statistically. The other thing to see is that the true value is zero but the estimated value isn't. And that is a facet of estimation always. Even if truly the parameter value is zero you would not obtain an estimate that is zero value, right. So that is why we require, I mean we need to perform hypothesis tests that is we perform best as follows. We say, well the first coefficient is zero, that is my hypothesis. The true value of it is zero against its alternative it isn't. And I have to do this for every perimeter. Then only your estimation is kind of complete because after all the purpose of estimation is to know, is to make some statements about the truth. I have with me only point estimates. The purpose of reporting the errors is to conduct these kind of hypothesis test. What kind of hypothesis test that every parameter, so the null against the alternative that it isn't. It didn't care whether it's positive or negative value. For each of the parameters I have to conduct that hypothesis test.

Now there is an, I hope you have taken some courses on statistics but if you haven't there is twelve hour course that I have offered an introduction to statistical hypothesis testing. If you are not familiar with it you should sit through those videos and familiarize yourself with the concepts. One of the standard ways of conducting hypothesis test is to construct what are known as confidence regents. Okay. Remember the hypothesis is on the parameters not on the estimates, why because the estimates will always almost turn out to be non-zero. We are not bothered about that. What we are given is estimate from the, we have given a point estimate and we are given the standard error in this estimate. So we have given this and the goal, the objective to conduct this hypothesis.

(Refer Slide Time: 06:14)



And as I just said, one of the ways of conducting this hypothesis tests, these are quite significant tests because the way, the of the null hypothesis, what we are asking is whether the true, whether the estimated parameter is significant or not, that is what it amounts to. How do you cancel the confidence region. Strictly speaking you should know. Remember your theta i hat had one of the things that you should keep telling yourself is, every estimate that you obtain from data is a random variable. Why? Because your estimate has being derived from data and your data has uncertainty and it has randomness in it. So that DNA of randomness in the data is passed on to the parameters. So, it's a child, coming, taking birth from the data. If not fully at least 10%, 20%, whatever it depends on the estimator it'll inherit some randomness. Therefore every parameter estimate is a random variable. Parameters need not be a random variable, in classical estimation we assume parameters are deterministic quantities. Estimates on the other hand are random variables. In order to construct the confidence regions for thetas, we need to actually know the distribution, the pdf of this theta hats. For now, don't worry about the theoretical details. We will learn more on this later on. But for now assume that this parameter estimates have a Gaussian distribution. Typically it holds good for large

observations. And from that you can derive these so-called 95% confidence intervals as theta I hat, plus or minus, this is a confidence interval for theta I, never ever say that I'm constructing a confidence interval for theta i hat. Why I am constructing a confidence interval for theta i because I do not know the truth and -- A region I believe contains a truth with ninety-five percent confidence. I can never say that this region definitely contains the truth. If that was so that region would be infinitely wide, right. Obviously, because of the uncertainty. So I'm saying with a fair degree of confidence I believe the true value of theta i is between theta i hat plus or minus 1.96 sigma theta hat. We'll prove this later or you'll see the proof even in the nptel online course. Now apply this. So how do we apply this thing for hypothesis testing. If the confidence interval contains zero then that means zero is also a possibility, since zero is also a possibility with a high degree of confidence. We say that the null hypothesis is not rejected. So the simple thing is, take theta i hat, you have computing sigma theta i hat, construct this ninety-five percent confidence region or ninety-nine percent confidence region and see if zero is contained in that interval. So apply this procedure to each of the parameter estimates in the third order polynomial for example. For which parameters is a null hypothesis not rejected because that's of interest, if null hypothesis is not rejected that means that corresponding parameters should not have been in the model. Is there any parameter in the third order polynomial for which the null hypothesis that theta I is zero is not rejected. What do you think? None.

