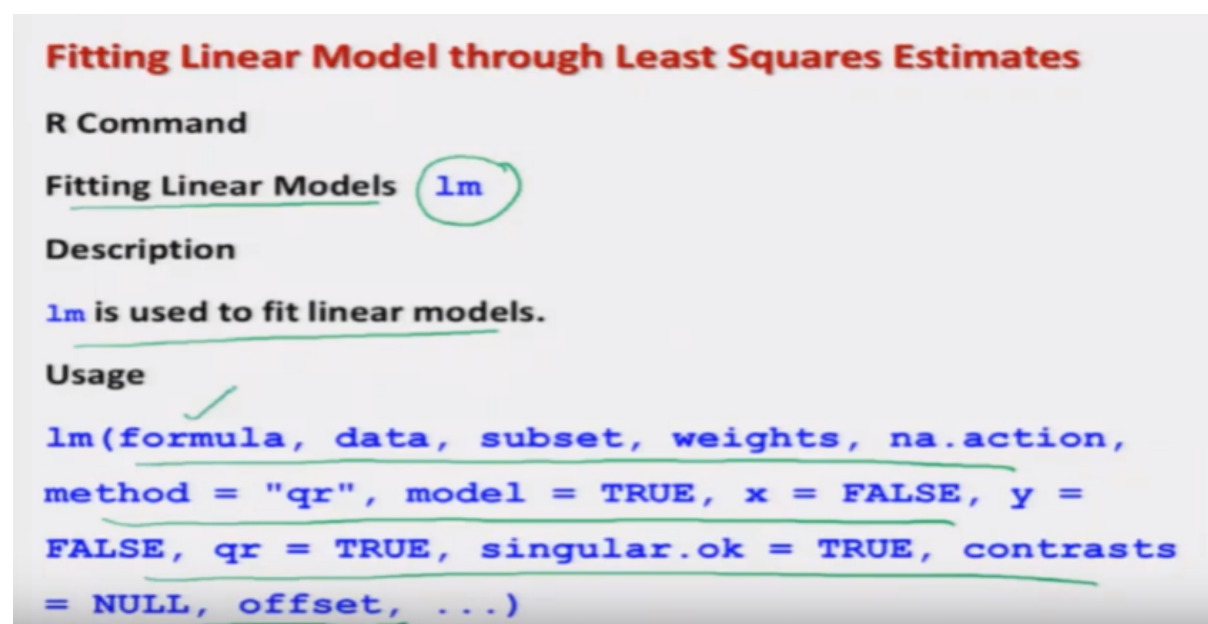


Lecture-34

Fitting of Linear Models: Least Squares Method – R Commands and More than One Variables

Welcome to the lecture on the course descriptive statistics with R, and welcome to the last lecture of the course. Yes! When ever the last lecture comes this gives us happiness for the students and for the teachers also. So we try to understand what are we going to do in this lecture you may recall that in the last lecture we had discussed the principle of least squares, and we also had understood that how one can obtain the estimates of the parameters or the values of the parameters on the basis of given set of data in case of a model y equal to α plus βx . We had taken an example and we had solved it manually the idea was to show you or to expose you with the basic concepts. Now in this lecture we will learn that how to obtain the same result using the R software, and after that considering only one input variable is not a very realistic thing so in case if you want to extend the principle of least square where, where you have more than one independent variables or more than one input variables then how to do it and how to use the R software to obtain the least square estimates of parameters in that model. So, this is what we are going to do in this lecture. So, the first thing comes what is the command for obtaining the least squares estimate or fitting a linear model using the R software.

Refer Slide Time: (2:05)



Fitting Linear Model through Least Squares Estimates

R Command

Fitting Linear Models `lm`

Description

`lm` is used to fit linear models.

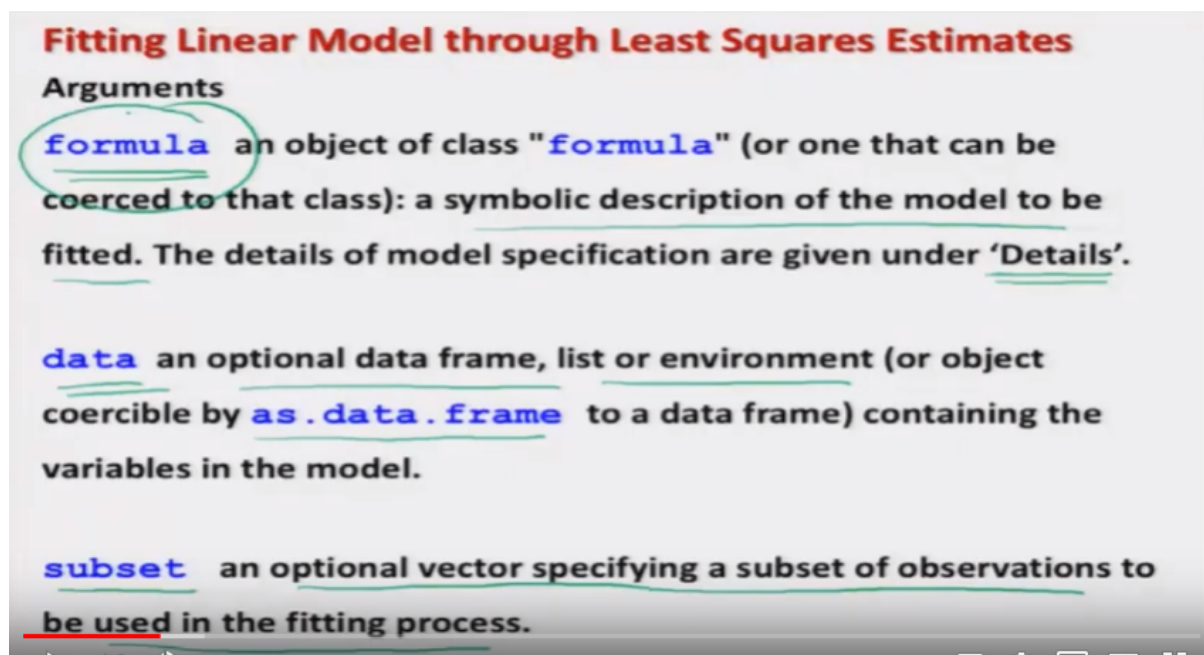
Usage

`lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, ...)`

So in case of our software the our command for fitting a linear model is `lm` these applets this is an abbreviation of 'l' means linear and 'm' for model, and this command 'lm' is used to fit the linear model, well you can see here that this 'lm' has various arguments and so on. But here we are going to use a very simple thing just basis on formula and you should know why? why I'm not going to discuss all the details I had explained you in the earlier lecture that whenever we try to fit a linear model this does not end the story after that you have to check how the model is going to perform with the entire population. There are different types of hysterial assumptions which are needed to expose the estimated values were different types of static a tool like a test of hypothesis goodness of fit and so

