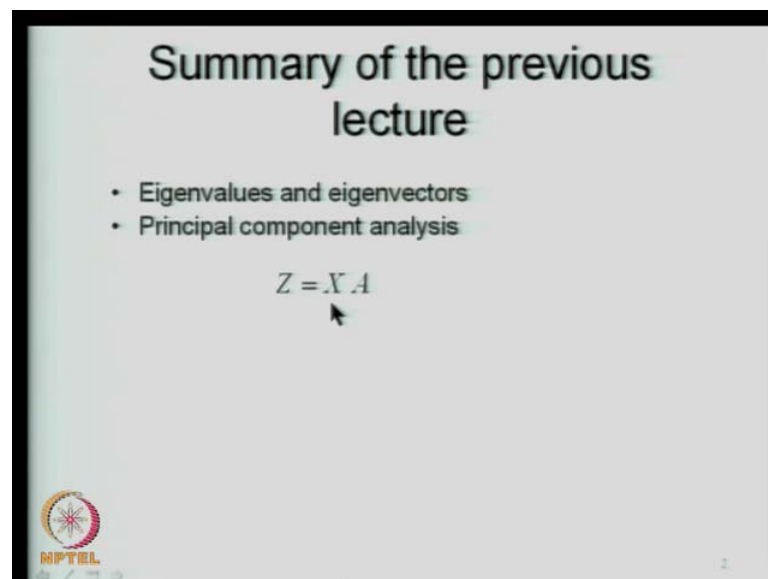


Stochastic Hydrology
Prof. P.P. Mujumdar
Department of Civil Engineering
Indian Institute of Science, Bangalore

Lecture No. # 32
Regression on principal Components

Welcome to this the lecture number 32 of the course stochastic hydrology. If you recall in the last lecture, we discussed essentially about the principal component analysis, but as a prelude to preparation to the principal component analysis, we discussed some basics of the matrix algebra especially, dealing with the eigen values and the eigen vectors. Now, for a square matrix, we saw that the eigenvalues can be got by determinant $A - \lambda I$ is equal to zero, where A is the square matrix for which you need the eigen values and λ are the eigen values and I is the identity matrix. Now, once you get the eigen values, you can also go on to get the eigen vectors by putting $A X = \lambda X$, where, X become the eigen vectors. If you have a the matrix A of size n by n , then you will have n eigen vectors corresponding to the n variables of the A matrix.

(Refer Slide Time: 01:36)



Then we went on to do the principal component analysis, where we express Z is equal to $X A$, where Z is a transform data X is your original data. If you have p variables and

each of the p variables having n observations, then this will be of the size n by p and A is the vector of A is a matrix consisting of the eigen vectors. So, you will have p by p size. So, that is how you get the transformed data Z . Now, this is what we did in the principal component analysis. Towards the end of the principal component analysis, in the previous lecture, I mentioned that there are advantages of dealing with the principal components, when we are regressing using the regression with multiple variables, large number of variables and the typical example that we provide in hydrology is, let say the runoff at a particular location is dependent on several rain gazes, several rainfall values in the catchment, and then apart from the rainfall values it may also depend on other variables.

For example, it may depend on the soil moisture, it may depend on the vegetation, it may depend on the area of the catchment and so on. So, there are several variables on which the runoff is dependent and you want to develop the regression, for the runoff in terms of all of these variables, there are two features in this particular exercise; one is the number of variables themselves number of variables itself is large let say you may be dealing with ten variables twelve variables and so on. With each of the variables having large amount of data, let say rainfall at the location, have you have fifty years of data similarly runoff, you have fifty years of data and rainfall, may be for several the locations at each rain gaze you may have fifty years of data. So, the size of the problem becomes large when you are dealing with large number of variables and large amount of data.

Additionally, many of these variables may be correlated among themselves, as I mentioned, if you are looking at soil moisture and rainfall together in the regression model as independent variables, the soil moisture and rainfall themselves will be highly correlated. So, to account for these correlations in essentially to remove these correlations among the independent variables and to reduce the size of the problem; that means, instead of dealing with ten variables I may want to deal with only three variables. To achieve this purpose, we carry out the principal component analysis. So, there are two major advantages of the principal component analysis; one is the original set of correlated variables are transformed to a set of uncorrelated components and this is a linear transformation of the original variables and in doing so, what we are also doing is that, we are identifying which of these components explains most of the variance present

in the process and therefore, we we can afford to choose only a few of the principal components and ignore the remaining.

So, what was a regression of on ten variables may, now be reduced to a regression only on two variables although these two variables that, we are now dealing with will not be directly related with the original variables. There some kind of a linear transformation of the original variables. So, the original variables would have lost their identity in the final regression, we may not be able to say that the first component is in fact, rainfall again component soil moisture and so on. Both these components will be a linear combination of all the variables that we considered earlier. So, we will progress now and see how we use the principal components in the regression. Remember, our idea is to develop a relationship between the dependent variable. Let say the runoff and the set of independent variables which we have identified based on the physical processes that governed the runoff process and we have the data on all of these variables on dependent variable. We have and concurrent values of the independent variables concurrent because we are saying that the rainfall during that period. Let say, you are talking about a month time period rainfall during that month produces the runoff, during that particular month and therefore, you should have concurrent values of all of these variables dependent variable as well as all the dependent variables.

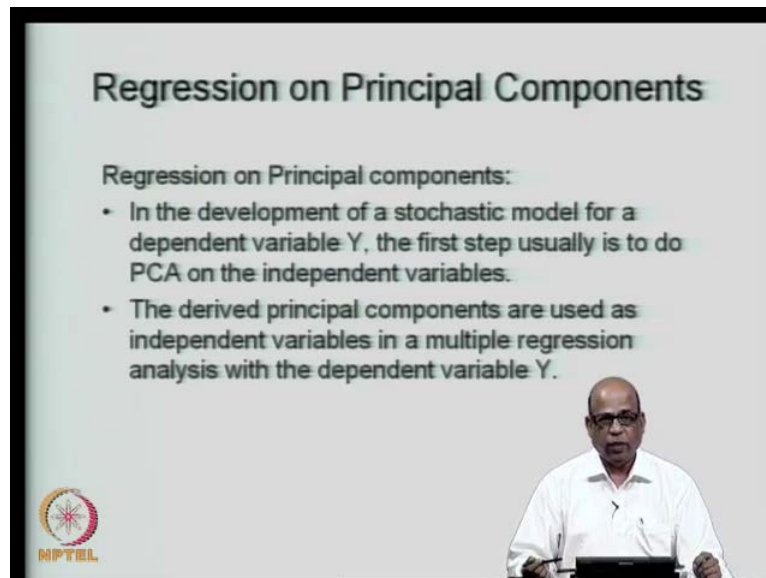
On the independent variables, we do the principal component analysis. In the last lecture, I introduced how to do the principal component analysis on any sets of variables. So, on the independent variables, now we carry put the principal component analysis. What is the idea? The idea is, again that we want to transform this set of independent variables independent in the sense, that they are independent of the dependent variable, but among themselves they may be correlated. So, we want to transform this set of independent variables $x_1, x_2, x_3, \dots, x_p$, into another set of variables which we call them as principal components, by using the transformation and that is what we do and then regress the dependent variable. Now y with respect to the principal components and not with respect to the original variables and that has several advantages let us see.

(Refer Slide Time: 08:25)

Regression on Principal Components

Regression on Principal components:

- In the development of a stochastic model for a dependent variable Y , the first step usually is to do PCA on the independent variables.
- The derived principal components are used as independent variables in a multiple regression analysis with the dependent variable Y .



So, essentially regression on principal components is necessary, because we want to develop a stochastic model on the dependent variables using all these independent variables and the derived principal components as we derive from the first set, we do the principal component analysis on the independent variables and these principal components are used as independent variables in the regression.

(Refer Slide Time: 08:38)

Regression on Principal Components

Procedure:

- Independent variables are standardized.

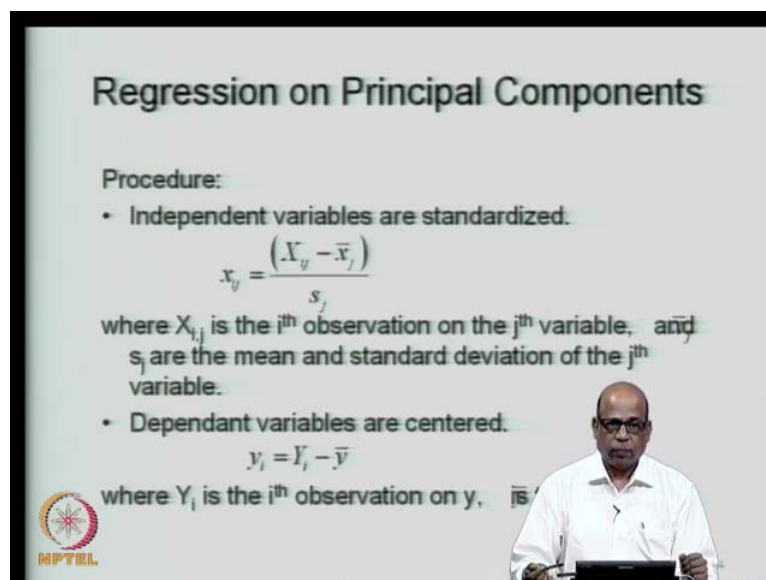
$$x_{ij} = \frac{(X_{ij} - \bar{x}_j)}{s_j}$$

where X_{ij} is the i^{th} observation on the j^{th} variable, and \bar{x}_j and s_j are the mean and standard deviation of the j^{th} variable.

- Dependent variables are centered.

$$y_i = Y_i - \bar{y}$$

where Y_i is the i^{th} observation on y , and \bar{y} is the mean of y .

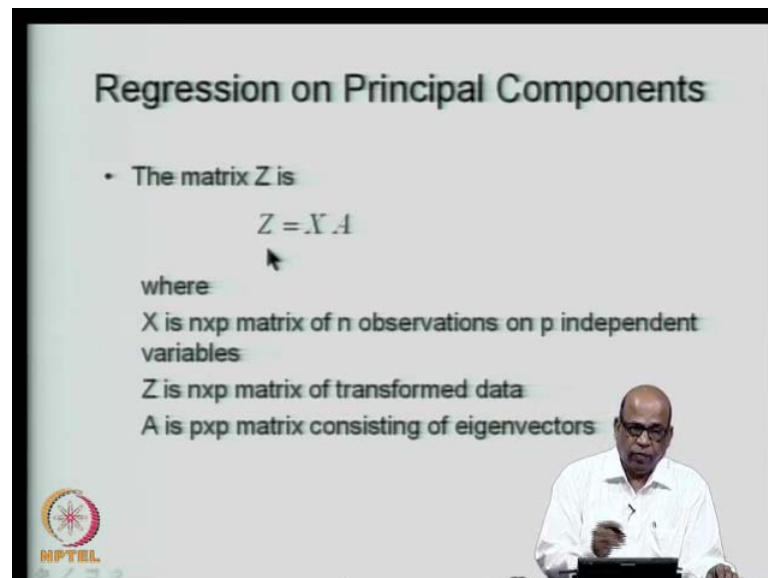


So, why now we regress with the principal components, that is what we call as regression on principal components. What do we do in this, there are p variables, p independent variables here and for each of the p variables you have number of values available with for example, i is the i th observation of the j th variable. So, you may have several

variables j is equal to one to p , p variables you may have the independent variables we standardize. Generally, rather than just centering because these independent variables may come with different units like, I said rainfall may be in depth units soil moisture may be in percentage area may be in area unit and so on. So, to account for the different variables that come with different units we standardize all of them. So, that you are writing it as $x_{ij} - \bar{x}_j$ over s_j . Now the capitals and small are used interchangeably in regression as I mentioned earlier.

