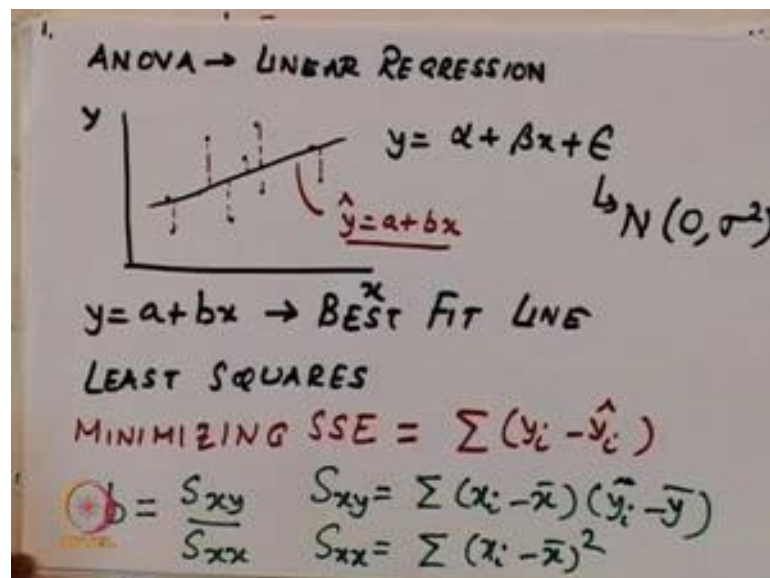


Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture – 40
ANOVA for linear regression, Block Design

Hello and welcome to today's lecture. So, we will continue today with what we had where we had left of an earlier class, which was discussing the applicability of ANOVA for linear regression, in linear regression.

(Refer Slide Time: 00:30)

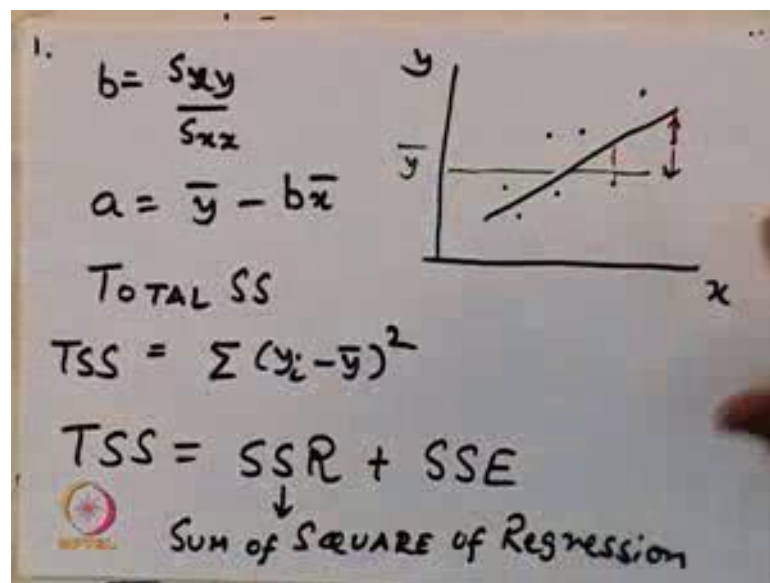


So, what is linear regression? You have data right and you wish to test whether there is a linear relationship between x and y, and you want to find out the expression of a line which best approximates that data. So, this data can be roughly written as alpha plus beta x plus some random error epsilon. So, this epsilon follows a normal distribution, with mean of 0 and variance of sigma square. So, it is as if that on a line you have added a random raider term at each point thereby generating this scatter. So, we want to generate a line y is equal to a plus b x, which is call the best fit line. And the approach is using the method of least squares approach. So, in this approach, what you do is for each of these points, you try to track the deviation of these errors.

So, this is nothing but, so this equation let us say if it is y is equal to a plus b x. So, you are in your least squares approach you are minimizing sum of squares of errors which is

given by the expression, summation of $y_i - \hat{y}_i$. What is \hat{y}_i ? \hat{y}_i is the expression is a linear equation of $a + b$. So, you want to minimize this particular term SSE and what you would do. So, what has what we have previously shown long back was using calculus you could determine the values of these coefficients a and b . So, b is given by S_{xy} / S_{xx} , where S_{xy} is given by summation of $(x_i - \bar{x})(y_i - \bar{y})$. And S_{xx} summation of $(x_i - \bar{x})^2$. So, note that from our earlier definition we had the divided by $n - 1$, which we have removed. So, this is the values of b and you can write down the value of a .

