

Stochastic Hydrology
Prof. P. P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Lecture No. # 30
Multiple Linear Regression

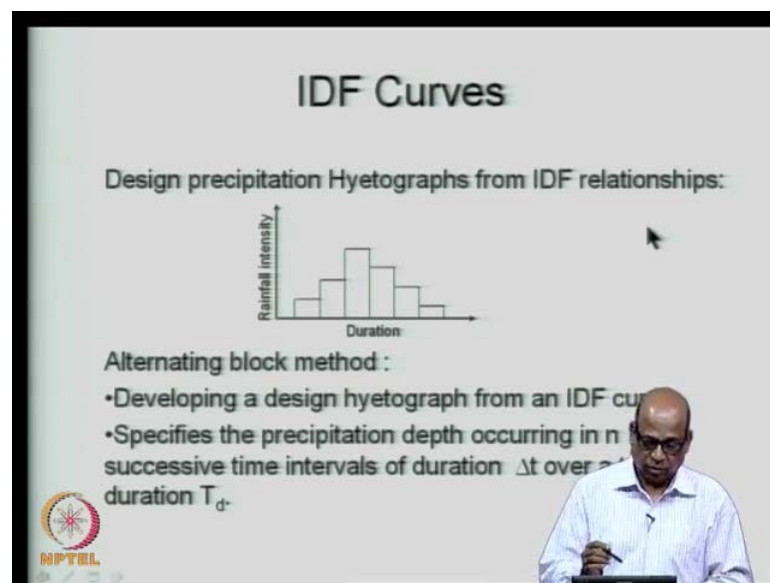
Good morning and welcome to this the lecture number 30 of the course stochastic hydrology. In the previous lectures we have been discussing about the intensity duration frequency relationships that is the IDF relationships, which are essentially used for flood controlled designs typically for urban flood drainage etcetera. From where you would like to pick up the design intensity associated with a return period and associated with the particular design duration. We also saw the methods by which we can construct the IDF relationships and we considered an example of the Bangalore city rainfall using which we construct the idea of the relationships for a given rainfall data series.

Now, today is lecture we will take this discussion forward and then look at once we look at what do we do with the design intensities that we have. So, obtained from the IDF relationships that is we fix a return period based on the type of design that we would like to make and then corresponding to that return period and for a given design duration we pick the intensity of the rainfall from the IDF relationships. Remember IDF relationships are obtained from use of extreme value distributions, typically the gumbel's extreme value type one distribution is what we used in that.

Now, starting with the intensities of the rainfall, we need to form what are called as the design hyetographs, which will give the distribution of the rainfall for the design duration. So, just do not lose sight of what we are doing from the design duration, we went to the IDF relationship for the given return period we picked up the intensity of the rainfall. Now we are coming back and then distributing that rainfall over the design duration, because the intensity that we picked is in fact, the maximum intensity. So, how this maximum intensity of rainfall is distributed over the design duration is what we will look at now.

So, this is the procedure by which we construct the design hyetographs, most of you who would have gone through a basic course in hydrology will know the difference between hyetograph and hydrographs. So, hyetograph is the time distribution of the rainfall intensities, whereas hydrographs are the time distribution where they show the graph of discharge versus time. So, we will start with the IDF relationship and then look at how we obtain the design hyetographs corresponding to the design intensities and these hyetographs. In fact, are used to construct the hydrographs and from the hydrographs we obtain the design dimensions.

(Refer Slide Time: 03:37)



The slide is titled "IDF Curves". Below the title, it says "Design precipitation Hyetographs from IDF relationships:". There is a histogram with "Rainfall Intensity" on the y-axis and "Duration" on the x-axis. The histogram has five bars of varying heights, with the tallest bar in the middle. Below the histogram, it says "Alternating block method :". There are two bullet points: "•Developing a design hyetograph from an IDF curve" and "•Specifies the precipitation depth occurring in n successive time intervals of duration Δt over a duration T_d ". In the bottom left corner, there is a logo for NPTEL. In the bottom right corner, there is a small inset image of a man in a white shirt, likely the presenter, looking at a tablet.

So, the idea of curves from the idea of curves we would now like to obtain the design precipitation hyetographs or design hyetographs from IDF relationship. As I mentioned the hyetograph is typically, it shows the rainfall intensity on the y axis and the duration on the x axis. We use what is called as a alternating block method, which is one of the methods to obtain the hyetograph.

So, this is a method for developing a design hyetograph from IDF curve or IDF relationship, IDF in intensity duration frequency relationship. It specifies the precipitation depth occurring in n successive time interval duration of duration Δt each. So, let us say this is 10 minutes, this is 10 minutes, 10 minutes and so on. So, of equal intervals Δt , how the precipitation depth is distributed with respect to time and this is the total duration of the rainfall event. So, we are talking about a storm of duration

from this point to this point, let us say it is of 1 hour duration, it is of 50 minutes duration and so on. And we have obtained the maximum intensity or the design intensity of the rainfall from the IDF relationship. How we distribute this across time for the duration of the rainfall is what we will see now. And the particular method we are discussing is called as the alternate, alternating block method.

(Refer Slide Time: 05:26)

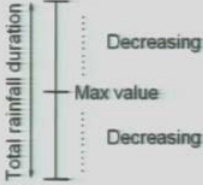
The slide is titled "IDF Curves". It contains a "Procedure" section with three bullet points: "Rainfall intensity (i) from the IDF curve for specified return period and duration (t_d)", "Precipitation depth (P) = $i \times t_d$ ", and "The amount of precipitation to be added for each additional unit of time Δt ". Below the procedure, the formula $P_{\Delta t} = P_{102} - P_{101}$ is shown. To the right of the formula, there is a diagram of a horizontal line segment representing a time interval Δt , starting from a point labeled t_{101} . A presenter is visible in the bottom right corner of the slide frame.


What we do in this is, we know the maximum intensity of rainfall from the intensity we get the precipitation depth corresponding to the duration t_d let say we get i into t_d as a precipitation depth i is intensity which has millimeters per hour or millimeters per time that is length per time is a units for that and time t_d is the time. So, as we are progressing in time let us say this is 10 minute duration, 20 minute duration, 30 minute duration etcetera. So, associated with these we get the depths and then this depth we distribute in this time interval Δt . So, the precipitation that is occurring in this duration will be simply the precipitation at this time minus precipitation at this time that is what we do here.

(Refer Slide Time: 06:20)

IDF Curves

- The increments are rearranged into a time sequence with maximum intensity occurring at the center of the duration and the remaining blocks arranged in descending order alternatively to the right and left of the central block to form the design hyetograph.

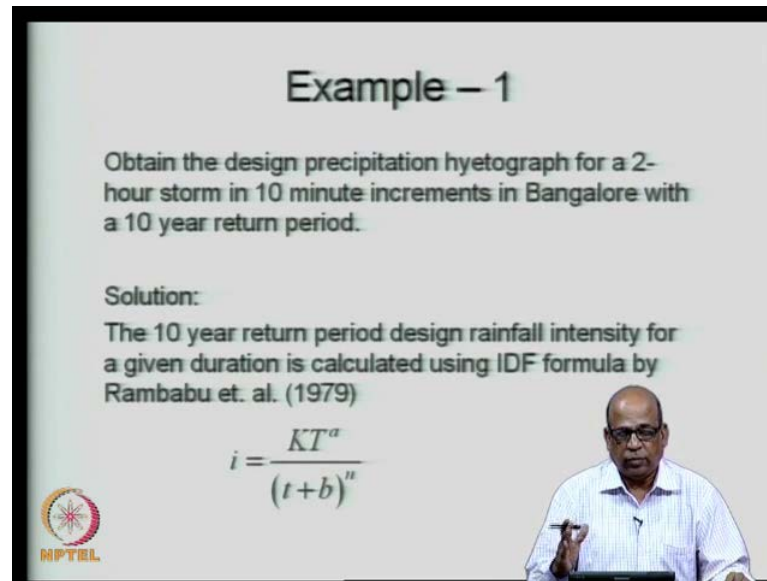


 5

Like this we get intervals of precipitation that is first 10 minutes there is a certain precipitation in the next 10 minutes there is another certain precipitation and so on. What we then do is we arrange these precipitation depths, such that the maximum precipitation depth occurs at the center of the total duration, you have the total duration let us say of one hour and then your Δt is of 10 minutes. So, 10 minutes, 20 minutes, 30 minutes, 40 minutes etcetera up to 60 minutes.

