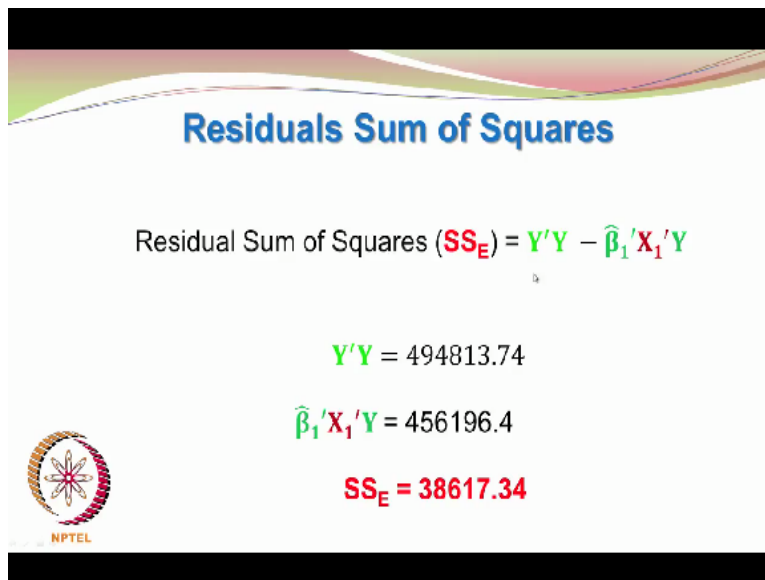


Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology - Madras

Lecture – 42
Regression Analysis: Example Set 8 Continued

Welcome back, we will be continuing with our regression analysis. We are currently doing example set 8 and we will continue with the problem we were discussing.

(Refer Slide Time: 00:29)




Residuals Sum of Squares

Residual Sum of Squares (SS_E) = $Y'Y - \hat{\beta}_1'X_1'Y$

$Y'Y = 494813.74$

$\hat{\beta}_1'X_1'Y = 456196.4$

$SS_E = 38617.34$



So we know that the residual sum of squares is an important component in the other analysis as it is tied down with the prediction capability of the developed regression model. So higher the residual sum of squares, higher would be the deviation between the experimental data and the model predictions. So we have the residual sum of squares as $Y'Y - \hat{\beta}_1'X_1'Y$ for this present case.


X_1 refers to the first regressor variable and the total sum of squares $Y'Y$ is 494813.74 transpose of Y multiplied by Y . So $\hat{\beta}_1'X_1'Y$ based on the matrix will give you 456196.4. So the difference between $Y'Y$ and $\hat{\beta}_1'X_1'Y$ is 38617.34.

(Refer Slide Time: 02:01)

Residuals Sum of Squares

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

$$= 13 - 2 = 11$$

$$MS_E = \frac{SS_E}{n - p} = 3510.67$$


Residual Sum of Squares (SS_E) = **38617.34**

So whenever we calculate sum of squares, we also identify the scaling factor or in other words the degrees of freedom. For the present case, the degrees of freedom would be n-p, we have 2 parameters. So $13 - 2 = 11$ degrees of freedom. The 2 parameters are beta 0 and beta 1. Beta 0 corresponding to the intercept and beta 1 corresponding to the regressor variable X1. So the mean square error would be sum of squares of error by n-p.


So we divide the sum of squares, the residual sum of squares which we found out as 38617.34 and that we divide by 11, we get 3510.67.

(Refer Slide Time: 02:58)

Residuals Sum of Squares

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

$$= 13 - 2 = 11$$

$$MS_E = \frac{SS_E}{n - p} = 3510.67$$


Residual Sum of Squares (SS_E) = **38617.34**


Now we will discuss about the calculation of the pure error. So we are trying to do something

different here. So what we do is when we ignore the mass of the powder, it is as if the experiments are being repeated. So the experimental settings corresponding to the mass of the powder are not taken into consideration and so it is as if we are doing some repeats. I request you to look at the table and confirm this for yourself.

(Refer Slide Time: 03:39)

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
8	40	9	116.5
9	50	9	81




So we are having the mass of power here and when we ignore the mass of the powder, then for example if you look at the 3 3 3 and 6 6 6, we are not caring about the mass of the powder and we have 30 40 50 30 40 50. So it is as if we are repeating the experiments at 30 40 and 50. Same thing for mass of powder of 9 grams.

(Refer Slide Time: 04:17)

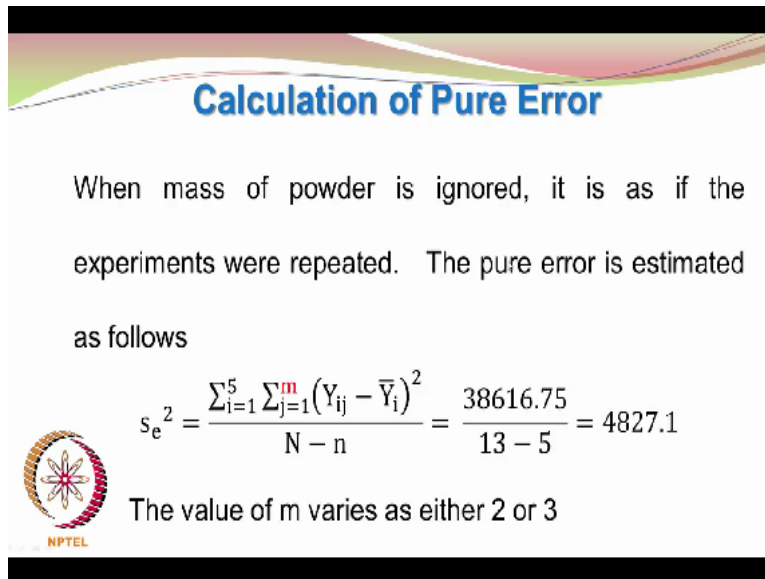
Actual Data

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 in this case we are having 45 and 35 at only 2 settings and then you also do at 45 and 35. So if you ignore the mass of the powder, then it is as if certain experiments have been repeated twice and certain other experiments have been repeated 3 times. For example, this 30 40 50 is one set of conditions, again 30 40 50, again 30 40 50. So we have 3 repeats. When you do the experiments at 45 and 35, it is as if we are having 2 repeats only. So with this background, let us go back to the regression analysis.