So, do not worry too much about the capitals and the small as long as you understand that this is the observed value of the i eth observation and j eth variable and therefore, we are talking about the mean of the j eth variable and the standard deviation of the j eth variable. So, that is how you standardize all the independent variables. The dependent variable being the single variable we can just use a centering. So, we call this as centering, this is observed value for the dependent variable i eth observation minus \bar{y} . \bar{y} is the mean of the dependent variable. So, the dependent variable we center and independent variables we standardize and then use the principal component analysis on the independent variables.

(Refer Slide Time: 10:48)



The slide is titled "Regression on Principal Components". It contains the following text:

- The matrix Z is

$$Z = X A$$

where:

- X is $n \times p$ matrix of n observations on p independent variables
- Z is $n \times p$ matrix of transformed data
- A is $p \times p$ matrix consisting of eigenvectors

The slide also features the NPTEL logo in the bottom left corner and a presenter in a white shirt in the bottom right corner.

Then the transform matrix Z , remember this is the this defines our X , which is the independent variables or the transformed independent variables or the standardized independent variables, if you wish and these are the principal components. So, this is

matrix consisting of the eigen vectors. So, on this we get the transform data Z. So, Z is transformed from the standardized data on the original independent variables, which is the p variables. So, this is a n, n by p matrix and this is p by p matrix, because we are talking about the eigen vectors and Z is again n by p matrix of the transformed data, then we look at the dependent variable.

(Refer Slide Time: 11:43)

Regression on Principal Components

- The regression model is:

$$Y = ZB \quad \text{or} \quad y_i = \sum_{j=1}^p \beta_j z_{ij}$$

Where:

- Y is nx1 vector of n observations of the centered dependent variable,
- Z is nxp matrix of n values for transformed data of p variables, and
- B is a px1 vector of unknown parameters

MPTCL

Let me explain that, you have the dependent variable let say, this is the y matrix which is n by one vectors because there are n observations and this is Z which is a n by p matrix. So, this is n by one and this is n by p and these are the coefficients of the regression this is a coefficient matrix and that is p by one vector of unknown parameters and our aim in the regression is to determine these unknown parameters. Recall from the lecture, before the previous one, where we discuss the multiple linear regression on how to obtain these unknown parameters for a multiple linear regression, exactly the same procedure we follow here and then obtain the betas. Now, in the scalar form it is written as y i is equal to j is equal to one to p summation beta j z i j, this is a scalar form of this matrix form. So, essentially we solve this now and then obtain beta Z is known Y is observed values of the dependent variable and therefore, we can obtain B the matrix B.

(Refer Slide Time: 13:20)

Regression on Principal Components

- The matrix B is estimated as

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

NPTEL

So, from your multiple linear regression, just compare this, how did we estimate B in that case. In this case, when we are writing $Y = ZB$, we write B cap, which is the estimated value of B as Z dash, Z inverse Z dash Y , where our Z is this Z now X A, this is Z and our Y is n by one. So, Z is let us look at the dimensions, now what is Z , Z is n by p matrix. So, Z is n by p . So, this is p by n Z dash and Z is n by p and therefore, you will get p by p and Z dash is again p by n and this is n by one. So, p by n , n by p . So, you get p by p here inverse p by p and then p by n therefore, you get n by n and then n by one. So, you will get n by one. I am sorry lets look at that this is, n by p Z is n by p minor matrix and this is Z dash is p by n . So, you will get p by p and this is Z dash is p by n . So, you will get p by p , p by n and n by 1. So, this will be p by 1 B dash will be p by 1.


(Refer Slide Time: 15:26)

Regression on Principal Components

- The matrix B is estimated as

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

Handwritten annotations in red ink: $Z'Z$ is a $p \times p$ matrix, $Z'Y$ is a $p \times 1$ vector, and \hat{B} is a $p \times 1$ vector. A vertical list of $\beta_1, \beta_2, \dots, \beta_p$ is shown to the right, with a $p \times 1$ label.




So, that is what you get here this is beta 1, beta 2, beta 3, etcetera up to beta p. This is p by 1 this is a matrix that you obtain from this expression here. So, essentially we write y is equal to z into B and then our aim is to obtain this B. And this is how we obtain B from this. Once these are fixed that is these are obtained or these are estimated beta 1, beta 2, etcetera beta p, then your regression equation is in place.

(Refer Slide Time: 16:06)

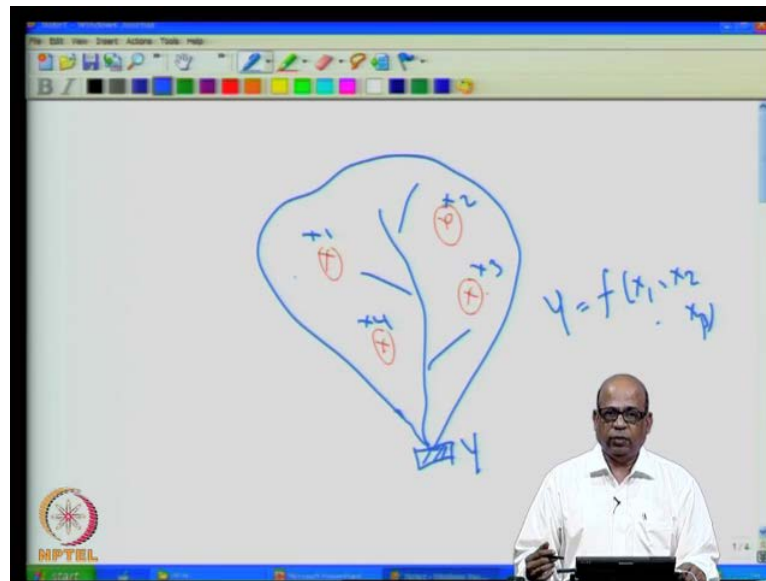
Example – 1

The annual yield of a basin is to be obtained from annual rainfall of 10 stations in and around the basin. The annual rainfall in mm for the 10 stations ($x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$ and x_{10}) and the observed annual yield (Y) in mm for 19 years is given.

Obtain the prediction model for calculating annual basin yield (Y) from annual rainfall using PCA.



(Refer Slide Time: 16:40)



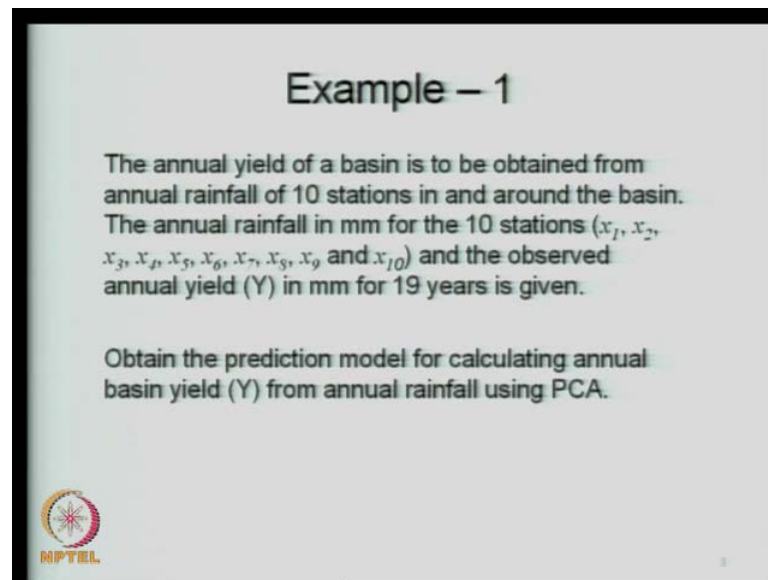
So, let us look at an example now, here what we are doing is that let say, we have a basin where let say, this is a basin and then you have the runoff values available here. And then these are several rain gauges. Let us look at this is rain gauge 1, rain gauge 2, rain gauge 3, 4 etcetera. A common method is, in the hydrology that you get some averaged rain value and then relate it with the y here. But it may so happen that, you would like to use all of these rain gauges independently let say, there is x_1 here x_2 x_3 and so on. You may want to use all of them and then relate with the runoff at this location. So, we may want to write y as a function of x_1 , x_2 , etcetera x_p . There are p independent variables and then we want to estimate the runoff, at this point which we also we call it as yield of this stream let say, there is a stream here and then these are all the stream locations. So, this rainfall is contributing to runoff at this location and this runoff, we want to estimate based on the values here x_1 , x_2 , x_3 , x_4 and so on.

There may be p such rain gauges. Let us do this exercise, the first point to be noted here is that, x_1 and x_4 may be correlated x_1 and x_2 , may be correlated because they may come from the same hydrologic region. We are talking about this watershed here and therefore, they may come from the same hydrologic region. This is a homogeneous hydrologic region and therefore, the rainfall here and rainfall here and at all these locations may have significant correlations. And the other thing is that you may have so many rain gauges. Here, that the size of the problem may become slightly unmanageable, if you are having let say, ten variables twelve variables and so on. Additionally, apart

from the rain gauges, you may also have some other variables being put into the runoff estimation. For example, you may want to put in a vapour transportation because from the run rainfall part of it also goes as a vapor transportation and therefore, you would like to account for that in the runoff that is observed. You have the observed values of y let say, for the last fifty years, every month you have the observed values. So, six hundred values, you may have, you have concurrent observed values on rainfall at this location which gives x_1 rainfall, at this location which gives x_2 and so on. For all the p variables, for all the p sites you have rainfall values observed concurrently with the runoff values.

Now, we want to fit a regression equation of the type y is equal to $f(x_1, x_2, \text{etcetera } x_p)$. We will use the principal component analysis and then fit this regression equation. So, on the principal component for the principal component analysis the first thing we do is to standardize the original data that is a on the independent variables and also center the dependent variable data then we have to carry out the principal component analysis how did you carry out the principal component analysis? Once you have the data matrix X , which is n values, corresponding to each of the p variables, which means a it is an n by p matrix. We calculate the covariance matrix from this; that means, covariance of p variables with p variables. So, we get a p by p matrix of covariances. On the p by p covariance matrix which is a square matrix, we get the eigen vectors and the eigenvalues for the covariance matrix and we use these eigen vectors as the principal components and then obtain the associated principal components for the covariance matrix and then use the regression equation on these principal components.


(Refer Slide Time: 16:03)



Example – 1

The annual yield of a basin is to be obtained from annual rainfall of 10 stations in and around the basin. The annual rainfall in mm for the 10 stations ($x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$ and x_{10}) and the observed annual yield (Y) in mm for 19 years is given.

Obtain the prediction model for calculating annual basin yield (Y) from annual rainfall using PCA.

 MPTEL

So, let us see how we do that in the example. So, if you have rainfall annual, rainfall at ten stations. So, these rainfall values we denoted as x_1, x_2, x_3 , etcetera up to x_{10} . So, there are ten stations. So, you have ten variables associated with it, the observed annual yield associated with these nineteen years of annual rainfall are also given; that means, this is annual rainfall and the associated annual yield by yield. We assume understand that it is a runoff at that particular location. Which means that is a yield of that water shade which is just the runoff that is observed. So, we have concurrent values for nineteen years on the annual yield as well as on these ten stations which are taken as independent variables here. So, we want to obtain the prediction model for the yield associated with the rainfall. So, we want to develop a relationship between the annual yield and the rainfall in all these ten stations.

