

**Dealing with Materials Data: Collection, Analysis and Interpretation**  
**Prof. Hina A Gokhale.**

**Department of Metallurgical Engineering and Materials Science**  
**Indian Institute of Technology, Bombay**

**Regression Analysis - 4**

Hello and welcome to the course on Dealing with Materials Data. In the past, we have seen various sessions of Regression, we are continuing with it. This is the fourth session on Regression Analysis. Let us have a quickly, quick review of what we have done in the past.

(Refer Slide Time: 00:39)

## Review

- Introduction to simple linear regression model
  - Parameter estimation
  - Inference: regression coefficient, mean response, prediction
- Coefficient of determination and correlation coefficient
- Residual analysis to check
  - Randomness of error
  - Normality of error
  - Common variance



We introduced simple linear regression model and its parameter estimation and its inference. Inference particularly on the regression coefficients that are called alpha and beta there, mean response and the predicted value. Then we also discussed the coefficient of determination which actually decides as to how much of the response variable is explained by the input variable  $x$  and we showed the relationship between coefficient of determination and correlation coefficient.

Coefficient of determination is always written as  $R^2$  is equal to the correlation coefficient square or in other words, the absolute value of correlation coefficient is equal to the square root of coefficient of determination.

Then we talked about very briefly, how to approach the problem of linear regression, that is once you get a data, how you go about doing it and finally how do you know that what you have done is correct and that is through checking the assumptions that we have made while going through

regression analysis on the errors and these assumptions are randomness of error, normality of error and common variance. So we have gone through this exercise before.

(Refer Slide Time: 02:13)

## Outline

- Transformation to Linear Regression model
- Multiple Regression Model
- Polynomial regression model
- Logistic Regression Model



In this session, we will see that certain models which are not directly a linear model can also be transformed through mathematical transformation to a Linear Regression Model. We will have an example; we will learn it through example. Then we will introduce Multiple Regression Model and how the parameter estimation can take place and what is the way to do the inference on the regression parameter.

Then we will introduce Polynomial Regression as a special case of multiple regression model and then finally just to give a taste that what happens when the assumption of normality fails. Let us recall that all this analysis that we have done, it actually takes into account the fact that we have assumed the response variable follows a normal distribution and suppose that assumption fails. Then one example we wish to give is the Logistic Regression Model.

We will not go into great details about it. However, it is a special case, a case in the class of generalize linear model, just to get a taste of it as to regression models are all not linear and there are treatments to be given, available in statistics for such generalize linear model.

(Refer Slide Time: 03:56)

### Transforming to Linearity

- Consider the case of fatigue crack growth
- The Paris relationship for crack growth rate per fatigue cycle under linear elastic fracture mechanics (LEFM) is given by

$$\frac{da}{dN} = C(\Delta K)^m$$

*ΔK independent*  
 *$\frac{da}{dN}$  = response.*

- Typical experimental data is obtained as  $\frac{da}{dN}$  and  $\Delta K$ .
- $\Delta K$  is independently fixed and corresponding  $\frac{da}{dN}$  is obtained.



So, let us begin. We take, as I said, transformation to linearity through a case study or a case of fatigue crack growth. The Paris relationship, which is an empirical relationship for fatigue crack growth rate per fatigue cycle under linear elastic fracture mechanics is given by this formula. Where  $a$  is the length of crack,  $N$  is the number of fatigue cycle,  $\Delta K$  is a rate of stress intensity factor.

So, this  $\Delta K$  is fixed and then you get a rate of change of crack per cycle. So, the data comes, let us look at it. The data comes as, you will have  $\Delta K$  as independent and  $\frac{da}{dN}$ , that is rate of change of crack growth per fatigue cycle becomes the response variable. Now as it can be seen that typical data has these two values with us and therefore as I had mentioned here, we have this relationship that  $\Delta K$  is independently fixed and corresponding  $\frac{da}{dN}$  is obtained.

(Refer Slide Time: 05:42)

## Transforming to Linearity

- Thus in Paris relationship  $\frac{da}{dN}$  is the response variable and  $\Delta K$  is independent input variable.
  - Paris Coefficients C and m are to be estimated.
  - Consider the log transformation
- $$\log\left(\frac{da}{dN}\right) = \log(C) + m \log(\Delta K) \quad \therefore \quad \textcircled{Y} = \alpha + \beta X$$
- log(da/dN) → Y, log(ΔK) → X*
- Note that above is a simple linear regression equation in  $\log\left(\frac{da}{dN}\right)$  and  $\log(\Delta K)$ .
  - log(C) and m can be estimated using LSE and further analysis can be carried out.