(Refer Slide Time: 03:42)

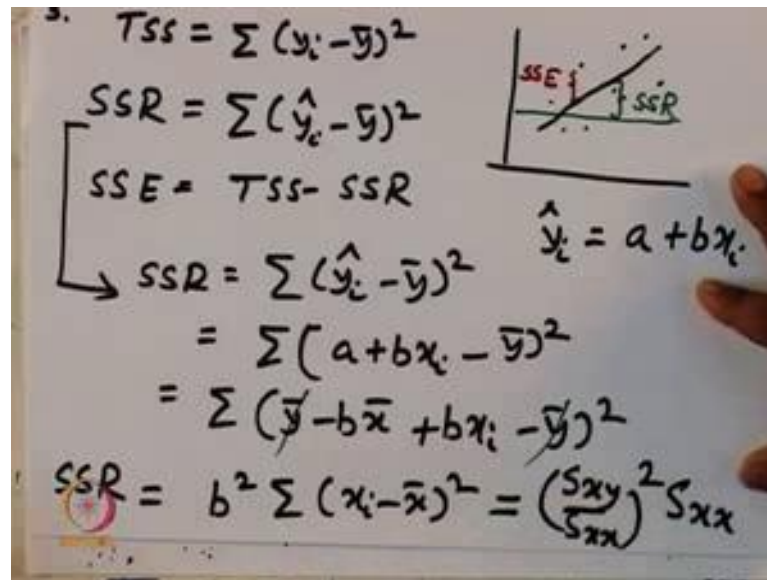


So, again if we write b is S_{xy} / S_{xx} . And a is given by $\bar{y} - b\bar{x}$. So, if this was true. So, my total SS total SS total sum of squares again let me draw. So, you have this line, and you have these points, is your x is your y . And let us assume that when you do your averaging, this is your \bar{y} value. So, your total sum of squares is going to be the deviation of these y values from \bar{y} . So, in ANOVA, this is your TSS value. So, you want to breakdown TSS into 2 components. So, TSS is component of sum of squares of regression.

So at each point, which is the sum of squares of regression? The sum of square of regression is essentially this component, your sum of square of regression is essentially this component, from this deviation what is the y term. This is your sum of squares of regression from this base. If the linear relationship did not exist, this as if imagining all

your data points where along the horizontal line then this point from this base, how much is the deviation of this line is your sum of squares of regression. So, how will I write down the expression for sum of squares of regression?

(Refer Slide Time: 05:58)



Handwritten mathematical derivations for TSS, SSR, and SSE, along with a small graph showing a regression line and data points.

$$TSS = \sum (y_i - \bar{y})^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$SSE = TSS - SSR$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

$$= \sum (a + b x_i - \bar{y})^2$$

$$= \sum (\bar{y} - b \bar{x} + b x_i - \bar{y})^2$$

$$SSR = b^2 \sum (x_i - \bar{x})^2 = \left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx}$$

The graph shows a scatter plot with a regression line. A vertical line from a data point to the regression line is labeled SSE (Sum of Squares Error). A vertical line from the regression line to the horizontal line representing the mean \bar{y} is labeled SSR (Sum of Squares Regression). The regression line is labeled $\hat{y}_i = a + b x_i$.

So, once again first let me write out TSS. TSS is summation of y_i minus \bar{y} whole square. SSR it is summation of \hat{y}_i minus \bar{y} whole square. So, once again this is your \bar{y} line, this is the line you fit and these are your points. So, this y_i what is this one at every point this difference is some of that is SSR. And SSE, SSE is this deviation from the line. So, SSE is given by TSS minus SSR. Now if we expand on this SSR expression. So, SSR is equal to summation of \hat{y}_i minus \bar{y} whole square. Now \hat{y}_i is given by the expression $\hat{y}_i = a + b x_i$.