(Refer Slide Time: 22:40)

The annual rainfall in mm for 10 stations and observed basin annual yield (Y) in mm

Year	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
1979	1948	4177	5496	2922	5713	3640	3203	2739	2167	2299	3255.2
1980	2261	3670	7797	3327	6934	4424	3692	3451	2866	2653	3682.7
1981	1989	4353	7392	2837	6275	4827	4476	4403	3568	3241	3921.9
1982	1999	3307	7061	3439	6641	4815	4256	4129	3447	3046	3909.3
1983	2086	4230	6564	2987	6675	3959	3900	3559	4078	3583	3768.9
1984	1717	2714	5919	3394	5605	3648	3085	2440	2631	2587	3106.4
1985	1383	2357	5053	2958	5144	3106	4052	3006	3049	2890	3069.4
1986	1470	3004	3951	2691	5116	3557	2775	1909	1952	1723	2940.2
1987	1350	2446	4280	2397	4722	3556	2818	2945	2931	2733	3015.3
1988	1602	4188	5910	3619	6869	5142	3190	3660	3964	3107	3953.2
1989	1417	3631	5145	3282	5226	3793	2663	3017	2579	3367	3172.4
1990	1662	4683	6384	6376	7313	4679	3037	3666	3142	2621	3791.0
1991	1955	4553	5679	6141	6068	3651	2601	2791	2148	2448	3344.8
1992	1974	3836	6021	5646	5876	4026	3037	3920	2583	2742	3650.3
1993	2094	4183	6733	6720	6044	6573	2465	3406	2410	2539	3878.7
1994	3149	6128	8151	9048	8384	7467	2888	3522	2496	2895	4606.2
1995	1471	2952	4151	4975	5149	4733	2603	3493	3396	3554	3498.8
1996	1691	3711	4200	4962	5359	3782	3185	3099	3381	2938	3241.0
1997	2373	4836	6704	6563	6197	5001	3902	3685	3636	3365	4013.5

So, let us look at how the data is organized we have nineteen years of data. So, these are the annual values. So, this is the year number and the annual yield is given here, for all the nineteen values this we take it as a dependent variable y and these are all the rainfall values given at the ten stations. So, at the ten stations you have the associated rainfall value now both are given in millimeter units. What do I mean by runoff wing in millimeters that you have observed the runoff. Let us say, you have observed the volume of runoff divided by the total area of the catchment, you will get runoff in the depth units. So, all of these are in depth units. Now, as you can see you may have a significant correlation between x_6 and x_9 . For example, x_6 and x_{10} , x_2 and x_4 and so on. So, among these variables, there may be some combinations of variables which have significant cross correlation; that means, they may be correlated with themselves.

Now, such problems in regression are called as those of multicollinearity; that means, there is a significant correlation among two or more variables among the independent variables. So, this correlation needs to be addressed when we are developing a regression relationship. As I said, the first step is that we standardize all the independent variables and generate the vector x, which consist of the matrix x which consist of the standardized independent variable that will be p by n which means ten by nineteen that is n by p that is nineteen into p variables.

So, let us look at the regression equation now. What we do is to demonstrate the utility or the usefulness of the principal components. First, we will use all of these ten variables as they are and develop a multiple linear regression ship linear regression between y and these ten variables as they are we will not do any principal component analysis to begin with.

(Refer Slide Time: 25:16)


Example – 1 (Contd.)

A regression equation is obtained using all the 10 stations annual rainfall data is as follows

$$Y_{(19 \times 1)} = X_{(19 \times 10)} B_{(10 \times 1)}$$

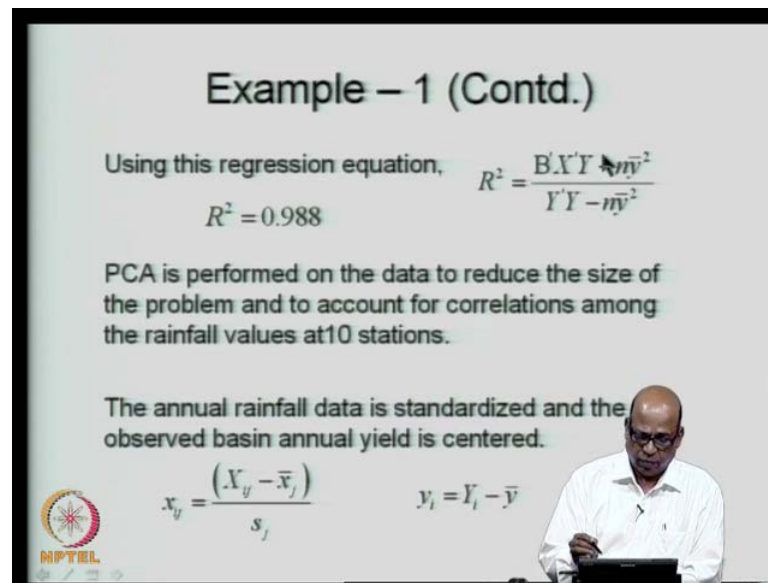
$$\hat{B} = (X'X)^{-1} X'Y$$

The multiple linear regression equation is as follows.

$$Y = 782.4 + 0.1861 x_1 + 0.0484 x_2 - 0.0198 x_3 + 0.0019 x_4 + 0.1196 x_5 + 0.1555 x_6 + 0.0232 x_7 + 0.1948 x_8 + 0.0799 x_9 - 0.0041 x_{10}$$


So, when we do that we express y as x b this is from your multiple linear regression y is a vector of 19 by n there are 19 values and there is only one variable. So, 19 by n x is a matrix of 19 values of ten variables. So, this is 19 by 10 and B is a vector of unknown parameters and there are ten such parameters one associated with each of the variables. So, this is 10 by 1. So, you get beta cap as x dash, x inverse, x dash y, this is from your original regression. And So, you will get or regression of this type, we also introduced the intercept there you just recall form your multiple linear regression, then you get expression of y as a function of these ten variables and you get the intercept as 782.4 to get the intercept what you would have done is you would have put the first values of all the betas as one, all the x one as one you just refer to the multiple linear regression to get this intercept and you get the coefficients beta one, beta two, etcetera up to beta ten based on this equation here.

(Refer Slide Time: 27:32)



Example – 1 (Contd.)


Using this regression equation, $R^2 = \frac{B'XY - n\bar{y}^2}{Y'Y - n\bar{y}^2}$

$R^2 = 0.988$

PCA is performed on the data to reduce the size of the problem and to account for correlations among the rainfall values at 10 stations.

The annual rainfall data is standardized and the observed basin annual yield is centered.

$x_j = \frac{(Y_j - \bar{y}_j)}{s_j}$ $y_i = Y_i - \bar{y}$



Now, this is what we obtain, when we regress the dependent variable on the independent variables as they are observed without doing any principle component analysis as you can see here, all the ten variables have been used and there are ten regression coefficients. Here, if we use this regression equation as obtained here and we get the R square value, this is what we discussed in the previous lecture R square value is given by $B'XY - n\bar{y}^2$, where \bar{y} is the mean of the Y data by $Y'Y - n\bar{y}^2$ n is the number of data X' is the transpose of your X matrix B' is the transpose of B vector.


(Refer Slide Time: 25:16)

Example – 1 (Contd.)

A regression equation is obtained using all the 10 stations annual rainfall data is as follows

$$Y_{(19 \times 1)} = X_{(19 \times 10)} B_{(10 \times 1)}$$
$$\hat{B} = (X'X)^{-1} X'Y$$

The multiple linear regression equation is as follows.

$$Y = 782.4 + 0.1861 x_1 + 0.0484 x_2 - 0.0198 x_3 + 0.0019 x_4 + 0.1196 x_5 + 0.1555 x_6 + 0.0232 x_7 + 0.1948 x_8 + 0.0799 x_9 - 0.0041 x_{10}$$


10

So, when we do that you get for this example, R square as point nine eight eight. So, this is this would have been nice and acceptable provided the size was not big by ten variables and provided we were sure that all of these are uncorrelated. For example, there is no correlation between x_1 and x_4 and so on. So, all of these, if they were uncorrelated and the size was not as large as this then it would have been acceptable, but we would like to express this regression in terms of the principle components, where we derive a set of variables which are all uncorrelated and then we should be able to choose less than ten number of variables may be six may be three may be two and so on.

(Refer Slide Time: 27:32)

Example – 1 (Contd.)

Using this regression equation, $R^2 = \frac{B'XY - n\bar{y}^2}{Y'Y - n\bar{y}^2}$

$R^2 = 0.988$

PCA is performed on the data to reduce the size of the problem and to account for correlations among the rainfall values at 10 stations.

The annual rainfall data is standardized and the observed basin annual yield is centered.

$x_j = \frac{(Y_j - \bar{y}_j)}{s_j}$ $y_i = Y_i - \bar{y}$

MPTEL

So, that is what we will try to do, try to achieve through the principal component analysis. So, what we do is, we perform the principal component analysis now on the independent variable that is $x_1, x_2, x_3, \dots, x_{10}$. We will do the principal component analysis on that then regress the y which is the dependent variable on the principal component analysis. So, the aim as I mentioned is to reduce the size of the problem and to account for correlations among the rainfall values at the ten stations. What do I mean by account for it is not as, if we are calculating the correlations and putting it into the regression, no we want to convert the original set of independent variables which are all correlated among themselves to another set of variables which are uncorrelated among themselves and that is what we mean by account for the correlation among the independent variable and also do not get confused between the usage of words independent variable.

They are independent in as much as they do not depend on the dependent variable whereas, the dependent variable depends on these variables and therefore, they are called as independent variables, but they may not be independent among themselves and that is what leads to multicollinearity, as I mentioned the multicollinearity by multicollinearity, I mean some of these variables that we are using in the regression are correlated with each other to do that, then we have ten stations and we have observations going from $i = 1$ to $i = 19$ in this particular case.

(Refer Slide Time: 22:40)

The annual rainfall in mm for 10 stations and observed basin annual yield (Y) in mm

Year	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	Y
1979	1948	4177	5496	2922	5713	3640	3203	2739	2167	2299	3255.2
1980	2261	3670	7797	3327	6934	4424	3692	3451	2866	2653	3682.7
1981	1989	4353	7392	2837	6275	4827	4476	4403	3568	3241	3921.9
1982	1999	3307	7061	3439	6641	4815	4256	4129	3447	3046	3909.3
1983	2086	4230	6564	2987	6675	3959	3900	3559	4078	3583	3768.9
1984	1717	2714	5919	3394	5605	3648	3085	2440	2631	2587	3106.4
1985	1383	2357	5053	2958	5144	3106	4052	3006	3049	2890	3069.4
1986	1470	3004	3951	2691	5116	3557	2775	1909	1952	1723	2940.2
1987	1350	2446	4280	2397	4722	3556	2818	2945	2931	2733	3015.3
1988	1602	4188	5910	3619	6869	5142	3190	3660	3964	3107	3953.2
1989	1417	3631	5145	3282	5226	3793	2663	3017	2579	3367	3172.4
1990	1662	4683	6384	6376	7313	4679	3037	3666	3142	2621	3791.0
1991	1955	4553	5679	6141	6068	3651	2601	2791	2148	2448	3344.8
1992	1974	3836	6021	5646	5876	4026	3037	3920	2583	2742	3650.3
1993	2094	4183	6733	6720	6044	6573	2465	3406	2410	2539	3878.7
1994	3149	6128	8151	9048	8384	7467	2888	3522	2496	2895	4606.2
1995	1471	2952	4151	4975	5149	4733	2603	3493	3396	3554	3498.8
1996	1691	3711	4200	4962	5359	3782	3185	3099	3381	2938	3241.0
1997	2373	4836	6704	6563	6197	5001	3902	3685	3636	3365	4013.5

(Refer Slide Time: 27:32)


Example – 1 (Contd.)