(Refer Slide Time: 05:07)




Calculation of Pure Error

When mass of powder is ignored, it is as if the experiments were repeated. The pure error is estimated as follows

$$s_e^2 = \frac{\sum_{i=1}^5 \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2}{N - n} = \frac{38616.75}{13 - 5} = 4827.1$$

The value of m varies as either 2 or 3



The pure error has to be estimated under this artificial situation where we do not account for the mass of the powder. Please note that as of now we do not know whether the mass of the powder is having a strong influence on the response of the process but we will just for demonstration purposes assume that the mass of the powder can be ignored and thereby we are having some repeats. So we have to calculate the pure error sum of squares. There are couple of ways of doing it.

(Refer Slide Time: 05:43)

Calculation of Pure Error

When mass of powder is ignored, it is as if the experiments were repeated. The pure error is estimated as follows

$$s_e^2 = \frac{\sum_{i=1}^5 \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2}{N - n} = \frac{38616.75}{13 - 5} = 4827.1$$



The value of m varies as either 2 or 3

One way is to take sigma i=1 to 5, 5 independent settings and then you have j=1 to m, where m is the number of repeat observations for every independent setting and then you have Yij-Y bar i whole squared/N-n where capital N is the total number of runs which is 13 and n is the number of independent settings. So you get after these calculations are over, 4827.1, the number of repeats as we just now saw may be either 2 or 3.

(Refer Slide Time: 06:28)

Calculation of Pure Error

Incidentally, the pure error is also estimated as follows

$$s_e^2 = \frac{\sum_{i=1}^5 v_i s_i^2}{N - n} =$$

$$\frac{2 * 1640.25 + 2 * 4970.25 + 2 * 10100.25 + 1 * 3655.125 + 1 * 1540.125}{13 - 5}$$



$$= 4827.1$$

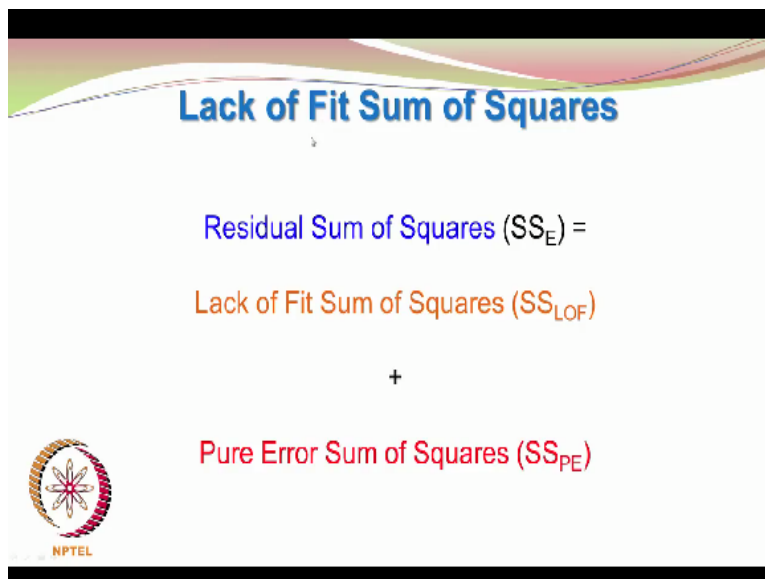
There is another way for calculating the pure error and that is given by sigma i=1 to 5 mu iSi squared by N-n. For every independent setting, you can calculate the Si squared and then weight it with the appropriate degrees of freedom, okay. So I request you to carry out this calculation for yourself and see whether your answers match with mine. So you have the case where you had 3

repeats.


You have 2 degrees of freedom and for that we had 3 independent settings and then you have 2 cases where you had 1 degree of freedom because there were only 2 entities in that independent set. So you have 1 and 1. Let me come again. So here i is running from 1 to 5, μ_i^2/N . So what you do is you calculate the standard deviation or the variance for every independent dataset, okay. You have 5 independent datasets. So for each dataset, you find out the variance S_i^2 for that particular dataset and then you weigh it with the degrees of freedom.

For the first independent dataset, we had 3 runs and so you have 2 degrees of freedom. For the second one also, you have 2 degrees of freedom. For the third one, you have 2 degrees of freedom. Then you had only 2 runs for the fourth dataset and only 2 runs for the fifth dataset and both of them would hence have only 1 degrees of freedom and then you have the total degrees of freedom as $2+2+4+2=6+1=7+1=8$ and that also leads to 4827.1. So you have 4827.1 by using this method and you also have the same answer by using a different method.

(Refer Slide Time: 09:01)



Lack of Fit Sum of Squares

$$\text{Residual Sum of Squares (SS}_E\text{)} =$$
$$\text{Lack of Fit Sum of Squares (SS}_{LOF}\text{)}$$
$$+$$
$$\text{Pure Error Sum of Squares (SS}_{PE}\text{)}$$


So we are now interested in finding the lack of fit sum of squares. We want to see whether our decision to exclude factor 2 from the modelling analysis was justified. So you have the residual sum of squares as a combination or additive combination of lack of fit sum of squares, sum of squares lacks of fit plus pure error sum of squares. This is a very important component and we

are going to compare the lack of fit sum of squares with the pure error sum of squares.
(Refer Slide Time: 09:37)

Lack of Fit Sum of Squares


38617.34

=

Lack of Fit Sum of Squares (=0.59)

+

38616.75



So you have 38617.34 and then you just now calculated the pure error as 38616.75, sum of squares as 38616.75 and so the difference between these 2 will give you a lack of fit sum of squares of 0.59.