So, if I plug in the values then what I get is $a + b x_i$ minus \bar{y} whole square. And \bar{y} so, we know that a is given by $\bar{y} - b \bar{x}$. We can add plus $b x_i$ minus \bar{y} whole square. So, your \bar{y} and \bar{y} terms cancel each other out, as a consequence of which you get SSR as b^2 summation of $(x_i - \bar{x})^2$. So, your SSR becomes b^2 summation of $(x_i - \bar{x})^2$. And so this you can rewrite as $\left(\frac{S_{xy}}{S_{xx}} \right)^2 S_{xx}$ is what is b times this is nothing, but S_{xy}^2 / S_{xx} .

(Refer Slide Time: 08:20)

4. $SSR = \frac{S_{xy}^2}{S_{xx}}$ $SSR = \sum (\hat{y} - b\bar{x} + b)^2 = b^2 \sum (x_i - \bar{x})^2$

$SSE = TSS - SSR$
 $= S_{yy} - \frac{S_{xy}^2}{S_{xx}}$

$df|_{TSS} = n - 1$ $df|_{SSE} = n - 2$

$df|_{SSR} = 1$ $MSE = s^2 = \frac{SSE}{n - 2}$
 \hookrightarrow unbiased estimator of σ^2

So finally, you can write SSR as S of x y square by S of xx. And your SSE is TSS minus SSR given by S yy minus S x y square by S xx. Now let us try to find out what is the degrees of freedom for each of those terms. So, degree of freedom of TSS, since they are in terms which contribute, so your TSS is summation by across from i equal to 1 to n. So, your df of TSS is n minus 1. What is of df of SSR? So, SSR has this expression which is b square into summation of x i minus x bar whole square right. So, all these x i and x bar are specified. So, you have only one degree of freedom coming from b. So, degree of freedom of SSR is equal to 1 which means that degree of freedom of SSE has to be n minus 2. And I can find out MSE which is equal to S square defined by SSE by n minus 2. So, this is the unbiased estimator of sigma square. This is the unbiased estimator of sigma square. So, if I write down the ANOVA table.

(Refer Slide Time: 10:11)

5. ANOVA

Source	df	SS	MS
Regression	1	$\frac{S_{xy}^2}{S_{xx}}$	$\frac{SSR}{1}$
Error	$n-2$	$S_{yy} - \frac{S_{xy}^2}{S_{xx}}$	$\frac{SSE}{n-2}$
Total	$n-1$		

TEST OF HYPOTHESIS

$H_0: \beta = \beta_0$

$H_a:$ ONE TAILED 2 TAILED TEST
 $\beta > \beta_0$ $\beta \neq \beta_0$
 $(\beta < \beta_0)$

So here ANOVA table should look something like this - source degree of freedom SS and MS. Your source is regression and error other 2 components of the total. So, regression has degree of freedom 1. Error has degree of freedom n minus 2 and the total is n minus 1. This is $S_{yy} - \frac{S_{xy}^2}{S_{xx}}$ this is SSR by 1. So, this is same as this SSR and this is SSE by n minus 2. So, if let us say we want to find out whether the linear regression is at all useful or not, we can write down the test of hypothesis, you can write down the test of hypothesis assign H_0 is $\beta = \beta_0$, let us say, or you can in general you can say $\beta = \beta_0$ and alternative hypothesis H_a , So, for a one tailed test.

You can write down as beta is either greater than beta naught or beta is less than beta naught, offer a 2 tail test you can find out beta as not equal to beta naught. So, for doing this we can use the students t test get to find out whether it is hypothesis is correct or not.

(Refer Slide Time: 12:17)

6. $t = \frac{b - \beta_0}{\sqrt{\text{MSE}/S_{xx}}} \rightarrow 't' \text{ Distr.}$
 $SE = \sqrt{\sigma^2/S_{xx}}$
 $t > t_{\alpha} \text{ or } |t| > t_{\alpha/2} \text{ or } t < -t_{\alpha/2}$

We can use the tth statistics given by the expression $b - \beta_0$ divided by the root of MSE by S_{xx} . So, this follows the t distribution, because your standard of error is root of sigma square by S_{xx} . And as we found that sigma square is the, your MSE is the unbiased estimate of sigma square. With this and once again as per the test once you calculate the t statistics, if t is greater than t alpha or t is greater than t alpha by 2. So, this is for the single tail test or t is less than minus t alpha by 2. You can find out the p value. And that would give us the final outcome whether your linear regression was at all useful or not. Now for the case of linear regression you can then find out how good is the approximation.

