Regression Analysis and Forecasting
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology-Kanpur

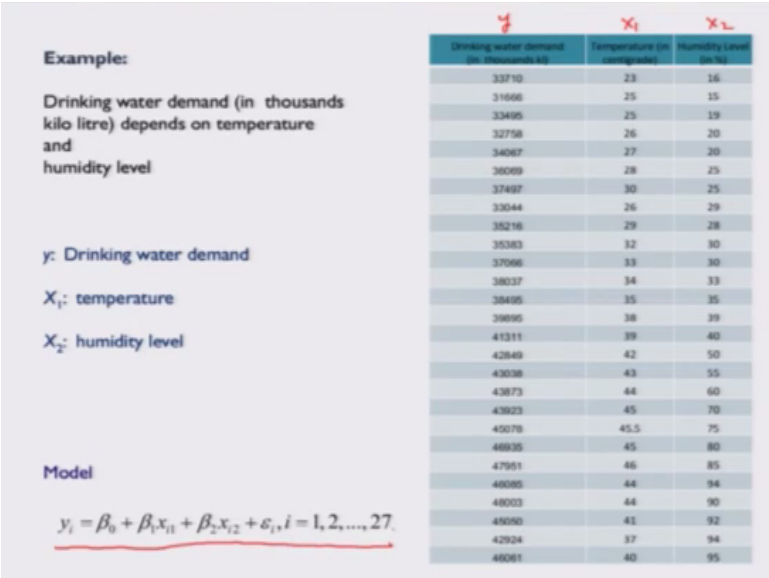
Lecture – 24
Software Implementation of Forecasting using MITNITAB

Welcome to the lecture you may recall that in the last two lectures we had constructed the predictors using the ordinary least s square estimator under the set up of a multiple linear regression model. We have used this predictor for within sample forecasting and say outside sample forecasting we had established the unbiasedness properties of the predictor, we have found their standard errors and we also constructed the prediction intervals under different types of condition.

Now the next question is how we can use these predictors and their corresponding properties when we are using a real dataset, this is what we are going to address in this lecture. What we are going to do that we are going to consider the same dataset that we had used in the case of multiple linear regression modelling and this was the dataset that was related to the demand of water supply which was depending on temperature and humidity.

We had considered the complete analysis and we had obtain the multiple linear regression model, now how to do the forecasting using that multiple linear regression model.

(Refer Slide Time: 01:41)

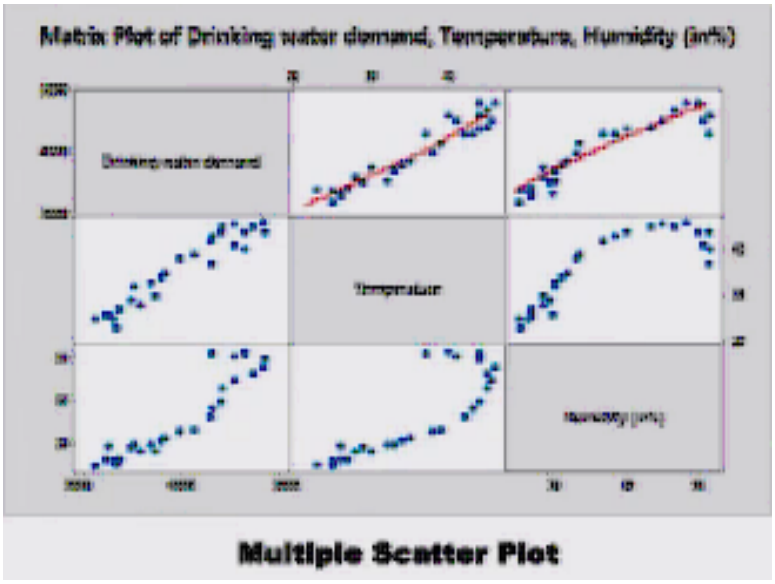


Under different types of settings like as within forecasting, outside sample forecasting and how to forecast for average value and actual value under these to respective categories. This

is what we are going to demonstrate in this lecture. You have to be cautious that sometime some softwares also provide forecasting in that case you have to be careful that what sort of forecasting they are providing whether it is within sample whether it is outside sample or this is for average or say actual value forecasting.