Using this regression equation, $R^2 = \frac{B_1 Y'Y}{Y'Y - n\bar{y}^2}$

$R^2 = 0.988$

PCA is performed on the data to reduce the size of the problem and to account for correlations among the rainfall values at 10 stations.

The annual rainfall data is standardized and the observed basin annual yield is centered.

$$x_j = \frac{(X_{ij} - \bar{x}_j)}{s_j} \quad y_i = Y_i - \bar{y}$$


So, nineteen years of data at ten stations. So, that data x_{ij} we deduct \bar{x}_j at the j th stations you take out the mean of that j th station data and divide it by the standard deviation of the j th station data.

(Refer Slide Time: 22:40)

The annual rainfall in mm for 10 stations and observed basin annual yield (Y) in mm

Year	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Y
1979	1948	4177	5496	2922	5713	3640	3203	2739	2167	2299	3255.2
1980	2261	3670	7797	3327	6934	4424	3692	3451	2866	2653	3682.7
1981	1989	4353	7392	2837	6275	4827	4476	4403	3568	3241	3921.9
1982	1999	3307	7061	3439	6641	4815	4256	4129	3447	3046	3909.3
1983	2086	4230	6564	2987	6675	3959	3900	3559	4078	3583	3768.9
1984	1717	2714	5919	3394	5605	3648	3085	2440	2631	2587	3106.4
1985	1383	2357	5053	2958	5144	3106	4052	3006	3049	2890	3069.4
1986	1470	3004	3951	2691	5116	3557	2775	1909	1952	1723	2940.2
1987	1350	2446	4280	2397	4722	3556	2818	2945	2931	2733	3015.3
1988	1602	4188	5910	3619	6869	5142	3190	3660	3964	3107	3953.2
1989	1417	3631	5145	3282	5226	3793	2663	3017	2579	3367	3172.4
1990	1662	4683	6384	6376	7313	4679	3037	3666	3142	2621	3791.0
1991	1955	4553	5679	6141	6068	3651	2601	2791	2148	2448	3344.8
1992	1974	3836	6021	5646	5876	4026	3037	3920	2583	2742	3650.3
1993	2094	4183	6733	6720	6044	6573	2465	3406	2410	2539	3878.7
1994	3149	6128	8151	9048	8384	7467	2888	3522	2496	2895	4606.2
1995	1471	2952	4151	4975	5149	4733	2603	3493	3396	3554	3498.8
1996	1691	3711	4200	4962	5359	3782	3185	3099	3381	2938	3241.0
1997	2373	4836	6704	6563	6197	5001	3902	3685	3636	3365	4013.5

That is, you take any station the mean of this and the standard deviation of this you use to standardize the data at station number six that is what we do here and so we obtain corresponding to each of the station. We obtain the n values nineteen values which are all standardize values. And then that dependent variable y, we simply center it we take the mean of this and then y minus y bar. So, we center all of this and obtain the dependent variable data.

(Refer Slide Time: 31:46)

Example – 1 (Contd.)

$$x_y = \frac{(Y_j - \bar{Y}_j)}{s_j}$$

$$y_i = Y_i - \bar{Y}$$

Station	Mean	Std. dev.
x_1	1873.2	434.3
x_2	3839.9	927.5
x_3	5925.8	1250.1
x_4	4436.0	1846.3
x_5	6068.9	914.9
x_6	4441.0	1091.3
x_7	3254.0	608.8
x_8	3307.3	599.4
x_9	2969.7	621.6
x_{10}	2859.6	462.4
Y	3569	-

The mean of all of these stations are given here and the standard deviation and of the dependent variable the mean is given here the standard deviation is not necessary for y. So, we will use these means at this particular location j is the station. So, j goes from one to ten here and i is the year of observation. So, you use for the j eth station you use x_j and s_j standard deviation. Similarly, for y you use this mean 3569 as \bar{y} and you get the standardize annual rainfall and centered observed basin annually.

(Refer Slide Time: 22:40)

The annual rainfall in mm for 10 stations and observed basin annual yield (Y) in mm

Year	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Y
1979	1948	4177	5496	2922	5713	3640	3203	2739	2167	2299	3255.2
1980	2261	3670	7797	3327	6934	4424	3692	3451	2866	2653	3682.7
1981	1989	4353	7392	2837	6275	4827	4476	4403	3568	3241	3921.9
1982	1999	3307	7061	3439	6641	4815	4256	4129	3447	3046	3909.3
1983	2086	4230	6564	2987	6675	3959	3900	3559	4078	3583	3768.9
1984	1717	2714	5919	3394	5605	3648	3085	2440	2631	2587	3106.4
1985	1383	2357	5053	2958	5144	3106	4052	3006	3049	2890	3069.4
1986	1470	3004	3951	2691	5116	3557	2775	1909	1952	1723	2940.2
1987	1350	2446	4280	2397	4722	3558	2818	2945	2931	2733	3015.3
1988	1602	4188	5910	3619	6869	5142	3190	3660	3964	3107	3953.2
1989	1417	3631	5145	3282	5226	3793	2663	3017	2579	3367	3172.4
1990	1662	4683	6384	6376	7313	4679	3037	3666	3142	2621	3791.0
1991	1955	4553	5679	6141	6068	3651	2601	2791	2148	2448	3344.8
1992	1974	3836	6021	5646	5876	4026	3037	3920	2583	2742	3650.3
1993	2094	4183	6733	6720	6044	6573	2465	3406	2410	2539	3878.7
1994	3149	6128	8151	9048	8384	7467	2888	3522	2496	2895	4606.2
1995	1471	2952	4151	4975	5149	4733	2603	3493	3396	3554	3498.8
1996	1691	3711	4200	4962	5359	3782	3185	3099	3381	2938	3241.0
1997	2373	4836	6704	6563	6197	5001	3902	3685	3636	3365	4013.5

So, this is centered values and these are standardized values. So, x_1 to x_{10} you get the standardized values and similarly these are the dependent variables, centered values now you focus on this sets of values x_1 to x_{10} you have 19 values associated with each of them.


(Refer Slide Time: 33:12)

Example – 1 (Contd.)

The covariance matrix for the standardized data

matrix

$$\text{cov}(X_1, X_2) = s_{x_1, x_2} = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n-1}$$

$$S = \begin{bmatrix} 1 & 0.79 & 0.81 & 0.64 & 0.78 & 0.71 & 0.18 & 0.38 & -0.03 & 0.08 \\ 0.79 & 1 & 0.65 & 0.72 & 0.80 & 0.68 & -0.02 & 0.40 & 0.03 & 0.12 \\ 0.81 & 0.65 & 1 & 0.38 & 0.84 & 0.64 & 0.44 & 0.61 & 0.17 & 0.19 \\ 0.64 & 0.72 & 0.38 & 1 & 0.55 & 0.71 & -0.35 & 0.25 & -0.15 & 0.04 \\ 0.78 & 0.80 & 0.84 & 0.55 & 1 & 0.70 & 0.21 & 0.51 & 0.21 & 0.13 \\ 0.71 & 0.68 & 0.64 & 0.71 & 0.70 & 1 & -0.10 & 0.48 & 0.08 & 0.18 \\ 0.18 & -0.02 & 0.44 & -0.35 & 0.21 & -0.10 & 1 & 0.52 & 0.59 & 0.37 \\ 0.38 & 0.40 & 0.61 & 0.25 & 0.51 & 0.48 & 0.52 & 1 & 0.64 & 0.64 \\ -0.03 & 0.03 & 0.17 & -0.15 & 0.21 & 0.08 & 0.59 & 0.64 & 1 & 0.79 \\ 0.08 & 0.12 & 0.19 & 0.04 & 0.13 & 0.18 & 0.37 & 0.64 & 0.79 & 1 \end{bmatrix}_{10 \times 10}$$


We will get the covariance we will obtain the covariance matrix for these ten variables in preparation for our co principal component analysis S p. How do we obtain the covariance let say, these are the variables. I write them as let say x 1, x 2 etcetera x 10. Similarly, x 1 x 2 etcetera up to x 10. So, these are the variables. So, x 1 to x 1 that is a covariance and x 1 to x 2 covariance x 1 to x 3 and this is what we obtain here on the standardized variables. Remember we are talking about the standardized variables and that is why you get the covariance x 1 x 1 as 1 itself and this will be all diagonal elements will be one here. So, this covariance is x 1 to x 2 let say I want to compute the covariance between x 1 and x 2 it is given by x 1 i i is the first first variable and for the i eth period i is equal to 1 to n minus x 1 bar x 2 i minus x 2 bar etcetera by n minus 1 n is nineteen in this case.


So, like this you form a ten by ten matrix. So, x 1 to x 10 similarly, x 10 to x 10. So, you can get a ten by ten matrix which is a square matrix and for this square matrix. We now obtain the eigenvalues and the eigen vectors for obtaining the eigen vectors first you have to get the eigenvalues. So, let us get the eigenvalue, if I denote this as the matrix s the covariance matrix we denote it as matrix s here and therefore, I should be able to get the eigen vectors and the eigenvalues corresponding to the matrix s.

(Refer Slide Time: 35:16)

Example – 1 (Contd.)

The eigenvalues and eigenvectors for the covariance matrix

Eigenvalues $ S - \lambda I = 0$									
4.945	2.631	1.047	0.364	0.307	0.257	0.205	0.140	0.063	0.042
Eigenvectors $(S - \lambda I)X = 0$									
0.390	-0.165	0.211	-0.191	0.451	-0.304	0.149	-0.043	-0.644	-0.079
0.381	-0.188	-0.053	-0.543	-0.127	0.215	-0.265	-0.574	0.189	0.157
0.393	0.029	0.382	0.235	0.074	-0.128	-0.328	0.319	0.227	0.600
0.298	-0.321	-0.390	-0.111	0.246	0.400	0.425	0.437	0.210	0.089
0.404	-0.065	0.179	-0.121	-0.589	-0.056	-0.093	0.393	-0.013	-0.522
0.371	-0.161	-0.229	0.546	-0.116	-0.394	0.301	-0.402	0.241	-0.087
0.122	0.462	0.521	-0.117	0.237	0.148	0.428	-0.136	0.393	-0.229
0.317	0.338	-0.122	0.444	0.069	0.603	-0.241	-0.134	-0.333	-0.135
0.136	0.529	-0.237	-0.201	-0.412	-0.110	0.388	0.031	-0.275	0.443
0.160	0.443	-0.477	-0.192	0.351	-0.358	-0.358	0.155	0.235	-0.234



How do I get that? So, the eigenvalues are obtained by determinant S minus λI is equal to 0, I set it as determinant S minus λI is equal to 0 and then obtain λ s here S is a ten by ten matrix and therefore, I will get ten λ values here. So, the eigenvalues this is obtained from the matlab routine, we get eigenvalues as four point nine five nine four five and. So, on. So, there are ten eigenvalues that are obtained for the covariance matrix. We use these eigen values and associated with each of the eigen values we get one eigen vectors, like that we get ten eigen vectors we use this S minus λI into X is equal to 0 this is a matrix and this is a matrix. So, we obtain the eigen vectors now this is eigen vector number one, eigen vector number 2, etcetera. So, this eigenvector is associated with this λ one this eigenvector is associated with this eigenvalue λ 2 and so on.