*simple linear regression model.*



## Transforming to Linearity

- Consider the case of fatigue crack growth
- The Paris relationship for crack growth rate per fatigue cycle under linear elastic fracture mechanics (LEFM) is given by

$$\frac{da}{dN} = C(\Delta K)^m$$

*ΔK independent, da/dN = response.*

- Typical experimental data is obtained as  $\frac{da}{dN}$  and  $\Delta K$ .
- ΔK is independently fixed and corresponding  $\frac{da}{dN}$  is obtained.



So, how do we transform this to linearity? Well, in Paris relationship,

$$\frac{da}{dN} = C(\Delta K)^m$$

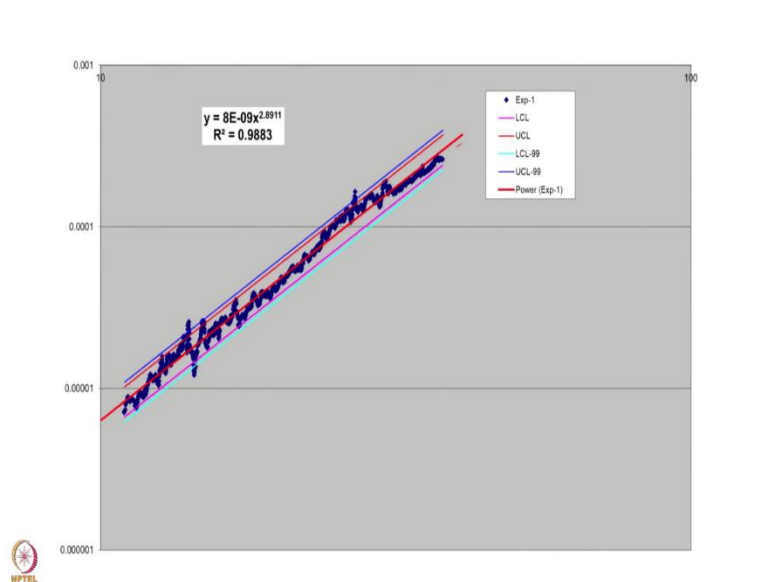
what we can do is take a log transformation on both sides. We again go back to the pen. If you take the logarithm of the Paris equation, it turns into a linear equation

$$\log\left(\frac{da}{dN}\right) = \log(C) + m \log(\Delta K)$$

So here, if you look at it, this can be seen as Y is equal to alpha plus beta x, where x is log of delta K, Y is log of da by dN.

So, what we are trying to show here is that, in the Paris relationship shown in the previous slide, in the Paris relation shown here, can be transformed into a linear relationship through log transformation and once having done that, you can follow the regression model in which log C, this is the game of pen and arrow, log C and m can be estimated through least squares estimate using linear regression model. Simple this is because there is only one element simple regression, linear regression. So, this is how you can approach it. What happens after that?

(Refer Slide Time: 08:16)



This is a plot, it actually shows you the plot. This shows you a plot, these are the data points, the blue are all the data points of, this side is log of da by dN and this is log of delta K, which is shown here, log of delta K which is show here and then these are the log of da by dN by versus log of delta K data points.

This is the straight line which we have estimated over it and we have on top of that, we have shown this two lines. Let us convert it into arrow. This will show the two lines which are actually you recall, we had the upper bound and lower bound. In other words, we had a confidence limit over

estimated value of the beta that is the parameters, regression parameters and we also had it for the mean response. So, this line actually gives a mean response line and these are the upper bound and lower bound. The outer one is a 99 percent upper and lower bound. While the inner one is a 95 percentage upper and lower bound.

What really this shows is that the data more or less except for a few points like this, this and this, most of the points are lie, lie between 95 percent confidence limit. It means that our model is correct and if you exactly put it without the logarithmic transformation, it says that, da by dN which is Y is equal to 8 times 10 to the power minus 9 multiplied by delta K to the power of 2.89 and the R square is almost 99 percent. So, it shows that you have a good fit for the data. Let us move on.