on. So well here I am NOT considering the course on linear regression analysis but I am simply trying to use one of the methods for finding out the model which is used in the case of linear regression analysis, and this command 'lm' in our software was developed in order to find out the further statistical tool related to the linear regression analysis. So that is why you will see that in the command 'lm' there are many, many arguments but I will not go into those details but I will request you that if you want to understand them first try to have a course on linear regression analysis and then try to understand those concept. The interpretation of all the terms inside the argument they can be, be studied from the help on 'lm' command. Right?

Refer Slide Time: (4:16)



Fitting Linear Model through Least Squares Estimates

Arguments

formula an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

data an optional data frame, list or environment (or object coercible by `as.data.frame` to a data frame) containing the variables in the model.

subset an optional vector specifying a subset of observations to be used in the fitting process.

So now my objective here is to show you that how to get the things but briefly I will try to give you the idea here what we are going to use that there is an option here 'formula'. So basically, we will be adding or using here the option of formula, so this is going to give us a sort of symbolic description of the model to be fitted. Right? and what are the details of the model they are given under as separate specifications and next the data this is also an optional argument where the data is given in the frame of in the structure of a data frame or list of environment is given by a as data dot frame and so on but we are not going into that detail and, similarly there is another option here upset and this gives you an option to use an optional vector and which is specifying a subset of observation to be used in the fitting process but anyway we are not going to use it we are simply going to use the option here 'formula'.

Refer Slide Time: (5:24)

Example

Data on marks obtained by 20 students out of 500 marks and the number of hours they studied per week are recorded as follows:

We know from experience that marks obtained by students increase as the number of hours increase.

Marks	337	316	327	340	374	330	352	353	370	380
Number of hours per week	23	25	26	27	30	26	29	32	33	34

Marks	384	398	413	428	430	438	439	479	460	450
Number of hours per week	35	38	39	42	43	44	45	46	44	41

So now how to use this formula I will try to show you with an example the same example we had considered earlier where we collected the data on 20 students on their marks and the number of hours in a week which they studied.

Refer Slide Time: (5:35)

Example

Solving it for the given data on marks and hours, we get the values of α and β as follows:

$$\bar{y} = \frac{1}{20} \sum_{i=1}^{20} y_i = 389.9, \quad \bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = 35.1$$

$$\hat{\beta} = \frac{\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{20} (x_i - \bar{x})^2} = 6.3,$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 168.65$$

manually.

Model: marks = 168.65 + 6.3*hours

And this data if you remember in the last lecture was obtained here that beta hat was obtained as 6.3 and alpha hat was obtained at 168.65 and model was in like it, but this all was done manually all the competitions were done by using the case simple calculator.

Refer Slide Time: (5:57)

Example

```
marks =  
c(337, 316, 327, 340, 374, 330, 352, 353, 370, 380, 384,  
398, 413, 428, 430, 438, 439, 479, 460, 450)  
  
hours =  
c(23, 25, 26, 27, 30, 26, 29, 32, 33, 34, 35, 38, 39, 42, 43,  
44, 45, 46, 44, 41)
```

Now I will try to show you how to do it on the R software. So, as we had discussed we already have stored this data in that two data vectors marks and hours and how to store this data that I discussed in the last lecture. Right? That this all this data is coming in the same order so that these observations are paired I mean the first observation 'x_iy_i' mean first observation of marks and first observation of hours.

Refer Slide Time: (6:26)

Fitting Linear Model through Least Squares Estimates

Example

R Command

```
> lm(marks ~ hours)
```

Call:
lm(formula = marks ~ hours)

Coefficients:
(Intercept) 168.647
hours 6.304

$lm(y \sim x)$
↓
Output variables
input variables

marks = $\alpha + \beta \times \text{hours} + \text{errors}$

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Now this is the most important part that how to give the command. First you try to write down the instruction 'lm' and then try to write down here the output where you will here 'y' and then use this equivalent sign this is present on your keyboard, the keyboard of the computer which you are using and after this you try it to use here the variable to give the input data. So why is your hair output variable and x is your here input variable. Right? So, and they have to be given in this format so in our case our output variable here is marks and input variable here is hours and they are given in this framework just joined by the sign equivalent. Right? and the sign I can make it here more bigger something like this you will see on the on your keyboard. Now once I try to do it this will give me this type of outcome so first we try to understand what is the meaning of this outcome, once I say that that

Im marks equivalents hours this is going to indicate that our model is marks is equal to see here alpha plus beta times see hours plus some error. Right? and this will inform the R software that we had obtained say beta had i goes from 1 to here n x_i minus \bar{x} and y_i minus \bar{y} upon summation i goes from 1 to n x_i minus \bar{x} whole square. So this command will inform the R software that the value of x_i 's are coming for example I will use a different colour, so this values of here x_i 's are coming from the data which is given here in hours, and similarly the values of y's are coming for computation of these things from the first theta vector marks. Right? and so it uses the formula is equal to marks, equivalents, hours and the outcome comes out to be like this so you can see here this is written as coefficients the coefficients means the value of alpha and beta, and this is specifying here the value of alpha that is also called as intercept. So, this value is coming out to be 168.647 and the value of the beta, which is coming here, here so this is here the value of alpha had, and this is here the value of beta had. So, this value is indicating that this is the intercept term, and this is indicating that this 6.304 is the coefficient associated with the variable hours and you can see here these are the same value which you had obtained manually you can compare here,

Refer Slide Time: (9:52)

Example

Solving it for the given data on **marks** and **hours**, we get the values of α and β as follows:

$$\bar{y} = \frac{1}{20} \sum_{i=1}^{20} y_i = 389.9, \quad \bar{x} = \frac{1}{20} \sum_{i=1}^{20} x_i = 35.1$$

$$\hat{\beta} = \frac{\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{20} (x_i - \bar{x})^2} = 6.3, \quad \checkmark$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 168.65 \quad \checkmark$$

manually.