What about the fourth one. The second one, here. That is the coefficient on uk. What about this one is Okay. This is out, that's also out. Well close. But the moment they find some parameter estimates being insignificant, so now you have-- Then I have to actually reject those parameters from the model and then estimate a fresh model. You cannot say, look take this hundredth order polynomial that are fit, ninety-five of them are ninety-seven of them or ninety-nine of them are insignificant. I will throw all of them and give you only the significant ones. In other words he that we have just figured out in the fourth order polynomial that the second coefficient and what else, that is the coefficient of u and the coefficient of u cube, right, their estimates are insignificant. So you should not report the model by saying, “Okay here is the model one-point-two-three plus point-one-four uk, plus point-zero-one-five to the power of four k."I have thrown all the insignificant ones. That's not a correct way of reporting the model. Why? Why shouldn't I report the model that way. I have done the correct test. I've done a test of significance. I've omitted the insignificant coefficients because they were supposed to have been actually zero valued and I'm only reporting the more significant ones. What would be wrong with that?

Different set of data, if some other coefficients may be.

That's okay. But that would also be true for any other rightly estimated model, that is like the third order polynomial also that's true. That cannot be the reason. Do you understand the question. Once I perform a significant test, I have figured out which parameter estimates are insignificant. Right. So I throw away those parameter estimates and then report the rest of the model. I said it is wrong and I want you to tell me, why it is wrong? Wrong in the sense it is not the correct thing to do. Why is that? Any idea? Even a simpler example, suppose I'm fitting a straight line with a slope and an intercept and then I compute the slope and estimate the slope and intercept, I also compute errors in the slope and intercept, I conduct a hypothesis test a significant test. I figured out that the intercept should not have been in the model. So you should not just throw away the intercept estimate and only report the slope estimate. That's not the appropriate way of doing it. Why? Why do you think, I'm saying this. Any idea?

Because, the answer is because, some degrees of freedom have been used up. Remember, if you don't like statistics you can think qualitatively. Data is a food for identification. When you included a

certain number of parameters in the model, estimating those parameters have taken up some food, have consumed some food. But in the first place they should not have been even invited for having food. Okay. So you should do a fresh job. You could actually again now invite the more significant guests and then feed them separately so that they now much better fed. Because I have devoted, I have actually fed someone who is not hungry, who did not really, let me say deserved to be invited for the occasion, right. As a result I have actually partaken some food and given to this guest who has not done anything. I mean in this sense, who did not, who is not suited for the occasion. Here also I have included parameters that are unnecessary, as a result apart of the information that is available in the data has been consumed by that parameter estimate. And therefore the errors in the estimate, the significant parameter estimates are going to be higher because there's not enough food for the rest of the significant parameters. When I perform, when I throw away the omitted, the insignificant ones, I should also do something else, I should re-estimate the model with the significant ones. So for example here, we have figured out the coefficient on uk and u square. I'm sorry, uk, in fact uk and u cube they are insignificant.

I should throw them away, refit the model with the most significant terms and then re-compute the point estimates and the errors and report that model. But before reporting, I should once again connect the significance test. Because their significance was determined in the presence of other parameters. So every time I estimate a model, I should perform a significance test and only report when I'm convinced that all parameter estimates in the model are significant, right. That means all parameters have passed have tested negative for this, or whichever way, negative or positive, the way you look at it.

So the bottom line is do not report a model whose parameter estimates are insignificant and claim that to be a good model. If that were to be the case, I would fit a fifth order model also. Because that will do a better job of fitting. So, over fighting occurs whenever the local chance variations are confused with the part to be a part of the global one. And a symptom of over fitting, a classical symptom of over fitting is large errors in parameter estimates or parameters estimates that are insignificant. Once you figured out that a model has insignificant, that is the estimates are turned out to be insignificant. The message should come to you that you have overfit or you-- It may not be over fitting, it mean it may be that you have included a wrong term, the mod, the parameters that should not have been a part of the model.