So let us try to consider the example, so you may recall that earlier we had considered this example where we had two independent variables x1 and x2 which are the temperature and humidity and we had recorded the drinking water demand as y and based on that we have considered this model earlier.

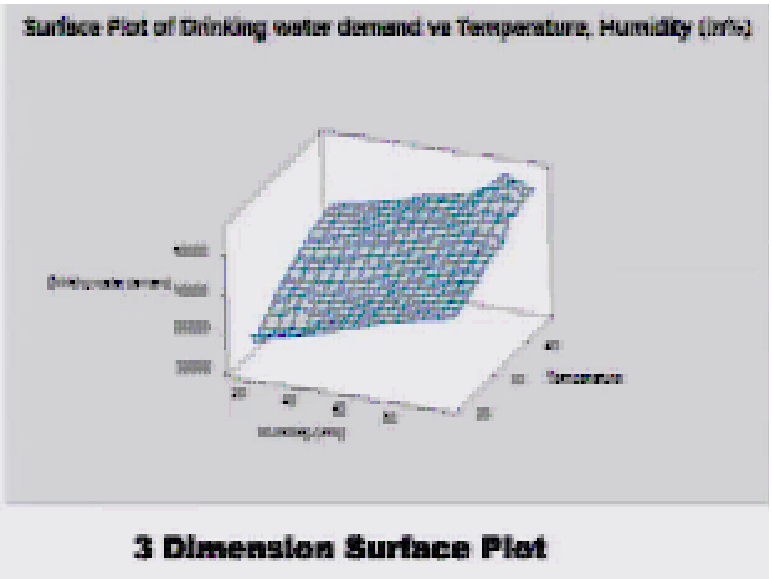
(Refers Slide Time: 03:00)



Multiple Scatter Plot

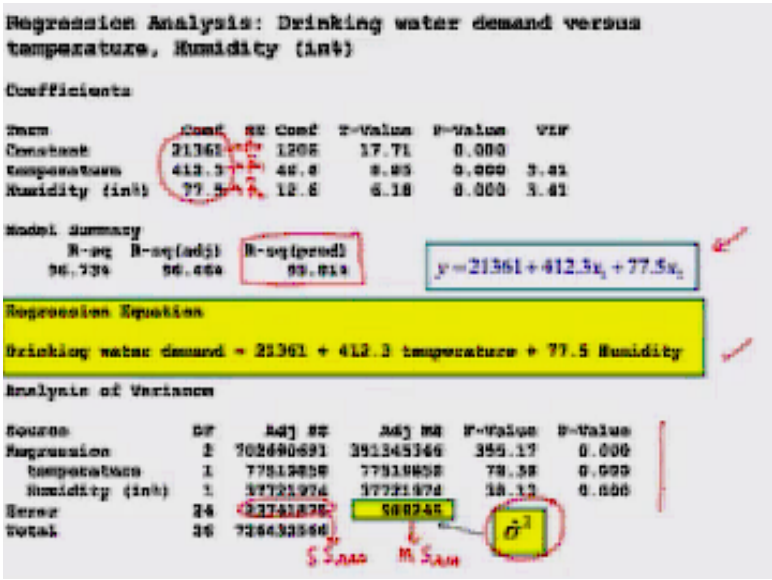
And just for the sake of quick revision we had plotted the variables y x1and x2 over a scatter plot and this is the matrix plot of all the three variables and one can see here that there scope of fitting a linear regression model, so now based on this conclusion

(Refer Slide Time: 03:24)



We also had verify this thing through three dimension plot, and three dimension plots are possible in this case because we here only three variables y x1 and x2 well once we have more than two variables than three dimension plots are not possible, but we are utilizing this opportunity here. So one can see here that there is a sort of increasing trend over here, so this again confirms that one can fit here a multiple linear regression model.

