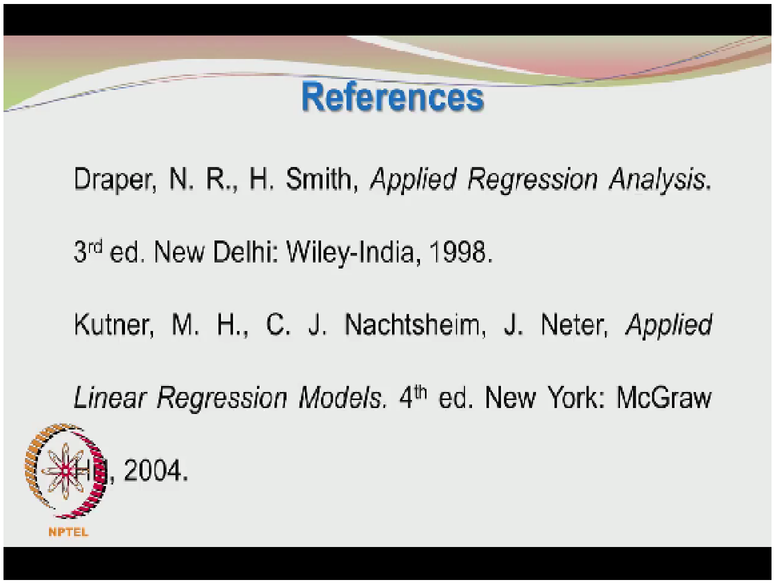
Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture – 41
Regression Analysis: Example Set 8

Hello. In today's class, we will be looking at the regression concepts through the working of an example problem.

(Refer Slide Time: 00:24)



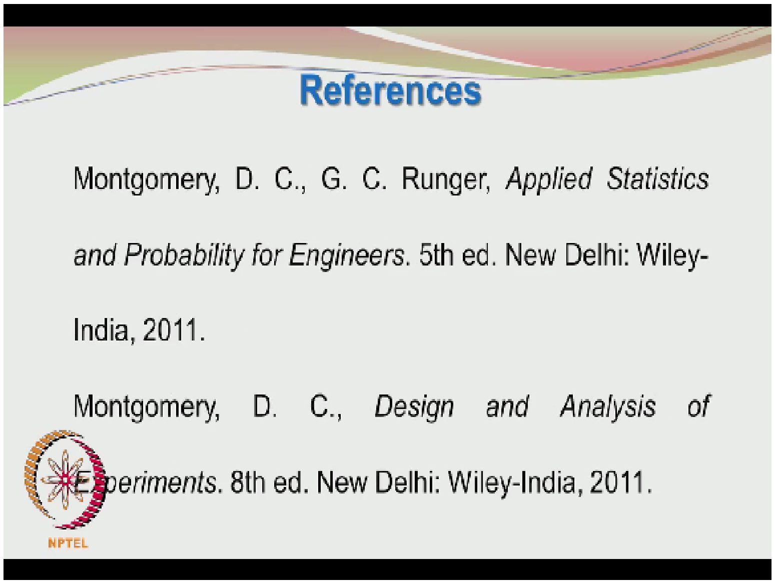
References

Draper, N. R., H. Smith, *Applied Regression Analysis.* 3rd ed. New Delhi: Wiley-India, 1998.

Kutner, M. H., C. J. Nachtsheim, J. Neter, *Applied Linear Regression Models.* 4th ed. New York: McGraw Hill, 2004.

The references are the book written by Draper and Smith titled Applied Regression Analysis, Kutner et al Applied Linear Regression Models, fourth edition, McGraw-Hill.

(Refer Slide Time: 00:40)



References

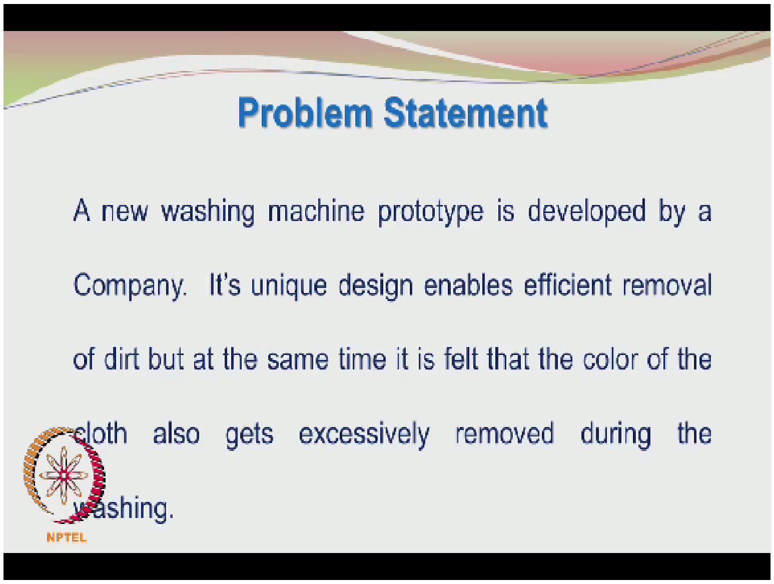
Montgomery, D. C., G. C. Runger, *Applied Statistics and Probability for Engineers*. 5th ed. New Delhi: Wiley-India, 2011.

Montgomery, D. C., *Design and Analysis of Experiments*. 8th ed. New Delhi: Wiley-India, 2011.

NPTEL

And then the prescribed textbook Montgomery and Ranger Applied Statistics and Probability for Engineers fifth edition and Montgomery's Design and Analysis of Experiments.

(Refer Slide Time: 00:54)



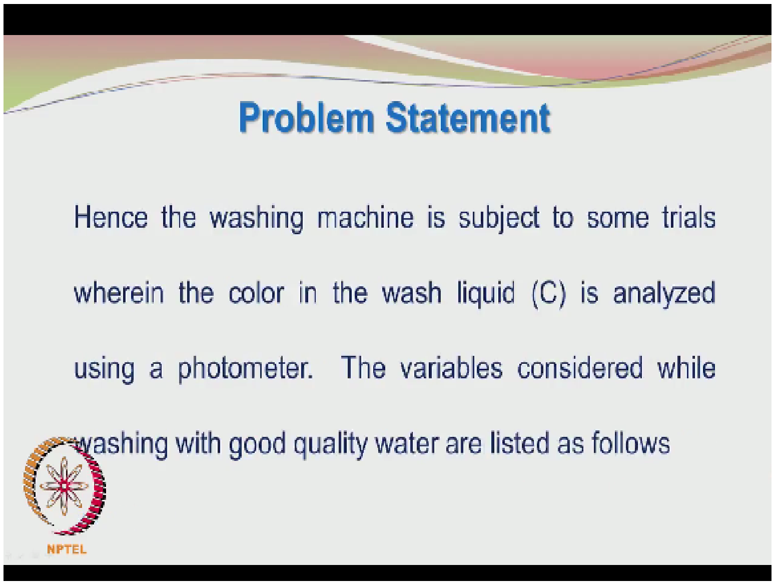
Problem Statement

A new washing machine prototype is developed by a Company. It's unique design enables efficient removal of dirt but at the same time it is felt that the color of the cloth also gets excessively removed during the washing.

NPTEL

The problem statement goes like this. A new washing machine prototype is developed by a company. It's unique design enables efficient removal of dirt but at the same time, it is felt that the colour of the cloth also gets excessively removed during the washing.

(Refer Slide Time: 01:13)

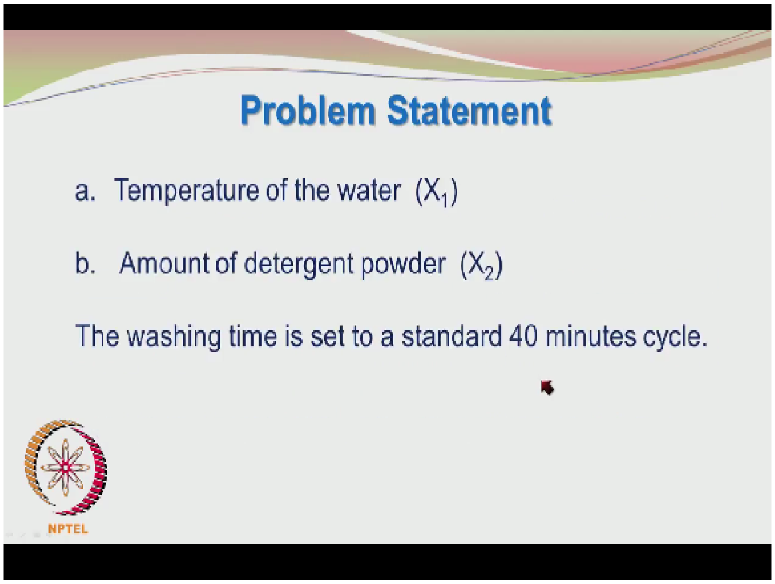


Problem Statement

Hence the washing machine is subject to some trials wherein the color in the wash liquid (C) is analyzed using a photometer. The variables considered while washing with good quality water are listed as follows

So the washing machine is subject to some trials where in the colour in the wash liquid C is analysed using a photometer. The variables considered while washing with good quality water are listed as follows.

(Refer Slide Time: 01:30)



Problem Statement

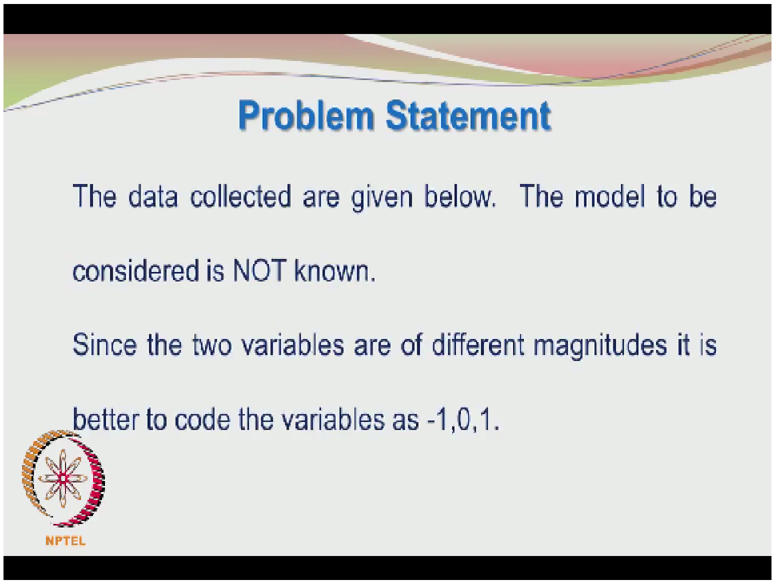
a. Temperature of the water (X_1)

b. Amount of detergent powder (X_2)

The washing time is set to a standard 40 minutes cycle.

Temperature of the water X1. Amount of detergent powder used X2. The washing time is set to a standard 40 minutes' cycle.

(Refer Slide Time: 01:43)



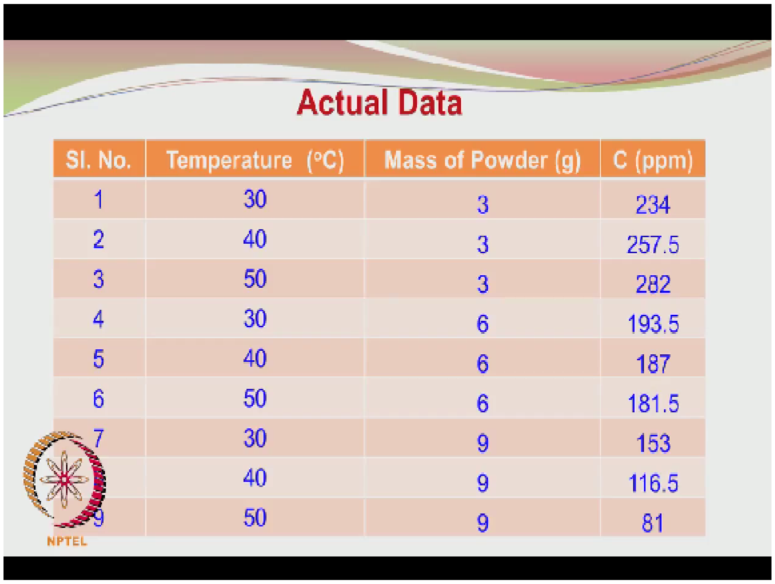
Problem Statement

The data collected are given below. The model to be considered is NOT known.

Since the two variables are of different magnitudes it is better to code the variables as -1,0,1.

The data collected are given in the form of a table below. We have to develop the model. The model is not given to us and if you look at the table, right table, I will first describe.

(Refer Slide Time: 02:04)

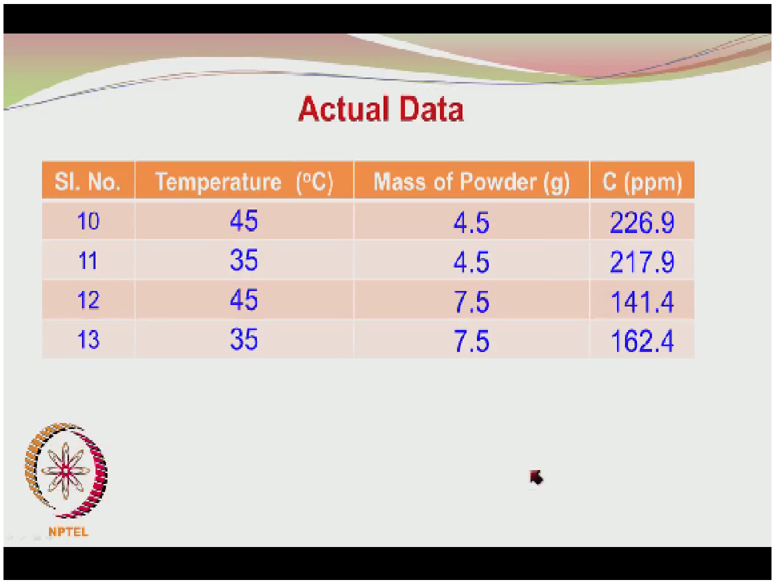


Actual Data

Sl. No.	Temperature (°C)	Mass of Powder (g)	C (ppm)
1	30	3	234
2	40	3	257.5
3	50	3	282
4	30	6	193.5
5	40	6	187
6	50	6	181.5
7	30	9	153
[illegible]	40	9	116.5
9	50	9	81

You can see the temperature of water is kept either at 30 degrees or 40 degrees or 50 degrees in the first 9 readings. The amount of powder used varies between 3 grams to 9 grams.