(Refer Slide Time: 10:03)

Regression Sum of Squares

d. Regression sum of squares:

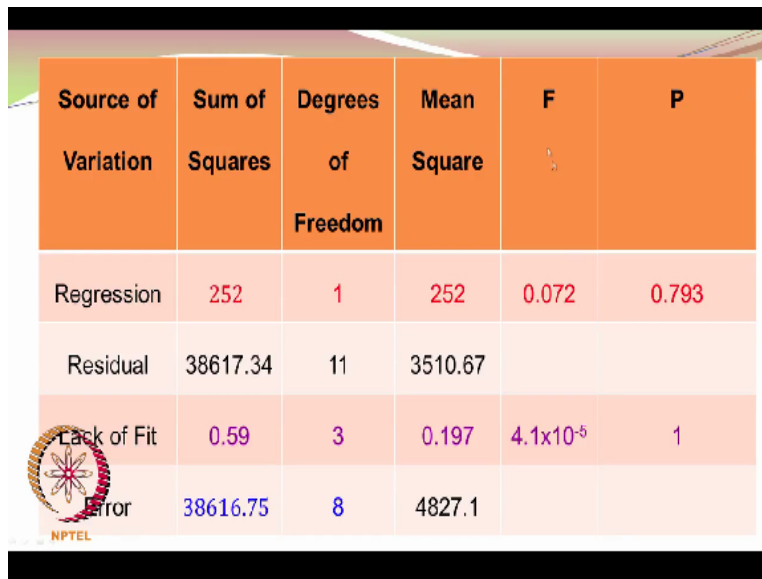
$$SS_{\text{Regression}} = \hat{\beta}_1' X_1' Y - \frac{(\sum_{i=1}^n Y_i)^2}{n}$$

$$SS_{\text{Regression}} = 456196.4 - \frac{2434.6^2}{13} = 252$$



So next we go on to the regression sum of squares. The regression sum of squares after removing the influence of the intercept beta 0 is given by beta hat 1 prime X1 prime Y-sigma i=1 to n Yi whole squared/n or this is also = nY bar squared. We have covered this in one of our previous lectures. So we have the sum of squares of regression as 252.

(Refer Slide Time: 10:36)



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	P
Regression	252	1	252	0.072	0.793
Residual	38617.34	11	3510.67		
Lack of Fit	0.59	3	0.197	4.1×10^{-5}	1
Error	38616.75	8	4827.1		

So we set up the ANOVA table and so you have regression, residual, lack of fit and error. The sum of squares whatever we calculated previously, we entered in this table 252 38617.34 0.59 38616.75 and so you have the degrees of freedom as 1 for the only parameter we are interested in, that is the parameter beta 1 and then you have 11 degrees of freedom for the residual. We saw it as $n-p$, where n is the total number of data points which is 13 and then p is the total number of parameters beta 0 and beta 1, which we are considering now.

So $13-2$ is 11 and then you have the lack of fit sum of squares and the pure error sum of squares. We saw in our calculation that the errors degrees of freedom was 8. Let me just go back. So you can see that is $13-5$ $2+2=4+2=6+1+1=8$. We also had this formulae $N-n$ $13-5$, total number of runs 13-number of independent settings 5 and so you have 8. So you can calculate the mean square and you find the regression mean square, residual mean square and the lack of fit mean square and pure error mean square.

To find these, all you have to do is divide the sum of squares by the degrees of freedom and you will get the corresponding F values and you can see that the p values are pretty high. So it means that the parameter of beta 1 is pretty much insignificant, okay. So it means that the X_1 is actually contributing little to the model. Let us move on.

(Refer Slide Time: 12:32)

Conclusions

- ❖ The parameter associated with X_1 is not at all significant and may be rejected.
- ❖ The regression sum of squares is very small and hence the residual sum of squares is very high



Let us discuss this further. So since the p value is pretty high, the parameter corresponding to X_1 , let me make a small correction here. The parameter associated with the X_1 is not at all significant and may be rejected and the regression sum of squares is very small and the residual sum of squares is pretty high. So this is a kind of very artificial situation where we have removed the effect of the mass of the powder which probably is having the important say in the process.

(Refer Slide Time: 13:07)

Conclusions based on LOF

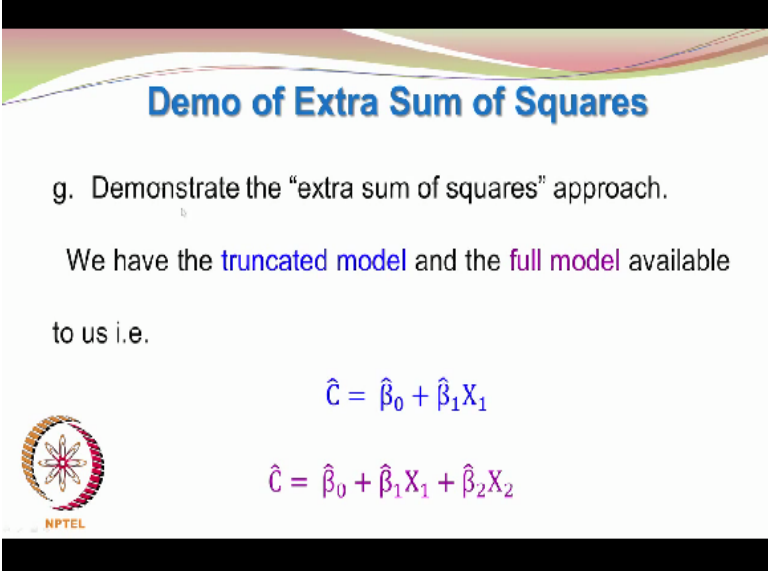
- ❖ We artificially created the repeats and hence we had a very high value of pure error sum of squares (SS_{PE})
- ❖ SS_{PE} is not genuine
- ❖ Genuine repeats are essential as insignificance of certain variables may not be assumed *a priori*.