(Refer Slide Time: 10:48)

### Weighted Least Squares

- Consider the simple linear regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i, i = 1, 2, \dots, n$$

- Where,  $E(\epsilon_i) = 0$  and  $Var(\epsilon_i) = \frac{\sigma^2}{w_i}, i = 1, 2, \dots, n$

- Then the estimators A and B should be chosen to minimise

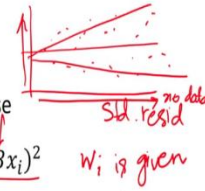
$$\sum \frac{[Y_i - (A + Bx_i)]^2}{Var(Y_i)} = \frac{1}{\sigma^2} \sum w_i (Y_i - A - Bx_i)^2$$

- Taking the partial derivatives following two simultaneous equations in A and B need to be solved

$$\left. \begin{aligned} \sum w_i Y_i &= A \sum w_i + B \sum w_i x_i \\ \sum w_i x_i Y_i &= A \sum w_i x_i + B \sum w_i x_i^2 \end{aligned} \right\}$$

$$Var(\epsilon) = Var\left(\frac{\epsilon}{\sqrt{w_i}}\right) = \sigma^2$$

$i = 1, 2, \dots, n$



There is another example in which we would like to talk about Weighted Least Squares. Now it would be nice if you recall in the previous session, we said that when you want to test the hypothesis, you want to test the assumption, not the hypothesis, you want to test the assumption that there is no heteroscedasticity.

In other words, we want to test the assumption that variance of epsilon, which is same as variance of response variable Y is sigma square.

$$Var(Y_i) = Var(\epsilon) = \sigma^2$$

So, even if the Y changes with different values of i, for i is equal to 1 2 etc, etc n, the sigma remains constant and if you recall we had also shown a plot like this in which we said that, if the standardized, standardized residuals. Please recall, standardized residuals.

$$\text{Var}(\epsilon_i) = \frac{\sigma^2}{w_i}, i = 1, 2, \dots, n$$

If these are the standardized residuals and this is the number of data point that is number of first data point, second data points, etc and if it shows a relationship as shown here. Then there is a chance of what is known as heteroscedasticity.

So, when such a thing happens, you take the variance of epsilon i. Suppose it is a constant divided by a weight w<sub>i</sub>. So, sigma square is still constant, the variation of i comes from a common factor. In other words, sigma square is a common factor and with every data point, i is equal to 1, 2, 3, 4, 5 the value changes proportionally.

In that case, the least squares estimator needs to minimize this particular equation,

$$\sum \frac{[Y_i - (A + Bx_i)]^2}{\text{Var}(Y_i)} = \frac{1}{\sigma^2} \sum w_i (Y_i - A - Bx_i)^2$$

that is the say you please recall, this is the same equation as before. But then we have to normalize it with variance of Y<sub>i</sub> and variance of Y<sub>i</sub> if we call this that this is divided by w<sub>i</sub>. Then it becomes, the 1 over sigma square comes out and this becomes the equation to be minimized and this relation to be minimized please remember w<sub>i</sub> is a given value, it is not to be estimated. The two things to be estimated are A and B and therefore you can follow again the process that we follow for linear, simple linear regression.

We take the partial derivatives with respect to A and with respect to B and we get these two equations and these two equations

$$\sum w_i Y_i = A \sum w_i + B \sum w_i x_i$$

$$\sum w_i x_i Y_i = A \sum w_i x_i + B \sum w_i x_i^2$$

need to be solved simultaneously as a to find a solution for A and B, which I think is a very simple algebraic question and we will not go into details of it. Let us move on. Let us move on to the next issue of Multiple Regression.

(Refer Slide Time: 14:34)

### Multiple Regression

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$   
 where  $\beta_0, \beta_1, \dots, \beta_k$  are some constants and  $\epsilon$  represents random error in this relationship, where  $E(\epsilon) = 0$  and  $Var(\epsilon) = \sigma^2$   
*Handwritten notes:  $\alpha$  above  $\beta_0$ ,  $\beta$  above  $\beta_1$ , and "regression coefficients" with an arrow pointing to  $\beta_1, \beta_2, \dots, \beta_k$ .*

- To find LSE, we need to minimize

$$\sum_{i=1}^n [Y_i - (B_0 + B_1 x_{i1} + B_2 x_{i2} + \dots + B_k x_{ik})]^2$$

Where,  $B_0, B_1, \dots, B_k$  are estimated of  $\beta_0, \beta_1, \dots, \beta_k$



So, let us recall that originally we considered this equation and we said that first let us deal with a simple regression, linear regression by taking only  $\beta_0$ , and  $\beta_1$  and at some point to make our life simple, notation simple, we call them alpha and beta. Let us consider the full case of the response variable Y which is equal to

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

They are also called regression coefficients. These are also known as regression coefficients.