(Refer Slide Time: 03:56)

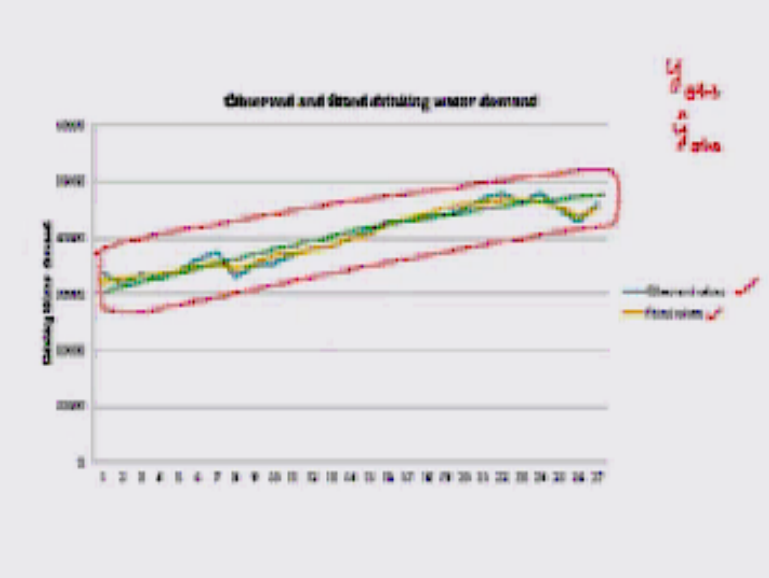


And after fitting the model using the software this type of outcome will come into picture, so now here I would like to use some particular part and I would like to have your attention. You see this you already have discussed that these three values they are the values of beta0 hat, this is the value of beta1 hat and this is the value of beta2 hat and then we have obtained this model. Now side Rs squares and R square adjusted we also have here R square prediction.

And this is giving us a confidence yes, if we use this model for forecasting then the model is going to behave well, okay. So now we have this regression model now another thing which we have to keep in mind is that this quantity here is the sum of square due to residuals and based on that this quantity is mean square due to residual and this is us the value sigma square hat.

So in the case of prediction interval wherever we are using the value of sigma square hat that can be obtain from this analysis of variance table, right as sigma square hat.

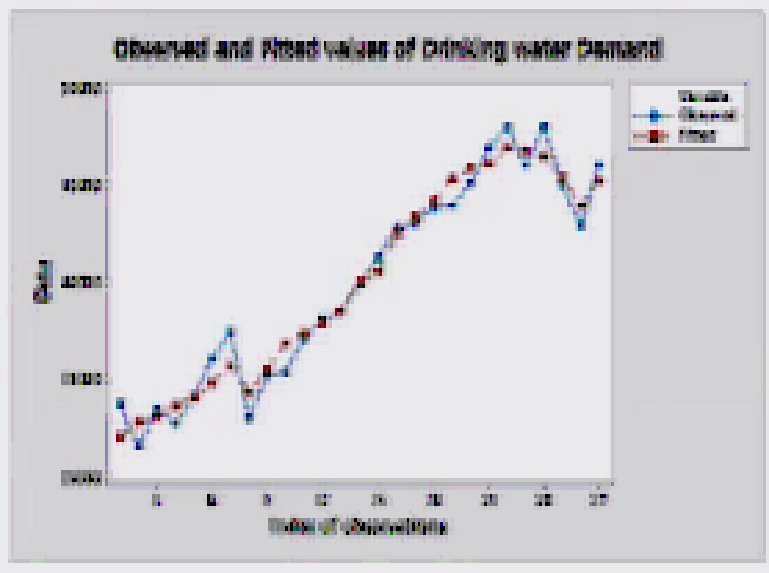
(Refer Slide Time: 05:39)



Now before doing the forecasting I would like to verify whether my model is going to work well or not, so we try to use as many as possible options. So here what I have done that, I have considered here the value of y observed those twenty seven observations and using the fitted model we also have obtained the fitted value, and I have plotted both the points at a same time. So here this green line is indicating the observed values and yellow line is indicating the fitted values.