(Refer Slide Time: 02:26)



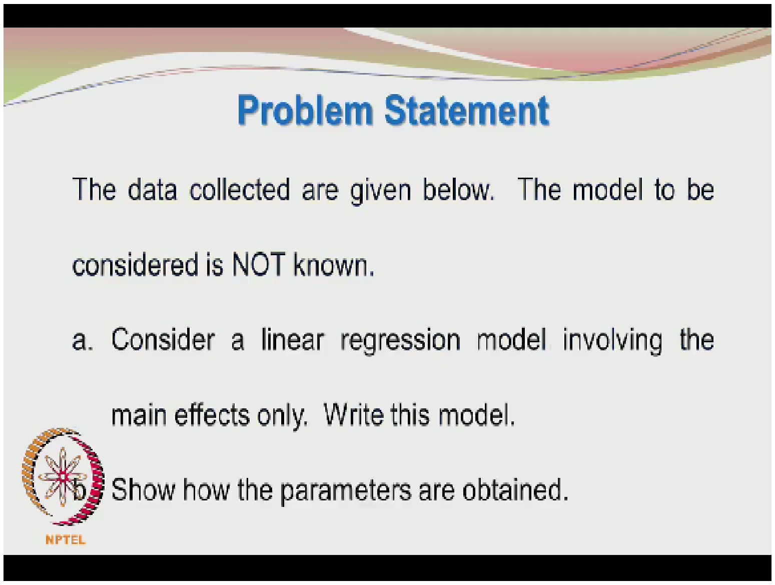
Actual Data

Sl. No.	Temperature (°C)	Mass of Powder (g)	C (ppm)
10	45	4.5	226.9
11	35	4.5	217.9
12	45	7.5	141.4
13	35	7.5	162.4

The data table continues and we have temperature at 45 degrees Centigrade and 35 degrees Centigrade. Mass of powder is 4.5 grams and 7.5 grams. The concentration in PPM is also given in the last column. Let us now go to the questions. Before we look at the questions, since the temperature was varying between 30 to 60 degrees Centigrade and the mass of the powder was between 3 and 9 grams.

There is an approximate one order of magnitude difference between the 2 variables. So it is better if we code the variables. So the variables are coded in the range between -1 to 1.

(Refer Slide Time: 03:23)



Problem Statement

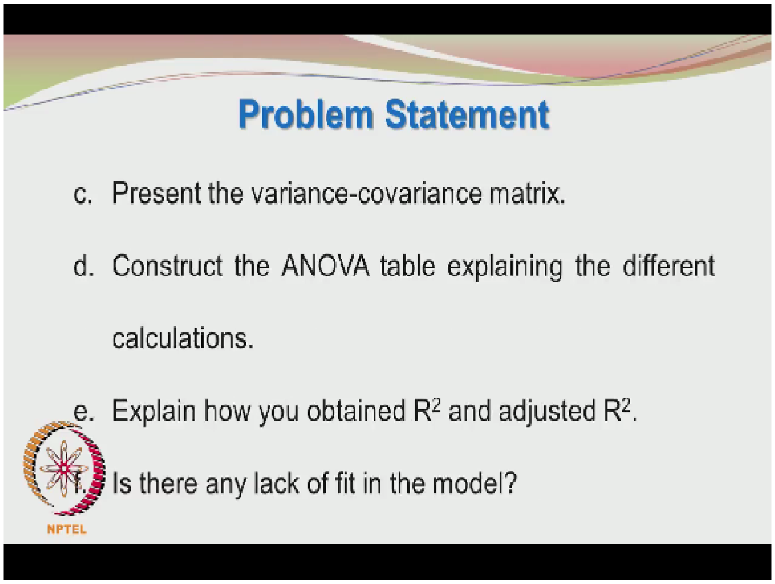
The data collected are given below. The model to be considered is NOT known.

a. Consider a linear regression model involving the main effects only. Write this model.

b. Show how the parameters are obtained.

So we have to consider a linear regression model involving the main effects only. Write down this model. Show how the parameters are obtained.

(Refer Slide Time: 03:34)



Problem Statement

c. Present the variance-covariance matrix.

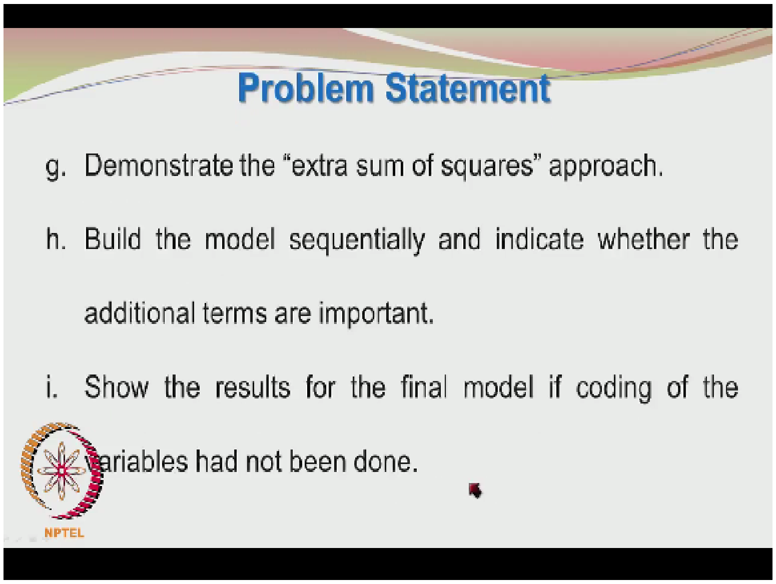
d. Construct the ANOVA table explaining the different calculations.

e. Explain how you obtained R^2 and adjusted R^2.

f. Is there any lack of fit in the model?

Present the variance-covariance matrix. Constructed the ANOVA table explaining the different calculations. Explain how you obtained R squared and adjusted R squared. Is there any lack of it in the model?

(Refer Slide Time: 03:47)



Problem Statement

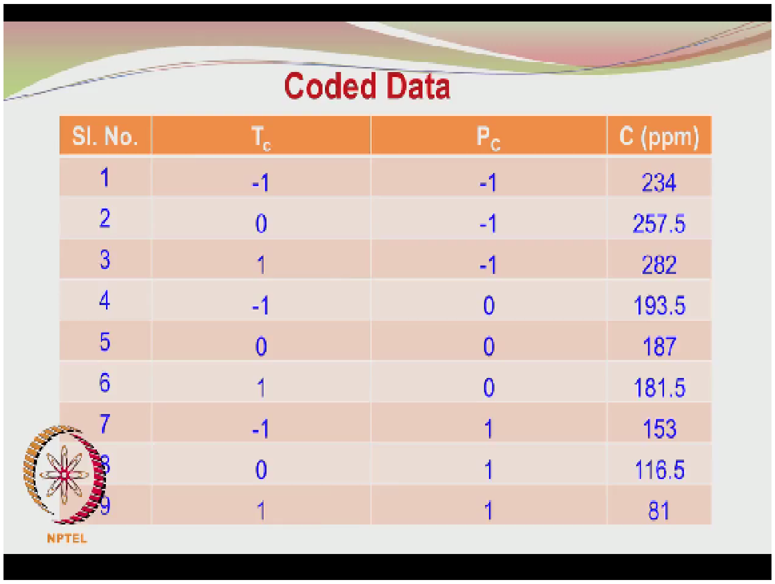
g. Demonstrate the "extra sum of squares" approach.

h. Build the model sequentially and indicate whether the additional terms are important.

i. Show the results for the final model if coding of the variables had not been done.

Demonstrative the extra sum of squares approach. Build the model sequentially and indicate whether the additional terms are important. Show the results of the final model if the coding of the variables had not been done. So we saw the data already.

(Refer Slide Time: 04:05)



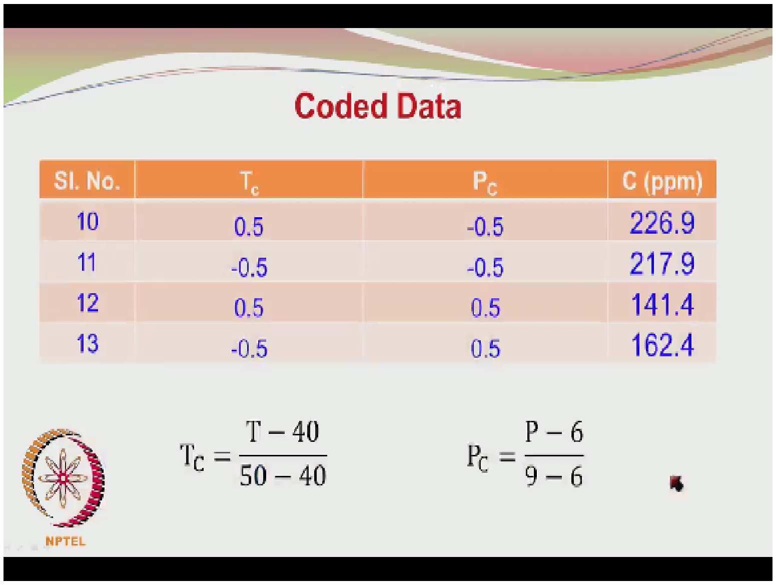
Coded Data

Sl. No.	T_c	P_C	C (ppm)
1	-1	-1	234
2	0	-1	257.5
3	1	-1	282
4	-1	0	193.5
5	0	0	187
6	1	0	181.5
7	-1	1	153
8	0	1	116.5
9	1	1	81

It is the actual data and when you express the data in the coded form, we get a table like this. The lowest setting of temperatures is kept at -1. The lowest setting of powder used, the weight of powder used rather is also set at -1 and the maximum value of temperatures is set at +1. So all these are minimum and the maximum amount of powder used which is 9 grams is coded at +1. So the intermediate setting is coded 0.

Similarly, the intermediate powder loading in the machine which is 6 grams is coded at 0. We do not touch the colour in the liquid. We keep the data as it is.

(Refer Slide Time: 04:56)



Coded Data

Sl. No.	T_c	P_C	C (ppm)
10	0.5	-0.5	226.9
11	-0.5	-0.5	217.9
12	0.5	0.5	141.4
13	-0.5	0.5	162.4

$$T_C = \frac{T-40}{50-40} \qquad P_C = \frac{P-6}{9-6}$$

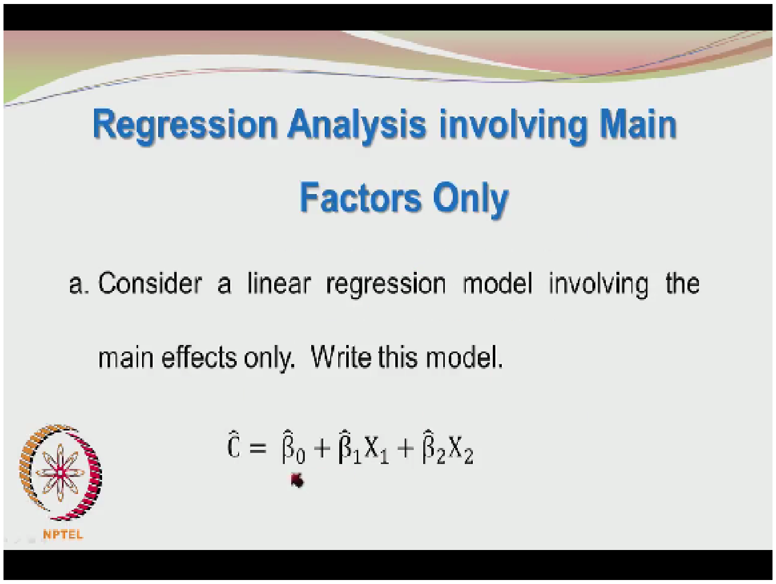
And since we are also doing additional 4 runs at intermediate settings at 45 degrees, gets a coding of 0.5 and 4.5. Let us see what is -0.5 correspond to that is 4.5 grams. So that is coming as -0.5 grams, sorry -0.5, the coded format and 7.5 grams is coming as +0.5. So 7.5-6 is 1.5 and 9-6 is 3. So 1.5/3 is 0.5. So this is how we have coded the different variables or the factors? Why do we do the coding.

The raw data was varying by about an order of magnitude. So the coefficient in the regression model associated with this raw data would be high and the regression coefficient associated with this raw data of temperature would be correspondingly low and this kind of order of magnitude difference may actually increase. In other words, one variable would be in the order of 1000s and the other variable would be in the order of fractions. So the difference between the 2 would be very considerable and the regression coefficient is also would correspondingly adjust.

So this may lead to wide variation in the estimated regression coefficients. The regression coefficient with the higher number may be in the order of 10 power -2 or 10 power -3, whereas the regression coefficient associated with the variable which is having very low values would correspondingly be quite high in the order of 10 power 2 or 10 power 3. And this may lead to a kind of an unwieldy model. The problem is also increased if the units are changed by mistake.