We obtain the maximum precipitation by looking at the $t_d/2$ and $t_d/1$ and so on. And then we arrange the precipitation in such a manner, that the maximum precipitation occurs at the center of the total rainfall duration let us say this is the total rainfall duration and we first fix the maximum value at the center. Then on either side we pick up the values from the precipitation depths and then put one value here, next value here, one value here, next value here, etcetera like this it keeps on decreasing in this direction also keeps on decreasing in this direction. So, let us say this was 1 hour duration, 10 minutes, 20 minutes, 30 minutes, 30 minutes may be here and 40 minutes, 50 minutes, 60 minutes like this, we put the maximum value of the precipitation at the 30 minutes and then keep decreasing on this direction and on this direction that is, how we obtain the precipitation hyetograph, design hyetograph.

(Refer Slide Time: 08:02)





Example – 1

Obtain the design precipitation hyetograph for a 2-hour storm in 10 minute increments in Bangalore with a 10 year return period.

Solution:
The 10 year return period design rainfall intensity for a given duration is calculated using IDF formula by Rambabu et. al. (1979)

$$i = \frac{KT^a}{(t+b)^n}$$

So, let us look at one example to drive home this point very simple procedure and except that you do not forget that the basis for all of this is the design duration and then from which we are picking up design intensity. So, let us look at a storm a 2 hour storm in 10 minute increments for Bangalore city rainfall which I have discussed earlier and we will again use the same rainfall data with a 10 year return period. So, you have 10 minute intervals and then it is a 2 hour storm. So, 120 minutes is the total duration of the storm and then we have 10 year return period. Now, we will just use the formula empirical formula you can also pick this up from the IDF relationship that we derived in the previous lecture corresponding to a 10 year return period and 10 minute 2 hour duration. So, 2 hour duration and associated with 10 year return period you can go to the IDF relationship and pick up the intensity.


(Refer Slide Time: 09:35)

Example – 1 (Contd.)

For Bangalore, the constants are:

$K = 6.275$
 $a = 0.126$
 $b = 0.5$
 $n = 1.128$

For $T = 10$ Year and duration, $t = 10$ min = 0.167 hr,

$$i = \frac{6.275 \times 10^{0.126}}{(0.167 + 0.5)^{1.128}} = 13.251 \text{ cm/hr}$$


However we will use this empirical relationship here i is equal to KT to the power a divided by t plus b to the power n . As I discussed in the previous lecture and for the Bangalore city they have given these constants. So, we will use those constants exactly the same way as we did in the previous class and then obtain for T is equal to 10 year duration t is equal to ten minutes and which is recall to 0.167 hours, 10 minutes is in hours remember the t that we are using here should be in hours. So, we get an i of 13.251 centimeters per hour. So, in this case the units are centimeters per hour because this is an empirical relationship. So, we must be always careful about the units that we are using.



(Refer Slide Time: 10:17)

Example – 1 (Contd.)

- Similarly the values for other durations at interval of 10 minutes are calculated.
- The precipitation depth is obtained by multiplying the intensity with duration.

Precipitation = $13.251 \times 0.167 = 2.208$ cm

- The 10 minute precipitation depth is 2.208 cm compared with 3.434 cm for 20 minute duration, hence 2.208 cm will fall in 10 minutes, the remaining 1.226 (= 3.434 – 2.208) cm will fall in the remaining 10 minutes.
- Similarly the other values are calculated and tabulated.



Now starting with this intensity what we do is, what is it that we did, we took 10 minute interval here sorry we took the 10 minute interval and for this 10 minute duration, we obtain the intensity here. So, this is 13.251 centimeters per hour is a intensity that is occurring in this duration. Similarly for 10 minutes to 20 minutes again, you obtain the intensity and so on. So, at every interval of 10 minutes duration you calculate the precipitation intensities. Now, the precipitation depth is obtained by simply taking intensity into duration. So, 13.251 centimeters multiplied by the 10 minute interval 0.167 which is 2.208.


There is a point that you need to note now, let us say when you go to 20 minutes duration you will put this as 20 the duration will be 20 minutes and it will not be uniformly distributed because there is a non-linear expression here and therefore, you keep putting t is equal to 10 minutes first 20 minutes, 30 minutes etcetera. So, you keep getting the accumulated precipitation depths from here you get the intensities and then you start getting the accumulated precipitation depths corresponding depths. So, you got for 10 minutes you got a precipitation depth of 2.208 centimeters that means, the first 10 minutes the precipitation is 2.208. Now we will go to t is equal to 20 minutes that means, you have a for precipitation of 20 minutes you have you get 3.434 centimeters.

When use it for 20 minutes and then multiplied by the time duration which is 20 minutes you get a precipitation depth of 3.434 which means what during the first 20 minutes the total rainfall is 3.434, but during the first 10 minutes already 2.208 has occurred and therefore, the remaining 10 minutes you will get a rainfall of 1.226 centimeters.

(Refer Slide Time: 13:17)

Example – 1 (Contd.)

Duration (min)	Intensity (cm/hr)	Cumulative depth (cm)	Incremental depth (cm)	Time (min)	Precipitation (cm)
10	13.251	2.208	2.208	0 - 10	0.069
20	10.302	3.434	1.226	10 - 20	0.112
30	8.387	4.194	0.760	20 - 30	0.191
40	7.049	4.699	0.505	30 - 40	0.353
50	6.063	5.052	0.353	40 - 50	0.760
60	5.309	5.309	0.256	50 - 60	2.208
70	4.714	5.499	0.191	60 - 70	1.226
80	4.233	5.644	0.145	70 - 80	0.505
90	3.838	5.756	0.112	80 - 90	0.256
100	3.506	5.844	0.087	90 - 100	0.145
110	3.225	5.913	0.069	100 - 110	0.087
120	2.984	5.967	0.055	110 - 120	0.055



Like this we keep on computing what is the incremental rainfall that occurs during every 10 minutes in this particular case. Once we get that this is how we get let us say these are the intensities 10 minute duration intensities 13.25, 20 minute 10. how did we get this, this we get from the empirical relationship by putting duration is equal to 10 minutes, 20 minutes, 30 minutes etcetera. So, t & d is put like this and then we obtain the intensities. Then corresponding to this intensity we start getting the cumulative depths. So, this is associated with the first 10 minutes, next 10 minutes, next 10 minutes etcetera. So, these are the cumulative depths.

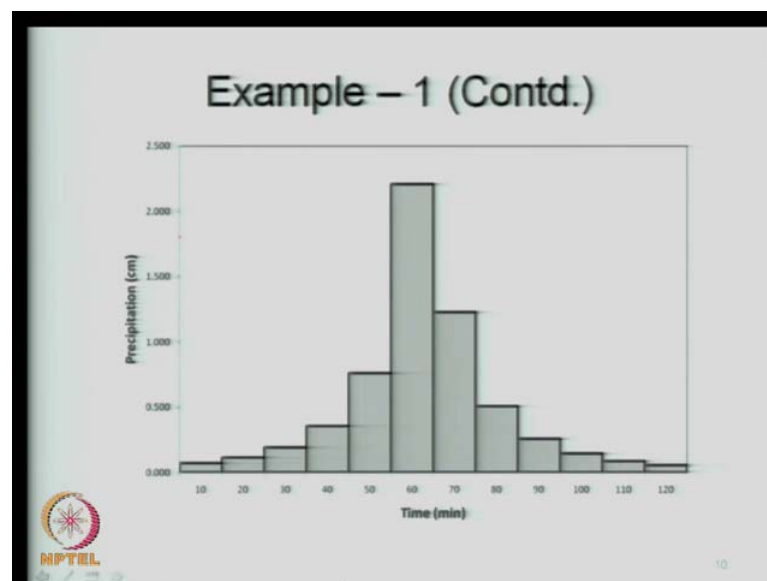
From the cumulative depth we get the incremental depth that means, during that ten minutes what has happened. So, let us say 0 to 10 minutes what is a incremental depth, now this is 2.208, 1.226 this will be this minus this, this will be this minus this and so on. Like this we keep getting incremental depth 4.699 minus 4.194 that is what you will get here and so on. Like this you get the incremental depth. Once we get this you look at the center point here, this is a center point put the maximum value that has occurred at the center. So, essentially what we are doing is that this value you look, this value is the maximum that has occurred here in the incremental depth. So, put this at the center, center of the duration.