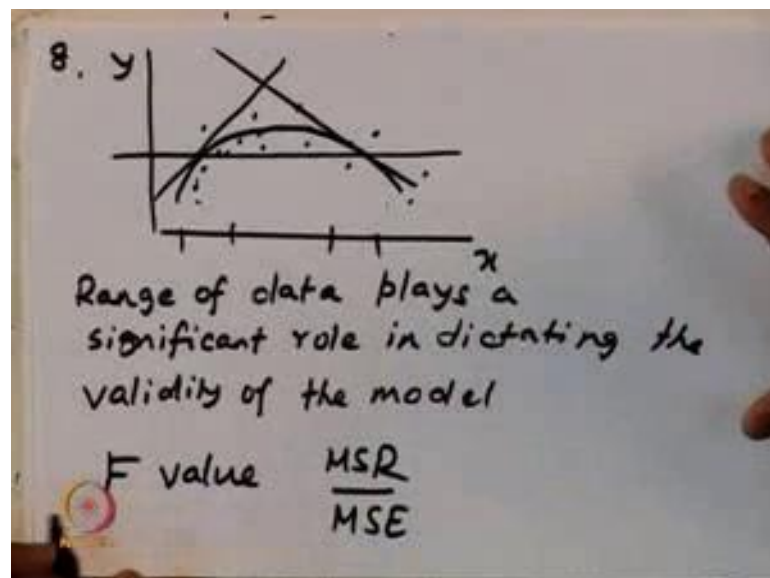
(Refer Slide Time: 13:36)

$$7. \quad r = \frac{S_{xy}}{s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$
$$\frac{SSR}{TSS} = \frac{S_{xy}^2 / S_{xx}}{S_{yy}} = r^2$$

$r^2 = \% \text{ Reduction in total variation obtained by using the regression line}$

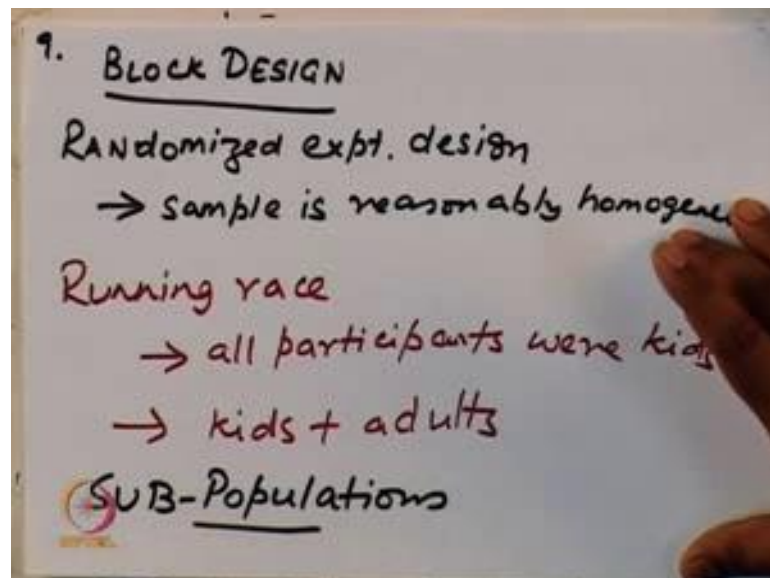
And we know that we use this term called r equal to S_{xy} by $S_{xx} s_y$ as a coefficient of determination. This in our current language can be written as r^2 . So, now, if you take the ratio SSR by TSS , what you get? You get S_{xy}^2 by S_{xx} divided by S_{yy} and what is this? This is nothing, but r^2 . So, you can say that r^2 is the percentage reduction in total variation obtained by using the regression line. So, r^2 is nothing, but square root of r^2 is nothing, but square root of SSR by TSS .

