Business Statistics
Prof. M. K. Barua
Department of Management Studies
Indian Institute of Technology - Roorkee

Lecture – 60
Examples of Multiple Regression

Hello friends, I welcome you all in this session, as you are aware in previous session we were discussing about multiple regressions where in one of the independent variable was categorical, right, so let us look at one more example.

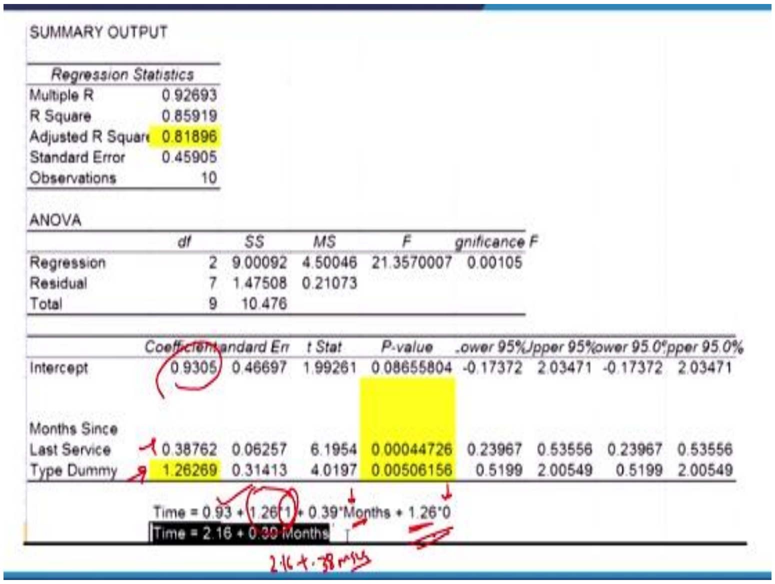
(Refer Slide Time: 00:50)

Type of Repair	Repair Time (hours)	Months Since Last Service	Type Dummy
elec	2.9	2	1
mech	3.0	6	0
elec	4.8	8	1
mech	1.8	3	0
elec	2.9	2	1
elec	4.9	7	1
mech	4.2	9	0
mech	4.8	8	0
elec	4.4	4	1
elec	4.5	6	1

So, we know that the repair time is a function of service since last or the time since last service, okay and we did find out a regression equation but it is not the months since last service affects the repair time but you need to consider what type of work it is. Whether it is electrical work or mechanical work, so this dummy variable are this independent categorical variable also affects repair time.

So, we have got a dummy variable here, it has got 2 levels right, 2 levels; mechanical work and electrical work, so you can 3, 4, 5, 6, 7, 8 different levels, right, so let us consider this dummy variable into account and see how regression equation looks like.

(Refer Slide Time: 02:00)



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.92693
R Square	0.85919
Adjusted R Square	0.81896
Standard Error	0.45905
Observations	10

ANOVA

	df	SS	MS	F	gnificance F
Regression	2	9.00092	4.50046	21.3570007	0.00105
Residual	7	1.47508	0.21073		
Total	9	10.476			

	Coefficient	andard Err	t Stat	P-value	.ower 95%	Jpper 95%	ower 95.0%	pper 95.0%
Intercept	0.9305	0.46697	1.99261	0.08655804	-0.17372	2.03471	-0.17372	2.03471
Months Since Last Service	0.38762	0.06257	6.1954	0.00044726	0.23967	0.53556	0.23967	0.53556
Type Dummy	1.26269	0.31413	4.0197	0.00506156	0.5199	2.00549	0.5199	2.00549

Time = 0.93 + 1.26*1 + 0.39*Months + 1.26*0

Time = 2.16 + 0.39 Months

2.16 + .38 mths

So, this is the output when you take dummy variable is one of the independent variables, right, so the adjusted R square is 81% approximately, 82% then the model is significant. you can see here the intercept coefficient is this, so y =.9305+.3876 months since last service, right +1.26 dummy variable, right, this is our dummy variable so now, dummy variable what we have done here is; if it is electrical work we are coding it as 1, otherwise it is 0, right.