Then the next value we pick up 1.226 you put it here. So, this is 1.226 the next smallest value which is 0.760 you put it here, put it here like this one by one you keep distributing

on either side of the maximum value until you reach the end points of the duration. So, I repeat the maximum value you put at the center and then keep distributing the remaining values in the decreasing order on either side of the maximum value and this is how you get the hyetograph. So, for the first duration of 0 to 10 minutes you will get a total precipitation of 0.069 as you can see the rainfall in precipitation depth is slowly increasing it reaches the maximum and then starts decreasing and reaches the minimum. So, that is how you obtain the design hyetograph.

So, this particular method is called as the alternating block method. We are alternating we are taking one value here, one value here, one next year, next year and so on. So, it is called as a alternating block method by far the simplest method of distributing the design intensity that you obtain from the IDF relationship.

(Refer Slide Time: 16:10)



So, this is the hyetograph that you get. So, precipitation in centimeters on the y axis time in minutes in the x axis this is the total duration is 2 hours. For the 2 hours at 60 minutes you get the maximum value which is 2.208 and then you are distributing the value. So, essentially this is the design hyetograph that you will use corresponding to the design intensity of the rainfall. So, this is what we do with the IDF relationships let us just quickly recapitulate what is the significance of the IDF relationship how we obtain and then what is the use of the IDF relationship. IDF relationships are essentially formed based on the extreme value distributions which means you are essentially looking at the

maximum precipitation depth, maximum flood volumes or the peak flood discharges and so on.

So, you are looking at maximum values now these maximum values are used to construct the hydrographs by some procedure you construct the hydrographs and then route the hydrographs through your channel systems or rivers systems etcetera to get hydrographs at various locations. So, any of your hydrologic designs essentially requires the hydrographs, which are obtained from the peak flows. Now from the IDF relationships you do not get the flows you what you get is the design intensities of the rainfall how did we get we fix the duration which is a design storm duration corresponding to the design storm duration corresponding to the return period you pick up the intensity.

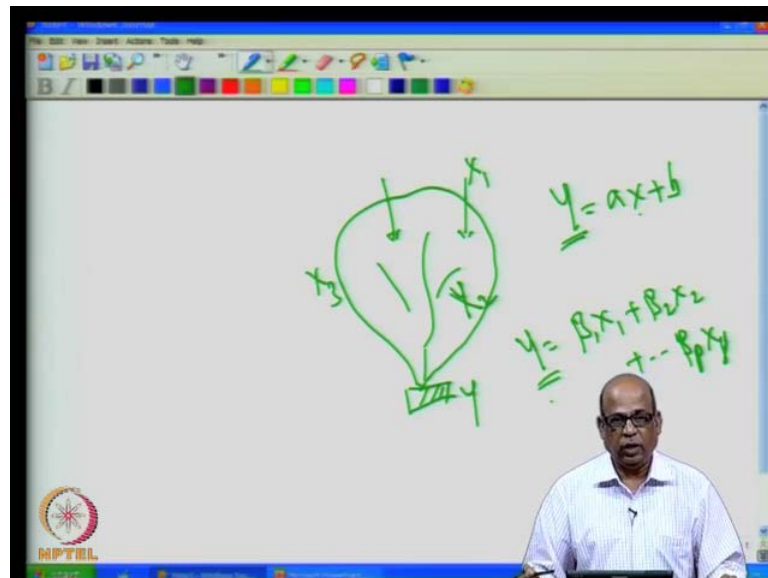
Now, from the intensity of the rainfall we also know now how to distribute the rainfall during the duration let us say that the storm duration was one hour and you picked up the intensity of 30 millimeters per hour or some such thing then you use the alternate block method and then construct the hyetograph, from the hyetograph you can move to hydrograph and so on. So, this once you know the hyetograph how to construct the hyetograph etcetera are not within the scope of this course stochastic hydrology we are essentially looking at only how to use the probability concepts in hydrologic designs. So, we will leave that part aside that will be covered in your basic hydrologic courses.

From the IDF relationship you get the peak intensities of the rainfall from the peak intensities of the rainfall you know how to construct the design hyetographs from the hyetographs you move on to hydrographs, hydrographs you route through your channel networks and then get your design criteria using your regular routing procedures and so on. Now, we will come back to some of the topics that we discussed earlier. So, for the time being we will leave aside the IDF relationships we have completed the portion on IDF relationship. Some time ago in some previous lectures may be lecture number 21, 22 etcetera. We also discussed about dependent and independent variables in hydrology and then how to formulate the relationships between the two types of variables.

For example, you may be considering runoff at a particular location, which is dependent on rainfall during the entire catchment. Now, we introduced the concept of linear regression at that time, where we build linear relationships between the dependent variable runoff in that case and the independent variable rainfall. So, typically we

constructed equations of the type y is equal to a x plus b where x is the independent variable and y is the dependent variable y can be runoff at location, at location and x can be rainfall in the catchment.

(Refer Slide Time: 21:13)



Now, that was a simple linear regression simple because there is only one independent variable and one dependent variable, linear because we are fitting linear equations there. Now in hydrology there are situations where the dependent variable depends not on one, but several independent variables let us look at the runoff itself. So, we were always talking about runoff having being dependent upon the rainfall, but runoff will also depend on let us say you look at a situation where you are looking at runoff at a particular location. Let us say you are looking at a catchment here and then you are interested in the runoff at this location and we were saying that the runoff will depend on the rainfall that is falling on this catchment. So, we use this as the dependent variable and this as the independent variable.

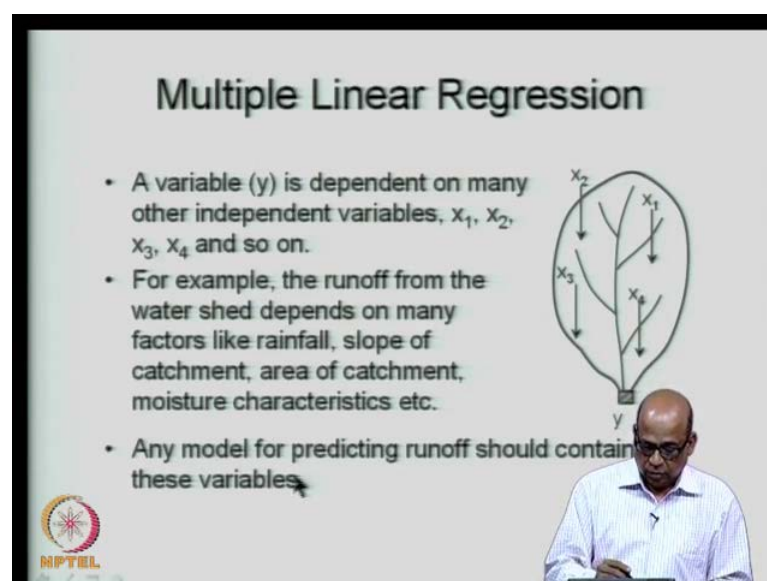
It is known from your basic hydrology that the runoff will not only depend on the rainfall, it will also depend on what kind of soil set up there in terms of let us say antecedent moisture content, how much rainfall has already fallen and how much soil moisture exists and what kind of vegetation is there, what kind of slope is there and so on. So, and also on the evapotranspiration that takes place and so on. So, the runoff at this location although simplistically we always relate it with only one variable x and

therefore, we were justified in using the linear regression also the simple linear regression. Once we start looking at the details of it, we know that it not only depends on the rainfall, but it depends on several other variables. For example, it may depend on x_2 here x_1 may be rainfall x_2 may be soil moisture and x_3 may be temperature in as much as it affects the evapotranspiration and so on.

x_4 may be the slope of the catchment itself and so on. So, y does not depend only on one variable, but it depends on several variables. The concept of the simple linear regression that we introduced earlier namely y is equal to $a x$ plus b this kind of equation, we now generalize and look at multiple linear regression, where the dependent variable y depends not on one, but on several variables x_1, x_2, x_3 etcetera. So, we may write y is equal to $\beta_1 x_1$ plus $\beta_2 x_2$ etcetera. $\beta_p x_p$ there may be p variables on which the dependent variable is dependent on. So, these are the p independent variables. From the simple regression now, simple regression we are now graduating into multiple regression where we are talking about p independent variables, but we still retain the linear structure of the regression equation.