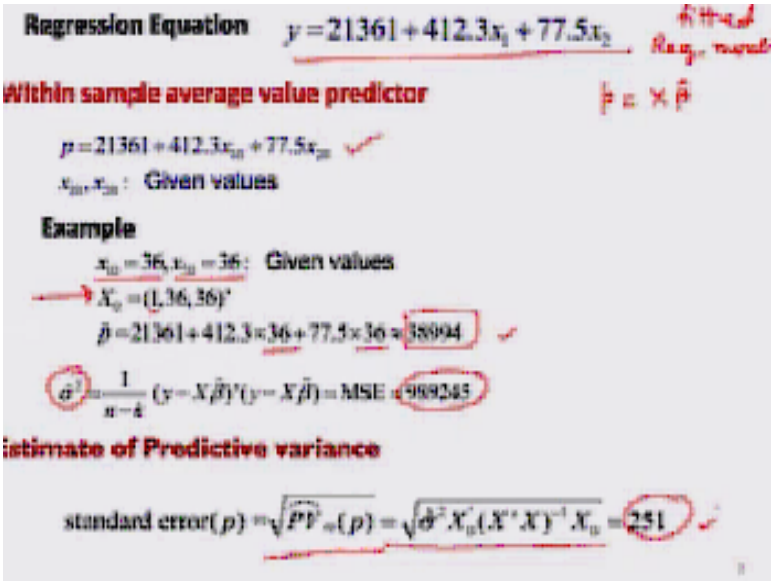
And one can see here in this segment that the difference between the observed and fitted values is not actually much and in most of the cases they are lying over the each other, even if I try to fit here a model possibly this may look like this, right. So this gives us a confidence yes, the model can be used for fore casting.

(Refer Slide Time: 06:49)



Now I also have plotted this individual points and here also one can see that there is not much difference between the observed and fitted values so this is again giving us a confidence, yes the model can be used for the forecasting.

(Refer Slide Time: 07:08)



Now let us try to consider first the within sample prediction and we consider the prediction of average value first, but before that remember that this was our regression model that was fitted. Now using this regression model we had discussed that a predictor can be constructed like as x beta hat. So using this model we have constructed here this predictor, right where x1 naught and x2 naught are some given values.

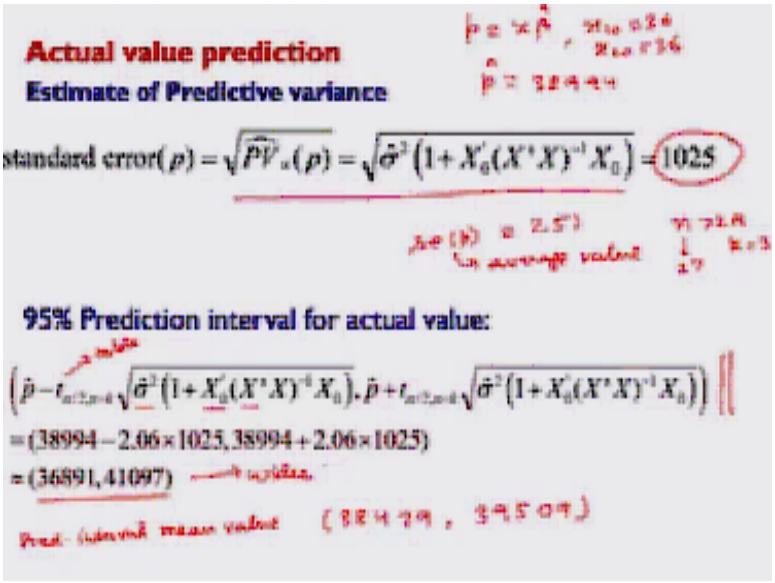
Now let us take an example and try to forecast, suppose we say that the value of x1 naught is 36 and value of x2 naught is 36 that means we would like to know the demand of water when the temperature is 36 degree centigrade and humidity is 36%, right quite dry. So as per our

symbolic notations x0 will be denoted by three cross 1 vector whose first element is one and other two elements are the values of x1 naught and x2 naught.

This one indicating the presence of intercept term in the model right, so now we substitute these values of x1 naught and x2 naught in the predictor, and like as here and see here and we try obtain the predicted value which is obtained here like this, okay. The value of sigma square hat this here, this is obtained directly from the analysis of variance table that we had obtain from the software.

Now we would like to find out the standard error of this predictor, so you may recall that we have derived the expression like this. So I tried to substitute here the values of x naught and the values of x they are obtain from the given set of data and we obtain the value here which is approximately 251. So I can say that when the temperature is 36 degrees and the humidity is 36% then in that case the demand of a water is going to be nearly 389 9400 kilo litre with the standard error of 251.