So, if you put hour here 1, it means electrical work, so electrical work will take this much more time, is not it compared to mechanical work, so this is our question right, so just see this, .93, 1.26 * one electrical work, .39 in two months since last service +1.26 * 0, right so if it is let us say mechanical work, we will not include this particular part of the equation, if it is electrical work, this would automatically become 0, right.

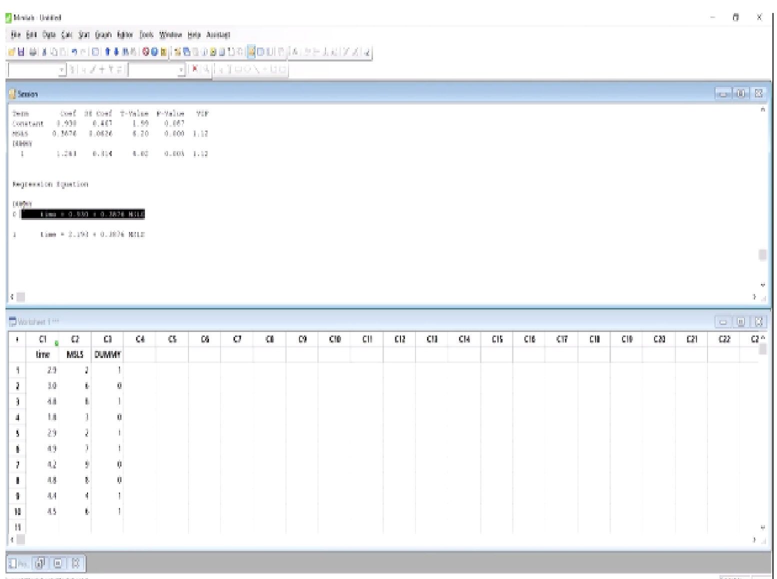
So, at the end of the day, you can solve this equation and it comes out to be it is 2.16 + 2.16 +.38 months; months since last service, right. So, what I am trying to say is that you need to consider categorical variables because they also affect dependent variable.

(Refer Slide Time: 04:16)

```
Analysis of Variance
Source          DF   Adj SS    Adj MS   F-Value  P-Value
Regression       2   9.0009   4.50046    21.36    0.001
  C2             1   8.0883   8.08826    38.38    0.000
  C3             1   3.4049   3.40489    16.16    0.005
Error            7   1.4751   0.21073
  Lack-of-Fit    6   1.4751   0.24585        *        *
  Pure Error     1   0.0000   0.00000
Total            9  10.4760
Model Summary
       S    R-sq  R-sq(adj)  R-sq(pred)
0.459048  85.92%     81.90%      71.61%
Coefficients
Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant    0.930    0.467     1.99    0.087
C2         0.3876   0.0626     6.20    0.000  1.12
C3          1.263    0.314     4.02    0.005  1.12
Regression Equation
C1 = 0.930 + 0.3876 C2 + 1.263 C3
```

Same equation, in fact if you look at these 2 equations, you can solve in fact this question just like any other multiple regression equation, so you will have a dependent variable, so this is these 2 are independent variables, right, so C2 is just this month since last service and this is the your categorical variable 1.26, right, now this can be either 0 or 1.So, let us solve this question using multiple regression and taking C1 means, x1 and x2 as independent variables, so first we will solve it, right, let us see what is the output, right.

(Refer Slide Time: 05:10)



So, here you have got your time; repair time months since last service, right and your type of work whether it is, you just write it dummy, right okay. so just enter the repair time 2.93, 4.8, 1.8, 2.9, 4.9, 4.2, 4.8, 4.4 and 4.5 months since last service now, dummy variable, just go up, this is second number value, you need to go off; yeah, so 1, 0, so 1 is electrical work, right and 0 is mechanical work.

So, first of all we will solve this question as we will take this variable as a regular predictor or independent variable, regular means metric right, so let us go to regression fit; fit regression models, so responses time, right and we will take both of these as continuous predictor months since last service and dummy, yeah okay so, this is your output, right just go down, yeah, so this what is your answer.

So, let us check whether we are getting .38 and 1.2, just see, same answer we are getting, right, .9 3, .38, 1.26 into dummy variable right but this is not the appropriate way of doing this question in fact, we need to take this variable as a categorical variable not as a metric variable, so what you should do is; we will go back to stat in fact that you can delete this one first, okay, now stat, regression, fit regression model.