Epsilon again represents the random error with the same assumption that the expected value of random error is 0 and it has a common variance sigma square. Then to find the solution the least squares estimates for this regression coefficient. We set up an equation which is,

$$\sum_{i=1}^n [Y_i - (B_0 + B_1x_{i1} + B_2x_{i2} + \dots + B_kx_{ik})]^2$$

if we assume that  $B_0, B_1, \dots, B_k$  are the estimates, least squares estimate for  $\beta_0, \beta_1, \dots, \beta_k$ . Then this is the estimated value of Y and this is the actual value of Y, we take the difference of the two, square it and then we minimize it. Then what happens, let us move on.

(Refer Slide Time: 16:37)

- This leads to solving following simultaneous linear equations, called normal equations.

$$\sum Y_i = nB_0 + B_1 \sum x_{i1} + B_2 \sum x_{i2} + \dots + B_k \sum x_{ik}$$

$$\sum x_{i1}Y_i = B_0 \sum x_{i1} + B_1 \sum x_{i1}^2 + B_2 \sum x_{i1}x_{i2} + \dots + B_k \sum x_{i1}x_{ik}$$

⋮

$$\sum x_{ik}Y_i = B_0 \sum x_{ik} + B_1 \sum x_{ik}x_{i1} + B_2 \sum x_{ik}x_{i2} + \dots + B_k \sum x_{ik}^2$$



There are k+1 unknowns and we have k+1 equations in k plus 1 unknowns. So, it is a case of solving a simultaneous linear equation

$$\sum Y_i = nB_0 + B_1 \sum x_{i1} + B_2 \sum x_{i2} + \dots + B_k \sum x_{ik}$$

$$\sum x_{i1}Y_i = B_0 \sum x_{i1} + B_1 \sum x_{i1}^2 + B_2 \sum x_{i1}x_{i2} + \dots + B_k \sum x_{i1}x_{ik}$$

$$\sum x_{ik}Y_i = B_0 \sum x_{ik} + B_1 \sum x_{ik}x_{i1} + B_2 \sum x_{ik}x_{i2} + \dots + B_k \sum x_{ik}^2$$

These linear equations look like this and this is again a problem of solving simple algebraic issue. This can also be presented in a different notation.

(Refer Slide Time: 17:54)

- In matrix notation


$$\bullet Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$\bullet$  The multiple regression model can be expressed as  $B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$

$$Y = X\beta + \epsilon \quad Y = XB$$

$$X'XB = X'Y \quad X'Y = X'XB$$

Where,  $X'$  is  $X$  transpose. Therefore,

$$B = (X'X)^{-1}X'Y$$


So, if we go to the matrix notation, you must have done it even to solve the simultaneous linear equation when you studied that in the, your previous degrees.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

So, we have the matrix equation to be solved and this can be solved by multiplying both the sides if you put, in place of beta you put a B. That is B is equal to your estimator betas  $B_0, B_1, \dots B_k$

$$Y = X\beta + \epsilon$$

$$X'XB = X'Y$$

Where,  $X'$  is  $X$  transpose. Therefore,

$$B = (X'X)^{-1}X'Y$$

Let us move on.

(Refer Slide Time: 21:10)

- $E(B) = E((X'X)^{-1}X'Y) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}X'X\beta = \beta$
- Similarly it can be shown that, if  $C = (X'X)^{-1}X'$  then  $CC' = (X'X)^{-1}$  and  $Cov(B) = \sigma^2(X'X)^{-1}$
- $SSR = \sum(Y_i - B_0 - B_1x_{i1} - B_2x_{i2} - \dots - B_kx_{ik})^2$ , and it can be shown that

$$\frac{SSR}{\sigma^2} \sim \chi_{n-k-1}^2$$

*n = # data points*  
*k+1 parameters estimated from n data*  
*∴ d.f. n - (k+1)*  
*= n - k - 1*

$$E\left(\frac{SSR}{n-k-1}\right) = \sigma^2$$



- In matrix notation

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}_{n \times k}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{k+1 \times 1}, \text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n \times 1}$$