So what we did was to artificially create the repeats and hence we had a very high value of the pure error sum of squares. Since the pure error sum of squares was very high, it sorts of made the regression sum of squares corresponding to the parameter beta 1 pretty much insignificant and also we note that the sum of squares of pure error is not genuine. So rather than being lazy and

not conduct repeats, it is important that we conduct genuine repeats. These are very essential and we cannot a priori or beforehand assume the insignificance of certain variables, okay.


(Refer Slide Time: 13:55)



Demo of Extra Sum of Squares

g. Demonstrate the “extra sum of squares” approach.

We have the **truncated model** and the **full model** available to us i.e.

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

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

So the next step is to demonstrate the extra sum of squares approach. We have the truncated model in blue and the full model available to us. The truncated model is given by $\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ and the full model is as of now involving only the main variables X_1 and X_2 . So we have the full model as of now written down as $\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$, okay.

We can of course develop the model by adding more and more terms like the second-order interaction terms, the third order interaction terms, the quadratic terms and so on but as of now to demonstrate the principal of extra sum of squares, we will stick with the model which is having only the main affects.

(Refer Slide Time: 14:51)

Extra Sum of Squares

$$SS_{\text{Regression}}(\hat{\beta}_1) = \hat{\beta}_1' X_1' Y - \frac{(\sum_{i=1}^n Y_i)^2}{n} = 252$$

$$SS_{\text{Regression}}(\hat{\beta}_2 | \hat{\beta}_1) = \left(\hat{\beta}' X' Y - \frac{(\sum_{i=1}^n Y_i)^2}{n} \right) - \left(\hat{\beta}_1' X_1' Y - \frac{(\sum_{i=1}^n Y_i)^2}{n} \right) =$$

$$35043.74 - 252 = 34791.74$$



We are now finding the sum of squares of regression due to the beta hat 1 and that is given by beta hat prime 1 X prime 1 Y - n Y bar squared or sigma i=1 to n Y_i whole squared/n that comes out to be 252. So this we saw earlier. Now we want to see the sum of squares of regression due to beta hat 2, the second parameter, the mass of the powder, given beta hat 1 and that is given as beta hat prime X prime Y - n Y bar squared - beta hat prime 1 X_1 or rather X prime 1 Y - n Y bar squared.

So you can either say it as Sigma i=1 to n Y_i whole squared by n or n Y bar squared. Small n is the number of data points and this particular summation is the sum of all the responses and then you square it. This can be shown to be = n * the average of the response squared, okay. So what we do is we are subtracting these terms in order to remove the influence of the intercept beta 0. This also we have seen earlier.

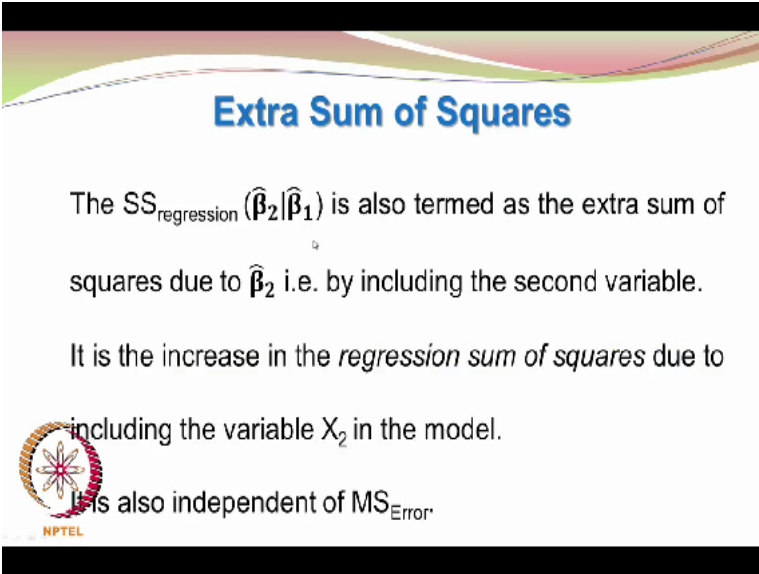
Now we are interested in finding out the sum of squares of regression due to beta hat 2 given that beta hat 1 is already present in the model. So we take the total regression sum of squares including all the parameters or the full model and then we subtract it with the regression sum of squares due to the model having only beta hat 1, okay. In both the cases, we are removing the influence of the intercept.

So in this full model also you have the intercept. In the truncated model also you have the

intercept. This we also saw. Even in the truncated model involving only X_1 you have the intercept parameter and even in the full model also, you have the intercept $\hat{\beta}_0$, okay. So when we do that, we essentially remove the effect of $\hat{\beta}_0$ from both the full model as well as the truncated model.

So this difference is the extra sum of squares brought in by the parameter $\hat{\beta}_2$ and that comes out to be quite significant. You can see that it is adding up to this term is 35043.74 and then we already found this term to be 252 and this comes out to be 34791.74.

(Refer Slide Time: 18:12)




Extra Sum of Squares

The $SS_{\text{regression}}(\hat{\beta}_2 | \hat{\beta}_1)$ is also termed as the extra sum of squares due to $\hat{\beta}_2$ i.e. by including the second variable.

It is the increase in the *regression sum of squares* due to including the variable X_2 in the model.

It is also independent of MS_{Error} .



So the sum of squares of regression $\hat{\beta}_2$ given $\hat{\beta}_1$ is also termed as the extra sum of squares due to $\hat{\beta}_2$, that is by including the second variable that is the increase in the regression sum of squares including the variable X_2 in the model that is also independent of mean square error.