So this becomes a dimensional equation and when you have a dimensional equation, the units are very important. Suppose somebody by mistake puts Kelvin instead of Centigrade, then the regression would give wrong answers. Similarly, if the powder, by mistake the person puts milligrams, then again the wrong results would be obtained but once you do the coding properly and the user is warned that the variables are coded, then the problem is solved. It is another advantage of coding. The regression variables become independent of units in the coded format.

(Refer Slide Time: 07:57)



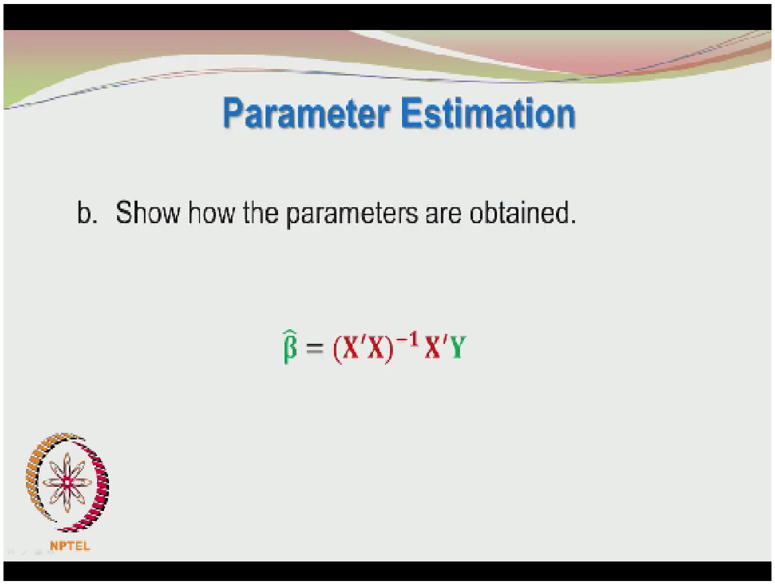
Regression Analysis involving Main Factors Only

a. Consider a linear regression model involving the main effects only. Write this model.

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

So let us look at the regression analysis involving the main factors only. So we will write down the model. The predicted colour is given in terms of beta0 hat+beta 1 hat X1+beta 2 hat X2. Just correct the typo here, right. This is a very important term in the regression model. It gives the value of the concentration when both X1 and X2 are 0. There is more to this beta0 hat which we will see in the coming slides.

(Refer Slide Time: 08:38)



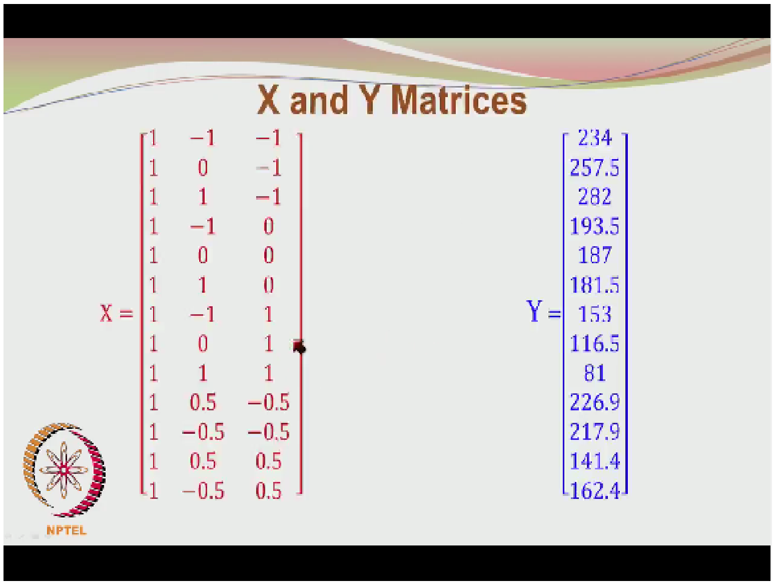
Parameter Estimation

b. Show how the parameters are obtained.

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

So the next part of the question is, show how the parameters are obtained? We first express the data in the matrix form which I will show in the next few slides and once the X prime X matrix is set up, then we take the inverse of the X prime X matrix and we also take the X prime Y matrix and we pre-multiply the X prime Y matrix with X prime X inverse and we get the vector of the estimated parameters.

(Refer Slide Time: 09:12)

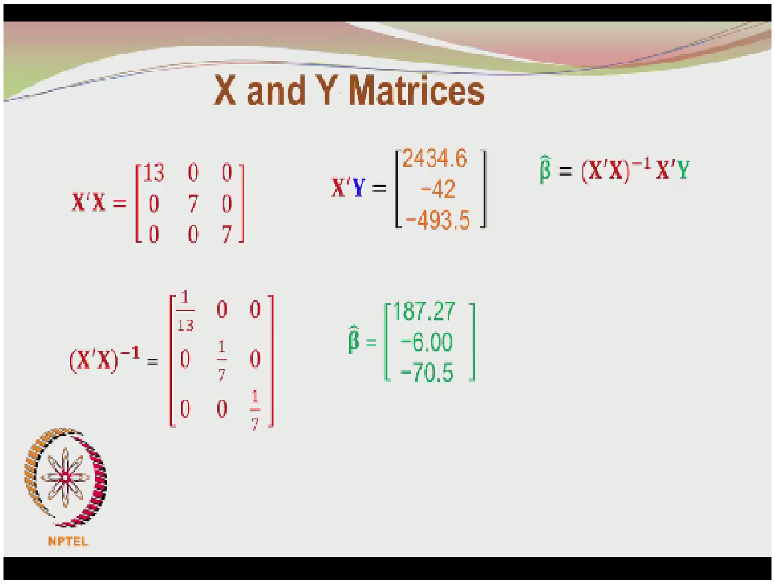


X and Y Matrices

$$X = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0.5 & -0.5 \\ 1 & -0.5 & -0.5 \\ 1 & 0.5 & 0.5 \\ 1 & -0.5 & 0.5 \end{bmatrix} \qquad Y = \begin{bmatrix} 234 \\ 257.5 \\ 282 \\ 193.5 \\ 187 \\ 181.5 \\ 153 \\ 116.5 \\ 81 \\ 226.9 \\ 217.9 \\ 141.4 \\ 162.4 \end{bmatrix}$$

So we have the X matrix. The first column is the column of ones and the second column is having -1 1 -1 1 -1 1 0.5 -0.5 0.5 -0.5 and the third column is having data like this. So all of them are coded and the Y matrix represents the concentrations recorded at each experimental setting. So when you add up the elements of the first column, we will get the number 13 because there are 13 settings in the experimental program 1 2 3 4 5 6 7 8 9 10 11 12 13. So we have 13 elements in the first column and we also have 13 observations here in the Y column vector. This shows that 13 experiments have been carried out.

(Refer Slide Time: 10:23)



X and Y Matrices

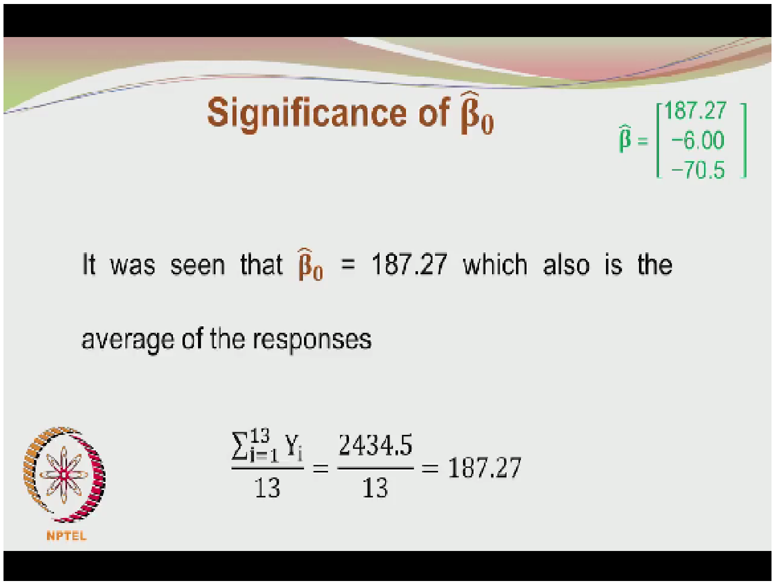
$$X'X = \begin{bmatrix} 13 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 7 \end{bmatrix} \qquad X'Y = \begin{bmatrix} 2434.6 \\ -42 \\ -493.5 \end{bmatrix} \qquad \hat{\beta} = (X'X)^{-1}X'Y$$

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{13} & 0 & 0 \\ 0 & \frac{1}{7} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} 187.27 \\ -6.00 \\ -70.5 \end{bmatrix}$$

So when you take X prime X, X prime here would correspond to changing columns into rows and rows into columns. So that would be the transpose of X. When you take X and pre-multiply it with the transpose of X X prime, we get X prime X and that comes to nicely 13 0 0 0 7 0 0 0 7. So this is very nice matrix because there are no off diagonal elements. The off diagonal terms are all 0 and inverting this matrix is a piece of cake. A matrix inverse is so defined that when you multiply a given matrix with its inverse, you should get the identity matrix.

So we obtain the inverse of this X prime X matrix by taking 1/13 1/7 and 1/7. You also have the X prime Y matrix which is given by these 3 row elements and we can also find out the beta hat. The parameters for the regression model which is X prime X inverse X prime Y and we get beta hat as these numbers. I request you to do these calculations on your own and make sure that the answers I am getting are matching with your answers.

(Refer Slide Time: 12:02)



Significance of $\widehat{\beta}_0$

$$\widehat{\beta} = \begin{bmatrix} 187.27 \\ -6.00 \\ -70.5 \end{bmatrix}$$

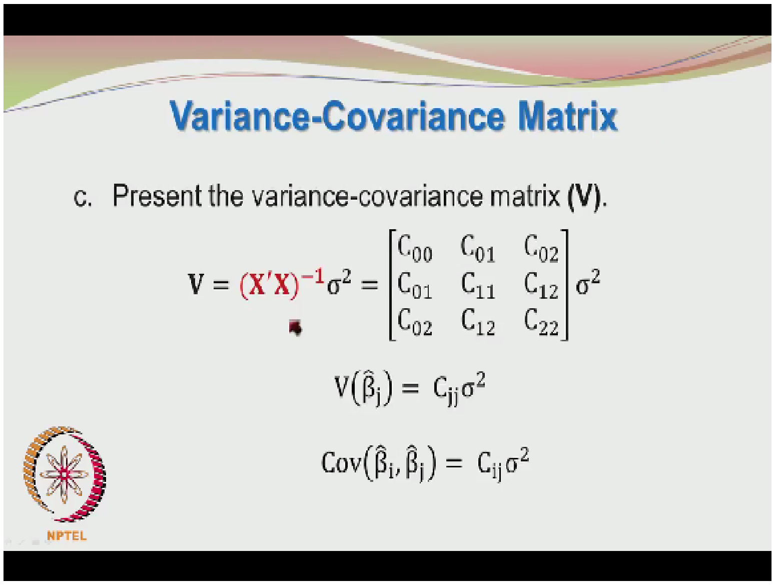
It was seen that $\widehat{\beta}_0$ = 187.27 which also is the average of the responses

$$\frac{\sum_{i=1}^{13} Y_i}{13} = \frac{2434.5}{13} = 187.27$$

And what is the significance of beta0 hat when the experimental observations are put in the coded format, experimental settings are rather put in the coded format. We are not putting the response in the coded format, only the experimental settings are put in the coded format. We carry out the regression parameter estimation and we get the different parameters. We get beta0 hat as 187.27. This means that the average of the responses is 187.27.

This is not true for the uncoded case. Please note the important distinction when we code it in the way I have shown, we get the average as the beta0 hat, that is the average response.

(Refer Slide Time: 12:58)



Variance-Covariance Matrix

c. Present the variance-covariance matrix **(V)**.

$$V = (X'X)^{-1}\sigma^2 = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{01} & C_{11} & C_{12} \\ C_{02} & C_{12} & C_{22} \end{bmatrix} \sigma^2$$

$$V(\hat{\beta}_j) = C_{jj}\sigma^2$$

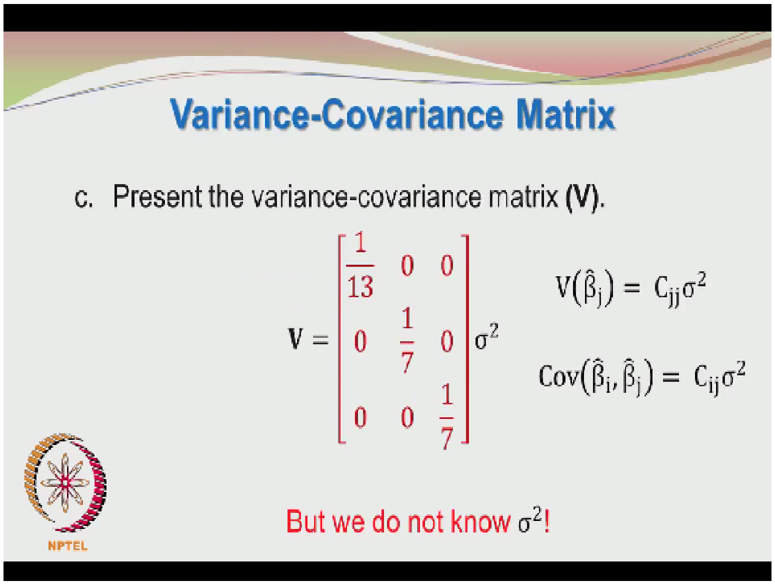
$$Cov(\hat{\beta}_i, \hat{\beta}_j) = C_{ij}\sigma^2$$

Now you can present the variance-covariance matrix V in the following form. This we saw in one of the previous lectures and the variance of the regression parameter or the regression coefficient is given by the diagonal terms of the variance-covariance matrix multiplied by sigma squared. So variance of beta hat j=Cjj sigma squared and then we have also the covariance between 2 regression coefficients or regression parameters, beta hat I and beta hat j is given by Cij sigma squared.