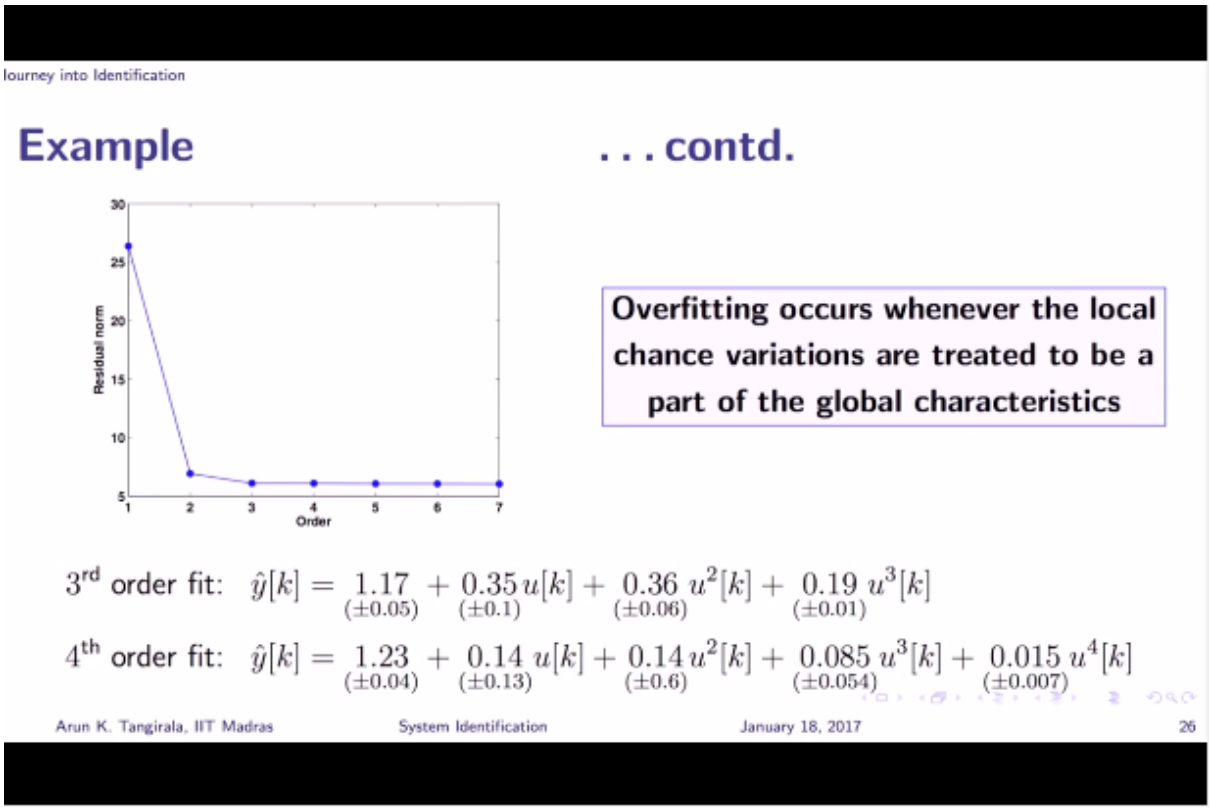
Now, here we are-- Remember we have argued that the fourth order model is an over fit without the knowledge of the true model, that's very important because in reality that's going to be the case. I do not know what the true model is. I have to rely solely on the data to figure out whether a model is over fit or not. And the way to do it is, to turn to parameter estimates. There are two mechanisms. One is to look at that errors in parameter estimates. If they are high then that means I have included unnecessary terms in my model and two the performance of the model and a fresh data set, cross validation. So these are the two things that one has to perform on the model before the model is deemed to be satisfactory. There are other steps but at least these are the two steps that you have to perform to ensure that over fitting has not occurred. There is also a possibility that under fit have occurred, right. For example here, if I had fit a linear straight line, that's an under fit. How do I know it's a under fit? How would we know, we have talked until now about over fitting and we will see this phenomenon of over fitting in the next case study, the liquid level case study as well.