So, associated with each of the eigen vector there is a eigen value. As I mentioned, they come in pairs there is an eigen value and there is associated eigen vector. So, this is what we obtain for the covariance matrix, then we use these eigen vectors and look at the percentage variance explained by each of the eigen vectors there are ten such principle components, eigen vectors consisting of the principal components here. Let us not get confused this is principle component number, one which is the eigenvector number one principle component number two number three and. So, on there are ten such values here, now we will see how much of percentage variance is explained by the first eigen vector the second eigen vectors third eigenvector and so on. To do that we look at the

eigen values associated eigen values. Let say, this is lambda one lambda two and so on. So, I will get the percentage variance explained by this eigenvector by using the corresponding eigenvalue.


(Refer Slide Time: 37:52)

Example – 1 (Contd.)

The eigenvalues and % variance explained: $\frac{\lambda_j}{\text{Trace}(S)}$

Eigenvalues	% variance explained
4.945	49.447
2.631	26.310
1.047	10.470
0.364	3.641
0.307	3.069
0.257	2.565
0.205	2.047
0.140	1.399
0.063	0.629
0.042	0.423

} > 95% variance explained by first 6 principal components



So, I will get lambda j by trace S where trace s is simply the summation of all the eigenvalues that is S in this case, will be this plus this plus this etcetera. So, I can write trace S if you recall from your last lecture, as we will write trace as is equal to simply lambda j over j all the eigen values summation of all the eigen values. And we get the percentage variance explained by a particular eigen vector y using this on the associated lambda j. So, this says that the first eigenvector explains 49.447 second one explains 26.310 and so on. This is arranged in descending order. So, this is how we obtain the percentage explained by this various eigen vectors.

Now, look at this, if we take the first six components that is 49.447 plus 26.310 and so on. If you take first six about 95 percent of the variance is explained by the first six component. So, we have the option of developing the regression only using the first six because we are satisfied that 95 percent of the variance is explained by this components. So, I can ignore the remaining four, if you want to include all the hundred percent of variance then all the ten principle components have to be included. So, let us look at first the six components, that is, I use only the first six eigen vectors and develop the relationship of the dependent variable on these six principle components, essentially this

is the advantage you know when we are using the principle components. We can see how much of percentage variance or how much of information content is in fact captured by using these six eigen vectors is the question that we are trying to answer. So, in this particular case, we are saying that 95 percent of the information can be captured in terms of the variance. In terms of the variance, we can capture about 95 percent of the information by using these six component and therefore, it is not necessary or you can afford to ignore the remaining four components that is the interpretation here. So, we will use these six components and then regress the dependent variable on the six eigen vectors now or six principle components. So, to say.


(Refer Slide Time: 40:47)

Example – 1 (Contd.)

First six components are considered in the analysis
and the modified data is obtained as $Z = XA$

$Z =$	<table style="width: 100%; border-collapse: collapse;"> <tr><td>0.17</td><td>0.36</td><td>-0.34</td><td>-0.82</td><td>-0.39</td><td>-0.73</td><td>-0.08</td><td>-0.95</td><td>-1.29</td><td>-1.21</td></tr> <tr><td>0.89</td><td>-0.18</td><td>1.50</td><td>-0.60</td><td>0.95</td><td>-0.02</td><td>0.72</td><td>0.24</td><td>-0.17</td><td>-0.45</td></tr> <tr><td>0.27</td><td>0.55</td><td>1.17</td><td>-0.87</td><td>0.23</td><td>0.35</td><td>2.01</td><td>1.83</td><td>0.96</td><td>0.82</td></tr> <tr><td>0.29</td><td>-0.57</td><td>0.91</td><td>-0.54</td><td>0.63</td><td>0.34</td><td>1.64</td><td>1.37</td><td>0.77</td><td>0.40</td></tr> <tr><td>0.49</td><td>0.42</td><td>0.51</td><td>-0.78</td><td>0.66</td><td>-0.44</td><td>1.08</td><td>0.42</td><td>1.78</td><td>1.57</td></tr> <tr><td>-0.36</td><td>-1.21</td><td>-0.01</td><td>-0.58</td><td>-0.51</td><td>-0.73</td><td>-0.28</td><td>-1.45</td><td>-0.54</td><td>-0.59</td></tr> <tr><td>-1.13</td><td>-1.60</td><td>-0.70</td><td>-0.80</td><td>-1.01</td><td>-1.22</td><td>1.31</td><td>-0.50</td><td>0.13</td><td>0.07</td></tr> <tr><td>-0.93</td><td>-0.90</td><td>-1.58</td><td>-0.95</td><td>-1.04</td><td>-0.81</td><td>-0.79</td><td>-2.33</td><td>-1.64</td><td>-2.48</td></tr> <tr><td>-1.20</td><td>-1.50</td><td>-1.32</td><td>-1.10</td><td>-1.47</td><td>-0.81</td><td>-0.72</td><td>-0.61</td><td>-0.06</td><td>-0.27</td></tr> <tr><td>-0.62</td><td>0.38</td><td>-0.01</td><td>-0.44</td><td>0.87</td><td>0.64</td><td>-0.11</td><td>0.59</td><td>1.60</td><td>0.53</td></tr> <tr><td>-1.05</td><td>-0.23</td><td>-0.82</td><td>-0.83</td><td>-0.92</td><td>-0.59</td><td>-0.97</td><td>-0.48</td><td>-0.63</td><td>1.10</td></tr> <tr><td>-0.49</td><td>0.91</td><td>0.37</td><td>1.05</td><td>1.36</td><td>0.22</td><td>-0.36</td><td>0.60</td><td>0.28</td><td>-0.52</td></tr> <tr><td>0.19</td><td>0.77</td><td>-0.20</td><td>0.92</td><td>0.00</td><td>-0.72</td><td>-1.07</td><td>-0.86</td><td>-1.32</td><td>-0.89</td></tr> <tr><td>0.23</td><td>0.00</td><td>0.08</td><td>0.66</td><td>-0.21</td><td>-0.38</td><td>-0.36</td><td>1.02</td><td>-0.62</td><td>-0.28</td></tr> <tr><td>0.51</td><td>0.37</td><td>0.65</td><td>1.24</td><td>-0.03</td><td>1.95</td><td>-1.30</td><td>0.16</td><td>-0.90</td><td>-0.69</td></tr> <tr><td>2.94</td><td>2.47</td><td>1.78</td><td>2.50</td><td>2.53</td><td>2.77</td><td>-0.60</td><td>0.36</td><td>-0.76</td><td>0.08</td></tr> <tr><td>-0.93</td><td>-0.96</td><td>-1.42</td><td>0.29</td><td>-1.01</td><td>0.27</td><td>-1.07</td><td>0.31</td><td>0.89</td><td>1.50</td></tr> <tr><td>-0.42</td><td>-0.14</td><td>-1.38</td><td>0.28</td><td>-0.78</td><td>-0.60</td><td>-0.11</td><td>-0.35</td><td>0.66</td><td>0.17</td></tr> <tr><td>1.15</td><td>1.07</td><td>0.62</td><td>1.15</td><td>0.14</td><td>0.51</td><td>1.08</td><td>0.63</td><td>1.07</td><td>1.09</td></tr> </table>	0.17	0.36	-0.34	-0.82	-0.39	-0.73	-0.08	-0.95	-1.29	-1.21	0.89	-0.18	1.50	-0.60	0.95	-0.02	0.72	0.24	-0.17	-0.45	0.27	0.55	1.17	-0.87	0.23	0.35	2.01	1.83	0.96	0.82	0.29	-0.57	0.91	-0.54	0.63	0.34	1.64	1.37	0.77	0.40	0.49	0.42	0.51	-0.78	0.66	-0.44	1.08	0.42	1.78	1.57	-0.36	-1.21	-0.01	-0.58	-0.51	-0.73	-0.28	-1.45	-0.54	-0.59	-1.13	-1.60	-0.70	-0.80	-1.01	-1.22	1.31	-0.50	0.13	0.07	-0.93	-0.90	-1.58	-0.95	-1.04	-0.81	-0.79	-2.33	-1.64	-2.48	-1.20	-1.50	-1.32	-1.10	-1.47	-0.81	-0.72	-0.61	-0.06	-0.27	-0.62	0.38	-0.01	-0.44	0.87	0.64	-0.11	0.59	1.60	0.53	-1.05	-0.23	-0.82	-0.83	-0.92	-0.59	-0.97	-0.48	-0.63	1.10	-0.49	0.91	0.37	1.05	1.36	0.22	-0.36	0.60	0.28	-0.52	0.19	0.77	-0.20	0.92	0.00	-0.72	-1.07	-0.86	-1.32	-0.89	0.23	0.00	0.08	0.66	-0.21	-0.38	-0.36	1.02	-0.62	-0.28	0.51	0.37	0.65	1.24	-0.03	1.95	-1.30	0.16	-0.90	-0.69	2.94	2.47	1.78	2.50	2.53	2.77	-0.60	0.36	-0.76	0.08	-0.93	-0.96	-1.42	0.29	-1.01	0.27	-1.07	0.31	0.89	1.50	-0.42	-0.14	-1.38	0.28	-0.78	-0.60	-0.11	-0.35	0.66	0.17	1.15	1.07	0.62	1.15	0.14	0.51	1.08	0.63	1.07	1.09	$\left[\begin{array}{l} 0.390 \\ 0.381 \\ 0.383 \\ 0.298 \\ 0.404 \\ 0.371 \\ 0.122 \\ 0.317 \\ 0.136 \\ 0.160 \end{array} \right]$	<table style="width: 100%; border-collapse: collapse;"> <tr><td>-0.165</td><td>0.211</td><td>-0.191</td><td>0.451</td><td>-0.304</td></tr> <tr><td>-0.188</td><td>-0.053</td><td>-0.543</td><td>-0.127</td><td>0.215</td></tr> <tr><td>0.029</td><td>0.382</td><td>0.235</td><td>0.074</td><td>-0.128</td></tr> <tr><td>-0.321</td><td>-0.390</td><td>-0.111</td><td>0.246</td><td>0.400</td></tr> <tr><td>-0.085</td><td>0.179</td><td>-0.121</td><td>-0.589</td><td>-0.056</td></tr> <tr><td>-0.161</td><td>-0.229</td><td>0.546</td><td>-0.116</td><td>-0.394</td></tr> <tr><td>0.462</td><td>0.521</td><td>-0.117</td><td>0.237</td><td>0.148</td></tr> <tr><td>0.338</td><td>-0.122</td><td>0.444</td><td>0.069</td><td>0.603</td></tr> <tr><td>0.529</td><td>-0.237</td><td>-0.201</td><td>-0.412</td><td>-0.110</td></tr> <tr><td>0.443</td><td>-0.477</td><td>-0.192</td><td>0.351</td><td>-0.358</td></tr> </table>	-0.165	0.211	-0.191	0.451	-0.304	-0.188	-0.053	-0.543	-0.127	0.215	0.029	0.382	0.235	0.074	-0.128	-0.321	-0.390	-0.111	0.246	0.400	-0.085	0.179	-0.121	-0.589	-0.056	-0.161	-0.229	0.546	-0.116	-0.394	0.462	0.521	-0.117	0.237	0.148	0.338	-0.122	0.444	0.069	0.603	0.529	-0.237	-0.201	-0.412	-0.110	0.443	-0.477	-0.192	0.351	-0.358
0.17	0.36	-0.34	-0.82	-0.39	-0.73	-0.08	-0.95	-1.29	-1.21																																																																																																																																																																																																																																										
0.89	-0.18	1.50	-0.60	0.95	-0.02	0.72	0.24	-0.17	-0.45																																																																																																																																																																																																																																										
0.27	0.55	1.17	-0.87	0.23	0.35	2.01	1.83	0.96	0.82																																																																																																																																																																																																																																										
0.29	-0.57	0.91	-0.54	0.63	0.34	1.64	1.37	0.77	0.40																																																																																																																																																																																																																																										
0.49	0.42	0.51	-0.78	0.66	-0.44	1.08	0.42	1.78	1.57																																																																																																																																																																																																																																										
-0.36	-1.21	-0.01	-0.58	-0.51	-0.73	-0.28	-1.45	-0.54	-0.59																																																																																																																																																																																																																																										
-1.13	-1.60	-0.70	-0.80	-1.01	-1.22	1.31	-0.50	0.13	0.07																																																																																																																																																																																																																																										
-0.93	-0.90	-1.58	-0.95	-1.04	-0.81	-0.79	-2.33	-1.64	-2.48																																																																																																																																																																																																																																										
-1.20	-1.50	-1.32	-1.10	-1.47	-0.81	-0.72	-0.61	-0.06	-0.27																																																																																																																																																																																																																																										
-0.62	0.38	-0.01	-0.44	0.87	0.64	-0.11	0.59	1.60	0.53																																																																																																																																																																																																																																										
-1.05	-0.23	-0.82	-0.83	-0.92	-0.59	-0.97	-0.48	-0.63	1.10																																																																																																																																																																																																																																										
-0.49	0.91	0.37	1.05	1.36	0.22	-0.36	0.60	0.28	-0.52																																																																																																																																																																																																																																										
0.19	0.77	-0.20	0.92	0.00	-0.72	-1.07	-0.86	-1.32	-0.89																																																																																																																																																																																																																																										
0.23	0.00	0.08	0.66	-0.21	-0.38	-0.36	1.02	-0.62	-0.28																																																																																																																																																																																																																																										
0.51	0.37	0.65	1.24	-0.03	1.95	-1.30	0.16	-0.90	-0.69																																																																																																																																																																																																																																										
2.94	2.47	1.78	2.50	2.53	2.77	-0.60	0.36	-0.76	0.08																																																																																																																																																																																																																																										
-0.93	-0.96	-1.42	0.29	-1.01	0.27	-1.07	0.31	0.89	1.50																																																																																																																																																																																																																																										
-0.42	-0.14	-1.38	0.28	-0.78	-0.60	-0.11	-0.35	0.66	0.17																																																																																																																																																																																																																																										
1.15	1.07	0.62	1.15	0.14	0.51	1.08	0.63	1.07	1.09																																																																																																																																																																																																																																										
-0.165	0.211	-0.191	0.451	-0.304																																																																																																																																																																																																																																															
-0.188	-0.053	-0.543	-0.127	0.215																																																																																																																																																																																																																																															
0.029	0.382	0.235	0.074	-0.128																																																																																																																																																																																																																																															
-0.321	-0.390	-0.111	0.246	0.400																																																																																																																																																																																																																																															
-0.085	0.179	-0.121	-0.589	-0.056																																																																																																																																																																																																																																															
-0.161	-0.229	0.546	-0.116	-0.394																																																																																																																																																																																																																																															
0.462	0.521	-0.117	0.237	0.148																																																																																																																																																																																																																																															
0.338	-0.122	0.444	0.069	0.603																																																																																																																																																																																																																																															
0.529	-0.237	-0.201	-0.412	-0.110																																																																																																																																																																																																																																															
0.443	-0.477	-0.192	0.351	-0.358																																																																																																																																																																																																																																															