So beta hat 0 or beta hat 0, beta hat 2 i=0 and j=2. So we are looking at the Cij C02 and C02*sigma squared, that gives a covariance between the 2 parameters and you can also see that C02 will match with C02 here. This should be actually written as C02 and this is C20 if you number the rows from 0, 1 and 2 but the matrix is symmetric, the variance-covariance matrix is symmetric and so C02 will match with C20.

Similarly, C01 here will match with C10, so they are pretty much written as C01 itself and so on. The important thing to note here is the diagonal elements of this matrix give the variances of the appropriate regression parameters whereas the off diagonal terms give the covariance between a pair of regression parameters.

(Refer Slide Time: 15:09)



Variance-Covariance Matrix

c. Present the variance-covariance matrix **(V)**.

$$\mathbf{V} = \begin{bmatrix} \frac{1}{13} & 0 & 0 \\ 0 & \frac{1}{7} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix} \sigma^2$$

$$V(\hat{\beta}_j) = C_{jj}\sigma^2$$

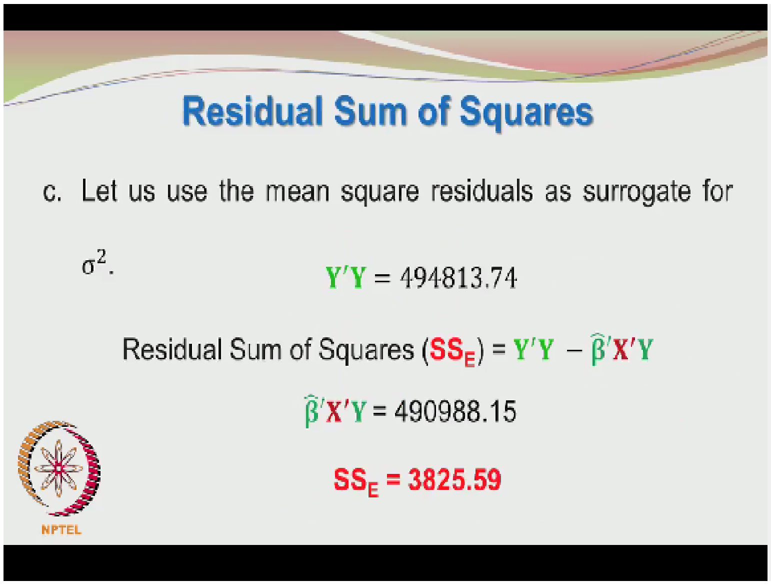
$$Cov(\hat{\beta}_i, \hat{\beta}_j) = C_{ij}\sigma^2$$

But we do not know σ^2!

So now we can present the variance-covariance matrix in terms of numbers. We found X prime X inverse as 1/13 0 0 0 1/7 0 0 0 1/7 sigma squared but we hit a roadblock here. We are unable to get a value for sigma squared because sigma squared is not provided. So we have to do with the best estimate of sigma squared. So we will use an estimated sigma squared. The sigma squared of course refers to the error variance which is assumed to be constant.

The errors are assumed to be normally distributed with 0 mean and constant variance sigma squared. So that sigma squared is known to us in many situations. So we have to have an estimate of the sigma squared.

(Refer Slide Time: 16:08)



Residual Sum of Squares

c. Let us use the mean square residuals as surrogate for σ^2.

$$\mathbf{Y'Y} = 494813.74$$

$$\text{Residual Sum of Squares } (\mathbf{SS_E}) = \mathbf{Y'Y} - \hat{\beta}'\mathbf{X'Y}$$

$$\hat{\beta}'\mathbf{X'Y} = 490988.15$$

$$\mathbf{SS_E = 3825.59}$$

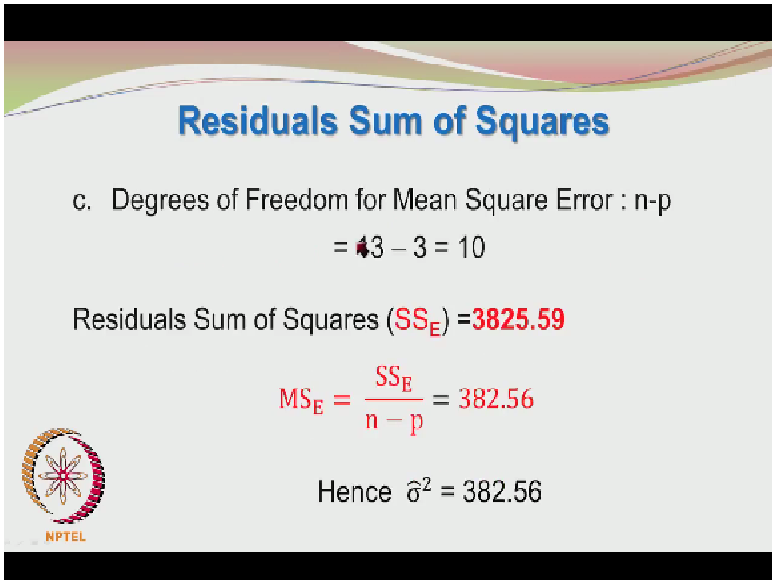
First let us find the residual sum of squares. We use the mean square residuals as surrogate for sigma squared. So to find the mean square residuals, we need to find the residual sum of squares, divide the residual sum of squares by the degrees of freedom for the residual sum of squares and we get the mean square residuals. So we have Y prime Y=494813.74. How did we get Y prime Y? We go back to the Y vector here. We take the transpose of this vector.

You can visualise that all these column elements becoming row elements when you take the transpose of Y. Then you multiply Y prime with Y. What will then happen is, you are essentially finding out 234 squared+257.5 squared and so on. So all the elements in the response vector are squared and added to get Y prime Y. You may verify this by doing the calculations yourself. So that the X care of Y prime Y, the residual sum of squares we saw is given by Y prime Y-beta hat prime X prime Y.

And sometimes the residual sum of squares may also be called as the error sum of squares because that represents the deviation between the experimental value and the model prediction and that represents the error. We also distinguish this residual sum of squares from the pure error sum of squares that we saw in the previous lecture. Pure error is obtained from repeated measurements. So we have beta hat prime X prime Y as 490988.15.

What is that beta hat prime X prime Y. We estimated the parameters beta hat, we take the transpose of this and then we multiply with X prime Y matrix and that gives us the regression sum of squares. It is important to note that this regression sum of squares includes the contribution from the beta 0 regression coefficient. So we subtract Y prime Y with beta hat prime X prime Y, we get sum of squares of the residuals as 3825.59.

(Refer Slide Time: 19:02)



Residuals Sum of Squares

c. Degrees of Freedom for Mean Square Error : n-p

= 13 – 3 = 10

Residuals Sum of Squares (SS_E) =**3825.59**

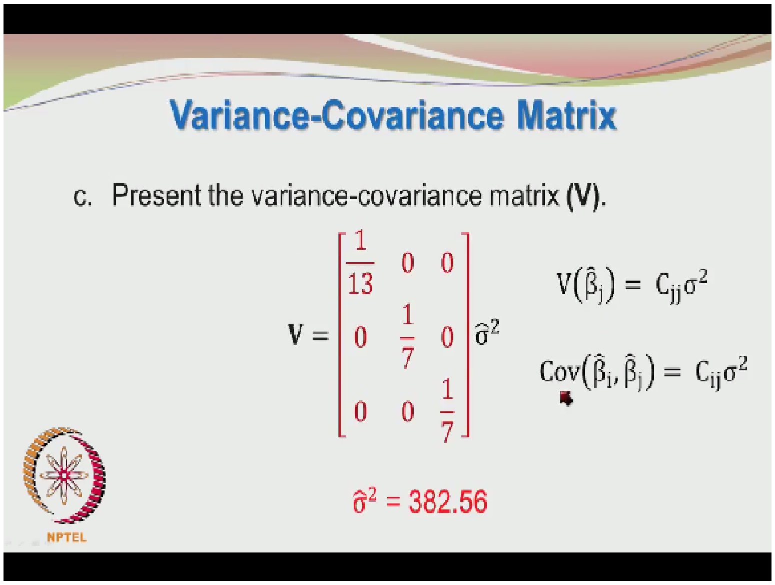
$$MS_E = \frac{SS_E}{n-p} = 382.56$$

Hence $\hat{\sigma}^2 = 382.56$

The degrees of freedom for the mean square error or mean square residuals is n-p which is 13-3. You have 3 parameters estimated from the year regression. You have 13 experimental observations. So we have 10 degrees of freedom for the mean square error and the residual sum of squares we saw from the previous slide as 3825.59 and that you divided by 10, the degrees of freedom attached to the residual sum of squares and we get 382.56.

So an estimate of the error variance given by sigma hat squared is the mean square of the residuals which is 382.56 or sigma hat=19.56.

(Refer Slide Time: 20:00)



Variance-Covariance Matrix

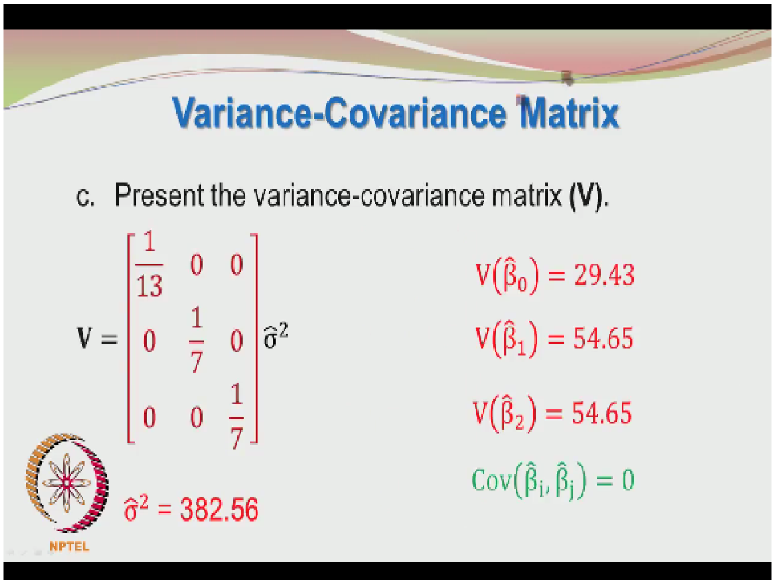
c. Present the variance-covariance matrix (**V**).

$$\mathbf{V} = \begin{bmatrix} \frac{1}{13} & 0 & 0 \\ 0 & \frac{1}{7} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix} \hat{\sigma}^2 \qquad V(\hat{\beta}_j) = C_{jj}\sigma^2 \qquad Cov(\hat{\beta}_i, \hat{\beta}_j) = C_{ij}\sigma^2$$

$\hat{\sigma}^2 = 382.56$

And now we can estimate the variance of the different regression parameters because we have plugged in instead of sigma squared, we have used sigma hat squared and sigma hat squared was 382.56. Now we can easily estimate the variance of the individual regression parameters. There is no problem with the covariance terms because they are all identically 0.

(Refer Slide Time: 20:31)



Variance-Covariance Matrix

c. Present the variance-covariance matrix (**V**).

$$\mathbf{V} = \begin{bmatrix} \frac{1}{13} & 0 & 0 \\ 0 & \frac{1}{7} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix} \hat{\sigma}^2$$

$\hat{\sigma}^2 = 382.56$

$V(\hat{\beta}_0) = 29.43$

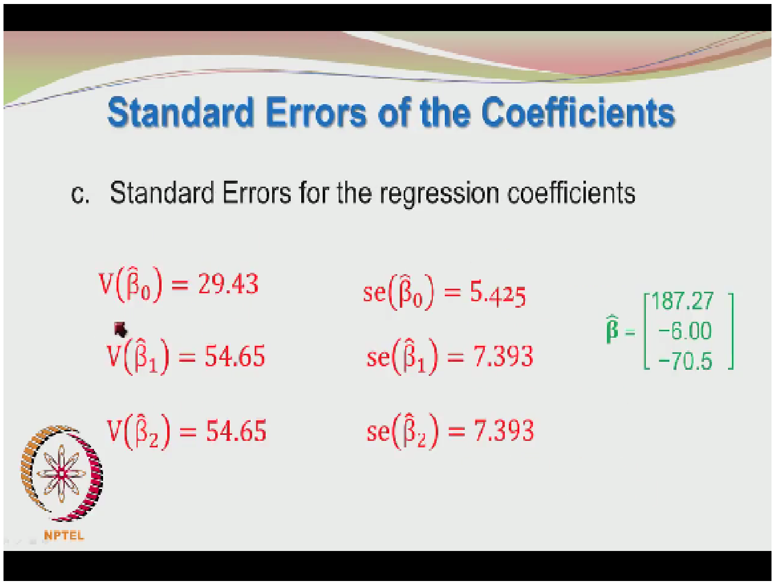
$V(\hat{\beta}_1) = 54.65$

$V(\hat{\beta}_2) = 54.65$

$Cov(\hat{\beta}_i, \hat{\beta}_j) = 0$

So once we have this 382.56/13 is approximately 30 and that is what we have here because 13*30 is 390 and then the variance of beta hat 1 is 382.56/7 which would be 54.65, that is fine and variance of beta 2 hat is also 54.65 because these 2 elements are identical and covariance between a pair of regression parameters is 0.