And this dummy variable in fact that should over here, categorical predictor, it is a categorical predictor right, okay now, this is the correct answer, so look at this so, the time; repair time if it is let us say, if it is electrical work then you just enter month since last service over here, let us take these value as 10 for example months since last service, so this becomes 3.8, so 3.8+2.2 almost 6, is not it?

But if it is let us say mechanical work then what will happen and this is 10 again, right, so this 3.8 so, it is approximately 4.7, so that is the difference, so the total time required; total repair time required is a function of categorical independent variable as well, it is not only the metric independent variable, right, so this is how you should solve a question wherein you are given dummy activity or dummy independent variable.

In fact, it is not the dummy independent variable, the independent variable is categorised as dummy variable, right.

(Refer Slide Time: 09:38)

```
Analysis of Variance
Source           DF   Adj SS    Adj MS   F-Value   P-Value
Regression        2   9.0009   4.50046     21.36     0.001
  MSLS            1   8.0883   8.08826     38.38     0.000
  DV              1   3.4049   3.40489     16.16     0.005
Error             7   1.4751   0.21073
  Lack-of-Fit     6   1.4751   0.24585         *         *
  Pure Error      1   0.0000   0.00000
Total             9  10.4760
Model Summary
       S     R-sq   R-sq(adj)   R-sq(pred)
0.459048   85.92%      81.90%       71.61%
Coefficients
Term          Coef   SE Coef   T-Value   P-Value    VIF
Constant     0.930     0.467      1.99     0.087
MSLS        0.3876    0.0626      6.20     0.000   1.12
DV
  1          1.263     0.314      4.02     0.005   1.12
Regression Equation
DV
0    RT = 0.930 + 0.3876 MSLS

1    RT = 2.193 + 0.3876 MSLS
```

So, let us look at solution we obtained, is not it, so the repair time is it is this, if it is mechanical work otherwise this if it is electrical work, right.

(Refer Slide Time: 09:50)

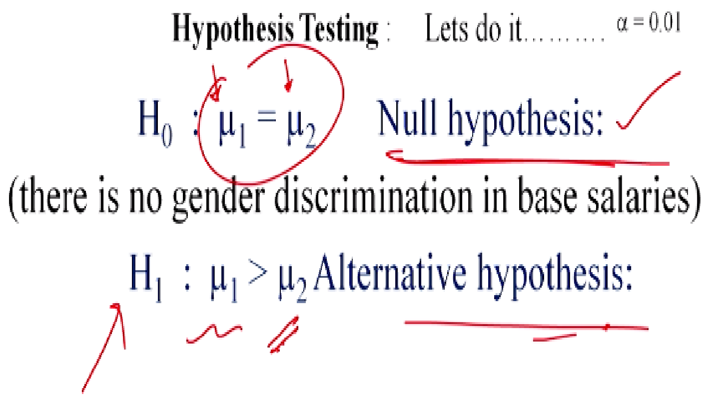
Salesmen		Saleswomen	
Months Employed	Base Salary ($ 1000s)	Months Employed	Base Salary ($ 1000s)
6	7.5	5	6.2
10	8.6	13	8.7
12	9.1	15	9.4
18	10.3	21	9.8
30	13.0		

Do women get less salary ?

Let us as look at one more interesting question; in a company there are sales men and sales women, so sales women are complaining that they are getting lower salary compared to salesmen, so they are complaining that there is a gender discrimination and because of that sales women are getting less salary compared to salesmen, so this is the question and the management decided to collect data on their salaries, right.

And number of months for which these people are working in the organisation, so months employed for men and their salaries, months employed for women and their salaries, right, so the question is do women gets less salary or do you think that women are being discriminated on the basis of gender, so this is the question, now how to proceed for this question, the first thing which we should come to your mind is that you should compare the means of these two, right means of their salaries, right.

(Refer Slide Time: 11:25)



Let us look at and let us try to solve this question by comparing their salaries. So what we will say initially that µ1 is the salary of men, µ2 is mean of salary of women, right. So initially, we will say that there one and be same thing and we will say that there is no gender discrimination in null hypothesis, alternative hypothesis, men get more salary than women, right, so it is upper tailed test, right, so one tailed upper tail test, right.