Let's talk about under fitting. When I invite guests to feed, I'm not only worried about over feeding them, I'm in fact mainly concerned about under feeding them. They should not go out hungry, "Atithi Devo Bhava," so we are supposed to actually feed them very well. If they eat more than what they're

supposed to be eating it's their problem. But here it is not, it's our problem if the model is over fit. But the most important concern is that I should not under fit. How do I figured out that I have under fit the model. Because I don't know the truth. In this example what would you do to figure out that the first order polynomial or even the second order polynomial is not satisfactory. What will you do with the residual?

Well, what is high for you maybe low for me, it depends on what you place on the table. Comes down. Yeah. It keeps going down, right. Yeah. So, one thing is to plot this when you're not sure of the model order. But this plot alone may not tell you, right? For example, this plot, tell me whether. Okay, I can say that the improvement from third to fourth is kind of marginal and I may arbitrarily stop at third. But in general, how do you resolve? Is this plot alone sufficient to figure out if you are under fit or not under fit.

(Refer Slide Time: 19:38)



Okay. Okay good. How do you figure out if the residual doesn't have a trend? Do you just visually inspect the residual? I'm not showing the residual here but you can try it out. What model? Good. What model will you fit?

Taking something as input, if we get the residual as the output of some model, you can-- if there is some unique model is exist then the residual will--

But modern needs an input and output, right?

Residual which we are getting from fitting this underfit model is the input for the model to get that next level of residual. If it has something then we can say that still it has some data which can destroy--

Yeah. I think you started off correctly but then you went off track. It's correct. We look at the residual, there is something called residual error model. Okay, or model error model. Sorry, model error model. What do you do is, you compute this epsilon, now you ask if epsilon itself can be explained again using the input, okay. What this means is, we are searching for any remnant effects of the input in the residual. If there are any remnant effects of the input then we conclude that we have under fit. And then go back and refine the model. That is exactly the logic that we will apply. Of course you have to do this on statistical grounds but that's exactly the approach that we will take in the next case study, where it will fit a model and we do not know what is the order. We will pretend we do not know what is order. Now we have to test or determine whether the fit model is adequate or not. All right. So it is important to analyze a residual. Other answer of searching for trends in the residuals is also correct but that is specific to some class of problems only, in these kind of applications where the model is relating and input and output ultimately what we want is to make sure that the map between input and output has been rightly obtained, that I have not left over something that could have actually improved my prediction.

So let me summarize this case study for you, given input output data and let's say we do not know the model order and so on. We start off with a guess and then we start refining the guess but there should be a basis for refinement. And the basis for refinement is always a residual test. What we mean by basis for refinement is, I should believe that the present model is an under fit. If I am satisfied that the model is not an under fit, then I should move on to the next stage which is to check for over fit. Typically if you follow a bottoms up approach, that is if you start with the simple model and proceed upwards the chance of over fitting is less because at every stage you're checking if it is underfit. The moment you have determined that the model is not an under fit, you have obtained overall as a correct model. But sometimes it may be necessary to go big beyond at least one order next and see if the next one is actually an over fit. Sometimes they maybe two or three models giving you very, like here you have third and fourth order for example, of course, in this case the fourth order will test positive for over fitting but in other situations it may be that the third, the one ordered may, one model order may be satisfactory but the next one also may be satisfactory or some other model structure maybe satisfactory. Here we are only restricting ourselves to polynomial models. I could have fit some other non-linear also. That is also valid. Since I do not know the truth, the only basis for determining whether a model is adequate or not is entirely dependent on residual analysis and estimation errors. So you see, without knowing the truth we are able to actually figure out or discover the true model. This may seem like a classroom example but that's also a fact that without knowing the truth and by performing rigorous tests in a certain sequence, first effort under fitting, then for over fitting, followed by cross validation and so on, will more or less guarantee that you end up with a decent model, that is a working model. That may not be the truth because the truth may be more complicated than what you imagine. But it is important therefore to follow a systematic procedure when it comes to model development. So let's get started on the case study and then of course we'll continue with this through tomorrow's class as well.