So, this is a linear expression y is equal to $\beta_1 x_1$ etcetera. β_1 to β_p are all constants therefore, this becomes a linear equation and therefore, we now introduce a multiple linear regression multiple because there are multiple independent variables linear because it is a linear equation. So, that is what we will do now.

(Refer Slide Time: 24:26)



Multiple Linear Regression

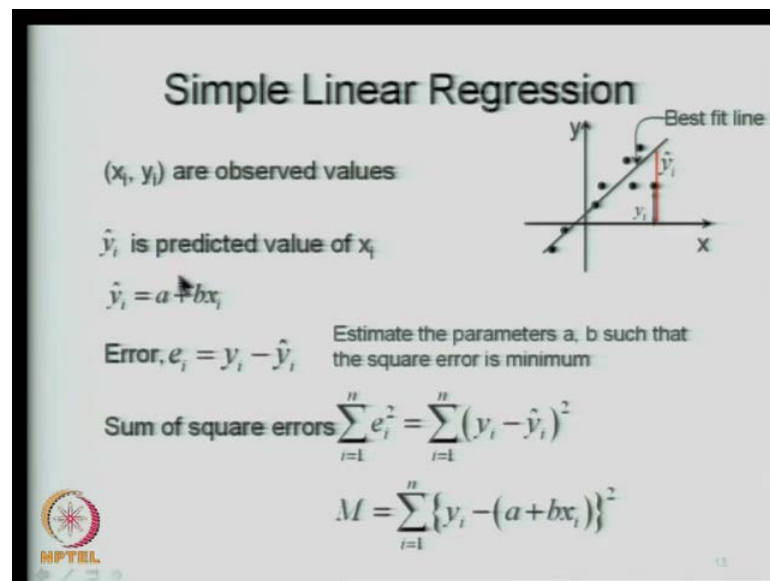
- A variable (y) is dependent on many other independent variables, x_1, x_2, x_3, x_4 and so on.
- For example, the runoff from the water shed depends on many factors like rainfall, slope of catchment, area of catchment, moisture characteristics etc.
- Any model for predicting runoff should contain these variables.

The diagram shows a tree with four arrows labeled x_1, x_2, x_3, x_4 pointing towards a box labeled y at the base of the tree.

NPTEL

So, we starting with this now we will go to multiple linear regressions. As I mentioned you have a dependent variable and you have several independent variables. Now we should use all these dependent variables, all these independent variables to model the independent variable at this location this can be runoff at this location which is dependent on several variables x_1, x_2, x_3, x_4 and so on. So, that is what we do. So, we develop a regression equation for y in terms of x_1, x_2, x_3, x_4 etcetera up to x_p .

(Refer Slide Time: 25:00)




How did we do the linear regression just let us recapitulate you had the observed data on x and y . And we wanted to fit a equation of this type y_i is equal to $a + b x_i$ and we put a y_i we put a cap to y_i to show that it is a predicted value there are observed data x_i and y_i let us say rainfall and runoff both the observed data are there. So, this is observed data, from the observed data you want to fit a best fit line here like this and get a and b . So, your objective there was to obtain the parameters a and b .

(Refer Slide Time: 26:32)

Simple Linear Regression

$$M = \sum_{i=1}^n \{y_i - a - bx_i\}^2$$
$$\frac{\partial M}{\partial a} = 0 \quad a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n}; \quad a = \bar{y} - b\bar{x}$$
$$\frac{\partial M}{\partial b} = 0 \quad b = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i)^2} \quad (x_i - \bar{x}) = x_i' \quad \text{and} \quad (y_i - \bar{y}) = y_i'$$
$$\hat{y}_i = a + bx_i$$



How did we do that we formulated error this is the observed value y_i and this is the estimated value \hat{y}_i . So, the error is observed minus estimated we considered the square errors and write it in terms of y_i minus \hat{y}_i using \hat{y}_i expression $a + bx_i$ and so on. So, we formulate this expression and then minimize this M which is a sum of square errors with respect to the parameters a and b and therefore, we differentiate this M with respect to a and differentiate this with respect to b and set it equal to zero to get the values a and b . So, that is what we did in our linear regression. So, we differentiate with respect to a and differentiate with respect to b and get finally, a and b essentially the principle remains the same.

Please look up your earlier lecture notes, lecture slides where we have discussed the simple linear regression we have also discussed some examples on that when we go to multiple linear regression the principle still remains the same except that we will start working with the matrices now, matrices and vectors. Y is the observed value let us say you are looking at runoff at a particular location and you have the observed series of runoff values and then you have x_1, x_2, x_3 , etcetera. These are independent variables for which the observed values are available let us say x_1 is rainfall, x_2 is soil moisture, x_3 is temperature and so on. So, all these variables which are the independent variables for these variables also you have the observed data.

Then you need to fix or estimate the parameters beta 1, beta 2 etcetera beta p such that, the errors between the observed values and the predicted values predicted values are predicted from the equation linear equation the error between the observed and the maximum and the predicted is minimized in some overall sense that is where we considered the squared errors and so on. So, precisely what we did in the simple linear regression we repeat it for the multiple linear regression the principle remains the same the mathematics also remain more or less the same.

(Refer Slide Time: 28:14)

Multiple Linear Regression

A general linear model of the form is

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_p$$

y is dependent variable,
 $x_1, x_2, x_3, \dots, x_p$ are independent variables and
 $\beta_1, \beta_2, \beta_3, \dots, \beta_p$ are unknown parameters

- 'n' observations are required on y with the corresponding 'n' observations on each of independent variables.

MPTEL

So, we write the general linear model for the multiple regression as y is equal to beta 1 x 1 plus beta 2 x 2 etcetera beta p x p. x 1, x 2, x 3 etcetera x p are the independent variables rainfall, soil moisture although rainfall and soil moisture may be related we will consider that how to handle such situations in next topic where we were discussing how do we handle the correlated variables right now we will assume that they are all uncorrelated. So, x 1, x 2, x 3 etcetera these are all independent variables let us say rainfall, temperature, catchment characteristics in terms of vegetation and so on.

The beta 1, beta 2 etcetera b p up to beta p are the unknown parameters for this particular model. When we use this model to predict the dependent variable we put a cap. So, we say that it is y cap and when we use that without the cap that is the observed values. So, we also have observed values for runoff and we also have observed values for x 1 which is rainfall, x 2 which is vegetation or let us say the catchment area x 3 which is

temperature and so on. So, we identify those particular independent variables which are influencing the dependent variable, we have the historical or the observed data on these variables we also have the historical or observed data on the dependent variable y .

Now, our task is to estimate the coefficients β_1 , β_2 , β_p etcetera β up to β_p . For every variable here y , β_1 , x_1 , x_2 etcetera. We have n observations available with us let us say runoff at a particular location we have for the last 30 years we have all the values available. Similarly concurrent values of rainfall are available, concurrent values of catchment characteristics are available and so on. So, whatever are the independent variables we have the observed values associated with the dependent variable.