(Refer Slide Time: 10:03)



Now we try to find out the 95% prediction interval, so you may recall that we had a derived the prediction interval under the case that sigma square is unknown like this obviously here we are trying to estimate the sigma square from the given set of data as sigma square hat so we are going to use this expression. Now in this expression if I try to substitute all these values over here, where values of this t alpha by 2 n-k they are obtain from the t tables.

And this is also obtained from the t tables, now once I try to do it here then the prediction interval turns out to be like this and you can observe that earlier we had obtained the p hat

that was 38994 and this is lying within this prediction interval, right. So now we have completed one task that forecasted the demand of water for a given temperature of 36 degrees and a given humidity level of 36%.

We have obtained its standard error as 251 in proper unit and the prediction interval is like this. So this completes the framework of the prediction of average value in this case. Okay, now we come to the aspect of actual value, so you may recall that that we had used the same predictor y=x beta hat for predicting the average value as well as actual value.

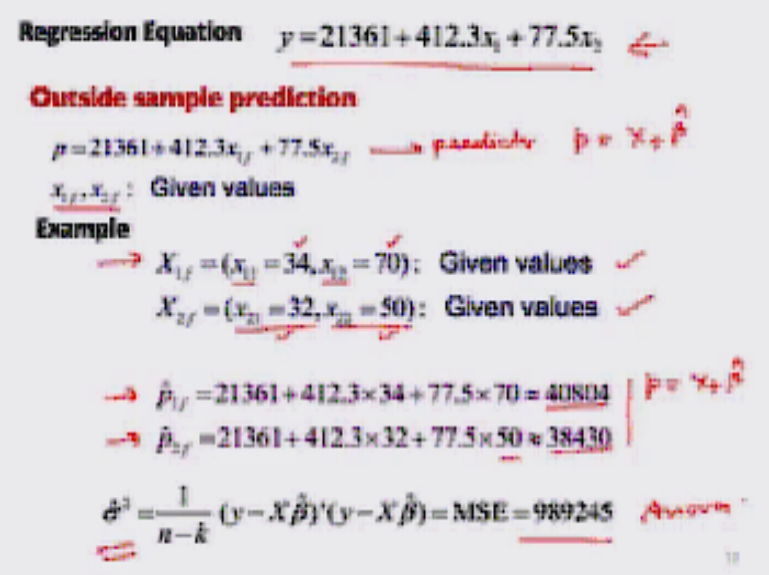
So when x one naught is given as 36 and x2 naught is given as thirty six again the predictor value will turn out to be the same like a 38994 as in the case of average value prediction we had obtained it, right. Now when we try to find out its standard error this we had obtained earlier, so now substituting the value of x naught x and sigma square hat we compute this standard error as 1025.

Now you may recall that the standard error of p in case of average value earlier was 251. So this indicates that the standard error of the same predictor when it is used for average value forecasting has lower standard error than the situation when it is used for actual value forecasting. You may also recall that we had derived a condition like n greater than two k to find whether the same predictor will behave better for average value prediction over the actual value prediction.

And this condition is also satisfied here because here n is27 and k here is three, so this confirms our theoretical finding also. Now we try to find out the 95% prediction interval for this actual value and we had derived the expression of prediction interval like this under the condition that sigma square is unknown and now once I try to substitute the value of x naught x and sigma square hat and obtaining that t value from the table of t probabilities we obtain the prediction interval here like this.

The point which we have to observe here that the prediction interval that we had obtain for the mean value prediction earlier was 38479 to 39509. You can see here that this prediction interval is wider, so obviously when we are trying to predict an actual value then in that case the predictor will have a wider confidence interval then when the predictor is used to forecast the average value. So this again confirms about theoretical finding.

(Refer Slide Time: 15:00)