(Refer Slide Time: 18:36)

Extra Sum of Squares

The null hypothesis $\beta_2 = 0$ may be tested by the statistic

$$F_0 = \frac{SS_{\text{Regression}}(\hat{\beta}_2 | \hat{\beta}_1) / 1}{MS_{\text{Error}}} = \frac{34791.74}{382.56} = 90.944$$



So now we can test the null hypothesis $\beta_2 = 0$ according to the following relationship. F_0 is sum of squares of regression β_2 given β_1 is already present in the model/the degrees of freedom due to β_2 which is nothing but 1 and then we divided it by the mean square error and we get $34791.74/382.56$. So you probably may want to see how you got the mean square error of 382.56, okay and this value comes out to be 90.944.

(Refer Slide Time: 19:24)

Extra Sum of Squares

$$F_0 = \frac{SS_{\text{Regression}}(\hat{\beta}_2 | \hat{\beta}_1) / 1}{MS_{\text{Error}}} = \frac{34791.74}{382.56} = 90.944$$

P-value is pretty much close to zero and hence the null hypothesis may be rejected. Hence the variable X_2



creates a significant contribution to the process

So the P value associated with this F_0 is pretty much close to 0 and hence the null hypothesis may be rejected. Hence the variable X_2 creates a significant contribution to the process.

(Refer Slide Time: 19:44)

Sequential Model Building

- h. Build models sequentially and indicate the important effects

Possibilities of additional terms in the model:

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_{12} X_1 X_2 + \hat{\beta}_{22} X_2^2$$



Now let us move on to the sequential Model Building. Here we build the models sequentially and indicate the important effects, okay. So you can add many terms to your already existing model but please make sure that you have adequate number of experimental data points. If you have only 5 data points, then you are pretty much saturated with this particular model. What I am trying to say is make sure that you have adequate degrees of freedom when you are calculating the regression sum of squares.

So you can see that this is 1 option, even though we are not going to look at this option. We will be looking at an addition to the original model having only the main affects. What is that addition? That addition is simply adding beta hat 12X1X2. So we are now only going to consider the effect of the interaction between X1 and X2.

(Refer Slide Time: 20:57)

Effect of Adding $\hat{\beta}_{12}X_1X_2$

- h. Build models sequentially and indicate whether the additional term is important.

Possibilities of additional terms in the model:

$$\hat{C} = \hat{\beta}_0 + \hat{\beta}_1X_1 + \hat{\beta}_2X_2 + \hat{\beta}_{12}X_1X_2$$



So we will be seeing the effect of adding beta hat 12X1X2 and our motivation or our aim is to build the model sequentially and indicate whether the additional term is important. So possibility of additional terms in the model is many and here we are considering the second-order interaction between 1 and 2 regressor variables.

(Refer Slide Time: 21:25)

Effect of Adding $\hat{\beta}_{12}X_1X_2$

h. $\hat{C} = \hat{\beta}_0 + \hat{\beta}_1X_1 + \hat{\beta}_2X_2 + \hat{\beta}_{12}X_1X_2$

To see the effect of the interaction term we simply add the column vector of X_1X_2 . We then simply treat it as a new model involving X_1X_2 and find its regression coefficient.



This may be done **ONLY** for orthogonal designs.

So when you want to do this in the matrix notation, we add a column vector of X_1 and X_2 . You have a column vector of settings of X_1 . You have a column vector of settings of X_2 . So all you have to do is multiply the 2 settings X_1 and X_2 and create a new column vector and this is going to be an additional column vector in your overall X matrix, okay.

(Refer Slide Time: 22:05)

Effect of Adding $\hat{\beta}_{12}X_1X_2$

$$h. \hat{C} = \hat{\beta}_0 + \hat{\beta}_1X_1 + \hat{\beta}_2X_2 + \hat{\beta}_{12}X_1X_2$$

To see the effect of the interaction term we simply add the column vector of X_1X_2 . We then simply treat it as a new model involving X_1X_2 and find its regression coefficient.



This can be done **ONLY** for orthogonal designs.

So we did not treat it as a new model involving X_1X_2 and find its regression coefficient, okay. So when you want to find the effect of adding $\hat{\beta}_{12}X_1X_2$, what you can do is you can forget the presence of the other terms, essentially you are considering a model which is having only $\hat{\beta}_{12}X_1X_2$. You may ask what happened to the other models? Is this correct? Okay. It is as if you are ignoring the presence and contributions of the other 3 entities in the present model and you are only considering $\hat{\beta}_{12}X_1X_2$, okay.

There is a rider or a condition to this way of doing the regression problem. The condition is it should be an orthogonal design, okay.

(Refer Slide Time: 23:00)

X_1X_2 and Y Matrices

$$X_1X_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \\ 0 \\ 0 \\ 0 \\ -1 \\ 0 \\ 1 \\ -0.25 \\ 0.25 \\ 0.25 \\ -0.25 \end{bmatrix}$$

$$Y = \begin{bmatrix} 234 \\ 257.5 \\ 282 \\ 193.5 \\ 187 \\ 181.5 \\ 153 \\ 116.5 \\ 81 \\ 226.9 \\ 217.9 \\ 141.4 \\ 162.4 \end{bmatrix}$$



So our X matrix which we are going to work with is simply the column vector X1X2. We are not having the columns of ones, we are not having the column corresponding to X1 settings and we are not having the columns corresponding to X2 settings. We are only having the column corresponding to X1X2 and you can see that we are having these values and you can see that when you add this column elements, they add up to 0.

So you have the Y column vector simply the same as before. It is the column vector of responses to the experiments at various settings.