(Refer Slide Time: 15:00)



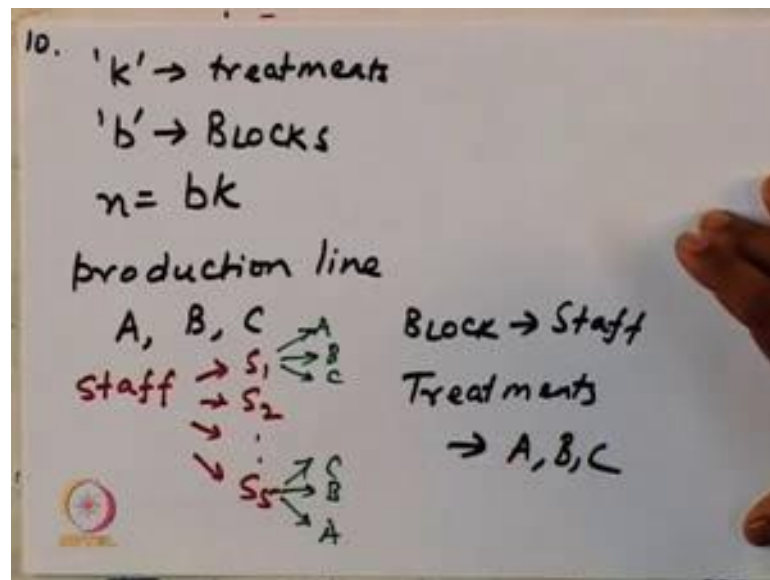
One other thing, so in the general case let us say imagine that x and y are not linearly related, but they are non-linearly related. Imagine that these are the points. So, let us say this is the fit and these are your points of x and y . So, if you were to choose only this domain, you would get a line which is something like this. If you have to choose a domain which is here you would get a line like this. And if you choose the middle you will get a line like this. So, your range of data plays a significant role in dictating the validity of the model which means that based on this if you are trying to estimate, if you are trying to extrapolate the data, you would incur huge amount of error. So, this line would predict a value here for this point, but as you can see that this is clearly not linear in nature. So, that concludes how you can make use of ANOVA, I forgot to mention one thing after you do all of this. You can calculate the F value and F value as MSR by MSE to determine whether your hypothesis of linear regression is good or not.

(Refer Slide Time: 16:51)



So, for the remaining time today I will discuss about one other concept which is the concept of block design. So, we need to randomize experimental design, this works very well when your sample is reasonably homogeneous. For example, if you take a running race, what you measure would be very different if all participants were kids versus participants had both kids and adults. So, this is what the aim of randomized block design is to do, is to take into account that you have various sub populations in your main population.

(Refer Slide Time: 18:18)



So, in randomized block design instead. So, you have k treatments you have. So, your entire population is divided into b blocks. So, for the previous cases you can imagine dividing your participants spaced on age. So, each age would be a block. So, they are b blocks and there are k treatment means or k treatments. So, there are n equal to b times k number of total sample size. So, if you take an example. So, imagine in a production line. So, the engineer is trying to compare between 3 methods for assembling a given product. So, what the engineer can do, he can choose he can choose staff where you can label them as S 1, S 2 dot dot S 5 and each of these each of these participants assembles using A B C A B C A B C, but in random order. So, in this case then you have your block or your staff and your treatments are A comma B comma C.

(Refer Slide Time: 20:06)

11. $TSS = SST + SSB + SSE$
 $TSS = \sum x_{ij}^2 - \frac{(\sum x_{ij})^2}{n}$ $n = bk$
 $SST = \sum \frac{T_i^2}{b} - CM$
 $SSB = \sum \frac{B_j^2}{k} - CM$
 $df|_{TSS} = bk - 1$ $df|_{SSB} = b - 1$
 $df|_{SST} = k - 1$ $\Rightarrow df|_{SSE} = (b-1)(k-1)$

So, in this approach you are TSS. So, earlier you would write TSS equal to SST plus SSE, but here we have one more time which is SSB. Sum of squares of blocks, so I can write the TSS as before the summation of x_{ij} square minus summation of x_{ij} whole square by n , where n is equal to b times k . You can write down the expression of SST is equal to T_i square by b minus CM . So, this is also called CM . So, T_i square minus b there are b blocks for each block you have these many treatments. So, you average over that and SSB is given by summation of B_j square by k where k is the number of blocks.

So, SSB becomes B_j square minus CM . So, the degrees of freedom for TSS is equal to b times k minus 1, the degrees of freedom of SST is equal to k minus 1, and degrees of freedom of SSB is equal to b minus 1. This would imply that df of SSE is b minus 1 times k minus 1. So, let us take a sample example and construct the ANOVA table for that.