Model: marks = 168.65 + 6.3*hours

with this, and here this.

Refer Slide Time: (9:54)

Fitting Linear Model through Least Squares Estimates

Example

R Command
`> lm(marks ~ hours)`

Call:
`lm(formula = marks ~ hours)`

Coefficients:
 (Intercept) 168.647
 hours 6.304

Handwritten notes:
 $lm(y \sim x)$
 ↓ output variables input variables
 $marks = \alpha + \beta \times hours + \text{errors}$

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

So, you can see here this is pretty straightforward to obtain such a result inside the R software.

Refer Slide Time: (10:02)

Fitting Linear Model through Least Squares Estimates

Example

```

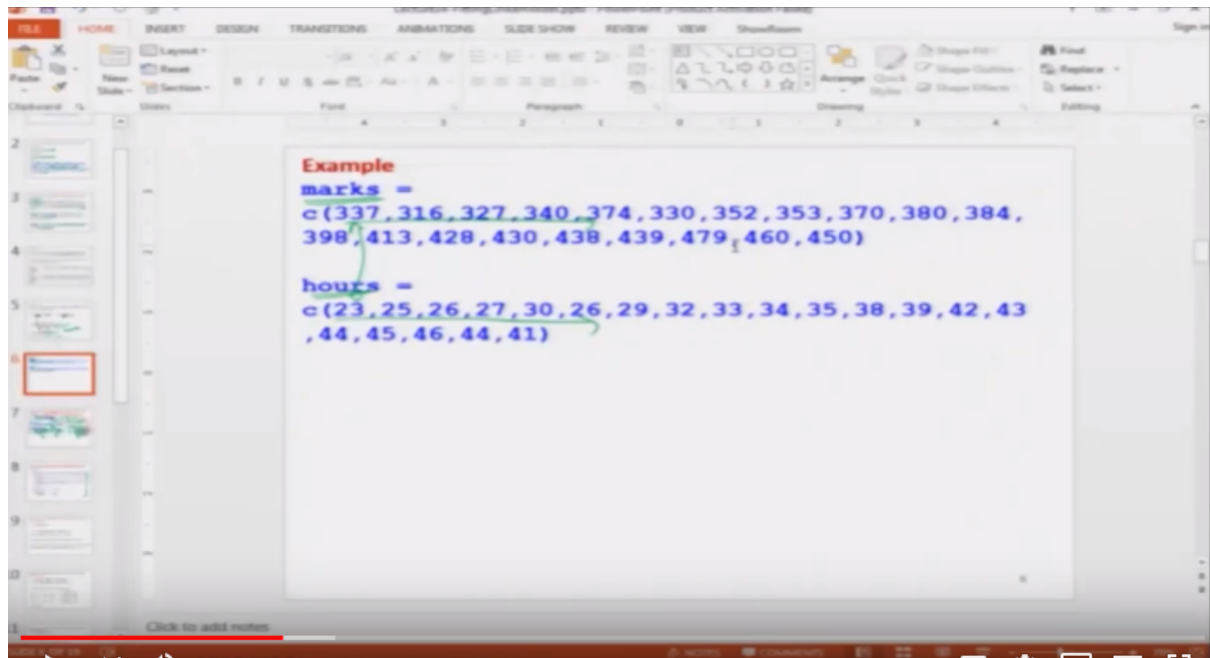
> hours
[1] 23 25 26 27 30 26 29 32 33 34 35 38 39 42 43 44
[17] 45 46 44 41
> marks
[1] 337 316 327 340 374 330 352 353 370 380 384 398
[13] 413 428 430 438 439 479 460 450
> lm(marks~hours)

Call:
lm(formula = marks ~ hours)

Coefficients:
(Intercept)      hours
  168.647         6.304
  
```

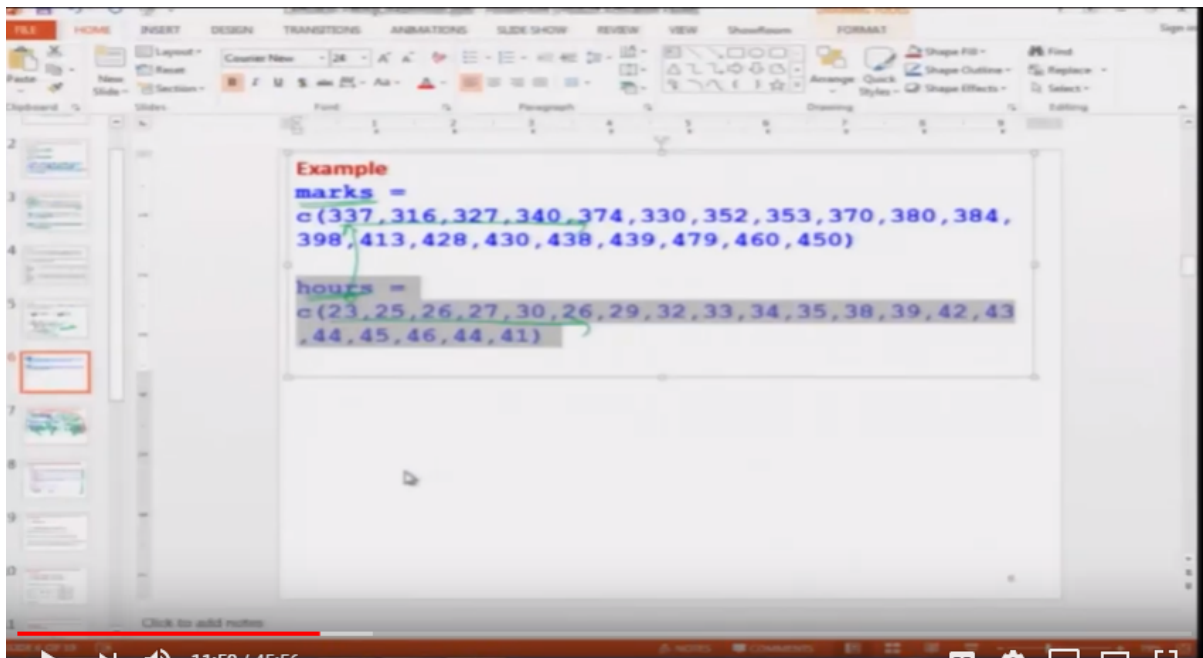
Now I will try to show it on the R console, and this is here the screenshot of the same thing.