(Refer Slide Time: 21:15)



Standard Errors of the Coefficients

c. Standard Errors for the regression coefficients

$V(\hat{\beta}_0) = 29.43$ $se(\hat{\beta}_0) = 5.425$

$V(\hat{\beta}_1) = 54.65$ $se(\hat{\beta}_1) = 7.393$

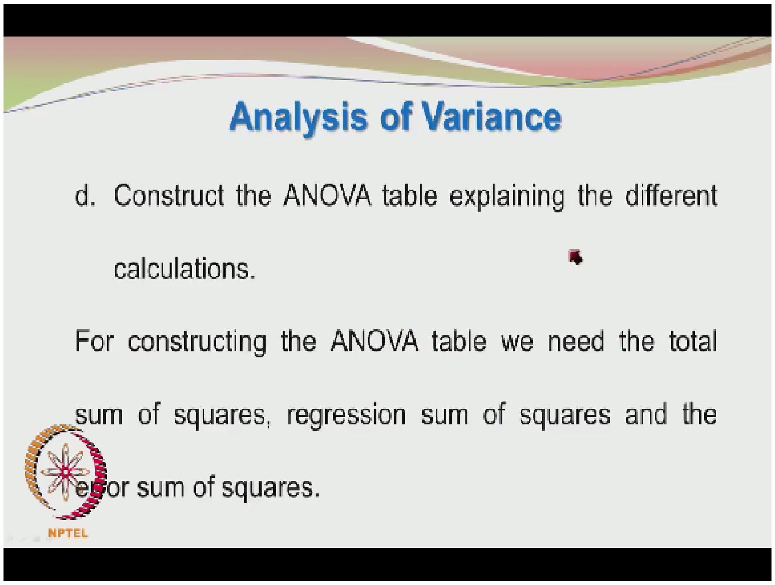
$V(\hat{\beta}_2) = 54.65$ $se(\hat{\beta}_2) = 7.393$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 187.27 \\ -6.00 \\ -70.5 \end{bmatrix}$$

Once you get the variance of the different parameters, we can find the standard error of the different parameters and we take the square root of these variances and we get 5.425, 7.4 and again 7.4. We can compare these estimated standard errors with the beta 0 hat, beta hat 1 and beta hat 2; beta hat 0, beta hat 1, beta hat 2. so we can see that these values are quite okay except for this particular term, here the parameter is -6.

And here we have 7.393 that is a cause for worry as for as the first regression coefficient is concerned. The others are looking okay. It is about 1/10th of the estimated parameter value.

(Refer Slide Time: 22:22)



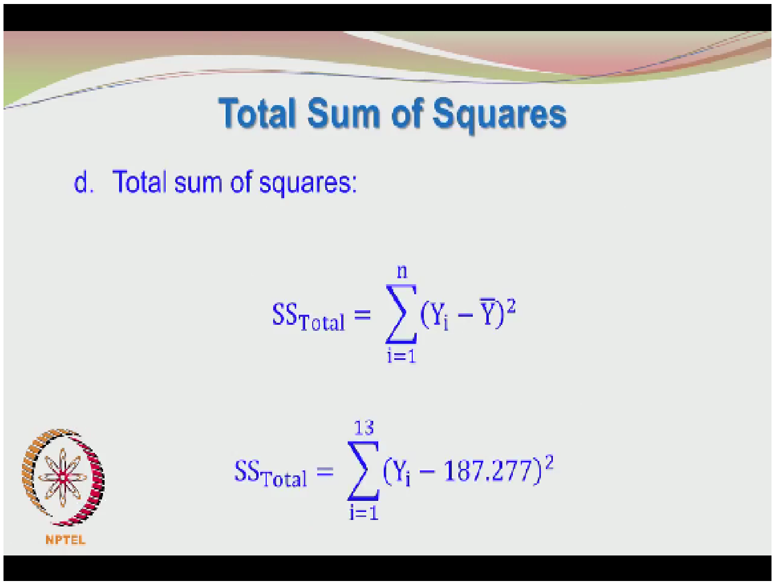
Analysis of Variance

d. Construct the ANOVA table explaining the different calculations.

For constructing the ANOVA table we need the total sum of squares, regression sum of squares and the error sum of squares.

So we can construct the ANOVA table explaining the different calculations. So for doing this, we need the total sum of squares, regression sum of squares and the error sum of squares.

(Refer Slide Time: 22:38)



Total Sum of Squares

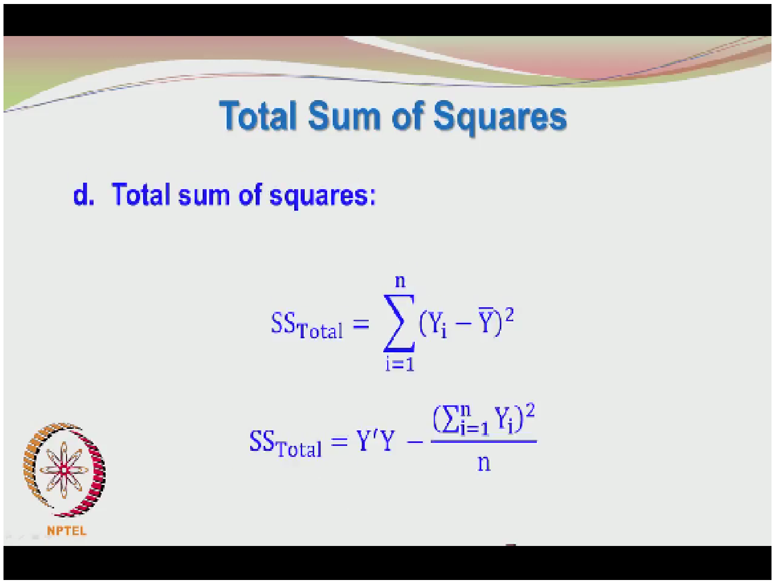
d. Total sum of squares:

$$SS_{Total} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$SS_{Total} = \sum_{i=1}^{13} (Y_i - 187.277)^2$$

The total sum of squares by now should be very easy for you to find. It is simply the deviation of Yi with respect to the average value of the responses that is Y bar. We take the deviation, square them. We take every deviation, square every deviation and then add them up. So we get i=1 to n Yi-Y bar whole squared where n is the number of experiments performed which is 13, Yi is the experimental response and Y bar is the average of the experimental responses. In this case, it comes to 187.277.

(Refer Slide Time: 23:28)



Total Sum of Squares

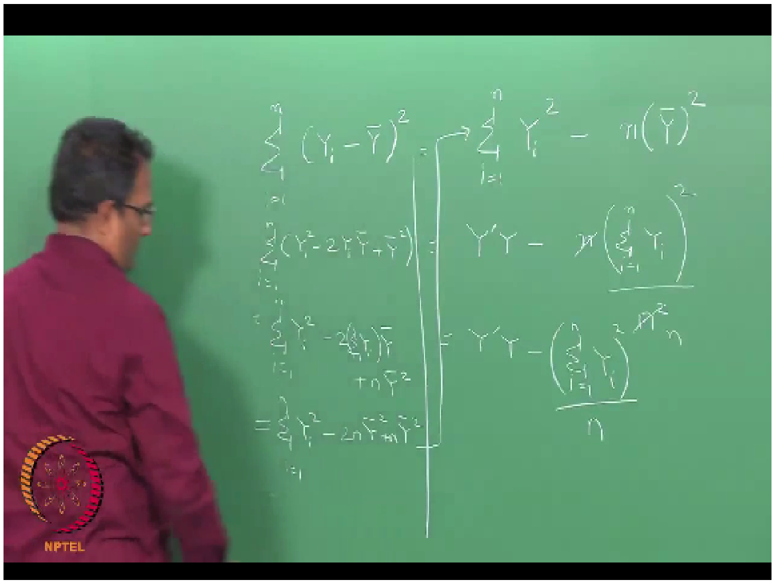
d. Total sum of squares:

$$SS_{Total} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$SS_{Total} = Y'Y - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

So we total this number and we can also write Yi-Y bar whole squared as Y prime Y-i=1 to n Yi whole squared/n. Actually you can show this term upon expansion as sigma i=1 to n Yi squared-n Y bar squared. Let me just use the board to expand it and then show what happens and why we get this particular term.

(Refer Slide Time: 24:03)


$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n(\bar{Y})^2$
$\sum_{i=1}^{n}(Y_i^2-2Y_i\bar{Y}+\bar{Y}^2)$
$= Y'Y - \cancel{n}\frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{}$
$=\sum_{i=1}^{n} Y_i^2 - 2\left(\sum Y_i\right)\bar{Y} + n\bar{Y}^2$
$= Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \cancel{n^2} n$
$=\sum_{i=1}^{n} Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2$
NPTEL

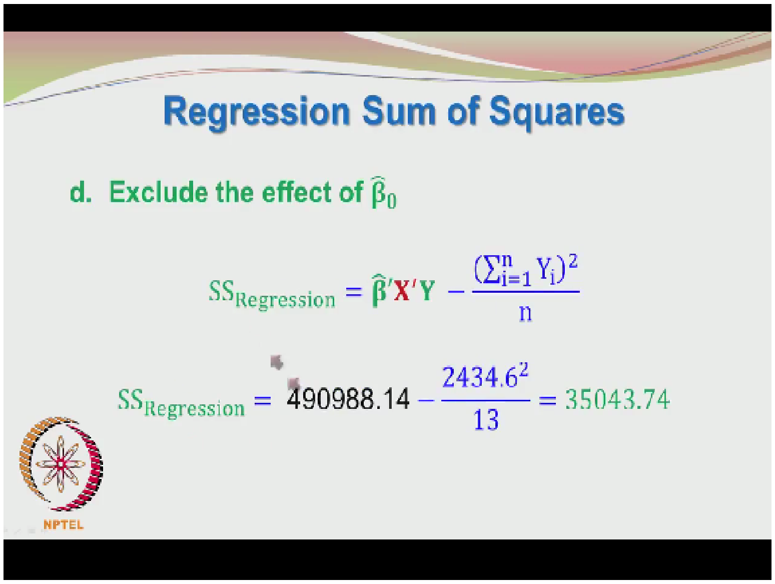
So we get this particular expression without any problem. Main thing is you may wonder how we got Yi squared-n Y bar squared, that is again very straightforward. We expanded sigma I=1 to n. This is one half of the story and this would be Yi squared-2YiY bar+Y bar squared and this becomes sigma Yi squared i=1 to n-2 sigma YiY bar+n Y bar squared, that is because we are adding a constant term n times and this becomes sigma Yi squared-, this can be written as nY bar+nY bar squared and so that is what we get here.

So the total sum of squares becomes 38869.34. Sometimes you may see a different value for total sum of squares. You may see the value as 494813.7 and that you may calculate as the total sum of squares but in some places, the total sum of squares may be reported as 38869.34. The reason is we are having an actual total sum of squares which is Y prime Y and after we adjust the total sum of squares for the beta hat 0, then the sum of squares becomes the lower value.

So to account for beta 0 parameter, we subtract the actual total sum of squares with the nY bar squared. We know that if beta 0 is the only regression coefficient considered, then that will become an average of the responses. So beta 0 will take on the value, will take up the value of the average of the responses and the sum of squares contribution from beta 0 would be nY bar squared where n is the number of experiments and to account for beta 0.

We subtract nY bar squared from the actual total sum of squares Y prime Y and that is why we get 38869.34. So we are removing the effect of beta 0 from our regression analysis.

(Refer Slide Time: 28:05)



Regression Sum of Squares

d. Exclude the effect of $\hat{\beta}_0$

$$SS_{Regression} = \hat{\beta}'X'Y - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}$$

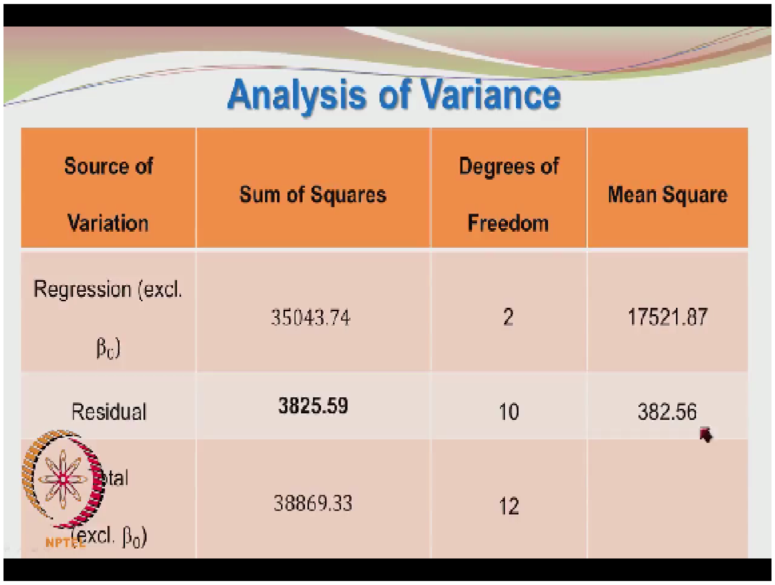
$$SS_{Regression} = 490988.14 - \frac{2434.6^2}{13} = 35043.74$$

Similarly, for the regression, the total regression sum of squares is given by beta prime hat X prime Y and to remove the regression contribution from beta 0 hat, we again subtract nY bar squared and the sum of square of regression due to parameters beta 1 hat and beta 2 hat is given by 35043.74. So this is where you have to be careful. You have to see in which situations beta 0 hat is present. Now we are coming to regression sum of squares.

Previously we were looking at total sum of squares. In the current analysis, we are removing the effect of beta hat 0 and the total sum of squares, we removed nY bar squared to remove the effect of beta hat 0 from the total sum of squares. Similarly, we also have to remove the influence of beta hat 0 from the regression sum of squares. We know that the regression sum of squares is given by beta hat prime X prime Y.