(Refer Slide Time: 30:39)

The slide is titled "Multiple Linear Regression". It contains the following text and equations:

- ' n ' equations are written for each observation as

$$y_1 = \beta_1 x_{1,1} + \beta_2 x_{1,2} + \dots + \beta_p x_{1,p}$$
$$y_2 = \beta_1 x_{2,1} + \beta_2 x_{2,2} + \dots + \beta_p x_{2,p}$$

...

$$y_n = \beta_1 x_{n,1} + \beta_2 x_{n,2} + \dots + \beta_p x_{n,p}$$

- Solving ' n ' equations for obtaining the ' p ' parameters.
- ' n ' must be equal to or greater than ' p ', in fact ' n ' must be at least 3 to 4 times large as ' p '.

In the bottom right corner of the slide, there is a small inset image of a man with glasses, wearing a light blue shirt, looking at a laptop screen. In the bottom left corner of the slide, there is a logo for NPTEL (National Programme on Technology Enhanced Learning) featuring a stylized sun or star symbol.

Now, from these observed values then what we do is because there are n observed values here we let us write n equations here. So, y_1 , y_2 etcetera y_n . What do we mean by that y_1 may be the runoff during the first month of first year, y_2 may be runoff during second month of first year and so on. So, like this we may have if you have 50 years of data monthly data we have 12 into 50 which is 600 value. So, n becomes 600 corresponding to each of these values we also have observed values on the independent variables. So, $x_{1,1}$ is the first value of the first variable, second value of the first variable, n -th value of the first variable similarly first value of the second variable, second value of the second variable etcetera like this you have $x_{1,p}$ up to $x_{n,p}$.

Let us say you are looking at 50 years of observed runoff data which means you had 600 values for monthly runoff and x_{11} is the first value of the first variable which is rainfall, second value of the rainfall and six hundredth value of the rainfall and this is the first value of let us say temperature, second value of temperature and so on. Like this you get for different variables you get these values and therefore, you write n equations of this form.

Now, what we need to do is we want those p variables β_1, β_2 etcetera β_p using this set of equations that we have so far. Now n must be at least equal to p , p is a number of parameters which is generally much smaller compare to n . So, a general guideline is that n must be at least three to four times larger than a p . Now p is the number of parameters you know you may be typically talking about two or three parameters in our hydrologic cases whereas, n is the number of data, data values which can be which is generally significantly larger.

(Refer Slide Time: 33:08)

Multiple Linear Regression

- If y_i is the i^{th} observation on y and $x_{i,j}$ is the i^{th} observation on the j^{th} independent variable, the generalized form of the equations can be written as

$$y_i = \sum_{j=1}^p \beta_j x_{i,j}$$

- The equation can be written in matrix notation as

$$Y_{(n \times 1)} = X_{(n \times p)} \times B_{(p \times 1)}$$

NPTEL

We write it in a slightly elegant form using the summation notation. So, we can write it as y_i is equal to $\sum_{j=1}^p \beta_j x_{i,j}$ that is, the i -th equation we write it in this form y_i is equal to $\sum_{j=1}^p \beta_j x_{i,j}$ then, we use the matrix notation because this will be comfortable for this notation also converting the set of equations in the matrix form will be convenient for us to derive the expressions for the constants β_1, β_2 etcetera β_p .


So, look at the matrix Y we have the n values. So, it is a column vector Y is equal to n into 1 it has n rows here is equal to X each of the p variables has n values. So, X is a matrix of n by p size into beta this is capital beta. So, I write it as B which is a vector which is a matrix in this case it is a row vector with I am sorry it is a column vector with p rows and one column. So, what do you get you will get n into 1 here Y. Y is equal to X into beta and capital B denotes the beta vector.

(Refer Slide Time: 34:51)

Multiple Linear Regression

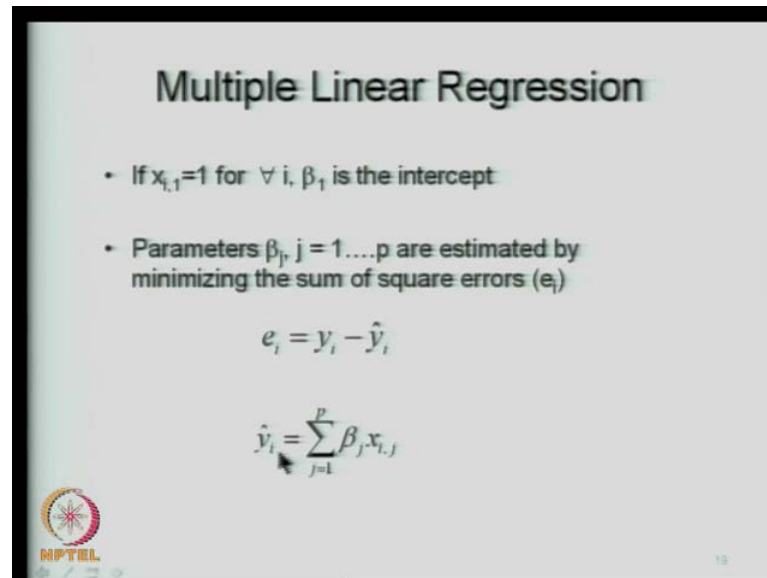
$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} x_{1,1} & x_{1,2} & x_{1,3} & \cdot & \cdot & x_{1,p} \\ x_{2,1} & x_{2,2} & x_{2,3} & \cdot & \cdot & x_{2,p} \\ x_{3,1} & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \\ x_{n,1} & x_{n,1} & & & & x_{n,p} \end{bmatrix}_{n \times p} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \cdot \\ \beta_p \end{bmatrix}_{p \times 1}$$

Y is an $n \times 1$ vector of observations on the dependent variable, X is an $n \times p$ matrix with n observations on each p independent variables, B is a $p \times 1$ vector of unknown parameters.



So, this is the expression that we use as the regression relationship. So, when we write it in a matrix form in the long form y_1 to y_n this is a n by 1 matrix 1 by n by one vector and then you have n by p matrix and then p by 1 vector here. So, this is what we write remember these are the observed values of the dependent variable and these are the observed values for the independent variable and these are the parameters β_1 , β_2 etcetera β_p .

(Refer Slide Time: 35:30)



Multiple Linear Regression

- If $x_{i,1}=1$ for $\forall i$, β_1 is the intercept
- Parameters $\beta_j, j = 1 \dots p$ are estimated by minimizing the sum of square errors (e_i)

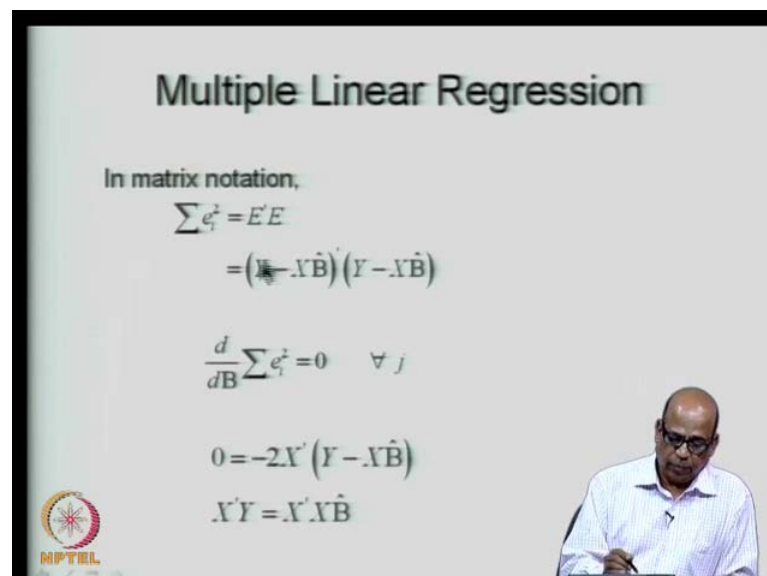
$$e_i = y_i - \hat{y}_i$$
$$\hat{y}_i = \sum_{j=1}^p \beta_j x_{i,j}$$

MPTEL

(Refer Slide Time: 33:09)

Now, when we use these equations we obtain the predicted values. So, we also write the predicted values from the same equation y_i is equal to $\beta_j x_{i,j}$ summation j is equal to 1 to p we write the predicted value as $\sum_{j=1}^p \beta_j x_{i,j}$ that is when we use the expression we get the predicted value corresponding to that. So, there is an observed value and there is a predicted value and therefore, the error is y_i minus \hat{y}_i . So, this becomes the error for the i -th value.