(Refer Slide Time: 11:59)

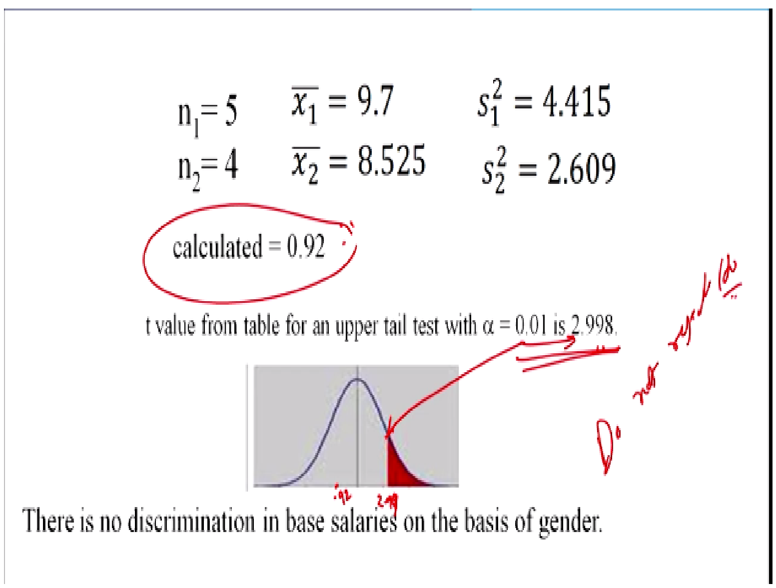
$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$\widehat{\sigma_{\overline{x_1} - \overline{x_2}}} = S_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sigma_{\overline{x_1} - \overline{x_2}}}$$

So, let us solve this question using our regular process, calculate pool variance calculate t statistics.

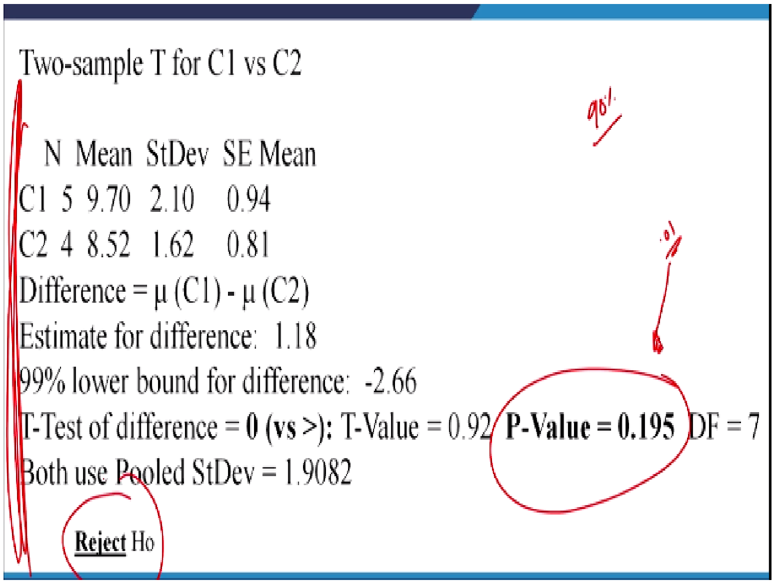
(Refer Slide Time: 12:02)



And then this is, what is your t statistics, you will be getting somewhere here, anyway, so this is what; is your calculated t statistics and the t value from table is this, right. So this is 2.98; 2.998, this point is this, 2.998 and this is .92 which is what your calculated right. So you are not rejecting null hypothesis, is not it, so you do not reject null hypothesis, right, when you do not reject null hypothesis, it means they are same, is not it.