(Refer Slide Time: 23:52)


X₁X₂ and Y Matrices

$$(X_1X_2)'X_1X_2 = [4.25] \quad (X_1X_2)'Y = [-127.5]$$

$$[(X_1X_2)'X_1X_2]^{-1} = \left[\frac{1}{4.25} \right] \quad \hat{\beta}_{12} = [-30]$$

NOTE: YOU CAN DO THIS ONLY FOR ORTHOGONAL DESIGNS

$$\hat{\beta}_{12} = (X_1X_2'X_1X_2)^{-1} X_1X_2'Y$$

 NPTEL

So we can now do the usual matrix manipulations. We can find the X prime X matrix. In our case, in the present case, it is X1X2 prime X1X2 that is coming out to be 4.25 and then X prime Y or X1X2 prime Y is -127.5 and you do the estimation of the beta hat 12 and that is X1X2 prime X1X2 inverse X1X2 prime Y. Let me make a small correction here. so here you go. We know that the parameter or parameters are obtained by X prime X inverse X prime Y, that is a general formula in matrix approach to linear regression.

In our present case, the X matrix is X1X2. So we have X1X2 prime X1X2 inverse of the entire product of the 2 matrices and then you have X prime Y which is nothing but X1X2 prime Y. So when you carry out this calculation, you will find that beta hat 12 is -30.

(Refer Slide Time: 25:16)

Regression Sum of Squares

Due to the X_1X_2 addition:

$$SS_{\text{Regression}}(\hat{\beta}_{12}|\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \hat{\beta}_{12}'(X_1X_2)'Y$$

$$SS_{\text{Regression}} = 3825$$

This is due to the contribution from X_1X_2 . Now add the contribution from the regression sum of squares due to other regression coefficients including the intercept $\hat{\beta}_0$.



So the sum of squares of regression due to beta hat 12 given that beta hat 0 beta hat 1 and beta hat 2 are already present in the model, is simply given by beta hat 12X1X2 prime Y which comes out to 3825. This is due to the contribution from X1X2 and now we have to add the contribution from the regression sum of squares due to other regression coefficients including the intercept beta hat 0 to get the regression sum of squares for the full model.

(Refer Slide Time: 25:49)

Error Sum of Squares

c. Degrees of Freedom for Mean Square Error : $n-p$
 $= 13 - 4 = 9$

Now Four parameters ($=p$) are found in the new model
viz.

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_{12}$$



And the degrees of freedom for the mean square error is $13-4$ which is $= 9$ and there are 4 parameters in the full model, okay.

(Refer Slide Time: 26:04)

Error Sum of Squares

Error Sum of Squares (SS_E) = $Y'Y -$

$$(\hat{\beta}_{12}'(X_1X_2)'Y + \hat{\beta}'(X'Y))$$

$$Y'Y = 494813.74$$

$$(\hat{\beta}_{12}'(X_1X_2)'Y + \hat{\beta}'(X'Y)) = 3825 + 490988.15$$



$$\text{Hence } SS_E = 0.59$$

So the error sum of squares is $Y'Y - \hat{\beta}_{12}'(X_1X_2)'Y - \hat{\beta}'(X'Y)$, okay. This is the sum of squares, regression sum of squares brought out by the regressor variable X_1X_2 and this is the regression sum of squares brought out by the original model involving only the main effects X_1X_2 , okay and this included beta not. So the regression sum of squares associated with this $\hat{\beta}_{12}'(X_1X_2)'Y$ corresponds to the model $\beta_0 + \beta_1X_1 + \beta_2X_2$, okay.

Only the main affects model. Then we are adding the X_1X_2 interaction term to the only main affects model, okay. So now this represents the full model including the parameter β_0 , β_1 , β_2 and β_{12} . β_{12} is the coefficient to X_1X_2 regression variable. This is the total sum of squares. This is the total regression sum of squares and so you get the sum of squares of error as 0.59. I request you to carry out these calculations independently and see if your answers are matching with mine.

(Refer Slide Time: 27:45)

Extra Sum of Squares

The null hypothesis $\beta_{12} = 0$ may be tested by the statistic

$$F_0 = \frac{SS_{\text{Regression}}(\hat{\beta}_{12} | \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) / 1}{MS_{\text{Error}}} = \frac{3825}{0.59/9} = 58347.5$$



This is obviously a high F value and hence this parameter β_{12} is significant.

So the null hypothesis $\beta_{12} = 0$ may be tested by the statistic. $F_0 =$ some of squares of regression β_{12} given $\beta_0, \beta_1, \beta_2$, only 1 parameter, independent parameter is being tested 1 degree of freedom and then you have the mean square error which is the sum of squares of the error/the error degrees of freedom and that comes out to be 58347.5. This is obviously very highly F value and hence the parameter β_{12} is significant. Please find the P value for this particular case corresponding to 1 and 9 degrees of freedom.

(Refer Slide Time: 28:30)

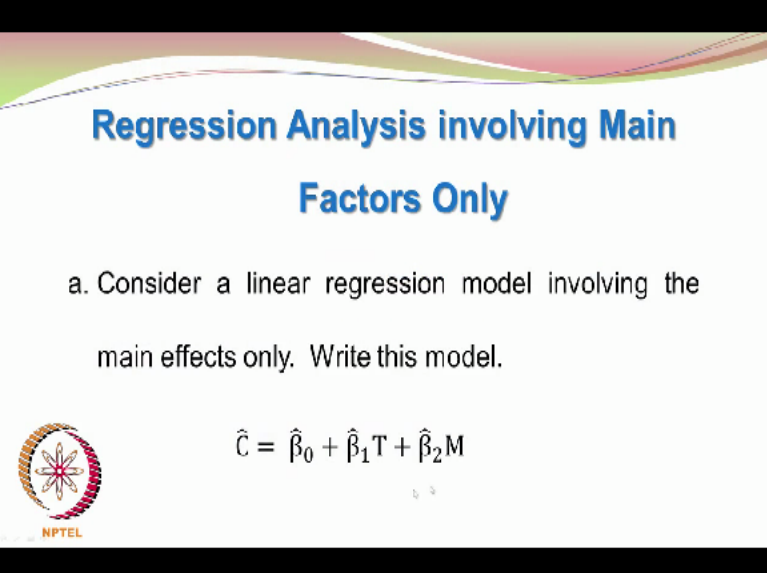
Regression Analysis without Coding



In the next analysis, we will be looking at the data if they had not been coded. For some of us, it would seem like coding is a pain and it is better to directly get a model without coding the data. Let us see what would happen in such a scenario? So this is the actual data, we have seen the


response to be obtained by varying temperature and mass of powder. So this is the complete dataset.