(Refer Slide Time: 36:10)



Multiple Linear Regression

In matrix notation,

$$\sum e_i^2 = E'E$$
$$= (Y - XB)'(Y - XB)$$
$$\frac{d}{dB} \sum e_i^2 = 0 \quad \forall j$$
$$0 = -2X'(Y - XB)$$
$$X'Y = X'XB$$

MPTEL

(A presenter is visible in the bottom right corner of the slide.)

Now, this is the error and then we consider the sum of square errors. So, this when we write it in matrix notation let us say I have this Y as the observed values and my predicted value is $X\beta$. So, $X\beta$ is a predicted value. So, I write this as $Y = X\beta + e$ which is the estimated values estimated value and then I take it as a transpose $e'e$ because I am squaring it. So, $e'e$ is what I take. So, $Y - X\beta$ now this is the observed value and $X\beta$ is the predicted value. So, this is a sum of square errors now all right. So, the sum of square errors I differentiate with respect to β the matrix β now and equated to 0 and that we will get it as when I equated to 0 this for all i for all j is not necessary here if I write it to with respect to individual j 's then this is necessary.

So, we are writing it in terms of the matrix notation. So, when I differentiate and equated it 0 we get the expression from here $-2X'Y + 2X'X\beta = 0$ this must be equal to 0. So, this we write it as $X'Y = X'X\beta$. So, this is what we get from the requirement that $\frac{d}{d\beta} \sum e_i^2 = 0$ $\sum e_i^2$ is written in matrix form as shown.

(Refer Slide Time: 38:13)

Multiple Linear Regression

- Multiplying with $(X'X)^{-1}$ on both the sides,

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

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

or

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

- $(X'X)$ is a $p \times p$ matrix and rank must be p for inverted.

Now, look at this expression now $X'Y = X'X\beta$ we are interested in getting β . So, from this I will write expressions for β to do that what we will do is use this expression and then bring it to a slightly more convenient form by multiplying it with $X'X$ inverse on both sides why that because on this side I have

$X^T X$. So, I will make it one by taking the inverse of that and multiplying it with this. So, $X^T X$ inverse I will take and then pre multiplying with $X^T X$ inverse on both the sides I get $X^T X$ inverse $X^T Y$ is equal to $X^T X$ inverse $X^T X$ beta cap from which I can write beta cap is equal to $X^T X$ inverse $X^T Y$.

(Refer Slide Time: 39:28)

Multiple Linear Regression

- Suppose if no. of regression coefficients are 3, then $(X^T X)$ matrix is as follows

$$(X^T X) = \begin{bmatrix} \sum_{i=1}^n x_{i,1}^2 & \sum_{i=1}^n x_{i,2} x_{i,1} & \sum_{i=1}^n x_{i,3} x_{i,1} \\ \sum_{i=1}^n x_{i,1} x_{i,2} & \sum_{i=1}^n x_{i,2}^2 & \sum_{i=1}^n x_{i,3} x_{i,2} \\ \sum_{i=1}^n x_{i,1} x_{i,3} & \sum_{i=1}^n x_{i,2} x_{i,3} & \sum_{i=1}^n x_{i,3}^2 \end{bmatrix}$$

This capital B here indicates. In fact, that beta is a vector beta is a vector and therefore, we are indicating this by B. Now $X^T X$ is a p by p matrix and its rank must be p for it to be inverted because we are interested in $X^T X$ inverse and therefore, its rank must be p for the matrix to be inverted. So, $X^T X$ matrix when we write this is X which is n by p matrix and we are getting a transpose of that and this is original n by p matrix, we write it as i is equal to 1 summation x_i^2 when you multiply this you will get it like this x_i^2 and then x_i to x_i one etcetera. So, these are the terms that you get for $X^T X$. So, once you get $X^T X$ matrix you can obtain your beta which is given by $X^T X$ inverse $X^T Y$.

(Refer Slide Time: 40:09)

Multiple Linear Regression

- A multiple coefficient of determination, R^2 (as in case of simple linear regression) is defined as

$$R^2 = \frac{\text{Sum of squares due to regression}}{\text{Sum of squares about the mean}}$$
$$= \frac{B'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}$$

MPTEL

Recall that in our linear regression, we also obtained the best fit or the goodness of the fit for the line that was given by the R square, which is the sum of square errors due to regression and sum of square errors about the mean. So, that is what you get in multiple regression also. So, we define some of squares of errors. So, we define R square as B dash X dash Y minus this can be shown to be this expression B dash X dash Y minus n y bar square where y bar is the mean of the y data. At X dash Y is what we have computed already X dash Y is what we would have computed here and we get y cap which y bar which is the mean of the observed data y is n by 1 matrix. So, you get one n by 1 column vector and you will get y bar from that and Y dash Y is simply y square you are squaring the matrix and then you get this value.

Now, this is a scalar number. So, R square you will get it as let us say 0.8, 0.9 and so on. So, this is a scalar number. So, when you do this it will indicate the goodness of fit in some sense the higher the R square value the better is the fit using those parameters. So, we know now how to estimate beta cap which is the set of parameters and we know how to estimate the goodness of fit in terms of the R square values. So, this completes the expressions for multiple linear regression. So, multiple linear regression we use when there are multiple independent variables all affecting the single dependent variable. So, you have a single dependent variable y and then x 1, x 2 etcetera x p as independent variables, we have the expression y i is equal to or y cap is equal to beta 1 x 1 1 plus beta


2×1 and so on. Like this we formulate the expressions and then β_1 , β_2 etcetera up to β_p .

(Refer Slide Time: 42:45)

Example – 2

In a watershed, the mean annual flood (Q) is considered to be dependent on area of watershed (A) and rainfall(R). The table gives the observations for 12 years. Obtain regression coefficients and R^2 value.

Q in cumec	0.44	0.24	2.41	2.97	0.7	0.11	0.05	0.51	0.25	0.23	0.1	0.054
A in hectares	324	226	1474	2142	430	45	38	363	77	84	46	38
Rainfall in cm	43	53	48	50	43	61	81	68	74	71	71	69



To understand this better we will just solve a simple example here, we will take only two independent variables and this can be generalized for any p number of independent variables. We look at the mean annual flood in a particular watershed, what I mean by that is that let us say we have collected data from several watersheds on the maximum discharge that has occurred over previous. We are picking up the maximum discharge and then we want to relate it with the area of the watershed and the rainfall. Remember these are different watersheds and these are the rainfall that have caused this discharges in this particular watersheds for example, a forty three centimeter rainfall occurring in a area of 324 hectares in a watershed of 324 hectares has caused a peak discharge of 0.44 cubic meters per second and so on. So, these are the interpretations.

(Refer Slide Time: 44:21)

Example – 2 (Contd.)


The regression model is as follows

$$Q = \beta_1 + \beta_2 A + \beta_3 R$$

Where Q is the mean flood in m³/sec.
A is the watershed area in hectares and
R is the average annual daily rainfall in mm

This is represented in matrix form as

$$Y_{(12:1)} = X_{(12:3)} \times B_{(3:1)}$$

 25


So, these are different watersheds we are trying to relate the discharge peak discharge with area of watershed and rainfall in centimeters. So, these two are independent variables rainfall and the area are independent variables and the peak discharge is the dependent variable. So, we formulate the expression Q is equal to beta 1 plus beta 2 A plus beta 3 R where A is my area in hectares and Q I represent it as R that is the discharge. So, this can be annual I am sorry Q is the dependent variable and A and R are independent variable. So, I will express Q as a function of A and R. To obtain a intercept we make the first variable as one that means, we wrote in our long form as beta 1 x 1 plus beta 2 x 2 etcetera.