- The multiple regression model can be expressed as

$$Y = X\beta + \epsilon$$

$$X'XB = X'Y$$

*Y = XB*  
*X'Y = X'XB*

$$B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Where,  $X'$  is  $X$  transpose. Therefore,

$$B = (X'X)^{-1}X'Y$$

$$y_i = \beta_0 + \beta_1x_{i1} + \epsilon$$

$$= \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \epsilon$$



In this case also, expected value of B. If you look at it, just as in the case of a linear equation. Just please, it is interesting that you look at it, this equation again, particularly this equation again. You look at this equation again, this sounds just like the equation we had written for the linear, simple linear regression equation.

So, simple linear regression say that  $Y_i$  is equal to  $\beta_0$  plus  $\beta_1 x_1$  plus epsilon. So, this you can write as  $\beta_0 \beta_1$ . Here you have to write  $x_1 x_2$ . This is the prime with this plus epsilon.

So, this equation and this equation are equivalent equations and going by that, you will see that again in the same fashion

$$E(\mathbf{B}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

Similarly, we can show that, if you define a constant, a matrix C as

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \text{ then } \mathbf{C}\mathbf{C}' = (\mathbf{X}'\mathbf{X})^{-1}$$

Then the covariance, variance covariance matrix of beta is sigma square times X Prime X inverse.

$$\mathbf{Cov}(\mathbf{B}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

I think this also you can work it out. The residual, our point of interest because we do all the testing using residuals. So, the residual sums of squares SSR is

$$SSR = \sum (Y_i - B_0 - B_1x_{i1} - B_2x_{i2} - \dots - B_kx_{ik})^2$$

We can show that sums of squares of residual divided by sigma square follows Chi square distribution with n minus k minus 1 degrees of freedom.

$$\frac{SSR}{\sigma^2} \sim \chi_{n-k-1}^2$$

Please remember n is number of data points, k plus 1 parameters estimated from n data and therefore the degree of freedom remains is n minus k plus 1 which is n minus k minus 1. So, this calculation should be clear to you and therefore expected value of sums of squares of residual divided by n minus k minus 1. Which is the degree of freedom is sigma square.

$$E\left(\frac{SSR}{n - k - 1}\right) = \sigma^2$$

(Refer Slide Time: 25:05)

## Polynomial Regression

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$$

where  $\beta_0, \beta_1, \dots, \beta_k$  are some constants and  $\epsilon$  represents random error in this relationship

$$E(\epsilon) = 0 \quad \text{Var}(\epsilon) = \sigma^2$$

Regression Coeff.

• This can also be written as

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_k z_k + \epsilon$$

multiple regression

X

Where,  $z_p = x^p, p = 1, 2, \dots, k$

• Thus it transforms itself to the case of multiple regression

• However, for  $k = 2, 3, \text{ or } 4$  it is simpler to solve simultaneous equations to estimate  $\beta_0, \beta_1, \dots, \beta_k$



With this let us move to Polynomial Regression. I want to show that polynomial regression is actually a one special case of multiple linear regression. So, here we write it down as

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \epsilon$$

$Y$  is equal to  $\beta_0$  plus  $\beta_1 x$  plus  $\beta_2 x^2$  plus and so on  $\beta_k x^k$  plus  $\epsilon$ . Now,  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  are some constants or we say that they are, regression coefficients and  $\epsilon$  represents the random error in the relationship with the same assumption that expected value of  $\epsilon$  is 0 and variance of  $\epsilon$  is  $\sigma^2$ .

Then we say that this can also be written as

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_k z_k + \epsilon; \text{ Where, } z_p = x^p, p = 1, 2, \dots, k$$

thus you can see that this is nothing but multiple regression equation. Multivariate, it transforms into multiple regression equation.

It transforms into multiple regression equation and it can be solved same as before. But please remember, the matrix  $X$  is going to be numerically heavy. Because it is going to have powers of  $x$

and therefore the numbers are going to be large and therefore this needs sometimes special treatment to solve the matrix equation, the simultaneous linear equations.

However, if you have  $k$  is equal to 2 3 or 4, it is simpler to solve the equations, the simple simultaneous equation and come to this solution without using matrix algebra. Now we come to the final case as I said. So, far we have been making assumption that  $Y$  follows a normal distribution,  $Y$  is a response variable and it follows a normal distribution with a variance, the mean value as a regression model and a variance as a sigma square.