Video Start Time: (10:10)



So first I will try to prepare my data so this is here marks data which I have entered and this is my here data on hours which is entered here so you can see here marks data is like this hours data is like this and now I'm trying to find out here the model 'lm' say marks equivalence say hours. So, you can see here that this is coming out to be like this. Right? and here you have to be careful that how you are going to specify the value of x and y for example in case if you make a mistake and if you say in place of marks hours and hours in place of marks, then this linear model will be fitted like hours and between hours and marks and you can see here that this is entirely different than the first one, you can see here the first value is 168 but here the value is .22 but this is a wrong thing to do because here in this case when you are trying to fit a model like this one you are trying to say that number of hours is your output variable and input variable is your marks which is not correct here in this case. So, this is how we you try to obtain this thing. Right?

Video End Time: (11:50)



Refer Slide Time: (11:52)

Fitting Linear Model through Least Squares: More than One Variables

$y = \alpha + \beta x + e$ → One input variable → x
 → more than one input variable
 p no. of input variables
 x_1, x_2, \dots, x_p
 $\beta_1 x_1 \rightarrow \beta_2 x_2 \rightarrow \dots \rightarrow \beta_p x_p$

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + e$

Relationship between y and x_1, x_2, \dots, x_p is linear. → error random

Matrix plots are useful in graphically verifying the linearity.

Conduct the experiment and obtain n tuples of observations on dependent variable (y) and independent variables x_1, x_2, \dots, x_p .

$(x_{1i}, x_{2i}, \dots, x_{pi}, y_i)$ (x_i, y_i)

So now let us try to come back to our slides, and now I try to give you one more concept now you see we have learned how to find out a model or the least square estimates of alpha and beta in the case of a linear model where you have only one variable but in practice you can imagine that any process is not going to be controlled only by one variable but there will be more than one variables and, we had discussed such example in the last lecture for example the yield of a crop will depend on several factors quantity of fertilizer quality of fertilizer temperature rainfall irrigation and so on. So now in case if you want to extend this principle of least squares to the case when you have more than one input variable then how to do it, well this is case of linear regression analysis under the topic multiple

linear regression model but here my objective is not to teach you the regression analysis my objective is to tell you or show you that how you can obtain the values of the parameters on the basis or given sample of data using the R software in a case when you have more than one input variables. So in this case I will try to take an example of the linear model and I will show you logically that how you can extend the model which you have considered here into a multiple framework when you more than one input variable but yeah this is not a very general technique that is valid for all the models miss every model has its own way to extend it to a multivariate situation but here you please try to learn that if I try to extend my linear model in this particular way then how the R software can be used to find out the values of the parameter estimates. Okay?

So now you may recall that we had considered the model here y equal to α plus βx now I'm trying to say that here you had only here one input variable, but now I'm trying to extend it and suppose I'm saying that there are more than one input variables. So, in the first case I have denoted the input variable by the notation x so now suppose I say there are p number of variable of input variables so I can denote them say here x_1, x_2 see here x_p . Right? miss I deal in that symbols and notation of regression analysis this is small x_1 is small x_2 is small x_p should be capital X_1 capital X_2 capital X_p but here my idea is something different than the objective in linear regression analysis so I try to extend it and since I am going to consider here and edit your model so I try to write down the terms like βx for each of the variable so for x_1 this will become say here $\beta_1 x_1$ for x_2 this will become here $\beta_2 x_2$ and similarly here for Expedia's will become here $\beta_p x_p$ and then I try to add them together so this is what I am writing here in the second line that y is equal to α plus $\beta_1 x_1$ plus $\beta_2 x_2$ plus up to here $\beta_p x_p$ plus e . So e is also here a header and this is random error involved in the observations no similar to the simple scatterplot that we had obtained in the last lecture here in this case you have more than one independent variable so matrix plots are more useful in verifying whether the relationship between y and x_1, x_2, x_p is linear or not, one thing which you have to keep in mind here that in the earlier case we try to establish the linear association between x and y but now you have here a group of x 's $x_1 x_2 x_p$ inside the one group and there is here one simple output variable y , so we are trying to verify the linearity of a single variable y with respect to a group of variable of x_1, x_2, x_p . This is not so straightforward one option is that I can make individual plots y versus x_1 y versus x_2 y versus x_3 and so on, and now what we can conclude that if all the relationship bit of y is with respect to each of the x_1, x_2, x_p is coming out to be linear then we can expect that their joint effect may also be linear, but this is a little bit tricky situation and you need some experience to handle this condition but here but here I would like to just inform you that how you can verify our how you can construct such plots in the R software.

So now the first I would try to show you that how to construct such mattress plot in this case if you try to see how you are trying to obtain the observation you are trying to conduct the experiment say in x and for every experiment you will have an observation like observation on the first variable

observation, on the second variable and observation on the p^{th} input variable and then here value of y_i , and if you remember earlier it was simply here x_i, y_i because there were only one variable so here the observations are going to be something $1, 2, \dots, n$ and y_i . So, this is going to represent the i^{th} and its tuple of observations on say y and x . Right? So now we assume that each set of observation will satisfy this equation.

Refer Slide Time (18:28)

Fitting Linear Model through Least Squares: More than One Variables

$$\begin{aligned}
 y_1 &= \alpha + \beta_1 x_{11} + \beta_2 x_{12} + \dots + \beta_p x_{1p} + e_1 \\
 y_2 &= \alpha + \beta_1 x_{21} + \beta_2 x_{22} + \dots + \beta_p x_{2p} + e_2 \\
 &\vdots \\
 y_n &= \alpha + \beta_1 x_{n1} + \beta_2 x_{n2} + \dots + \beta_p x_{np} + e_n
 \end{aligned}$$

First set of obs
n obs

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

p+1

So we can write here for the first set of observation we can write that this set of observations satisfies the equation like this so instead of here x_1 I have x_{11} instead of here x_2 I have here x_{12} instead of here P I have here x_{1p} , and associated random error here is even and the obtained value of y here is y_1 and similarly we have obtained here say n observation which are satisfying this equation. Now using a very simple theory of mathematics based on vectors and matrix this entire set of equation can be expressed in the form of vectors and matrices. So this I'm trying to give it here I am not explaining the theory of metrics here but I assume that you know it so all this observation they are contained inside the n cross 1 vector here y_1, y_2, y_n and all those parameters $\alpha, \beta_1, \beta_2, \beta_p$ they are contained in say another vector here consisting of P plus 1 elements $\alpha, \beta_1, \beta_2, \beta_p$ and the Associated matrix of input variables on the data on say here x_1, x_2 and say here X_p this is given here like this, and the first column is indicating the presence of intercept M 1111 . So now I can express this y vector here as say small y and this entire matrix here to be here x this entire matrix to be here β and this entire e_1, e_n vector to be here small e . Right?