To remove the influence of beta hat 0, we simply subtract nY bar squared again from beta hat prime X prime Y and the sum of squares of regression becomes 35043.74.

(Refer Slide Time: 29:31)



Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square
Regression (excl. β_c)	35043.74	2	17521.87
Residual	**3825.59**	10	382.56
Total (excl. β_0)	38869.33	12	

Now we can write down the different contributions in the ANOVA table and we have the regression source of variation excluding the effect of beta 0 as 35043.74. The degrees of freedom correspond to beta hat 1 and beta hat 2, both of which are independent and so we have 2 degrees of freedom here and the mean square is obtained by dividing the sum of squares with the degrees of freedom.

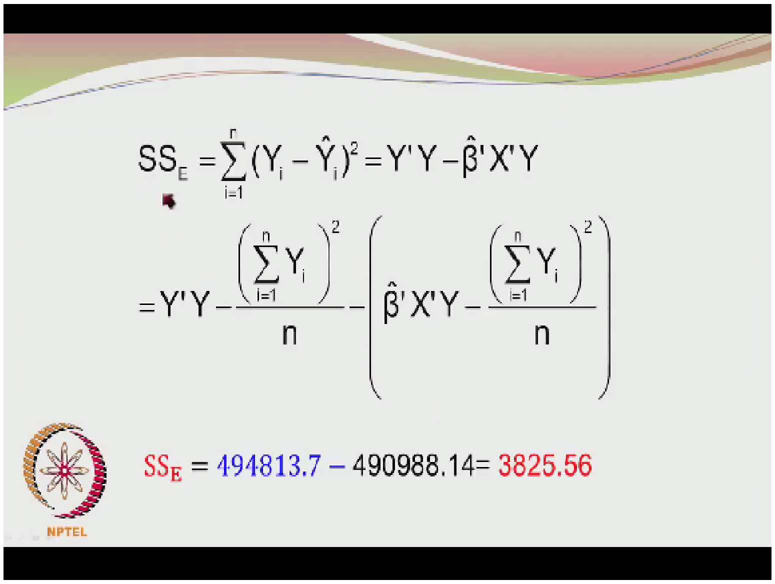
The residual sum of squares is also computed. The residual sum of squares is given in the second row. It is the difference between the total sum of squares and the regression sum of squares. If you are computing the total sum of squares without removing the contribution from beta hat 0, then you should also compute the regression sum of squares without excluding the effect of beta hat 0.

On the other hand, if you are computing the total sum of squares by excluding the effect of beta hat 0, then from the regression sum of squares also you should remove the effect of beta hat 0. So it is a question of whether you are subtracting nY bar squared from the total sum of squares and regression sum of squares. If you are subtracting it from the total sum of squares, you also subtract it from the regression sum of squares.

The difference between the regression sum of squares and the total sum of squares or rather it is the other way round, the difference between the total sum of squares and the regression sum of squares will give you the residual sum of squares which in this case is 3825.59. The degrees of freedom for residual sum of squares would be n-p where n is the number of experimental observation which is 13 and p is the total number of parameters beta hat 0, beta hat 1 and beta hat 2.

So we have 3 parameters. P=3 and so n-p will be equal to 10, 13-3, so 10 and so the mean square is obtained by residual sum of squares divided by 10, we get created 382.56. So we have the total sum of squares excluding beta 0 based on adding these 2 and we have the total degrees of freedom which is 12. We had 13 data points. Since we excluded the effect of beta 0 regression coefficient, we have 12 degrees of freedom.

(Refer Slide Time: 32:03)



$$SS_E = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = Y'Y - \hat{\beta}'X'Y$$

$$= Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} - \left(\hat{\beta}'X'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}\right)$$

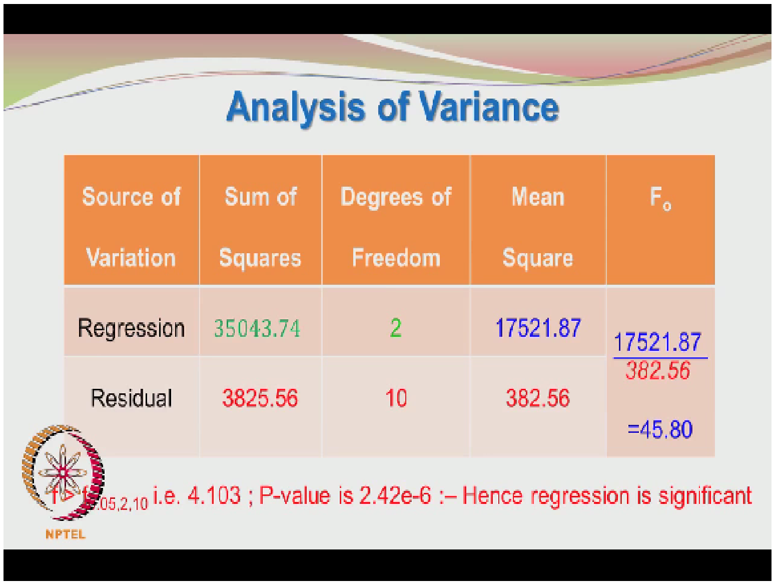
$SS_E = 494813.7 - 490988.14 = 3825.56$

We have the residual sum of squares given as the difference between the experimental observation and the predicted values and that is squared and added. So every experimental observation is subtracted with the corresponding predicted value, that deviation is squared and all such deviations are added to give the total residual sum of squares and that may be shown to be equal to Y prime Y-beta hat prime X prime Y.

And here we are subtracting nY bar squared in both these terms and that is also another way of finding the residual sum of squares where you are correcting for beta 0 parameter. Anyway it does not matter with the residual sum of squares and we get 3825.56 whether you add and subtract this nY bar square or do it directly but one thing you have to be very careful is when you are calculating the residual sum of squares, you cannot use the total sum of squares uncorrected for beta 0 and regression sum of squares corrected for beta 0, then you will get a wrong residual sum of squares.

So you correct total sum of squares for beta 0, you also correct regression sum of squares for beta 0, the difference between the 2 will give you the correct residual sum of squares or you take the total sum of squares directly, you take the total regression sum of squares, the difference between the 2 will give you the residual sum of squares. So correcting for one term for the beta 0 and not correcting the other term for beta 0 will lead to a wrong estimate of residual sum of squares. So this is to be carefully kept track of.

(Refer Slide Time: 34:10)



Analysis of Variance

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	35043.74	2	17521.87	$\frac{17521.87}{382.56}$ =45.80
Residual	3825.56	10	382.56	

$f > f_{0.05,2,10}$ i.e. 4.103 ; P-value is 2.42e-6 :– Hence regression is significant

So now we have the numbers in the ANOVA table. We have the source of variation, regression and residuals. Sum of squares are also noted here and the degrees of freedom are given here and the mean square values are put here. So the regression sum of squares is 35043.74 that we saw previously, 35043.74 and the residual sum of squares is 3825.56. Another thing again I am warning you here is the regression sum of squares is excluding the contribution from the beta 0 parameter.

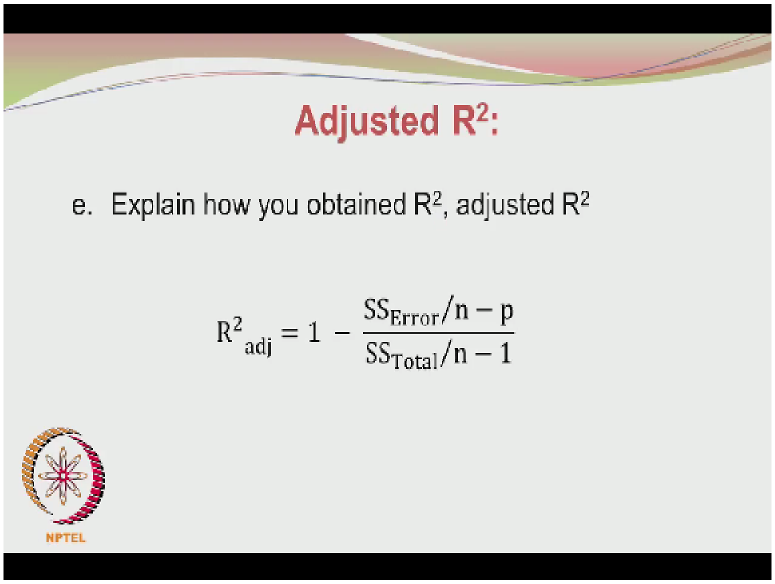
So that is what we did when we calculated the regression sum of squares and the residual sum of squares are given here. We have 2 parameters. Since beta 0 is excluded, we have parameters beta 1 and beta 2. So you have 2 degrees of freedom here and you have 10-p degrees of freedom which is 13-3 and that is 10 and the mean square is obtained by dividing the sum of squares with degrees of freedom.

So you get 17521.87 and 382.56. The ratio of these 2 will give 45.8. Without any further testing, we can be reasonably sure that this is lying in the rejection region. This statistic is lying in the rejection region because the regression sum of squares is about 50 times higher than the residual sum of squares. So we cannot really say that the 2 contributions are similar. If it is, let us say, in the order of 382.56, so this is around 400 and if you have got the regression sum of squares also as 500 or 600, then the 2 would have been quite comparable but this is 50 times more.

The regression sum of squares is 50 times more than the residual sum of squares. So the contribution to the total variation from the regression sum of squares is 50 times more than the residual sum of squares. So we have reasons to believe that the regression sum of squares is definitely contributing towards explaining the variation in the observed responses and the critical value is only 4.103.

And hence this F0 statistic is much higher than that and so it is lying in the critical, in the rejection region rather and so we reject the null hypothesis which says that regression contribution is 0.

(Refer Slide Time: 37:01)



Adjusted R²:

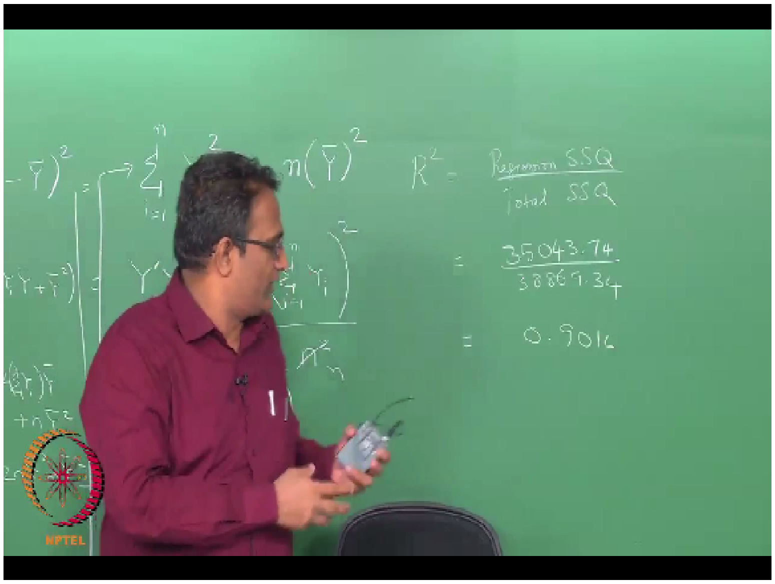
e. Explain how you obtained R^2, adjusted R^2

$$R^2_{adj} = 1 - \frac{SS_{Error}/n-p}{SS_{Total}/n-1}$$

Is first easy to find the R squared, the coefficient of determination. R square is obtained by seeing the fraction of regression sum of squares to the total sum of squares. So the regression sum of squares is 35043.74 and the total sum of squares would be the contribution of 35043.74 and the contribution of 3825.56, that number we have already that is 38869.34. So the sum of squares total is 38869.34 and the regression is 35043.74.

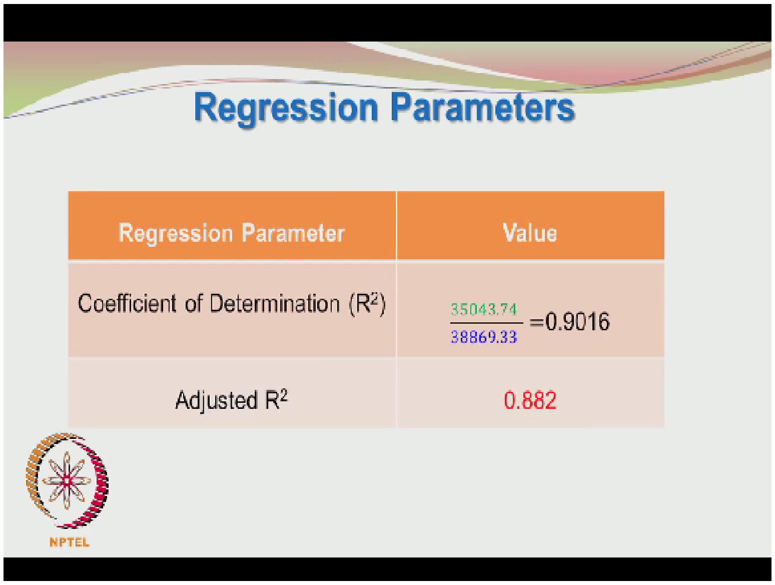
So we can take regression sum of squares with the total sum of squares and that would give us R squared, that can be verified. So let me write down the R squared value.

(Refer Slide Time: 38:11)


$R^2 = \frac{\text{Regression SSQ}}{\text{Total SSQ}}$
$= \frac{35043.74}{38869.34}$
$= 0.9016$

So this is what we get as R squared, the coefficient of determination.