(Refer Slide Time: 21:54)

12. Pricing plan by cellphone companies for users of diff. ranges (usages)

USAGE LEVEL	A	B	C	D	Totals
Low	27	24	31	23	$B_1 = 105$
Moderate	68	76	65	67	$B_2 = 276$
High	308	326	312	300	$B_3 = 1246$
Total T_t	403	426	408	390	$G = 1627$

So we had discussed the possibility of a pricing plan by cell phone companies for users of different ranges or usages. So, imagine there are 4 companies. A B C and D, and as per the usage level can be low moderate or high. And let us say this is you have sample rates at which they charge.

So as before you can calculate the totals, so you can calculate T 1 as 403, T 2 as 426 T 3 as 408 T 4 as 390. And your G is the sum total of all these values which comes out to be 1627. But what you can also calculate here. So, you can also make another panel called totals here and this is B. So, this is a block average B 1 is 105, B 2 is 276, B 3 is 1246. So, this gives you, so you have 3 blocks as per usage levels, low moderate and high and 4 treatments or which are your companies and their pricing strategies.

(Refer Slide Time: 24:16)

Low	27	24	31	23	$B_1 = 105$ $B_2 = 276$ $B_3 = 1246$
Moderate	68	76	65	67	
High	308	326	312	300	
Total T_j	403	426	408	390	$G = 1627$

$$CM = \frac{T_1^2 + T_2^2 + T_3^2 + T_4^2}{3} - CM$$

$$SSB = \frac{B_1^2 + B_2^2 + B_3^2}{4} - CM$$

So, for this you can ask. So, your B, so your B is a number of blocks is 3 k is a number of companies A B C D is 4 n is 12. So, you have this total combination of 12 observations. You can calculate the CM which is given by G square by n. So, g we calculated the global the grand total of all observations. So, we can write down as 1627 square by 12. Sum of squares is just is all these squares put together. So, you take the squares of all the observations. You can calculate SST. So, SST is T 1 square plus T 2 square plus T 3 square plus T 4 square by 3, minus CM. This is T 1 is 403 so on and so forth. You can calculate SSB which is given by, SSB is these values right you have B 1 square plus B 2 square plus B 3 square by 4 minus CM and based on these values.

(Refer Slide Time: 25:55)

14. ANOVA

SOURCE	df	SS	MS
USAGE	2	189335	94667.6
COMPANY	3	222	74.1
Error	6	242	40.2
Total	11	189799	

Test for equality of treatments
 H_0 : no difference in 'k' treatment means
 H_a : no " " " 'b' block means

You can calculate the following ANOVA table.

Your source you have 2 treatments now, usage and company. Then you have error, and then the total. This is 2 this is 3 this is 6 this is 11. So, in calculate the values 5. So, based on this you can then prove the test of hypothesis for equality of treatments. So, your H_0 is either no difference in k treatment means or you can also have no difference in b block means you can have these 2 cases.

(Refer Slide Time: 27:57)

15.

$$MSE = \frac{SSE}{(b-1)(k-1)}$$

→ unbiased estimate of σ^2

$$F_T = \frac{MST}{MSE} = 1.84$$

↓
 H_0 is true

$$F_B = \frac{MSB}{MSE} = 2357.99$$

↓
 $H_0 \neq$ true

So, you can calculate your MSE as $SSE / (b - 1)$ into $k - 1$, as an unbiased estimate of sigma square. And you calculate the f statistics which is given by mean square of treatments by mean square of error for calculating your hypothesis for treatments or this is F of T or F of B is given by MSB / MSE .

See if you calculate the value of F of T this comes out to be 1.84 and F of B comes out to be 2357.99. So, this would say that you are H_0 is true for treatments your H_0 is true; that means, that there is no difference in the pricing strategy used by the different companies. But what you can clearly see every value is way higher than whatever confidence interval α you choose. So, this really means that H_0 is not true.

With that I conclude my session today. So, what you saw was, but we discussed was how ANOVA can be used to test the whether linear regression can at all be used for fitting your experimental data. And then towards end we discussed about how you can make use of block design or randomized block design for quantifying or probing analysis of variance where you have multiple subpopulations which exist in the population.

Thank you for your attention.