Refer Slide Time: (20:14)

Fitting Linear Model through Least Squares: More than One Variables

How to find parameters?

$$y = x\beta + e$$

$$S = \sum_{i=1}^n e_i^2 = e'e = (y - x\beta)'(y - x\beta)$$

$$\frac{\partial S}{\partial \beta} = 0 \Rightarrow \hat{\beta} = (x'x)^{-1}x'y$$

Use principle of least squares

matrix of obs on input variable

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

Least squares estimator

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$\hat{\beta} = \begin{pmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}$

$(p+1) \times 1$

So now we have this model here y equal to $x\beta$ plus e . Now in case if you try to use the principle of least squares then principle of least squares if you remember in the last lecture we defined the sum of square of this e_i 's and this were defined as s summation i goes from 1 to n e_i^2 so this quantity can be written as say here e transpose e where is now a vector so e is now here y minus $x\beta$ transpose y minus $x\beta$ like this. Now in case if you try to differentiate this quantity with respect to here β put it equal to 0 then after solving you get here say β equal to $\hat{\beta}$ which is equal to x transpose x whole inverse x transpose x which I am writing here. Right? I am NOT giving you here the details but if you try to see this is a simple extension of the least square estimate that you obtain in the model with one input variable so now these things are replaced by matrix so you can see here that X is the matrix of X here is matrix of observations on say input variable so this is known to us similarly here Y this is a vector of observations on the observed values $y_1 y_2 y_n$ so this x and y are known so I can compute the value of $\hat{\beta}$ and this will be called as least squares estimator of parameter β , and $\hat{\beta}$ will look like if you see β is looking like this then $\hat{\beta}$ will look like say $\hat{\alpha}$ $\hat{\beta}_1$ $\hat{\beta}_2$ up to here $\hat{\beta}_p$. So, this also has the same order as of β which is of order P plus 1 Cross 1. Right?

Refer Slide Time (22:25)

Example with Two Variables

Following data is obtained on the delivery time taken in delivering the parcels and corresponding distance travelled by a courier person.

Delivery Time Data				Delivery Time Data			
Obs. number	Delivery time(in minutes) (y)	Number of parcels (x_1)	Distance (in meters) (x_2)	Obs. number	Delivery time(in minutes) (y)	Number of parcels (x_1)	Distance (in meters) (x_2)
1	16.68	7	560	13	13.5	4	255
2	11.5	3	220	14	19.75	6	462
3	12.03	3	340	15	24	9	448
4	14.88	4	80	16	29	10	776
5	13.75	6	150	17	16.35	6	200
6	18.11	7	330	18	19	7	132
7	8	2	110	19	9.5	3	36
8	17.83	7	210	20	35.1	17	770
9	79.24	30	1460	21	17.9	10	140
10	21.5	5	605	22	52.32	26	817
11	40.33	16	688	23	18.75	9	450
12	21	10	215	24	19.83	8	635
				25	10.75	4	450

Now I would try to show you the computation of beta hat using example and I have just taken two input variables in this example to make it more simple and in this example 25 observations have been collected on the time taken by a courier person in delivering the parcels. So, it is recorded that how much time the courier person takes in delivering the parcels and obviously this time is going to be dependent that how many parcels are being delivered by this courier person and also how much time the courier person has to travel so this number of parcels are going to be denoted by x_1 and distance traveled by the courier person is to denoted as here say x_2 and whatever the time is taken that is denoted by here y . So, the interpretation of the data goes like that there is a courier person who has to deliver 7

Parcels, and suppose the courier person travels 560 meters and that total time in doing this job is 16.68 minutes. Similarly, there is another courier person who delivers three parcels and percent travel 220 meters and person takes 11.5 minutes and so on. So, this is how we have obtained this 25 set of data and all this data on y , x_1 and x_2 has been stored inside the data vectors inside the R software. Right? As we have done it many times in the past.

Refer Slide Time: (24:18)

Example with Two Variables

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, 2, \dots, 25$$

deltime =

c(16.68, 11.5, 12.03, 14.88, 13.75, 18.11, 8, 17.83, 79.24, 21.5, 40.33, 21, 13.5, 19.75, 24, 29, 16.35, 19, 9.5, 35.1, 17.9, 52.32, 18.75, 19.83, 10.75)

parcelno =

c(7, 3, 3, 4, 6, 7, 2, 7, 30, 5, 16, 10, 4, 6, 9, 10, 6, 7, 3, 1, 7, 10, 26, 9, 8, 4)

distance =

c(560, 220, 340, 80, 150, 330, 110, 210, 1460, 605, 688, 215, 255, 462, 448, 776, 200, 132, 36, 770, 140, 817, 450, 635, 450)

So essentially now we are going to fit here a model y_i equal to α plus $\beta_1 x_{1i}$ plus $\beta_2 x_{2i}$ plus e_i where all these observations are denoted by here I , and all observation will satisfy this model y equal, equal to α plus $\beta_1 X_1$ plus $\beta_2 X_2$ so all this data in the same order has been stored in three vectors the time of delivery and del time in the number of parcel in parcelno they have been a parcel numbers and traveled distance inside the data vector whose name is the distance. Right? So once again I will explain you that this first observation correspond to in the first observation of parcel number and this correspond to the first observation in that distance which is given here,

Refer Slide Time: (25:06)

Example with Two Variables

Following data is obtained on the delivery time taken in delivering the parcels and corresponding distance travelled by a courier person.

Delivery Time Data				Delivery Time Data			
Obs. number	Delivery time (in minutes) (y)	Number of parcels (x_1)	Distance (in meters) (x_2)	Obs. number	Delivery time (in minutes) (y)	Number of parcels (x_1)	Distance (in meters) (x_2)
1	16.68	7	560	13	13.5	4	255
2	11.5	3	220	14	19.75	6	462
3	12.03	3	340	15	24	9	448
4	14.88	4	80	16	29	10	776
5	13.75	6	150	17	16.35	6	200
6	18.11	7	330	18	19	7	132
7	8	2	110	19	9.5	3	36
8	17.83	7	210	20	35.1	17	770
9	79.24	30	1460	21	17.9	10	140
10	21.5	5	605	22	52.32	26	817
11	40.33	16	688	23	18.75	9	450
12	21	10	215	24	19.83	8	635
				25	10.75	4	450

I say observation number here one. Right? So, this data is given over there so that we already had understood in the earlier example.