(Refer Slide Time: 29:13)



Regression Analysis involving Main Factors Only

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

 $\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 M$

And so what we will do is consider a linear regression model involving only the main affects only. We are just now doing this for demonstration purposes. We can of course do it for larger model or a model with more terms but as of now, let us consider it for only a model with main affects alone. Main affects are temperature and mass of the powder. So let us call it as $\hat{C} = \hat{\beta}_0 + \hat{\beta}_1 T + \hat{\beta}_2 M$. Let me correct a typo, okay. So these are regressor variables, temperature and mass of the powder expressed as they are, okay.

You can see that the mass of the powder is in the range of 3 to 7.5 whereas the temperature is in the range of 30 to 50, okay. So an order of magnitude difference is there between the 2 factors. In other experiments, this difference between the factors can be quite significant. For example, some factor may be ranging only between 0 and 1 whereas there may be other factors which are running into the order of 100s or 1000s.

So you can see that there is going to be a big order of magnitude difference between the various factors. In such a situation, it is better to do scaling but in the present exercise, we will not do the scaling or coding and making the values range only between -1 and +1. We will take the values as they are and see what happens to model.


(Refer Slide Time: 31:08)

X and Y Matrices

$$X = \begin{bmatrix} 1 & -1 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0.5 & -0.5 \\ 1 & -0.5 & -0.5 \\ & 0.5 & 0.5 \\ & -0.5 & 0.5 \end{bmatrix}$$

$$XUC = \begin{bmatrix} 1 & 30 & 3 \\ 1 & 40 & 3 \\ 1 & 50 & 3 \\ 1 & 30 & 6 \\ 1 & 40 & 6 \\ 1 & 50 & 6 \\ 1 & 30 & 9 \\ 1 & 40 & 9 \\ 1 & 50 & 9 \\ 1 & 45 & 4.5 \\ 1 & 35 & 4.5 \\ 1 & 45 & 7.5 \\ 1 & 35 & 7.5 \end{bmatrix}$$

$$Y = \begin{bmatrix} 234 \\ 257.5 \\ 282 \\ 193.5 \\ 187 \\ 181.5 \\ 153 \\ 116.5 \\ 81 \\ 226.9 \\ 217.9 \\ 141.4 \\ 162.4 \end{bmatrix}$$



So we have the X matrix for the coded case. You can see that all of them are in the same range between -0.5 to +1, okay. So it looks pretty neat here but when you look at the uncoded X matrix, this is the matrix, the column vector of ones and this is the temperature column and this is the mass of powder column. So you can see that the values are differing by about 10 and this is the response vector. We do not really code this response vector. We will just take it as it is.

(Refer Slide Time: 31:56)

Comparison of Coded and Uncoded Results

$$X'X = \begin{bmatrix} 13 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 7 \end{bmatrix}$$


$$XUC'XUC = \begin{bmatrix} 13 & 520 & 78 \\ 520 & 21500 & 3120 \\ 78 & 3120 & 531 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{13} & 0 & 0 \\ 0 & \frac{1}{7} & 0 \\ 0 & 0 & \frac{1}{7} \end{bmatrix}$$

$$(XUC'XUC)^{-1} = \begin{bmatrix} 2.9341 & -0.0571 & -0.0952 \\ -0.0571 & 0.0014 & 0 \\ -0.0952 & 0 & 0.0159 \end{bmatrix}$$

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

$$\hat{\beta}_{uc} = \begin{bmatrix} 352.277 \\ -0.60 \\ -23.5 \end{bmatrix}$$



And so what we are now doing in the slide is to compare the coded and uncoded results. So you have X prime X matrix which is a nice diagonal matrix easy to calculate and easy to invert whereas the uncoded case X prime X matrix is quite ugly and you cannot easily invert it here. To

find the inverse, I can do it with my eyes close. I will just put 1/13 1/7 1/7 here. It is not that easy, especially if you do not have access to a computer.

You will have to do the inverse calculations on your own and these numbers are also looking pretty much different and the off diagonal terms have not vanished and the off diagonal terms are also present and it is a symmetric matrix A_{ij} is A_{ji} , okay and you can see that there is a big order of difference between the values. So in the coded case, the parameters were estimated to be 187.27, -6 and -70.5 whereas in the uncoded case, 352.277 -0.6 and -23.5 and these values are pretty much different, okay.

So when you do linear regression, you have to see how the experimenter has treated the data. Whether he has coded it and got the parameters or whether he has used the data in the raw form and you cannot expect the same model to be obtained in the coded and uncoded cases.

(Refer Slide Time: 33:40)

X and Y Matrices

$$X'Y = \begin{bmatrix} 2434.6 \\ -42 \\ -493.5 \end{bmatrix} \quad XUC'Y = \begin{bmatrix} 2434.6 \\ 96964 \\ 13127.1 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 187.27 \\ -6.00 \\ -70.5 \end{bmatrix} \quad \hat{\beta}_{UC} = \begin{bmatrix} 352.277 \\ -0.60 \\ -23.5 \end{bmatrix}$$

NPTEL

So you have the X prime Y matrix and here also you have the X prime Y matrix. They also look quite different and the parameters are obviously again different.