If you make X 1 as 1 which means for all the n values you put X 1 as 1 then you get the intercept associated with the first variable and that variable is beta 1 that parameter will be beta 1. So, we get Q is equal to beta 1 plus beta 2 a plus beta 3 R. Now the task is to obtain the values beta 1, beta 2, beta 3. So, we will express this in our usual notation as Y as the dependent variable and X as independent variable matrix and then beta this is a capital beta. So, B as the parameter vector. So, we have these values now these are the observed values. So, this is y and this is x 1 and this is x 2. So, that is what we write here.

(Refer Slide Time: 45:46)

Example – 2 (Contd.)

To obtain coefficients this equation is to be solved

$$\begin{bmatrix} 0.44 \\ 0.24 \\ 2.41 \\ 2.97 \\ 0.7 \\ 0.11 \\ 0.05 \\ 0.51 \\ 0.25 \\ 0.23 \\ 0.1 \\ 0.054 \end{bmatrix}_{12 \times 1} = \begin{bmatrix} 1 & 324 & 43 \\ 1 & 226 & 53 \\ 1 & 1474 & 48 \\ 1 & 2142 & 50 \\ 1 & 420 & 43 \\ 1 & 45 & 61 \\ 1 & 38 & 81 \\ 1 & 363 & 68 \\ 1 & 77 & 74 \\ 1 & 84 & 71 \\ 1 & 46 & 71 \\ 1 & 38 & 69 \end{bmatrix}_{12 \times 3} \times \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}_{3 \times 1}$$




These are the observed values of y and as I said the first variable will make it as 1, we are writing essentially in terms of three variables now the first variable we make it as 1. So, that we get beta 1 as the intercept itself and these are the second variable which is area in this case and this is a third variable which is the rainfall. So, we are writing y is equal to x into beta this is 12 by 1 and this is 12 by 3 and this is 3 by 1 size of those matrices.

(Refer Slide Time: 46:46)

Example – 2 (Contd.)

The coefficients are obtained from

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

$$(X'X) = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i3}x_{i1} \\ \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \sum_{i=1}^n x_{i3}x_{i2} \\ \sum_{i=1}^n x_{i1}x_{i3} & \sum_{i=1}^n x_{i2}x_{i3} & \sum_{i=1}^n x_{i3}^2 \end{bmatrix}$$




So, you look at this expression now we need to get $X^T X^{-1} X^T Y$. So, first let us get independently all these variables all these matrices. So, we get $X^T X^{-1} X^T Y$. So, $X^T X$ is as I mentioned this is $\sum_{i=1}^n x_i^2$ when do this multiplication you can write it in this summation form we use this summations this is i is equal to 1 to n of the first variable, i is equal to 1 to n and then you are multiplying here the second variable and the first variable, i is equal to 1 to n you are multiplying here third variable and the first variable and so on.

(Refer Slide Time: 47:42)

Example – 2 (Contd.)

$$(X^T X) = \begin{bmatrix} 12 & 5277 & 732 \\ 5277 & 7245075 & 269879 \\ 732 & 269879 & 46536 \end{bmatrix}$$

The inverse of this matrix is

$$(X^T X)^{-1} = \begin{bmatrix} 3.35 & -6.1 \times 10^{-4} & -0.05 \\ -6.1 \times 10^{-4} & 2.9 \times 10^{-7} & 7.9 \times 10^{-6} \\ -0.05 & 7.9 \times 10^{-6} & 7.5 \times 10^{-4} \end{bmatrix}$$


So, like this you get the summations you can use any of the spreadsheet programs to obtain this and you get $X^T X$ matrix. So, for this particular case you get the values $X^T X$ as 12, 5277 and so on then we take the inverse $X^T X^{-1}$ you can use mat lab program for this and then get the inverse of that. So, this is $X^T X^{-1}$ from this you obtain this.

(Refer Slide Time: 48:01)

Example - 2 (Contd.)

$$(X'Y) = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i,2} y_i \\ \sum_{i=1}^n x_{i,3} y_i \end{bmatrix} = \begin{bmatrix} 8.06 \\ 10642 \\ 417 \end{bmatrix}$$

Then you get X dash Y similar to what we obtain for X dash X you can express this in terms of the summations. So, X dash Y will be obtained as summation i is equal to 1 to n y i and so on like this. So, the X dash Y in this case turns out to be 8.06, 10642 and 417.

(Refer Slide Time: 48:31)

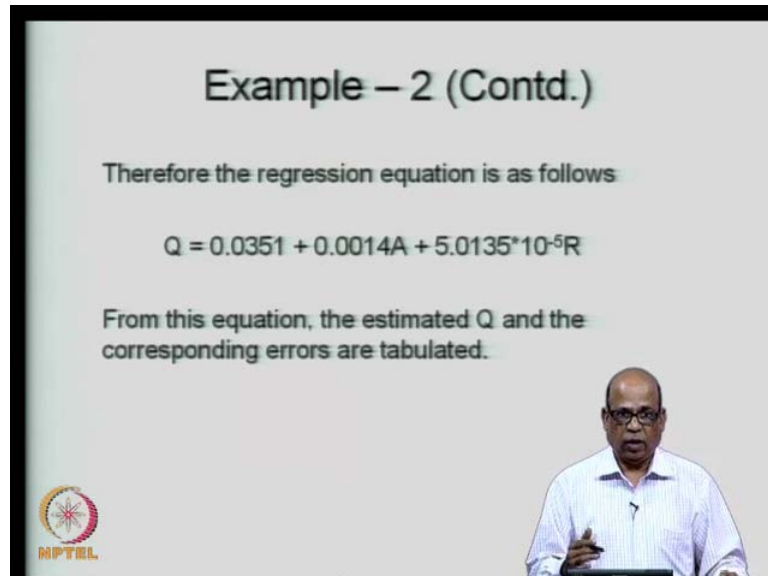
Example - 2 (Contd.)

$$\hat{B} = (X'X)^{-1} X'Y$$
$$= \begin{bmatrix} 3.35 & -6.1 \times 10^{-4} & -0.05 \\ -6.1 \times 10^{-4} & 2.9 \times 10^{-7} & 7.9 \times 10^{-6} \\ -0.05 & 7.9 \times 10^{-6} & 7.5 \times 10^{-4} \end{bmatrix} \times \begin{bmatrix} 8.06 \\ 10642 \\ 417 \end{bmatrix}$$
$$= \begin{bmatrix} 0.0351 \\ 0.0014 \\ 5.0135 \times 10^{-5} \end{bmatrix}$$

Once you get both X dash X inverse as well as say X dash Y you get B cap which is your parameter vector. So, the parameter vector is obtained by this is X dash X inverse which we obtain just now this is X dash X inverse this matrix and X dash Y, this is X dash Y. So, this is 3 by 3 matrix, this is 3 by 1 matrix. So, you get a 3 by 1 matrix. So, beta cap is

obtained as 0.0351, 0.0014 and this 10 to the power minus 5.0135 into 10 to the power minus 5 which means this is a beta 1, beta 2 and this is beta 3.

(Refer Slide Time: 49:23)





Example – 2 (Contd.)

Therefore the regression equation is as follows

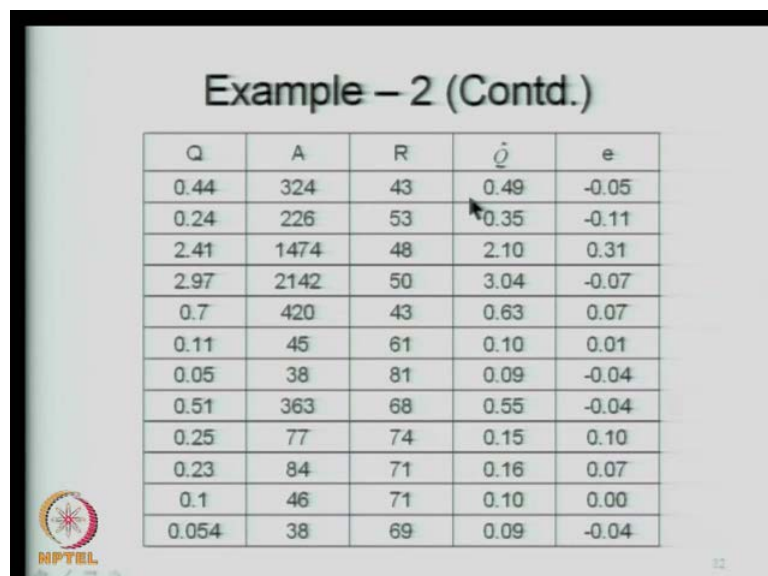
$$Q = 0.0351 + 0.0014A + 5.0135 \cdot 10^{-5}R$$

From this equation, the estimated Q and the corresponding errors are tabulated.