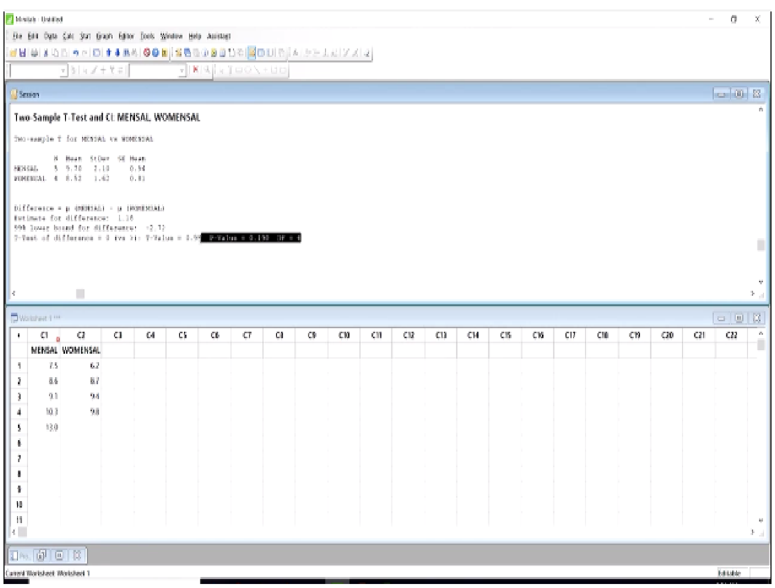
So, this is one way of solving this question but the point is in this method we have not considered gender or the independent variable which has got 2 levels; men and women, right, so let us move on to this question once again.

(Refer Slide Time: 13:19)



So, let us find out what is the; in fact, this is the output of this question using Minitab in fact, you can get, so p value is when you test this hypothesis at let us say 90% significance level, so this is .01, right, so this value is more than .01, so we will not reject null hypothesis, in fact you can solve this question using Minitab.

(Refer Slide Time: 13:48)

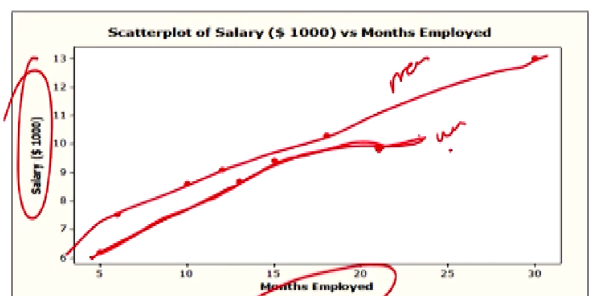


Let us solve this question using Minitab, so we will enter salaries of men and women, right, so men and women salary, right, so 7.5, 8.6, 9.1, 10.3, 13, 6.2, 8.7 and like this you are supposed to enter remaining values, right, so we will go to stat, we will go to basic statistics, then it is a 2 sample means stats, right, so this is the one, so both samples are not in one column but we are in 2 different columns, right, so this is simple one, yeah one column, so men, sample 1, just see like this, similarly for women salary options we set 99% significance level.

So, this is your alternative hypothesis has to be greater than type, right, this is greater than type, okay, yeah, so this p value is .190, right, this is what I have shown over here, it is .195, right, okay, so we will reject null hypothesis right and we will say that there is no discrimination based on gender or their salaries are same or equal, right. Now, let us look at the one more approach to this question.

(Refer Slide Time: 15:59)

Salesmen		Saleswomen	
Months Employed	Base Salary ($ 1000s)	Months Employed	Base Salary ($ 1000s)
6	7.5	5	6.2
10	8.6	13	8.7
12	9.1	15	9.4
18	10.3	21	9.8
30	13.0		



Let us find out relationship between months employed and salary so generally, we say that the higher the salary if the months employed are more, right so, higher the months of employment higher the salary, right. so you can have this as a dependent variable and this as the independent variable, okay and when you plot months and salaries, you can find out these points, so there are total 9 points, right.

So, 5 for men and for 4 for women, right, so if you look at these points, then you should just see this, this is the salary of men, all these points and this is these points are salary of women, now this plot tells you many things, is not it. So you can see that there is some difference, right just scatter plot is giving you some information that there is something wrong, they are not same, right.