If sigma square varies proportionally with the data. Then we treated it as weighted least squares. But otherwise we assume that there is no heteroscedasticity and they are the same, this is what we have assumed in our course.

(Refer Slide Time: 28:28)

## Logistic Regression Model

- Experiments are performed at various levels of input variable  $x$
- Response is binary:
  - Success or failure
  - Defective or Non defective
  - ...
- Consider the case where, probability of success can be expressed
$$p(x) = \frac{e^{a+bx}}{1 + e^{a+bx}} \text{ for } -\infty < x < \infty$$
- Such a model of experiments is called Logistic Regression Model.



Now, consider the case where the experiments are performed at various levels of input and your response  $Y$  is either success or failure or it is defective or non-defective. In such cases, if you can express the probability of success in this equation.

$$p(x) = \frac{e^{a+bx}}{1 + e^{a+bx}} \text{ for } -\infty < x < \infty$$

If you, if you are able to express your probability success in this equation, then such a model of experiment is called regression, logistic regression model. Please remember this is a very specific case, this is just to give you a taste that life is not all linear regression models, there are generalized models and this is one of them. Let us go to the next slide.

(Refer Slide Time: 29:26)

### Estimation of Parameters

- Let  $Y_i$  be the response of experiment from logistic regression model with  $p(x_i)$  as its probability of success,

- Then, as per Bernoulli density function, we have

$$P[Y_i = y_i] = p(x_i)^{y_i} * (1 - p(x_i))^{1-y_i}$$

$$P[Y_i = y_i] = \left( \frac{e^{a+bx_i}}{1 + e^{a+bx_i}} \right)^{y_i} \left( \frac{1}{1 + e^{a+bx_i}} \right)^{1-y_i}$$

- a and b can be estimated here by Maximum Likelihood estimation

- Log likelihood function is given by

$$\log(P(Y_i; i = 1, 2, \dots, k)) = \sum_{i=1}^k y_i(a + bx_i) - \sum_{i=1}^k \log(1 + e^{a+bx_i})$$



How do you estimate the parameters? We say that let  $Y$  be the response of experiment from logistic regression model and then you can express it by a Bernoulli trial. You see it is a Bernoulli trial, it is a success and failure. So, if there are

$$P[Y_i = y_i] = p(x_i)^{y_i} * (1 - p(x_i))^{1-y_i}$$

$$P[Y_i = y_i] = \left( \frac{e^{a+bx_i}}{1 + e^{a+bx_i}} \right)^{y_i} \left( \frac{1}{1 + e^{a+bx_i}} \right)^{1-y_i}$$

Then this is the equation which you need to use to estimate a and b and this a and b can estimated, best estimated by using Maximum Likelihood estimation and the log likelihood of the function is given by this.

$$\log(P(Y_i; i = 1, 2, \dots, k)) = \sum_{i=1}^k y_i(a + bx_i) - \sum_{i=1}^k \log(1 + e^{a+bx_i})$$

You can see that this is a nice linear part. This introduces a little difficulty. So, you can use any method like gradient decent or one such method of estimation and get the approximate value of, estimated value of a and b. So with this, we come to an end of this session.

(Refer Slide Time: 31:23)

## Summary

- Transformation to linear regression model
  - Paris equation and log transformation
- Multiple regression model
  - In matrix notation
- Polynomial Regression model
  - As a special case of multiple regression model
- Logistic Regression
  - Parameter estimation by MLE



Let us quickly summarize. We worked on first transformation to linear regression model and we took the case of Paris equation where the log transformation transformed the Paris equation into a simple linear regression model and you can work out the analysis. Estimate the Paris coefficients and do your further analysis. The multiple regression model, we saw that in matrix notation it looks, it is very similar to the simple linear model and you have to use the matrix algebra to solve the matrix equation and come up with the least squares estimates.

We also saw that polynomial regression model can be transformed into a multiple regression model and can be solved numerically using matrix algebra. However, because the terms are going to be the powers of  $x$ , the independent variable, numerically it can become bit challenging. However if you are the powers are 2, 3 or 4, the simple linear models can be, system of linear equations can be solved. Finally we saw an example, a simple example of what happens when the normality assumption is not true. We took the case of a logistic regression model, which is an example of a generalized regression model and generally how it is approached, thank you.