So, the expression that we can write in terms of our Q, A and R is Q is equal to beta 1 plus beta 2 into a plus beta 3 into R, R is the rainfall, A is the area in hectares. So, from this expression we can estimate Q for any given A and R. So, let us look at how these values look, that is using this expression now I want to examine what kind of errors we are getting and so on individually.

(Refer Slide Time: 50:13)



Example – 2 (Contd.)

Q	A	R	\hat{Q}	e
0.44	324	43	0.49	-0.05
0.24	226	53	0.35	-0.11
2.41	1474	48	2.10	0.31
2.97	2142	50	3.04	-0.07
0.7	420	43	0.63	0.07
0.11	45	61	0.10	0.01
0.05	38	81	0.09	-0.04
0.51	363	68	0.55	-0.04
0.25	77	74	0.15	0.10
0.23	84	71	0.16	0.07
0.1	46	71	0.10	0.00
0.054	38	69	0.09	-0.04



Although remember when we apply the regression equations, we are interested in overall error and that is why we are looking at the maximum fit which minimizes the sum of squared errors, but let us also apply this equation for the actual data and then see how much error we are getting. So, this is the observed value and this is the predicted value by predicted value I mean I apply this equation for the given A and R and get the predicted value and these are the errors. So, this is the type of error that I get if I use the expression that we just derived. On of the Q, Q which is the peak discharge and these are the kind of errors that we get.

(Refer Slide Time: 50:48)

Example - 2 (Contd.)

Multiple coefficient of determination, R^2 :

$$R^2 = \frac{B'X'Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}$$

$$= \frac{15.64 - 5.42}{15.77 - 5.42}$$

$$= 0.99$$

$\bar{y} = 0.672, n = 12$

$$B' = [0.0351 \quad 0.0014 \quad 5.0135 \times 10^{-5}]$$

$$(X'Y) = \begin{bmatrix} 8.06 \\ 10642 \\ 417 \end{bmatrix}$$

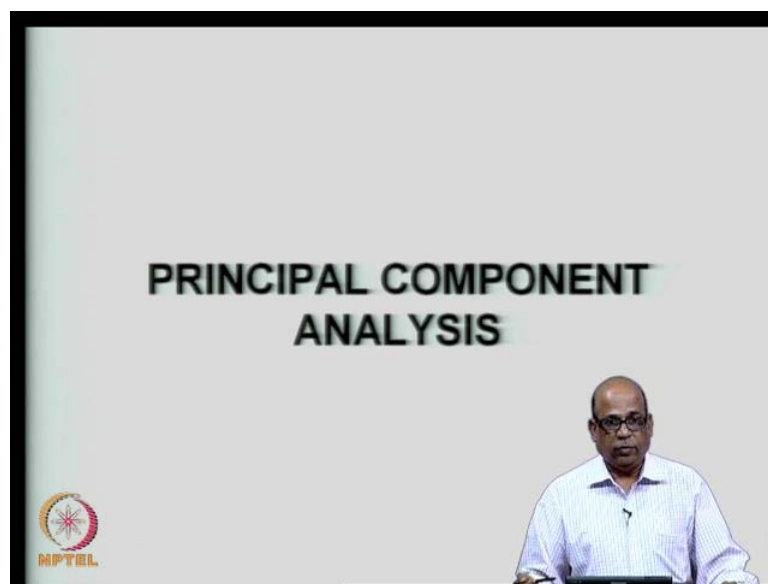
$$Y'Y = 15.77$$

Then we go next and then look at the goodness of fit in some sense that how good is this particular regression. So, we will obtain R square which is given by B dash X dash Y minus n y cap square by Y dash Y minus n y cap y bar square y bar is the mean. So, in this case y bar is simply the mean of these values observed dependent variable values. So, that is 0.672 and then in our case is twelve number of values. So, n y cap you can y, y bar you can get and B dash we have obtained it, B we have obtained it as this vector 0.0351 etcetera. And B dash is the transpose of that. So, I write B dash as the transpose of that and X dash Y we have obtained earlier and Y dash Y is simply Y transpose into Y which comes out to be as scalar quantity which is 15.77. So, we know all of these values here, that is B dash is given X dash Y is given you multiply these two you get 15.64 and when n y cap y bar square.

So, you will get 5.42 and so on. So, R square transfer to be 0.99. Remember R square the closer to one the better is the fit. So, in this particular case a good linear fit exist between the peak discharge, the area of the watershed and the rainfall in the watershed. So, this is how we fit multiple linear regression equations. Typically these situations arise when we are dealing with number of variables for example, rainfall in a particular location dependent on several climatic variable for example, mean sea level pressure it may depend on the geo potential height, it may depend on the land pressures, it may also depend on the wind speed and so on. So, when we are relating the hydrologic variables with climatic variables which quite often arises when we are doing dealing with the climate change impacts.

We need to relate the hydrologic variables with the climate variables there are a large number of variables which are affecting the hydrologic variables when we start looking at the climatic variables. And therefore, we need to have statistical relationships between the hydrologic variables and the climate variables and that is where we typically use the multiple linear regressions and also some non-linear transforms also we use hopefully in the towards the end of this course I will give some background on down scaling of climatic variables. In that lecture we will discuss how we use multiple linear regression for relating the rainfall in a particular location with the climatic variables and so on.

(Refer Slide Time: 54:16)



So, these are the situations where we use the multiple linear regression and this is the background for that, the way we obtain the parameters and the way we assess how good is a fit. Now a question arises, where we are dealing with a large number of variables in the multiple linear regression. Let us say p is quite large let us say ten variables, fifteen variables, twenty variables etcetera. And then some of them may be correlated with each other let us say that one of the variables was soil moisture that you are using and then another variable is evapotranspiration. Now evapotranspiration and soils moisture cannot be treated as independent the evapotranspiration is in fact, dependent on the soil moisture and then the more the evapotranspiration takes place the more the depletion in the soil moisture. So, they are mutually dependent.

So, if you are taking these kind of variables, kinds of variables, where they are mutually dependent there is a correlation that exist between correlation existing between the two variables and also you would not like to handle large sizes. What do I mean by that let us say you have fourteen variables, fifteen variables and each of them have data for hundreds of years then the size that you have to deal with in the regression will be quite large. So, first to handle the dependence among the variables and next to reduce the size of the problem. We divides certain methods and then make both these possible and the one of the methods that I will be discussing in the next lecture is the principle component analysis, which is a very powerful method for most of the multiple linear regression techniques where you are interested in reducing the size as well as addressing the dependence among the several variables.

So, in today is lecture essentially we saw how to start, how to obtain the rainfall hyetographs starting with the IDF relationships. From the IDF relationships you obtain the intensity and then we use the alternating block method to distribute this intensity across the duration and , we want went on to the next topic which is the multiple linear regression starting with the linear regression we have introduced several dependent variables in the linear regression expression. And saw a method by which we obtain the parameters β_1 , β_2 , β_p etcetera. And also to obtain R^2 square which gives the goodness of the fit for this particular equation.

So, in the next lecture we will continue this discussion on multiple linear regression and introduce the principle component analysis by which we will address the dependence of several variables and also address the issue of the size of the problem itself can we

reduce the size of the regression equations. So, thank you very much for your attention will continue the discussion next time.