(Refer Slide Time: 17:34)

Regression Analysis: Salary ($ 1000) versus Months Employed

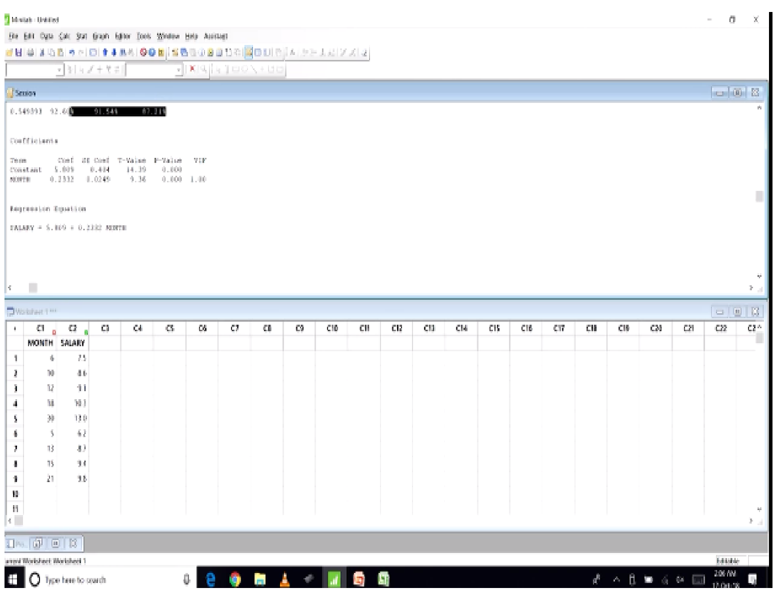
The regression equation is
Salary ($ 1000) = 5.81 + 0.233 Months Employed

Predictor	Coef	S E Coef	t-ratio	P
Constant	5.8093	0.4038	14.39	0.000
Months Employed	0.23320	0.02492	9.36	0.000

S = 0.549393 R-Sq = 92.6% R-Sq(adj) = 91.5%

So, what you are doing is let us find out the effect of months employed on salary, so you can just do it for these two variables; dependent variable salary and independent variable months employed, right okay.

(Refer Slide Time: 17:52)



So, we will just solve this question, so with salary now, okay. We can delete these columns in fact, you have just, so these are salaries now for women, this is months employed, right, so 6, yeah 6, 10. So we will just enter all these data as far as months employees are concern, we will have just regression equation, so regression, so salary is dependent variable or response, right, so C2 is response and C1 is continuous predictor, right, okay.

So, this is 5.8, this is the equation, right, so this is the estimation line which you can see here, right, 5.8 and .233 months employed, right, R square 92.6, here as well, R square, 92.6, right, this value, okay.

(Refer Slide Time: 19:48)

Table of residuals

S.No.	Salary	Fits1	Resi 1
1	7.5	7.2085	0.2914
2	8.6	8.1413	0.4586
3	9.1	8.6077	0.4922
4	10.3	10.0069	0.2930
5	13	12.8054	0.1946
6	6.2	6.9753	-0.7752
7	8.7	8.8409	-0.1409
8	9.4	9.3037	0.0926
9	9.8	10.7066	-0.9065

Salary ($ 1000) = 5.81 + 0.233 Months Employed = 5.81+0.233 (6) = 7.20

Any special observation

Now, if you look at the table of residuals for this question, then you need to look at the residuals. let us find out residuals for this question, so this is your equation, 5.8+.23 months employed if you put let us say independent variable 6 over here right, months employed 6 because this is the months employed. So one months employed is 6, this is your estimated value or y hat value, 7.2 right.

So, this is what is your fitted value, right, 7.2. Similarly, if you put let us say in this equation instead of six, if you put 10, right. so that would be the fitted value, right, 8.1 now, you need to find out residual which is difference of these two. so 7.5 minus is this, 8.6 minus this is this, these are residuals, if you look at residuals then all these are positive, all these are; almost all these are negative.

So, there is a pattern, so residuals are not randomly distributed, right so this is a special observation in residual, right, so it means there is something is missing in this question, right.