(Refer Slide Time: 39:21)



Regression Parameters

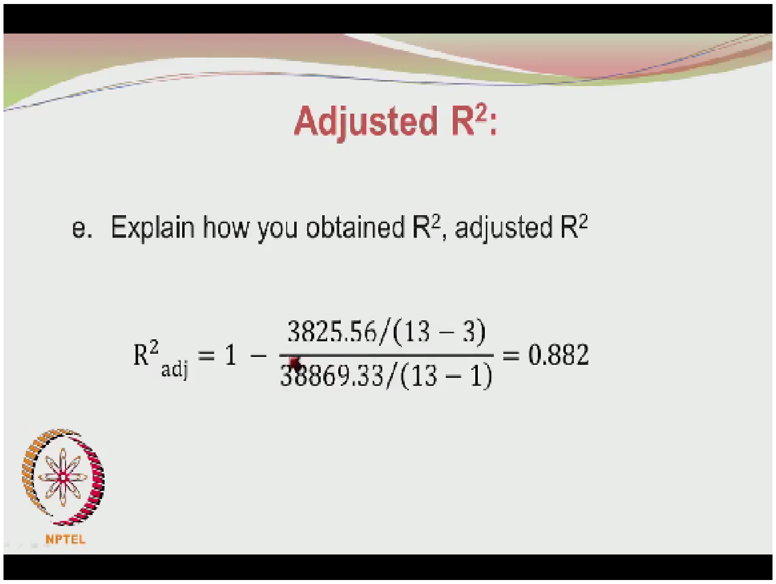
Regression Parameter	Value
Coefficient of Determination (R^2)	$\frac{35043.74}{38869.33}$ =0.9016
Adjusted R^2	0.882

NPTEL

Then we have to find the adjusted R squared. The adjusted R squared is obtained by using the mean square error and the mean square total. We get the mean square error by dividing the sum of squares of the residuals by n-p and the mean square total is obtained by dividing sum of squares of total with n-1.

So we are penalising the R squared for adding more parameters and you can see that when more parameters are added, the degrees of freedom for the error would actually decrease and the term in the numerator will increase and that would reduce the actual R squared value and so we are adjusting the R squared value for adding more parameters and that is why it is called as adjusted R squared.

(Refer Slide Time: 40:35)



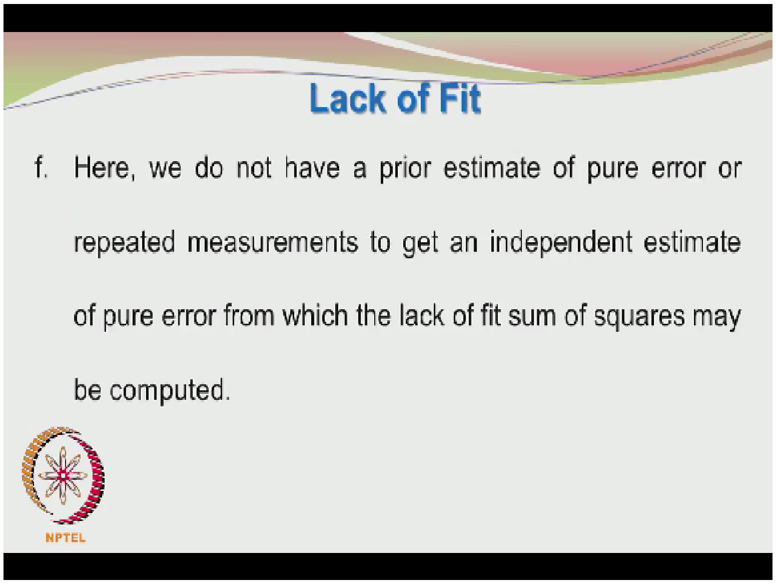
Adjusted R^2:

e. Explain how you obtained R^2, adjusted R^2

$$R^2_{adj} = 1 - \frac{3825.56/(13-3)}{38869.33/(13-1)} = 0.882$$

And so we have 1-3825.56, that is the residual sum of squares/n-p 13-3 that is 10 and this is the total sum of squares/the degrees of freedom which is not 13 because 1 is used for finding the mean of the observations. So we have this ratio and we get 0.882 and these are the values which are tabulated in the slide and you can see that adjusted R square is less spectacular when compared to the coefficient of determination, it is 0.882 which is lower than 0.9016. Sometimes the discrepancy between these 2 may be quite high.

(Refer Slide Time: 41:29)



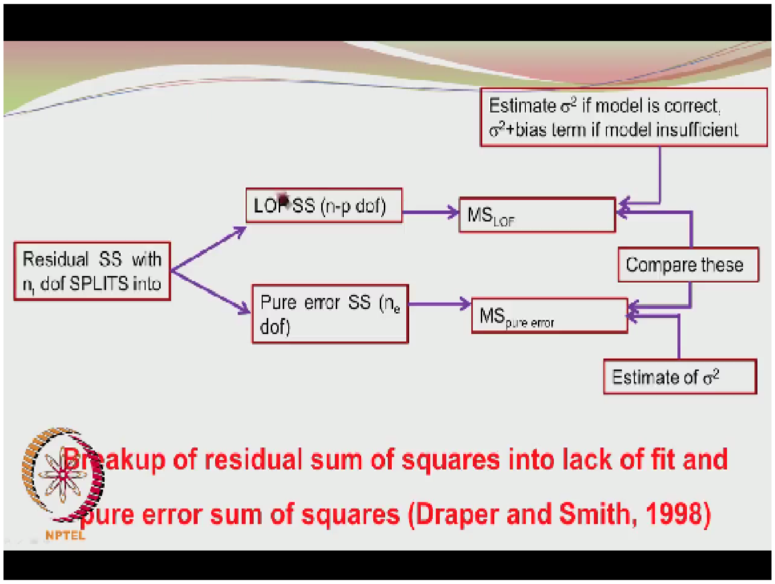
Lack of Fit

f. Here, we do not have a prior estimate of pure error or repeated measurements to get an independent estimate of pure error from which the lack of fit sum of squares may be computed.

So now we have to do the lack of fit analysis. We do not have a prior estimate of pure error or repeated measurements to get an independent estimate of pure error from which the lack of fit sum of squares may be computed. So for computing the lack of fit, we have to actually partition the residual sum of squares into lack of fit sum of squares and pure error sum of squares.

So for doing the partitioning exercise, we need to have an idea about the pure error. Unfortunately, in this experimental sequence, the repeats were not carried out. May be the people who are running the experiments thoughts they will get identical values if they repeat the experiment, so they did not carry out the repeats. So an independent estimate of the pure error could not be obtained.

(Refer Slide Time: 42:24)



Estimate σ^2 if model is correct, σ^2+bias term if model insufficient

LOF SS (n-p dof)

MS_{LOF}

Residual SS with n_r dof SPLITS into

Compare these

Pure error SS (n_e dof)

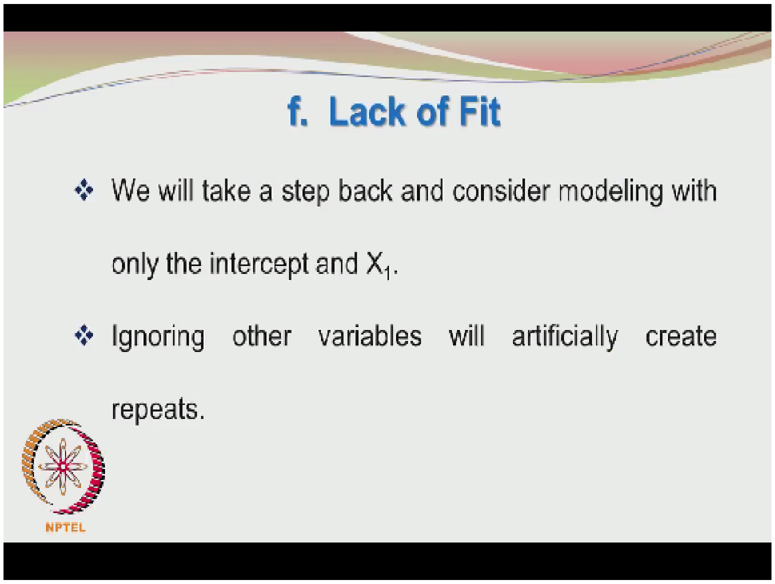
$MS_{pure\ error}$

Estimate of σ^2

Breakup of residual sum of squares into lack of fit and pure error sum of squares (Draper and Smith, 1998)

So as I was explaining in the previous lecture, we have the residual sum of squares split into lack of fit sum of squares and pure error sum of squares. We get the mean square lack of fit and mean square pure error and we compared the mean square lack of fit with mean square pure error and make the appropriate conclusions. If mean square lack of fit is much higher than mean square pure error, then the model developed is inadequate but if these 2 are comparable, then the model is adequate and both of them are independent estimates of sigma squared.

(Refer Slide Time: 43:08)

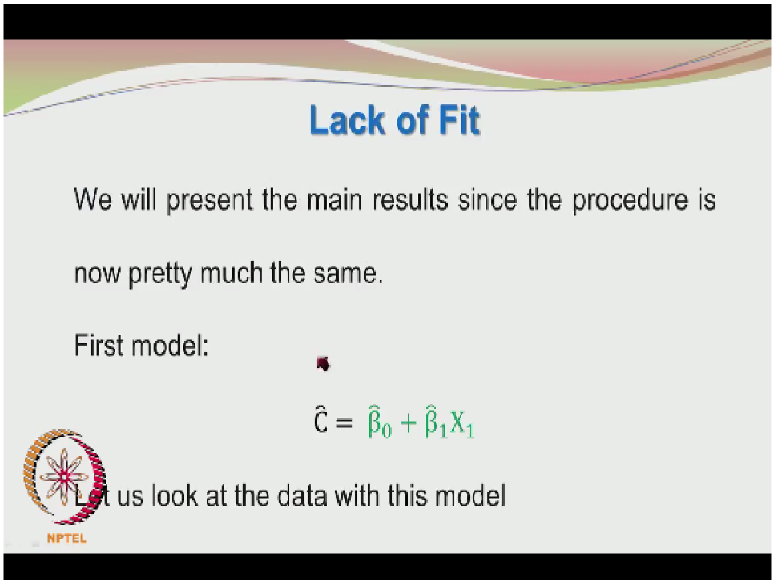


f. Lack of Fit

- We will take a step back and consider modeling with only the intercept and X_1.
- Ignoring other variables will artificially create repeats.

So what we may do is we can artificially create repeats and let us see the trouble created by artificially creating the repeats. We will consider modelling only the intercept and X1. So we are ignoring the effect of X2. We are ignoring the effect of the amount of powder used in the machine and if we ignore that, let us see what happens?

(Refer Slide Time: 43:43)



Lack of Fit

We will present the main results since the procedure is now pretty much the same.

First model:

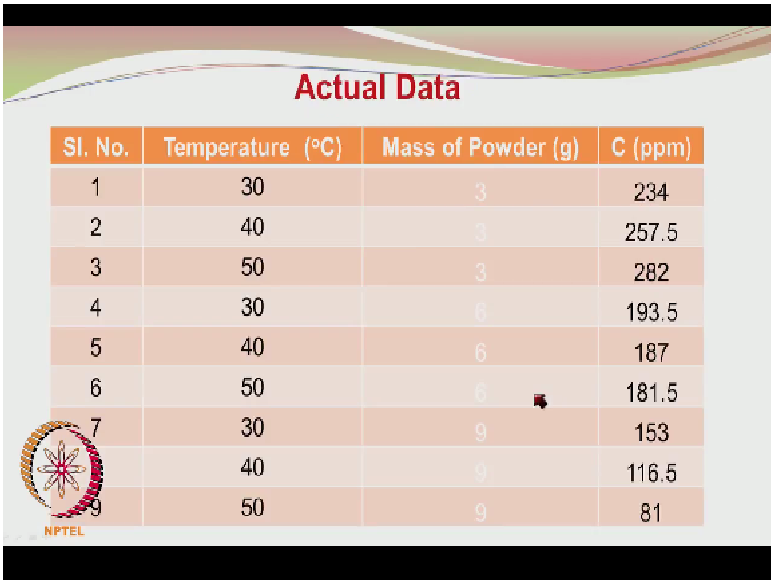
$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

Let us look at the data with this model

Then the model we are testing is C hat=beta hat 0+beta hat 1X1. This entire procedure looks very suspicious and that is going to be justified in the next set of calculations but what it does show us, suppose you had started with this model, what would have been the lack of fit sum of squares. Even though this procedure is based on the assumption that X2 is not significant which we do not know yet, okay.

So what we have to do is see what is going to happen if we assume a priori that the X2 is not going to have an effect, so we will only test with this model.

(Refer Slide Time: 44:37)

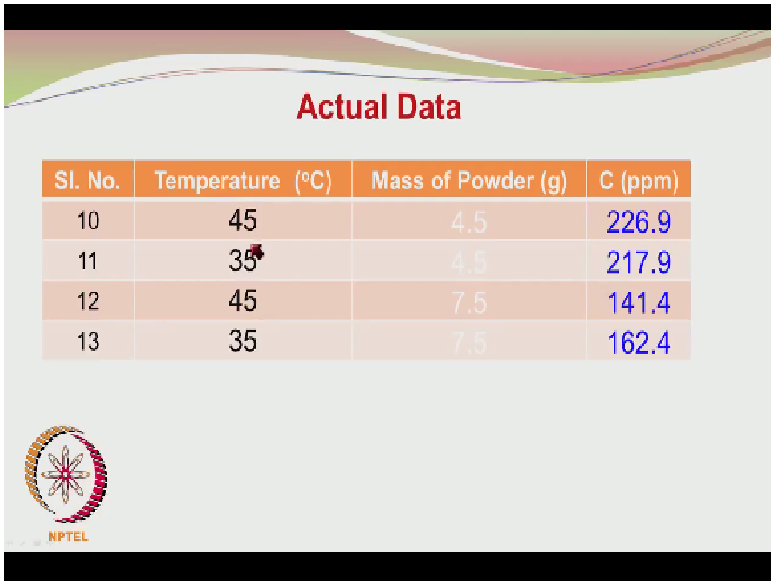


Actual Data

Sl. No.	Temperature (°C)	Mass of Powder (g)	C (ppm)
1	30	3	234
2	40	3	257.5
3	50	3	282
4	30	6	193.5
5	40	6	187
6	50	6	181.5
7	30	9	153
	40	9	116.5
9	50	9	81

Then it is as if we are hiding this particular column totally and it then appears that we have repeats here. For example, this is 30 40 50 30 40 50 30 40 50. So it will appear as if we have repeated the experiment not once but 2 times, that means we are having 3 repeated observations.