Refer Slide Time: (25:16)

Matrix Plot
`pairs(x, ...)` produces a matrix of scatterplots.
`pairs(formula, data = NULL, ..., subset, na.action = stats::na.pass)`

Arguments:
x coordinates of points given as numeric columns of a matrix or data frame.
formula a formula, such as $\sim x + y + z$. Each term will give a separate variable in the pairs plot, so terms should be numeric vectors.
data a data.frame (or list) from which the variables in **formula**

Now I would try to first show you that how to create a matrix plot. Right? what is a matrix plot you see when we had an idea or when we had discussed the simple case where I had only one independent variable, or one input variable, and one output variable, then I can make a scatter plot and a scatter plot will give us an idea that how the things are going to look like whether there is a linear trend or a nonlinear trend, but definitely when we have more than one input variables then what we are looking we are looking between the joint effect of one variable y with respect to a group of variable x_1, x_2, x_p . So finding outliers the Kutchka curve is more difficult so what we try to do we try to create the scatter plots pairwise between y and x one y and x_2 y and x_3 and so on, and then what we interpret is that if all the relationship that means the relationship of y with respect to each of the x_1, x_2, x_p is linear so we can expect at the joint effect or joint association between y and all x_1, x_2, x_p is also expected to be linear. Well you need some experience in interpreting search results but here I would like to show you that how to create such matches plot. Okay?

So that command for creating the matrix plot here is `pairs` 'pairs' and inside the argument what we try to give here the data vectors and with some Orion form and another mode general structure is say disappears and inside the argument we give the formula just in that case of say linear models the operator 'lm' and after that we there are different options to be given data subsets any action and so on. So here you can see here I have given the details here that x is trying to give us the coordinates of a point given as numeric columns of a matrix or a data frame and similarly here formula that this is the same thing what we discuss in the case of linear model so we have to write the formula as say with an

equivalent science given by the variables and separated by plus sign. Right? So, each of this term will give a separate plot with respect to y, and then data is a data frame in which the data for the formula is has to be used. Right?

Refer Slide Time: (27:49)

Matrix Plot

Arguments

subset an optional vector specifying a subset of observations to be used for plotting.

data a data.frame (or list) from which the variables in **formula** should be taken.

Then there is an option here subsets and data exactly in the same way as we did in the case of linear model. So, I am not discussing it here.

Refer Slide Time: (27:57)

Example with Two Variables

Matrix Plot

```
> pairs(~deltime + parcelno + distance,
main="Matrix Plot of Delivery Time data")
```

Matrix Plot of Delivery Time data

deltime

parcelno

distance

deltime vs. deltime

distance vs. deltime

deltime vs. distance

distance vs. distance

deltime vs. parcelno

parcelno vs. deltime

parcelno vs. distance

distance vs. parcelno

32:02 / 45:56

Now I would try to create here a matrix plot of the given data set just by writing the formula 'formula' you see I am writing here now pairs and then inside the argument formula for that I write here an equivalent sign and then I need to find out the scatter plot of three variables delivery time, parcel

number, distance so I'm not discriminating here what is my input variable and what is my output variable and it means you can get more information about this command pairs from the help but here I am using one option which is the 'main' to give the title of the graph. So, you can see here this is my graph and this title is the same here matrix plot of delivery time data which is printed here. I will try to show you on the our console also but first we try to understand this picture this matrix plot will look like this so you can see here first try to understand this graphic so you can see here what is there on the y-axis, what is here on the y-axis this is the same thing which written here, and what is here on the x-axis I will use a different color so that you can observe this is the same thing which is written here. So in this box the x-axis is denoting the parcel number and y axis is denoting that the delivery time, and this is the scatter plot of two variables parcel number versus delivery time or the number of parcels versus the delivery time and you can see here that distant is nearly the say this a linear trend and a positive trend rather similarly if you come to this block again now what is being mentioned here, here this is the same thing which is written here say distance so now this is the plot between distance, and what is happening on the y axis here this is that delivery time so you have to take the y axis from the y axis side and x axis from the x axis side, so this is a graphic between the distance travelled by the courier percent versus delivery time, and you can see here that this curve is also coming out to be a sort of a having a linear trend and similarly if you try to look here in the third case, see here so you can see here on the x axis here this will be here distance and on the y axis we will have here number of parcels.

So this is a graph between the distance versus parcels parcel number or the number of parcels, and you can see here, here also the if you try to look at the trend this is again shows a nearly linear trend and then if you observe direction of my pen it is here this crossing this diagonally. Now what I have shown you here is the plots in the upper diagonal in this side, now what is happening in the lower diagonal this is the same thing for example these two will match this, two will match and these two will match. So usually we try to look in either in the upper diagonal or on the lower diagonal, so they are going to give us the similarity information. Okay? and yeah! means some of these blocks which I am indicating here by cross these are not used why because this is a plot between for example in this case this, is the plot between say here delivery time versus - delivery time which has no meaning. Right? So, by looking at this matrix plot we can have an idea that whether the joint relationship of y with respect to x_1 and x_2 is a linear or not.