19 x 10 10 x 6




(Refer Slide Time: 35:16)

Example – 1 (Contd.)

The eigenvalues and eigenvectors for the covariance matrix

Eigenvalues $ S - \lambda I = 0$									
4.945	2.631	1.047	0.364	0.307	0.257	0.205	0.140	0.063	0.042
Eigenvectors $(S - \lambda I).X = 0$									
0.390	-0.165	0.211	-0.191	0.451	-0.304	0.149	-0.043	-0.644	-0.079
0.381	-0.188	-0.053	-0.543	-0.127	0.215	-0.265	-0.574	0.189	0.157
0.393	0.029	0.382	0.235	0.074	-0.128	-0.328	0.319	0.227	0.600
0.298	-0.321	-0.390	-0.111	0.246	0.400	0.425	0.437	0.210	0.089
0.404	-0.065	0.179	-0.121	-0.589	-0.056	-0.093	0.393	-0.013	-0.522
0.371	-0.161	-0.229	0.546	-0.116	-0.394	0.301	-0.402	0.241	-0.087
0.122	0.462	0.521	-0.117	0.237	0.148	0.428	-0.136	0.393	-0.229
0.317	0.338	-0.122	0.444	0.069	0.603	-0.241	-0.134	-0.333	-0.135
0.136	0.529	-0.237	-0.201	-0.412	-0.110	0.388	0.031	-0.275	0.443
0.160	0.443	-0.477	-0.192	0.351	-0.358	-0.358	0.155	0.235	-0.234



So, what we are doing now Z is equal to $X A$ is, what we are saying and A is the principle components and we are using only the six principle components. So, A is a vector of ten into six there are six principle components look at this now, the first column is the first eigenvector which is the first principle component second principle component third principle component etcetera look at the eigen vectors, here this is the first principle component, second principle component up to six we go one, two, three, four, five, six and these are the six principle components we have considered. And these ten are associated with the ten variables. So, you had initial ten variables. So, you have got ten values corresponding to that remember eigen vectors this dimension is p by p which is ten by ten out of that we are taking up to six therefore, we are taking ten by six matrix here from this and then using the principle using the regression relationship and this is your original X value by original X value. I mean it is a transformed X value standardize X values. So, this comes from the standardized X values here.

(Refer Slide Time: 40:47)

Example – 1 (Contd.)

First six components are considered in the analysis
and the modified data is obtained as $Z = X A$

$Z =$

0.17	0.35	-0.34	-0.82	-0.39	-0.73	-0.08	-0.95	-1.29	-1.21
0.89	-0.18	1.50	-0.60	0.95	-0.02	0.72	0.24	-0.17	-0.45
0.27	0.55	1.17	-0.87	0.23	0.35	2.01	1.83	0.96	0.82
0.29	-0.57	0.91	-0.54	0.63	0.34	1.64	1.37	0.77	0.40
0.49	0.42	0.51	-0.78	0.66	-0.44	1.08	0.42	1.78	1.57
-0.36	-1.21	-0.01	-0.58	-0.51	-0.73	-0.28	-1.45	-0.54	-0.59
-1.13	-1.60	-0.70	-0.80	-1.01	-1.22	1.31	-0.50	0.13	0.07
-0.93	-0.90	-1.58	-0.95	-1.04	-0.81	-0.79	-2.33	-1.64	-2.48
-1.20	-1.50	-1.32	-1.10	-1.47	-0.81	-0.72	-0.61	-0.06	-0.27
-0.62	0.38	-0.01	-0.44	0.87	0.64	-0.11	0.59	1.60	0.53
-1.05	-0.23	-0.82	-0.83	-0.92	-0.59	-0.97	-0.48	-0.83	1.10
-0.49	0.91	0.37	1.05	1.36	0.22	-0.38	0.60	0.28	-0.52
0.19	0.77	-0.20	0.92	0.00	-0.72	-1.07	-0.88	-1.32	-0.89
0.23	0.00	0.08	0.66	-0.21	-0.38	-0.36	1.02	-0.62	-0.28
0.51	0.37	0.85	1.24	-0.03	1.95	-1.30	0.16	-0.90	-0.69
2.94	2.47	1.78	2.50	2.53	2.77	-0.60	0.38	-0.78	0.08
-0.93	-0.96	-1.42	0.29	-1.01	0.27	-1.07	0.31	0.89	1.50
-0.42	-0.14	-1.38	0.28	-0.78	-0.60	-0.11	-0.35	0.66	0.17
1.15	1.07	0.62	1.15	1.14	0.51	1.08	0.63	1.07	1.00

0.390	-0.165	0.211	-0.191	0.451	-0.304
0.381	-0.188	-0.053	-0.543	-0.127	0.215
0.033	0.029	0.382	0.235	0.074	-0.128
0.298	-0.321	-0.390	-0.111	0.246	0.400
0.404	-0.085	0.179	-0.121	-0.589	-0.058
0.371	-0.161	-0.229	0.546	-0.116	-0.394
0.122	0.462	0.521	-0.117	0.237	0.148
0.317	0.338	-0.122	0.444	0.069	0.603
0.136	0.529	-0.237	-0.201	-0.412	-0.110
0.160	0.443	-0.477	-0.192	0.351	-0.358

19×10 10×6

This is point one seven two minus one point two one. So, we use this entire data here, after transforming after standardizing and then put the regression relationship. So, this is point one seven two minus one point two nine. So, this is your nineteen by ten values the matrix is nineteen by ten and this is ten by six. So, you obtain a matrix of nineteen by six, for Z that will be n by p and now this p is restricted to six now.

(Refer Slide Time: 42:54)

Example – 1 (Contd.)

$Z = X A =$

-1.283	-1.278	1.259	-0.492	0.139	0.045
1.133	0.192	1.776	0.367	-0.068	-0.381
1.825	2.512	0.974	0.410	0.245	0.386
1.273	1.972	0.998	0.826	0.040	0.127
1.176	2.403	0.134	-1.034	-0.232	-0.649
-1.907	-0.547	0.724	0.067	0.087	-0.705
-2.395	1.519	0.673	0.048	0.429	0.160
-3.779	-2.327	1.050	-0.058	-0.483	-0.172
-3.115	0.333	-0.480	0.475	-0.042	-0.194
0.830	1.247	-0.735	0.055	-1.484	-0.236
-1.700	0.094	-1.054	-0.154	0.348	-0.373
1.345	-0.585	-0.305	-0.126	-1.214	1.016
-0.430	-2.241	0.012	-0.817	0.203	0.574
0.243	-0.433	-0.169	0.391	0.593	1.067
1.336	-2.171	-0.751	1.326	0.157	-0.178
5.527	-2.835	-0.199	-0.183	0.169	-0.643
-1.202	0.857	-2.516	0.433	0.246	-0.259
-1.219	0.366	-0.975	-0.743	0.059	0.311
2.341	0.921	-0.416	-0.792	0.808	0.081

19×6

So, this is the matrix Z, we get this is the transformed data as after I do the multiplication X Y A, when I do this, I get the matrix Z as nineteen by six now, this one has to be

regressed on the dependent variable Y. What did we do now, we use the principle component analysis and transform the original data into Z now we use this Z to regress on Y, Y will be regressed on this transformed data Z.

(Refer Slide Time: 43:35)

Example – 1 (Contd.)