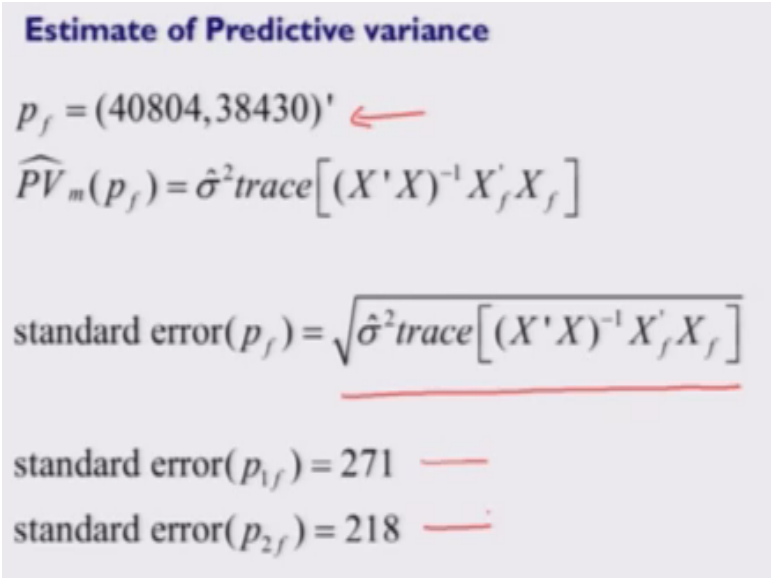
Now the next we consider the outside sample prediction remember that this was our regression equation that we had obtain earlier and our predictor was like this p=xf beta hat, right now the values of xf are given to us and since we have here only two variables x1 and x2, so I need two values x1 f and x2 f to forecast the value of y. So suppose I try to illustrate here that, suppose we want to forecast more than one values at the same time then how to do it?

So here I consider here two values means I say capital x1 f consist of two values x1 one x1 two and similarly the second set of values for the observation is given by x2 1=32 and x2 2=50, so in the first set we want to know the value of y when x1 takes value 34 and x2 takes value 17, and in the second case we would like to find out the value here y when x1 takes value 32 and x2 takes value 50.

So now using the predictor y=xf beta hat I have obtained these two values over here after making some elementary calculation. So I can say now from the first value that when the temperature is 34 degrees and the humidity is high say up to 70% then the demand for the water is going to be 4080 4000 kilo litre and when the temperature is 32 and the humidity level is 50% then the demand of water is going to be reduced.

And in that case this is going to be 38430 kilo litre and if you see this confirms the real life story also when the temperature is less humidity is less people try to consume less water. Okay now the value of sigma square we have obtained directly from the analysis of variance stable that we had obtained in the software.

(Refer Slide Time: 17:42)

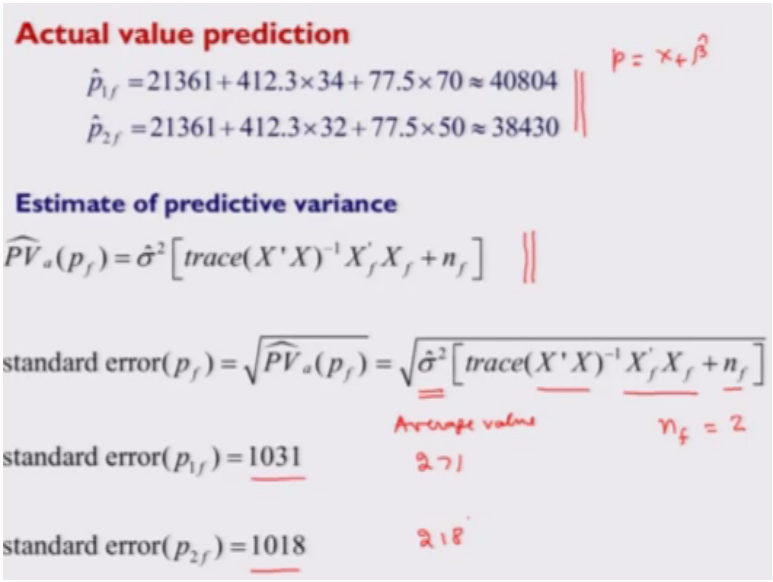


Now based on that we have here two values of predictor like this, and we calculate their standard errors and in order to calculate the standard errors you may recall that we had obtained the predictive variance like this, and once try I to substitute the values of x1 f x2 f and the earlier given values of x matrix, we obtain the standard error of the first forecasted value as 271 and the standard error of second forecast has 298.

So you can see that it is not difficult to calculate such standard errors and given the set of data, so now we try to obtain the prediction intervals for this average value. So in this case you again recall that we had obtained the prediction interval under the case that sigma square is unknown to us by this expression. So now again this value of t that is going to be obtain from the table of t probability, and now we all these values x matrix and x f matrix also.