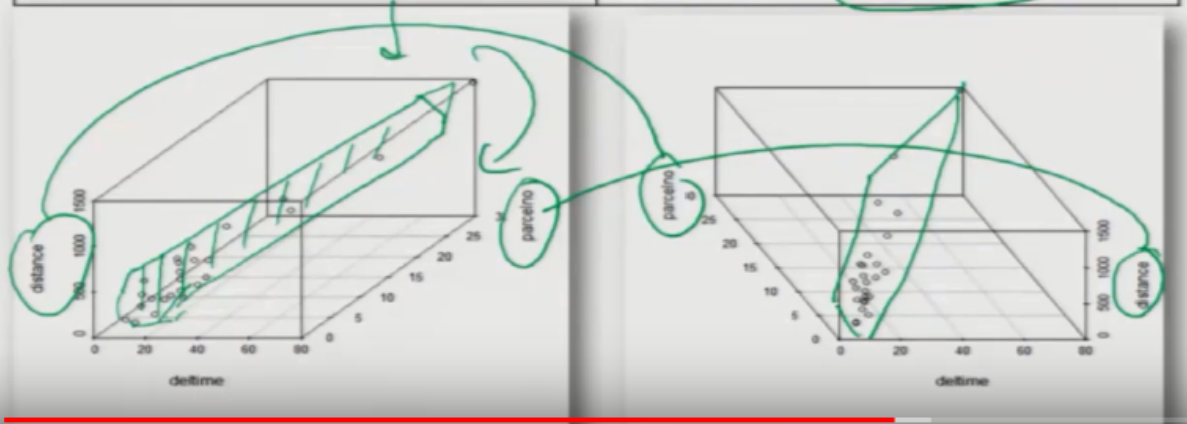
Refer Slide Time: (32:07)

Example with Two Variables

```
> library(scatterplot3d)
```

```
scatterplot3d(delttime, parcelno, distance)
```

```
scatterplot3d(delttime, parcelno, distance, angle=120)
```



So in this case I can take a call yes it is approximately linear well in this case I will try to show you one thing more that is since we are dealing here only with two variables and I want to see the joint effect of y with respect to x_1 and x_2 . So there are three variables y, x_1, x_2 so I can also take the help of this three dimensional plots and if you remember we had discussed earlier a three dimensional plot which is by using the command scatter plot 3d. So, we try to use it here, but it is always not possible because sometimes the data is more than three directions, but this is your judgment what you want to do. So if you try to use the command here the scatter plot 3d then first you need to upload the library so use this command library inside the argument a scatter plot 3d, and now I try to plot this scatter plot 3d with these three variables delivery time number of parcels and distance and this comes out to be like this. So, this is trying to give you here the sort of here panel, which is here in this case. So you can see here this panel is looking like yes! there is a that most of the points are lying close to the panel now you can also use the option to change the direction of this cuboid for example, I use here the option in the next command as angle equal to 120 and I try to rotate this figure by handed 20 degree. So you can see here this that the excesses are change these are change and this is see here changed, but now I looking at different structure now this is something like this so just by looking at different types of picture you can finally conclude whether you want to have a linear model fitted to this data or not.

Refer Slide Time: (34:07)

Example with Two Variables

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, 2, \dots, 25$$

```
>lm(delttime ~ parcelno + distance) + ... + ...  
      ~ x1 + x2 + x3 + x4
```

```
Call:  
lm(formula = deltime ~ parcelno + distance)
```

Coefficients:

```
(Intercept)  parcelno  distance  
2.19579      1.67803      0.01311
```

Model:

```
delttime = 2.196 + 1.68 * parcelno + 0.013 * distance
```

model with
2 input variables.

Now we have concluded on the basis of given set of data yes! we are confident that a linear model can be fitted so now we use the R command. So now you can see here we are interested in this model where I have two variables x_1 and x_2 and coefficients are α , β_1 and β_2 . So I need to estimate three parameters α , β_1 and β_2 so I will try to use the same command that we use earlier but now I am making here a small change I will use the same command `lm` now whatever is my here output variable I am trying to give it an a here as such, and now there is an equivalent science which is trying to indicate the formula that now the variables to be used as input variables are starting. So now I'm using here two variables named parcel number and distance so they are given in this format that they are separated by this plus sign so if you have more variable suppose if the model is $\beta_1 x_1$ plus $\beta_2 x_2$ plus $\beta_3 x_3$ or so plus $\beta_4 x_4$. So, all those variables will be added here something like say here x_1 plus x_2 plus x_3 plus x_4 and so on. So, this formula will remain the same and if you try to operate it on the R software you will get this type of outcome, so this is giving us the formula. Right? So, this is essentially writing that y is equal to something like $\beta_1 x_1$ and say here α plus $\beta_1 x_1$ plus $\beta_2 x_2$, and now this outcome has to be read like this that these three values are giving us the values of coefficient associated with these values. So, 2.19579 it is the value of intercept term 1.67803 this is the value of the coefficient associated with parcel number which is here actually $\hat{\beta}_1$, and this is the value of the coefficient associated with the variable distance. So, in our symbolic this is the value of $\hat{\beta}_2$, and yeah! this intercept term this 2.19579 this is the value of $\hat{\alpha}$.

So, my model becomes here del time this is the delivery time is equal to 2.196 plus 1.68 times parcel number, or the number of parcel plus .013 times distance. So, you can see here now we have obtained a model with two variable or model with two input variable and the same story you can continue with more than one variables, and I will try to show you these things on the R console also.

Refer Slide Time: (36:47)

Example with Two Variables

```
> deltime
[1] 16.68 11.50 12.03 14.88 13.75 18.11 8.00 17.83 79.24 21.50 40.33
[12] 21.00 13.50 19.75 24.00 29.00 16.35 19.00 9.50 35.10 17.90 52.32
[23] 18.75 19.83 10.75
> parcelno
[1] 7 3 3 4 6 7 2 7 30 5 16 10 4 6 9 10 6 7 3 17 10 26 9
[24] 8 4
> distance
[1] 560 220 340 80 150 330 110 210 1460 605 688 215 255
[14] 462 448 776 200 132 36 770 140 817 450 635 450
> lm(deltime~parcelno+distance)

Call:
lm(formula = deltime ~ parcelno + distance)

Coefficients:
(Intercept)      parcelno      distance
  2.19579         1.67803         0.01311
```

And you can see here this is the screenshot what I'm going to show you now.