(Refer Slide Time: 33:53)

Significance of $\hat{\beta}_0$

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

$$\hat{\beta}_{uc} = \begin{bmatrix} 352.277 \\ -0.60 \\ -23.5 \end{bmatrix}$$

It was seen that $\hat{\beta}_0 = 187.27$ which was also the average of the responses in the coded case but **NOT** so in the uncoded case



So it was seen that beta hat 0 is 187.27 in the coded case which was also the average of the responses, okay but it is not so in the case of the uncoded case. So the average of responses is same in both the coded and uncoded cases. There is no doubt about that but the parameter beta hat 0 in the coded case match to the average of the responses but in the uncoded case, it did not.


(Refer Slide Time: 34:25)

Variance-Covariance Matrix

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

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

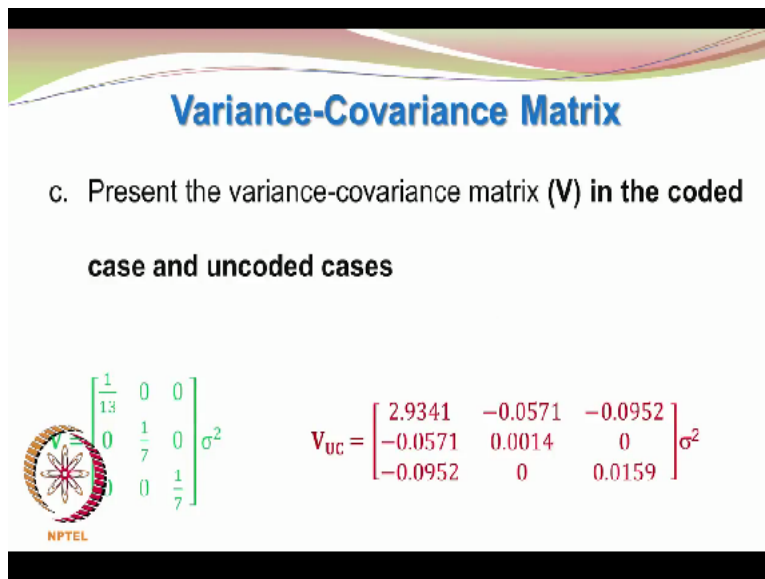
$$V(\hat{\beta}_j) = C_{jj}\sigma^2$$

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


And you how to present the variance-covariance matrix. This also we have seen. We know that the variance-covariance matrix. The variance of the parameters are given by the diagonal terms multiplied by the sigma squared whereas the covariances between the parameters are given by the off diagonal terms and this is a symmetric matrix where C01 is same as C or rather C12 same as C21.


Actually we will call this matrix as V. So V12 will be = V21, V13 will be = V31 and so on and that is why I have put C01 and C01 here, C02 and C02 here and C12 is matching with C12 here. So it is a symmetric matrix. So we have the variance given by the diagonal elements multiplied by sigma square and the covariance given by the off diagonal elements multiplied by sigma squared where sigma squared is the error variance.

(Refer Slide Time: 35:19)



Variance-Covariance Matrix

c. Present the variance-covariance matrix (**V**) in the coded case and uncoded cases



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

$$V_{uc} = \begin{bmatrix} 2.9341 & -0.0571 & -0.0952 \\ -0.0571 & 0.0014 & 0 \\ -0.0952 & 0 & 0.0159 \end{bmatrix} \sigma^2$$

So this is the variance-covariance matrix for the coded case. As I said earlier, it is a nice diagonal matrix. The off diagonal terms are vanishing. There is no covariances between the estimated parameters which is a big plus whereas in the uncoded case, you can see that there is a covariance between the estimated parameters.

(Refer Slide Time: 35:41)

Error Sum of Squares

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

$$\text{Error Sum of Squares (SS}_E) = Y'Y - \hat{\beta}_{UC}'XUC'Y$$

$$Y'Y = 494813.74 \text{ same as in coded case}$$

$$\hat{\beta}_{UC}'XUC'Y = 490988.15 \text{ (same as in coded case)}$$

$$\text{SS}_E = 3825.593 \text{ (same as in coded case)}$$



So now we have to calculate the error sum of squares, that is because we usually use the error sum of squares as a surrogate for sigma squared. So the error sum of squares is nothing but $Y'Y - \hat{\beta}_{UC}'XUC'Y$. The $Y'Y$ is the same as it was in the coded case. You are simply squaring the responses and adding them up and then you also do $\hat{\beta}_{UC}'XUC'Y$ and that comes out to be 490988.15.

This is for the model including X_1X_2 in the main effect form only. We do not consider X_1X_2 in this model, in this uncoded case and this 490988.15 is same as in the coded case. So whether you do coding or not coding, the regression sum of squares and the total sum of squares and the error sum of squares do not change, okay. So you have sum of squares of error as 3825.593.

(Refer Slide Time: 36:47)

Residual Sum of Squares

c. Degrees of Freedom for Mean Square Residuals : $n-p$
 $= 13 - 3 = 10$

Error 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 is $n-p$ $13-3$ which is 10 and so the mean square error is 382.56. So an estimate of the error variance is given by the mean square error which is sigma hat squared, we take it as 382.56.

(Refer Slide Time: 37:04)

Error Sum of Squares

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

Or

$$\hat{\sigma} = 19.56$$



So we get Sigma hat which is the error standard deviation estimate as 19.56. So this completes our discussion on the example set 8 for now and the important thing to understand here is the extra sum of squares concept and the matrix approach to linear regression. You can see that the matrix approach linear regression made everything elegant, compact and easy to execute, especially for large datasets.

It is recommended that you do the calculations pertaining to this example problem on your own and compare your answers with mine. This will give you good training and also make you confident in analysing problems that may crop up in your own industrial experience or when doing research. Thank you for your attention.