So I simply try to substitute these values over here and we can obtain the prediction interval in the first case for x1 f like this and in the second case like this. Just for the sake of information you may recall that the value p1 f was 40840 and the value of p2 f was 38433, so again we can see that these two values are lying within this prediction interval.

(Refer Slide Time: 20:08)



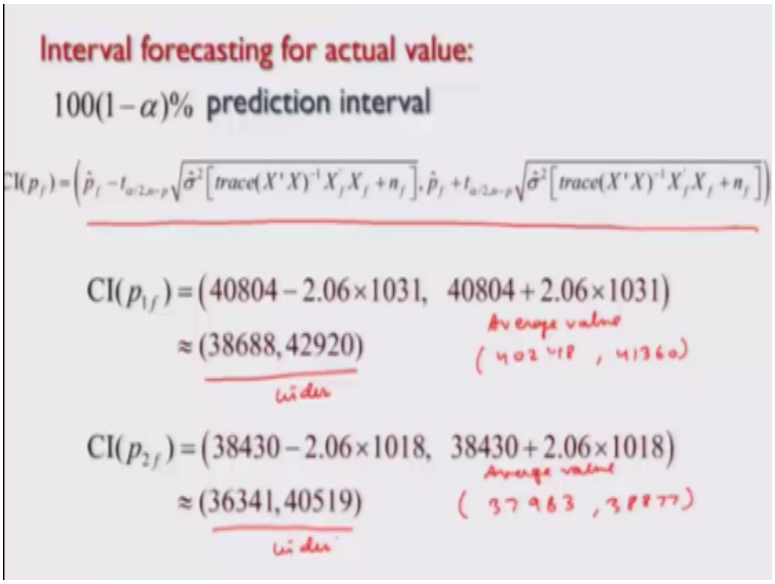
Now next we try to make the actual value prediction up to now we have considered the average value predictions. So now again as we have seen that the predictor remains the same in both the cases whether we are interested in average value prediction or say actual value prediction, so these two values they will remain the same as earlier, but the difference will come when we are trying to find out their standard errors.

So we had earlier derived the expression of predictive variance of pf like this and now when I try to substitute the values of sigma square hat, x matrix and x f matrix and using the value here n f here n f we are interested in making the forecast for two values then I calculate the standard errors of the first prediction as 1031 and the standard error of the second prediction as 1018.

I would like to draw your attention here that when we were interested in the average value forecasting then we had obtained this standard error for the first predicted value as 272 and for the second value as 218. So now you can again observe that the standard errors for the actual value predictions they are higher than the standard error for the average value prediction, and this confirms our theoretical finding also.

So now we consider the prediction interval for these two cases you may kindly recall that we had obtained the expression for the prediction interval earlier like this.

(Refer Slide Time: 22:06)



And now we have here two values for corresponding to p1 f and p2 f, and we substitute these values over here in this expression and we obtain the prediction interval for p f like this for the first observation and the prediction interval for the second set of observations like this. Okay, now you may also observe one thing over here that earlier when we try to forecast the average value then in that case corresponding to the first set of observation this prediction interval was 40248 240360.

And in the case of second set of observation this average value prediction interval was 37983, 238877, so again you see that the prediction interval for the actual values here they are wider in comparison to their prediction interval for the average value and this again confirms theoretical finding that when we are trying to forecast the actual value the standard errors are going to be higher and the prediction intervals are going to be wider.

So now here you can see it is possible that some software may have such facility to obtain these expressions directly but any way they only need a one line requirement we have seen that forecasting is very important, but that depends on the modelling. In case if the fitted is statistical model is good you are going to get a better forecast. Now when we are obtaining the regression model we have to keep in mind that this is not a wonder step procedure.

We start with several sets of variable and then there are some other aspects which are to be taken care like as variable selection, model selection and after finding out the model we try to see how we can improve them for example by taking a suitable transformation or by adding or deleting sum variables and so on and after that we again try to fit the model and we try to check the model performance and finally after several iteration usually we get a good model.

Now once we get a model forecasting is not difficult that is what we have illustrated in this lecture we have taken more time in explaining the basic concepts of modelling, and then forecasting was not so difficult. So with this lecture I would like to conclude this course and wish you all the best, thank you.