Regression analysis is performed on these components:

$$Y = ZB$$

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

Y =



-314.2
113.3
352.5
339.9
199.5
-463.0
-500.0
-629.2
-554.2
383.7
-397.0
221.6
-224.6
80.9
309.3
1036.7
-70.6
-328.4
444.0

(Refer Slide Time: 42:54)

So, we write Y is equal to Z B, where Y is the dependent variable and Z is this transformed data now this matrix is obtained and B is the matrix, the vector of unknown parameters which we want to estimate. So, being a multiple regression, multiple linear regression we have seen how to estimate B. So, we call it as B cap is equal to Z dash Z inverse Z dash Y much the same way as we did for our multiple linear regression with the original variable except that we are now using the transformed data the transform data is obtain from transforming the standardized values, standardized vector of standardized matrix of standardized values of the independent variables. And multiplying that matrix with the principle component and we have used six principle components, associated with the six eigen vectors which explain about 95 percent of the variance in Y. So, we know Z therefore, we can get Z dash Z inverse and Z dash is obtained and we know Y, Y is the centered values of the the vector of the centered values of the dependent variable. So, we know Y and therefore, we can get B dash.

(Refer Slide Time: 45:24)

Example – 1 (Contd.)



$$(Z'Z)^{-1} = \begin{bmatrix} 89.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 47.36 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 18.85 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 6.55 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 5.52 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 4.62 \end{bmatrix}_{6 \times 6}$$
$$\hat{B} = (Z'Z)^{-1} Z'Y = \begin{bmatrix} 192.2 \\ 13.3 \\ -33.1 \\ 73.9 \\ -64.1 \\ -15.7 \end{bmatrix}_{6 \times 1}$$


Now, B dash is the estimated values of the parameters beta one, beta two, etcetera beta p, in this case we get six beta values. So, Z, Z dash Z inverse this matrix Z dash Z inverse. I will give you directly this is a six by six matrix and this is a symmetric matrix diagonal diagonal symmetric matrix and this is what you get here and using this we get B cap which is the estimated values for betas as six by one matrix as one ninety two point two and so on.

(Refer Slide Time: 46:06)

Example – 1 (Contd.)

The regression equation is

$$y = 192.1569 P_{c1} + 13.29536 P_{c2} - 33.1304 P_{c3} \\ + 73.92323 P_{c4} - 64.0569 P_{c5} - 15.6921 P_{c6}$$
$$R^2 = 0.978$$


So, this is your beta one this is beta two and so on. So, we have now estimated the coefficients for the regression relationship with beta one, beta two, etcetera beta six and we can write the regression relationship. Now, using this as if we do not have the intercept directly we use all the six species, we write this as one ninety two point one five six P c one, etcetera P c two, P c three and so on. This is your y.

(Refer Slide Time: 43:35)

Example - 1 (Contd.)

Regression analysis is performed on these components:

$$Y = ZB$$

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

Y =

-314.2
113.3
352.5
339.9
199.5
-463.0
-500.0
-629.2
-554.2
383.7
-397.0
221.6
-224.6
80.9
309.3
1036.7
-70.6
-328.4
444.0

(Refer Slide Time: 46:06)

And the associated R square value we obtain as point nine seven eight, which is slightly smaller than what we obtain earlier which was point nine eight or some such thing. Now, there are two important things, here important aspects we are writing Y is equal to Z B and that in the long form, we are writing this as y is equal to these are the beta one, beta two, beta three, etcetera values the P c one, P c two, etcetera are in fact, the eigen vectors that we have obtained these are the eigen vectors let say, we are looking at six values. I am sorry.

(Refer Slide Time: 40:52)

Example – 1 (Contd.)

First six components are considered in the analysis
and the modified data is obtained as $Z = XA$

$Z =$

0.17	0.36	-0.34	-0.82	-0.39	-0.73	-0.08	-0.95	-1.29	-1.21
0.89	-0.18	1.50	-0.60	0.95	-0.02	0.72	0.24	-0.17	-0.45
0.27	0.55	1.17	-0.87	0.23	0.35	2.01	1.83	0.96	0.82
0.29	-0.57	0.91	-0.54	0.63	0.34	1.64	1.37	0.77	0.40
0.49	0.42	0.51	-0.78	0.66	-0.44	1.08	0.42	1.78	1.57
-0.36	-1.21	-0.01	-0.56	-0.51	-0.73	-0.28	-1.45	-0.54	-0.59
-1.13	-1.60	-0.70	-0.80	-1.01	-1.22	1.31	-0.50	0.13	0.07
-0.93	-0.90	-1.58	-0.95	-1.04	-0.81	-0.79	-2.33	-1.64	-2.46
-1.20	-1.50	-1.32	-1.10	-1.47	-0.81	-0.72	-0.61	-0.06	-0.27
-0.62	0.38	-0.01	-0.44	0.87	0.64	-0.11	0.59	1.60	0.53
-1.05	-0.23	-0.82	-0.83	-0.92	-0.59	-0.97	-0.48	-0.83	1.10
-0.49	0.91	0.37	1.05	1.36	0.22	-0.36	0.60	0.28	-0.52
0.19	0.77	-0.20	0.92	0.00	-0.72	-1.07	-0.86	-1.32	-0.89
0.23	0.00	0.08	0.66	-0.21	-0.38	-0.36	1.02	-0.62	-0.28
0.51	0.37	0.85	1.24	-0.03	1.95	-1.30	0.16	-0.90	-0.69
2.94	2.47	1.78	2.50	2.53	2.77	-0.60	0.36	-0.76	0.08
-0.93	-0.96	-1.42	0.29	-1.01	0.27	-1.07	0.31	0.89	1.50
-0.42	-0.14	-1.38	0.28	-0.78	-0.60	-0.11	-0.35	0.66	0.17
1.15	1.07	0.62	1.15	0.14	0.51	1.08	0.63	1.07	1.00

0.390	-0.165	0.211	-0.191	0.451	-0.304
0.381	-0.188	-0.053	-0.543	-0.127	0.215
0.333	0.029	0.382	0.235	0.074	-0.128
0.299	-0.321	-0.390	-0.111	0.246	0.400
0.404	-0.085	0.179	-0.121	-0.589	-0.056
0.371	-0.161	-0.229	0.546	-0.116	-0.394
0.122	0.462	0.521	-0.117	0.237	0.148
0.317	0.338	-0.122	0.444	0.069	0.603
0.136	0.529	-0.237	-0.201	-0.412	-0.110
0.160	0.443	-0.477	-0.192	0.351	-0.358

19×10

10×6

(Refer Slide Time: 46:06)

So, let us look at the eigen vector there these are the eigen vectors. So, that is how we obtain regression relationship on Pc's that is the principle components, which are essentially the eigen vectors. Now, there are some interesting features on this particular regression equation you see you obtain beta one as one ninety two point one six beta two and this value and so on, by using six principle components. Because we said that 95 percent of the variance has to be explained by this model and therefore, I go up to six values let say that, I sacrifice some more information and not go up to 95 percent. But I will restrict myself to first three principle components alone. So, I may not be able to explain 95 percent, but may be slightly smaller than that let say 85 percent or 86 percent and soon and then discarding some other principle components, I redevelop the regression relationship.

Let say, out of this six, I redevelop discarding this three, now an interesting point here is that, when we do that these coefficients still remain the same this is because they are orthogonal to each other. Remember, the principle components are orthogonal to each other and that feature brings to the fore the fact that, if you discard some of the beta values or some of the principle components and then redo the regression your coefficients still will remain the same and this is quite interesting.


(Refer Slide Time: 37:52)

Example – 1 (Contd.)

The eigenvalues and % variance explained: $\frac{\lambda_j}{\text{Trace}(S)}$

Eigenvalues	% variance explained
4.945	49.447
2.631	26.310
1.047	10.470
0.364	3.641
0.307	3.069
0.257	2.565
0.205	2.047
0.140	1.399
0.063	0.629
0.042	0.423

} > 95% variance explained by first 6 principal components



Let us look at that, these are your percentage explain from your earlier table, we used these are the percentage explained we have arranged them in a descending order. In fact, in this particular it comes in the descending order. So, when we used 95 percent of the variance we went up to the sixth principle component. So, we used all the six principle components we use the same table now and look at only the first three principle components. So, the first three principle components explain 85 percent of the variance approximately let say that I am satisfied with using 85 percent of the information; that means, I want to reproduce 85 percent of the information in terms of the variance and therefore, I restrict my regression relationship only to these three principle components and develop the regression relationship, again much the same way as I did just a while ago using all the six principle components, I do this now on these three principle components.

(Refer Slide Time: 40:52)

Example – 1 (Contd.)


First six components are considered in the analysis
and the modified data is obtained as $Z = XA$

$Z =$

0.17	0.36	-0.34	-0.82	-0.39	-0.73	-0.08	-0.95	-1.29	-1.21
0.89	-0.18	1.50	-0.60	0.95	-0.02	0.72	0.24	-0.17	-0.45
0.27	0.55	1.17	-0.87	0.23	0.35	2.01	1.83	0.96	0.82
0.29	-0.57	0.91	-0.54	0.63	0.34	1.64	1.37	0.77	0.40
0.49	0.42	0.51	-0.78	0.66	-0.44	1.08	0.42	1.78	1.57
-0.36	-1.21	-0.01	-0.58	-0.51	-0.73	-0.28	-1.45	-0.54	-0.59
-1.13	-1.00	-0.70	-0.80	-1.01	-1.22	1.31	-0.50	0.13	0.07
-0.93	-0.90	-1.58	-0.95	-1.04	-0.81	-0.79	-2.33	-1.64	-2.48
-1.20	-1.50	-1.32	-1.10	-1.47	-0.81	-0.72	-0.61	-0.06	-0.27
-0.62	0.38	-0.01	-0.44	0.87	0.64	-0.11	0.59	1.60	0.53
-1.05	-0.23	-0.82	-0.83	-0.92	-0.59	-0.97	-0.48	-0.83	1.10
-0.49	0.91	0.37	1.05	1.36	0.22	-0.38	0.60	0.28	-0.52
0.19	0.77	-0.20	0.92	0.00	-0.72	-1.07	-0.88	-1.32	-0.89
0.23	0.00	0.08	0.66	-0.21	-0.38	-0.36	1.02	-0.62	-0.28
0.51	0.37	0.85	1.24	-0.03	1.95	-1.30	0.16	-0.90	-0.69
2.94	2.47	1.78	2.50	2.53	2.77	-0.60	0.38	-0.78	0.08
-0.93	-0.96	-1.42	0.29	-1.01	0.27	-1.07	0.31	0.89	1.50
-0.42	-0.14	-1.38	0.28	-0.78	-0.60	-0.11	-0.35	0.66	0.17
1.15	1.07	0.62	1.15	0.14	0.51	1.08	0.63	1.07	1.09

0.390	-0.165	0.211	-0.191	0.451	-0.304
0.381	-0.188	-0.053	-0.543	-0.127	0.215
0.033	0.029	0.382	0.235	0.074	-0.128
0.298	-0.321	-0.390	-0.111	0.246	0.400
0.404	-0.085	0.179	-0.121	-0.589	-0.056
0.371	-0.161	-0.229	0.546	-0.116	-0.394
0.122	0.462	0.521	-0.117	0.237	0.148
0.317	0.338	-0.122	0.444	0.069	0.603
0.136	0.529	-0.237	-0.201	-0.412	-0.110
0.160	0.443	-0.477	-0.192	0.351	-0.358

19×10 10×6




(Refer Slide Time: 42:54)

Example – 1 (Contd.)