Video Start Time: (36:50)

The screenshot shows a presentation slide with the following content:

Example with Two Variables

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, i = 1, 2, \dots, 25$$

deltime =
c(16.68, 11.5, 12.03, 14.88, 13.75, 18.11, 8, 17.83, 79.24, 21.5, 40.33, 21, 13.5, 19.75, 24, 29, 16.35, 19, 9.5, 35.1, 17.9, 52.32, 18.75, 19.83, 10.75)

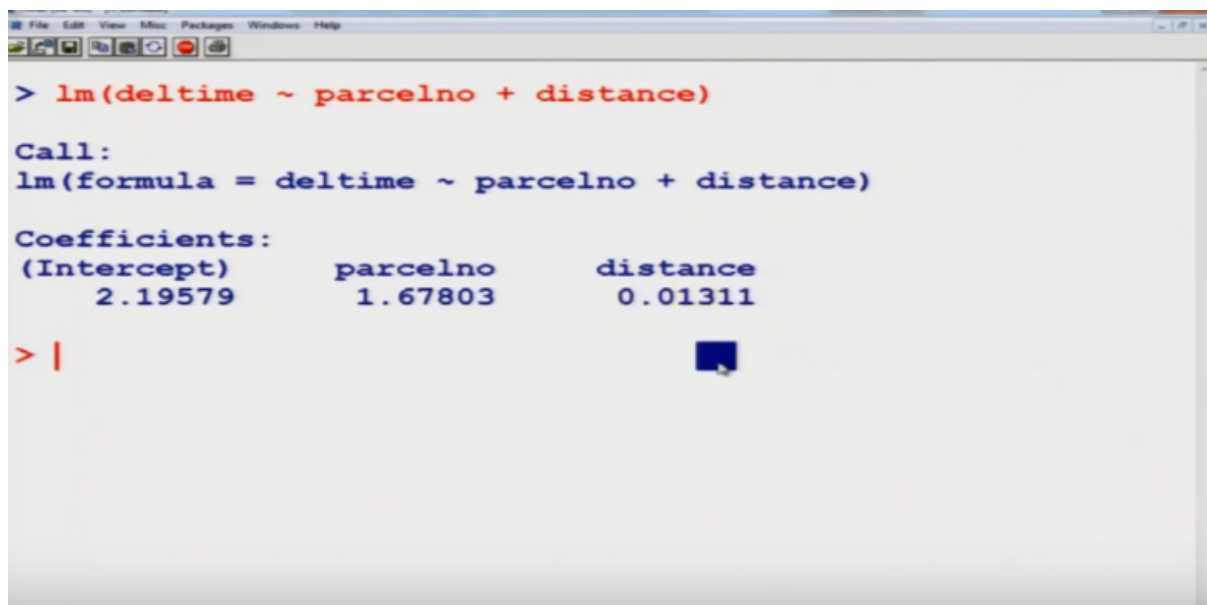
parcelno =
c(7, 3, 3, 4, 6, 7, 2, 7, 30, 5, 16, 10, 4, 6, 9, 10, 6, 7, 3, 17, 10, 26, 9, 8, 4)

distance =
c(560, 220, 340, 80, 150, 330, 110, 210, 1460, 605, 688, 215, 255, 462, 448, 776, 200, 132, 36, 770, 140, 817, 450, 635, 450)

So, let me first try to create this data vector here, so you can see here and then similarly I try to create the data vector on parcel number it is here, and similarly I try to create the data on distance table like this. So, you can see here this is my data del time like here like this parcel numbers number of parcel it is say like this and distance here like this. Now I first clear the screen so that you can now understand what is going to happen so I will first try to create here a three dimensional scatter plot and for that you know that first you need to create here, copy this command here so you can see here what I'm doing that I try to copy and paste this command and you get here this type of so this is the same

plot which I just explained you inside the slides, and now after this I try to obtain the scatterplot in three dimension so first I need to load the library here so I load the library of scatter plot 3d we are we already had installed this library earlier so this is already there because you need to install this package only once and after that whenever you want you just use the library. So you can see here this is the scatter 3d plot and if you try to change the angle here say angle is equal to suppose 120 so it will give us this change picture, and suppose if you try to make it here see angle equal to see her 90 so it will give you a different picture like this one. So, by using these things you can have an idea that what really you are going to do. Okay? Now next I try to fit here a linear model with this data set. Right? I close this pictures and I clear the screen by control L, and I paste the command over here and this gives me the this command so you can see here these are the values of the coefficients that you have obtained so this value 2.19 this is the least square estimate of intercept term this 1.67803 quickest is the least square estimate of the coefficient associated with the variable number of parcel, and this 0.01311 is the value of the least square coefficient or the value of the least square estimate of the coefficient associated with the variable distance travel.

Video End Time: (39:58)



```
> lm(delttime ~ parcelno + distance)

Call:
lm(formula = deltime ~ parcelno + distance)

Coefficients:
(Intercept)      parcelno      distance
  2.19579      1.67803      0.01311

> |
```

And this same screenshot has been given in the slides also. So now I would like to stop in this lecture I have tried to give you the idea of principle of least square and how to implement it inside the R software. Sometimes it is possible that there are some nonlinear curves, and in case if it is possible to make some transformation to change those nonlinear form into a linear form say by taking log or by taking exponential then you can use the command 'lm' to find out the least square estimate in that transformed linear model, that was originally nonlinear. But in that case you have to keep in mind that if the original data is given as y and in case if you transform the data into are the variable into log of y then the curve becomes linear then you need to input the data on log of y so these things say you have

to keep in mind, and using this technique you can estimate the least square estimate very easily using the R software. But definitely you will need more things to learn if you want to use it further, and which is usually not possible for me to, to do in the same course there is a usually there is a entire course on the topic of linear regression analysis. But here now we come to an end to this course and this was the last lecture of this course I hope you enjoyed it you understood it well I am saying that this is the end of the course but practically for you this is the beginning of the new course. I have given you only the basic fundamentals I have told you very basic things but believe me these are the things which are going to create the foundation for further things.

In case if you want to conduct a Monte Carlo simulation you want to find out a good model or you want to do any data mining or anything, the first step is the tools of descriptive statistics and even inside those things for example in data mining and other things you use the tools which we have discuss here under the course descriptive statistics, and you have seen that one tool will not give you the complete information about the entire data sets. There are different types of features which are hidden inside the data, data is so now that it is not telling you that I have this I have this I have this you are the only one who has to use different types of tools, and get the information from the data, and it is very important that you try to use the correct tool on the correct Decatur in case if you try to use anything else like as wrong data on correct tool or correct tool on wrong data you will not get a correct aesthetical outcome, and then people try to make different types of stories that aesthetics is not telling the truth, because they don't know what is the appropriate tool for an appropriate data. So my request is that please try to understand these topics in more detail before you try to apply them on a real data set definitely you need to also study a book which will give you more properties of these tools and more application of these tools before you become eligible to use the tools in a real situation, and in a real situation you cannot say that I can use only the means of central tendency or the variation or the Association all the tools can be applied to any data set but this is only you who is going to take a call take a correct decision that which of the tools have to be used in a given situation in their graphical or say analytical or a combination of them. So, you learn more in statistics and enjoy the course I will take a leave and see you sometime soon once again till then Goodbye, and God bless you.