(Refer Slide Time: 45:05)



Actual Data

Sl. No.	Temperature (°C)	Mass of Powder (g)	C (ppm)
10	45	4.5	226.9
11	35	4.5	217.9
12	45	7.5	141.4
13	35	7.5	162.4

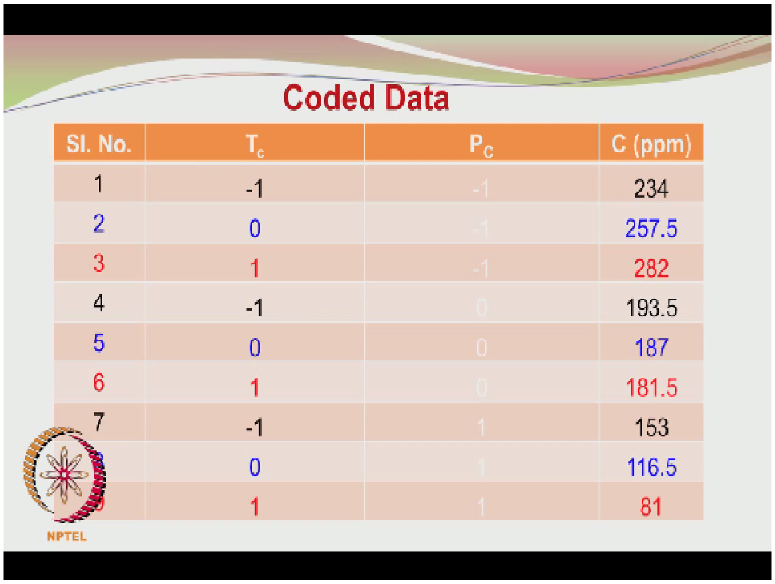
So it looks very nice and if you go further, you have 2 repeats, 45 and 45 here, 35 and 35 here. So it looks as if we have solved the problem by ignoring the effect of the mass of powder used but we do not know whether that is the correct procedure because the mass of powder used may have an important implication in the process. So just let us take a look at the repeated sets. We will take a look at it here.

We will compare these two 45 degrees and you are seeing on one hand it is 226.9, on the other hand, it is 141.4. So when you do repeats, you are getting 141 and then you are getting 226 or 227. So almost 2 times variations there between these 2 readings and even if you compare this 35 with this 35, 162 appears to be very far from 218, okay. So if this was 140 and this was 160, it is okay. If this was 160 and this was 170 or 150, it is okay, that may be because of random fluctuations.

But this looks to be too large a random fluctuation, too large a fluctuation to be ascribed to random phenomena. Let us go back to the table and we will see that for conditions of 30, you have 234 and you have 194 and another 30 is 150. So for the same experimental condition of 30 degrees centigrade, the colour varies from 153 to 234, that is a huge difference. So by just looking at the data itself we can see that ignoring the mass of powder was not such a good idea.

What it is doing is it is exaggerating the role played by experimental error plus there is also a moral in this story, if you are doing experiments and you find when you repeat the experiments that you get larger variations in the repeats, then there is probably another factor influencing the process response which you have not identified. So you should rather than doing more and more repeats and getting more and more variability, see what is the factor in the experiment which is actually causing these different variabilities that should be your focus.

(Refer Slide Time: 47:45)

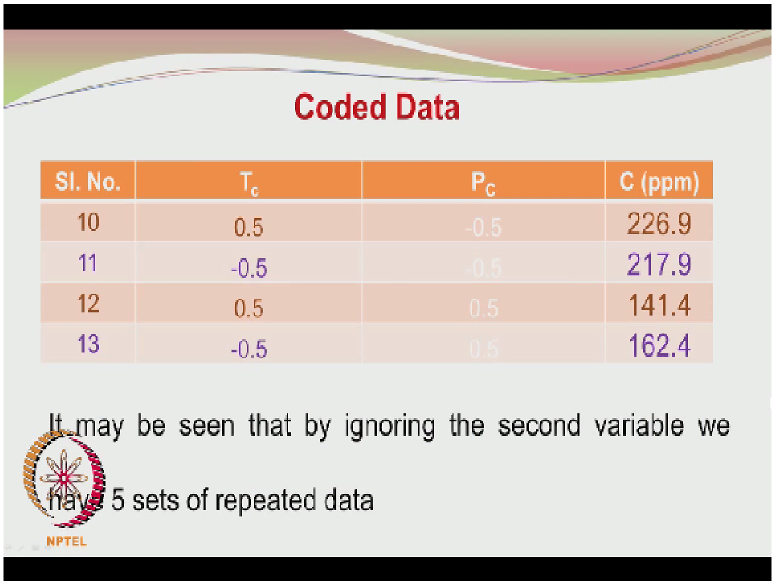


Coded Data

Sl. No.	T_c	P_C	C (ppm)
1	-1	-1	234
2	0	-1	257.5
3	1	-1	282
4	-1	0	193.5
5	0	0	187
6	1	0	181.5
7	-1	1	153
8	0	1	116.5
9	1	1	81

So we will put the data in the coded format. You can see that the black colours are all -1 -1 -1. They represent a one set of conditions. 0 0 0, they represent another set of conditions. 1 1 1, which are coded in red, they correspond to another set of repeated conditions, pseudo-repeated conditions and other column is whitened out.

(Refer Slide Time: 48:10)



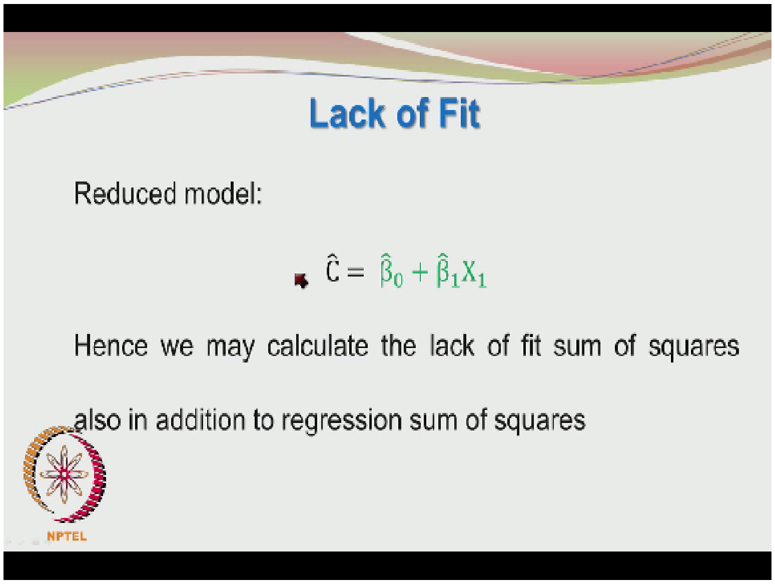
Coded Data

Sl. No.	T_c	P_C	C (ppm)
10	0.5	-0.5	226.9
11	-0.5	-0.5	217.9
12	0.5	0.5	141.4
13	-0.5	0.5	162.4

It may be seen that by ignoring the second variable we have 5 sets of repeated data

And similarly you have Tc which is 0.5 and 0.5 and you also have Pc -0.5 and -0.5. So you have 2 repeat conditions here. So you have 5 sets of repeated data. These are 2 sets of repeated data and then in this slide, you have 3 sets of repeated data. So totally we have 5 sets of repeated data.

(Refer Slide Time: 48:35)



Lack of Fit

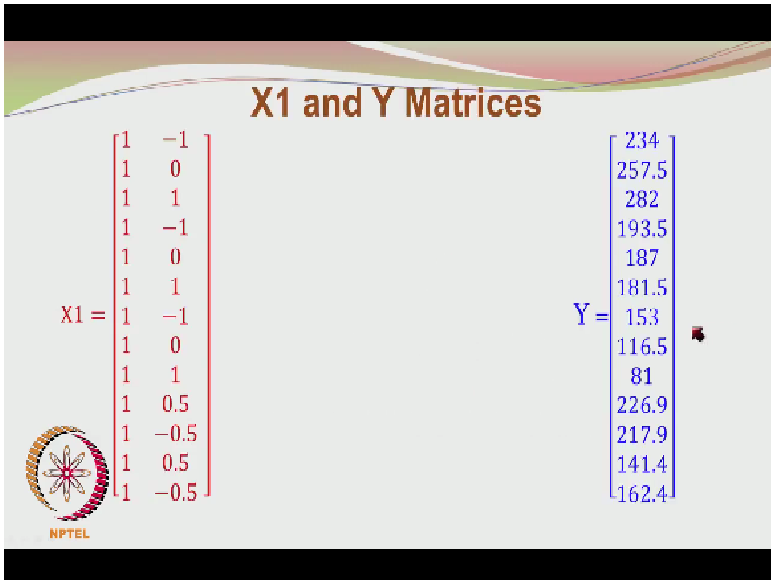
Reduced model:

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

Hence we may calculate the lack of fit sum of squares also in addition to regression sum of squares

So you have the reduced model, C hat=beta hat 0+beta hat 1X1 and we can again calculate the parameters regression sum of squares, residual sum of squares and so on.

(Refer Slide Time: 48:51)



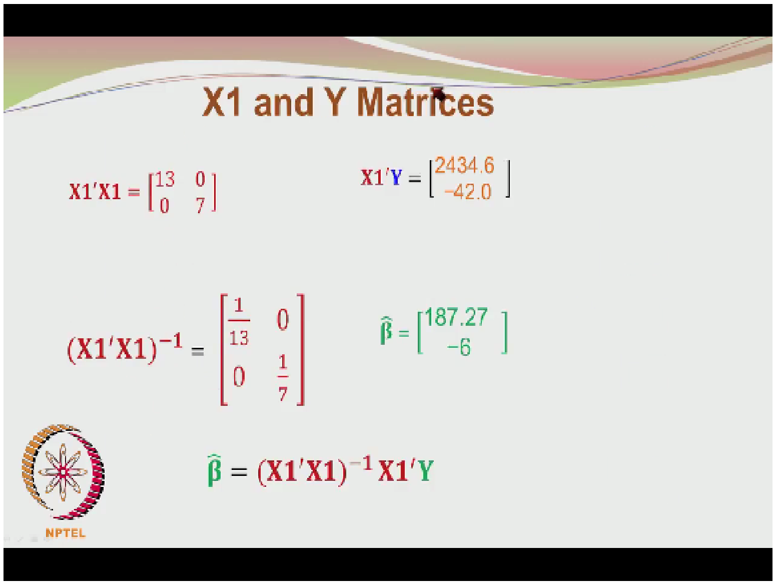
X1 and Y Matrices

$$X1 = \begin{bmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & -0.5 \\ 1 & 0.5 \\ 1 & -0.5 \end{bmatrix} \qquad Y = \begin{bmatrix} 234 \\ 257.5 \\ 282 \\ 193.5 \\ 187 \\ 181.5 \\ 153 \\ 116.5 \\ 81 \\ 226.9 \\ 217.9 \\ 141.4 \\ 162.4 \end{bmatrix}$$

So we have the new X matrix where we have contribution only from the X1 regressor variable or we have contributions only from temperature expressed in the code format. We have removed the column containing the coded values for the mass of detergent powder used. When we look at the column vector Y, it can be seen that the responses are the same as before. We have not made any modification to the responses column vector.

Only thing is, we have removed the second column from the X matrix. From the original X matrix, we get the new X1 matrix which contains the vector of once and then the coded values for temperature.

(Refer Slide Time: 49:41)



X1 and Y Matrices

$$X1'X1 = \begin{bmatrix} 13 & 0 \\ 0 & 7 \end{bmatrix} \qquad X1'Y = \begin{bmatrix} 2434.6 \\ -42.0 \end{bmatrix}$$

$$(X1'X1)^{-1} = \begin{bmatrix} \frac{1}{13} & 0 \\ 0 & \frac{1}{7} \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} 187.27 \\ -6 \end{bmatrix}$$

$$\hat{\beta} = (X1'X1)^{-1} X1'Y$$

So we can carry out the exercise and that is not going to be very difficult. We can see that the X1 prime X1 matrix would be 13 0 0 7, X1 prime Y would be 24 34 .6 -42 and then we have the inverse of this matrix which is 1/13 and 1/7 and then we get the parameters 187.27 and -6. It is very interesting to note that even though we have removed the effect of the mass of detergent, we are getting the same, otherwise same X1 prime X1 matrix, X1 prime Y matrix.

And then the inverse is also 1/13 1/7, we had an additional 1/7 as the diagonal element that is no longer present and then we also have the parameters which are the same as before. We have 187.27 and -6. Let us just go back a few slides and we will see that these values are exactly the same for the first 2 entries. So in the X1 prime X1 matrix, we do not have the last 7. When you look at the X1 prime Y matrix, we do not have the last entry here. We have 2434.6 and -42.

Then X prime X inverse matrix, we have 1/13 1/7 and beta hat, we have 187.27 and -6. We do not have -70.5. So the values are exactly the same as before and this shows the advantage of doing the analysis in the coded format. We have an orthogonal design which I will talk about in my next lecture. So we have the data as I had presented, right. So we will do further analysis of these results in the coming lecture. Thank you for your attention.