$Z = XA =$

-1.283	-1.278	1.259	-0.492	0.139	0.045
1.133	0.192	1.776	0.367	-0.068	-0.361
1.825	2.512	0.974	0.410	0.245	0.386
1.273	1.972	0.998	0.826	0.040	0.127
1.176	2.403	0.134	-1.034	-0.232	-0.649
-1.907	-0.547	0.724	0.087	0.087	-0.705
-2.395	1.519	0.673	0.048	0.429	0.160
-3.779	-2.327	1.050	-0.058	-0.483	-0.172
-3.115	0.333	-0.480	0.475	-0.042	-0.194
0.830	1.247	-0.735	0.055	-1.484	-0.236
-1.700	0.094	-1.054	-0.154	0.348	-0.373
1.345	-0.585	-0.305	-0.126	-1.214	1.016
-0.430	-2.241	0.012	-0.817	0.203	0.574
0.243	-0.433	-0.169	0.391	0.593	1.067
1.336	-2.171	-0.751	1.326	0.157	-0.178
5.527	-2.835	-0.199	-0.183	0.169	-0.643
-1.202	0.857	-2.516	0.433	0.246	-0.259
-1.219	0.366	-0.975	-0.743	0.059	0.311
2.341	0.921	-0.416	-0.792	0.808	0.081

19×6



So, we do the same thing our x matrix remains the same and our A matrix now is restricted only three variables, these are the three principle components. So, I will get a ten by three matrix instead of the ten by six that I got earlier, I use this transformed data now and then Z is equal to X into A. So, this is a z matrix this will be nineteen by three, instead of the nineteen by six, that I got earlier.

(Refer Slide Time: 51:03)


Example – 1 (Contd.)

Regression analysis is performed on these components

$$Y = ZB$$

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

$Y =$	$\begin{bmatrix} -314.2 \\ 113.3 \\ 352.5 \\ 339.9 \\ 199.5 \\ -463.0 \\ -500.0 \\ -629.2 \\ -554.2 \\ 383.7 \\ -397.0 \\ 221.6 \\ -224.6 \\ 80.9 \\ 309.3 \\ 1036.7 \\ -70.6 \\ -328.4 \\ 444.0 \end{bmatrix}$	$Z =$	$\begin{bmatrix} -1.283 & -1.278 & 1.259 \\ 1.133 & 0.192 & 1.776 \\ 1.825 & 2.512 & 0.974 \\ 1.273 & 1.972 & 0.998 \\ 1.176 & 2.403 & 0.134 \\ -1.907 & -0.547 & 0.724 \\ -2.395 & 1.519 & 0.673 \\ -3.779 & -2.327 & 1.050 \\ -3.115 & 0.333 & -0.480 \\ 0.830 & 1.247 & -0.735 \\ -1.700 & 0.094 & -1.054 \\ 1.345 & -0.585 & -0.305 \\ -0.430 & -2.241 & 0.012 \\ 0.243 & -0.433 & -0.169 \\ 1.336 & -2.171 & -0.751 \\ 5.527 & -2.835 & -0.199 \\ -1.202 & 0.857 & -2.516 \\ -1.219 & 0.366 & -0.975 \\ 2.341 & 0.921 & -0.416 \end{bmatrix}$
-------	---	-------	--



(Refer Slide Time: 51:24)


Example – 1 (Contd.)

Considering first three components Considering first six components

$$(Z'Z)^{-1} = \begin{bmatrix} 89.01 & 0.00 & 0.00 \\ 0.00 & 47.36 & 0.00 \\ 0.00 & 0.00 & 18.85 \end{bmatrix}$$

$$(Z'Z)^{-1} = \begin{bmatrix} 89.01 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 47.36 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 18.85 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 6.55 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 5.52 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 4.62 \end{bmatrix}$$

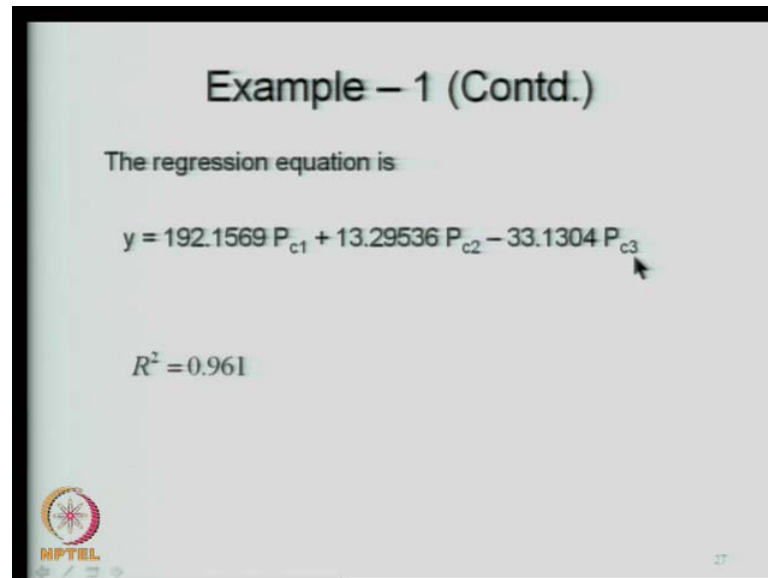
$$\hat{B} = (Z'Z)^{-1} Z'Y = \begin{bmatrix} 192.2 \\ 13.3 \\ -33.1 \end{bmatrix}$$

$$\hat{B} = (Z'Z)^{-1} Z'Y = \begin{bmatrix} 192.2 \\ 13.3 \\ -33.1 \\ 73.9 \\ -64.1 \\ -15.7 \end{bmatrix}$$


Then we develop the regression relationship which is y is equal to Z into B and y is this this remains unchanged and z is modified, now z is ten by three now. So, I get beta cap as Z dash Z inverse Z dash Y . So, I get the beta cap Z dash Z inverse the intermediate values are given, look at this now, if we took only the first three components, I am getting the matrix up to this point this is my first three components whereas, if I took all the six components I would have got this matrix. So, this is just a subset of the earlier matrix. So, I get this principle component and then obtain the B cap which is the estimated values for the parameters b as 192.2 13.3 minus 33.3 point three, look at the

earlier estimates these three components still remain the same. So, when I change my dimension these have not got changed these are just the same and I have just discarded discarded these other remaining three.

(Refer Slide Time: 52:38)



The slide displays the following information:

Example – 1 (Contd.)

The regression equation is:

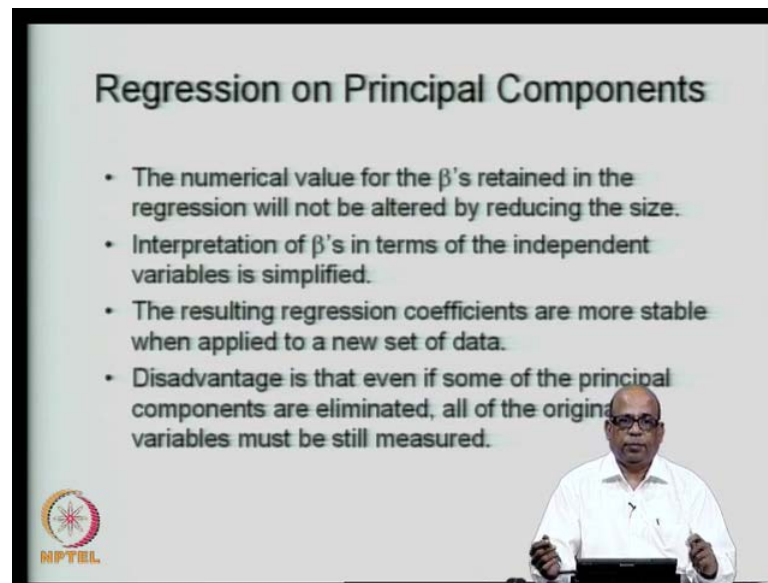
$$y = 192.1569 P_{c1} + 13.29536 P_{c2} - 33.1304 P_{c3}$$

$R^2 = 0.961$

The slide also features a logo in the bottom left corner and a small number '27' in the bottom right corner.

Now, that is the interesting feature of the principle component. So, we can just ignore these three and first use the first three in our regression relationship. So, the equation now is the first three coefficients only will come into the picture and these three coefficients are exactly the same as we obtained earlier using the six components and I obtain r square value of 0.961 which is smaller than if I had used six components. So, if I use all the ten variables in their original shape, I got some other r square value I got some ah regression relationship r square value higher than the other two then I reduce the size I used six components I got R square value of 0.97 or some such thing I reduced it further and then I get a R square value of 0.961. Now, this in some sense indicates amount of information that can be reproduced by using this particular regression relationship.

(Refer Slide Time: 53:38)



Regression on Principal Components

- The numerical value for the β 's retained in the regression will not be altered by reducing the size.
- Interpretation of β 's in terms of the independent variables is simplified.
- The resulting regression coefficients are more stable when applied to a new set of data.
- Disadvantage is that even if some of the principal components are eliminated, all of the original variables must be still measured.

MPTEL

So, this is what we do in regression using the principle components, now there are certain advantages of doing this; as I just mentioned, let us say that you took 95 percent of the variance, and then that is a that is those eigen vectors which would reproduce 95 percent of the variance, in our example it was six. So, six of them you took and then obtain the regression relationship. On that, if you discard the three the remaining three coefficients still remain the same that is the beauty of this. So, the numerical values of the betas as you obtain using large number of principle components will remain the same, even if you discard some of them some of the later betas and then the interpretation of the betas in terms of the independent variables is simplified. Now, because we are saying that these betas are these principle components that we are using in the regression relationship are explaining so much of the variance.

But you know when we do this principle component analysis, even if we use let say only three of principle components in the regression relationship still the complete observed data on all the ten variables are necessary. This, I keep on repeating that the betas or the principle components that we are using. Now, should not be related one to one with the original variable these principle components are linear transformation of combination of all these ten variables and therefore, even if we use only three principle components the original data must be available for all the ten variables. So, that is minor disadvantage. So, to conclude now in today's lecture, essentially we picked up on the principle component analysis that, we discussed in the previous lecture and saw how we use the

principle components in the regression relationship. So, essentially if you recall the principle component analysis is done on the principle components are the eigenvectors which are obtained in the covariance matrix.

So, if you had ten variables you have a ten by ten covariance matrix and on this, you obtain the eigen vectors and these eigen vectors are in fact, the principle components. So, for a ten by ten matrix, you get ten eigen vectors and each of these eigen vectors represent a principle component and instead of doing your regression of the dependent variable on the original independent variables, we do the regression of the dependent variable on the principle components or the eigen vectors. This has the advantage that the eigen vectors are uncorrelated. So, the first eigenvector has no correlation with the second eigen vectors and so on. And also we have the advantage that, we can reduce the size of the regression instead of using ten principle components or ten variables original variables. We may use just six in the example that we just saw we used only six because the six eigen vectors explained or contributed to 95 percent of the variance in the original problem original dependent variable. And thus, we can reduce the size depending on how much percentage the particular set of principle component can explain. In fact, in many realistic situations we may have twenty variables thirty variables and.

So, on associated with that we may have twenty eigen vectors, out of which only the first three may explain up to ninety percent or ninety five percent of the variance and therefore, you can afford to ignore all the remaining principle components and developing the relationship just based on the first three principle components. So, these are the advantages. Next, we move on to multivariate stochastic models; that means, we now know how to develop regression relationships based on multiple variables. We will extend this, specifically to hydrologic problems where we are dealing with multiple variables and then we want to develop stochastic models on that. So, we will continue the discussion in the next lecture. Thank you for your attention.