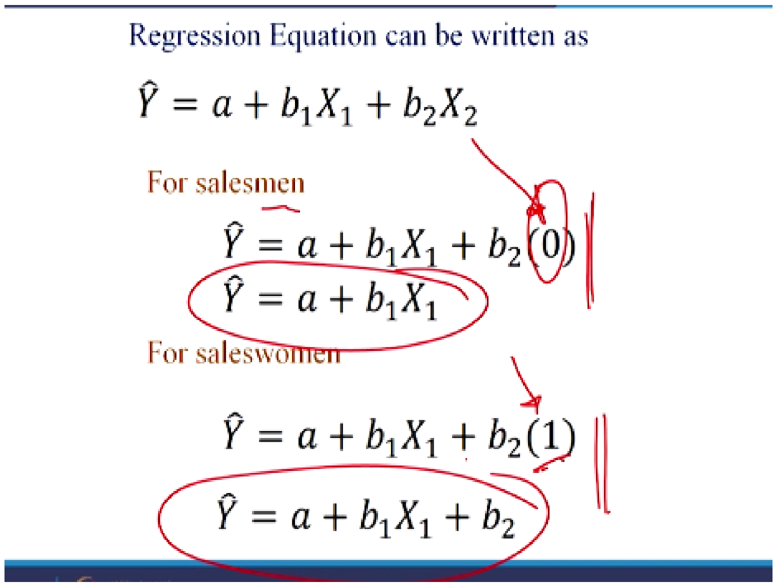
(Refer Slide Time: 21:42)

Input data for gender discrimination regression

Gender	X1	X2	Y
	Months employed	Gender code	Base Salary($ 1000)
Men	6	0	7.5
	10	0	8.6
	12	0	9.1
	18	0	10.3
	30	0	13
Women	5	1	6.2
	13	1	8.7
	15	1	9.4
	21	1	9.8

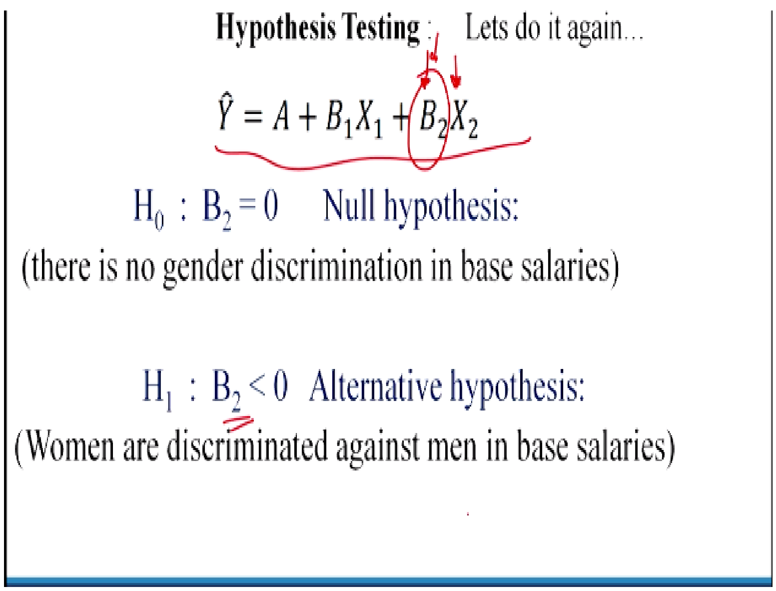
So that can now be further analysed, we will take gender as one of the independent variables, right. so months employed gender and let us call all these minus 0 or give them as code = 0 and to women 1, right, it is up to you, you can have 1 over here and 0 over here there won't be any change in the solution except the sign of the coefficient, right. So, let us look at the solution to this part of the question.

(Refer Slide Time: 22:19)



So, this is how you can again rewrite equation, so for salesmen this x2 = 0, right, we are taking x2 = 0 and for women x2 = 1, right, so this is the final equation for salesmen, this finally equation for sales women, right.

(Refer Slide Time: 22:35)



So, let us now reframed null hypothesis and we will say that let us say if in this equation, if this is negative, the coefficient of the B2 is negative. It means because this is a coefficient for women right for independent variable x2 not for women, right. So initially, we will say that this is zero, the moment we say it is zero, it means x2 is whether it is men or women, it is not affecting the salary, right but if B2 is negative, it means women are getting less salary, okay.

(Refer Slide Time: 23:25)

Regression Analysis: Base Salary(versus Months employed, Gender code

The regression equation is

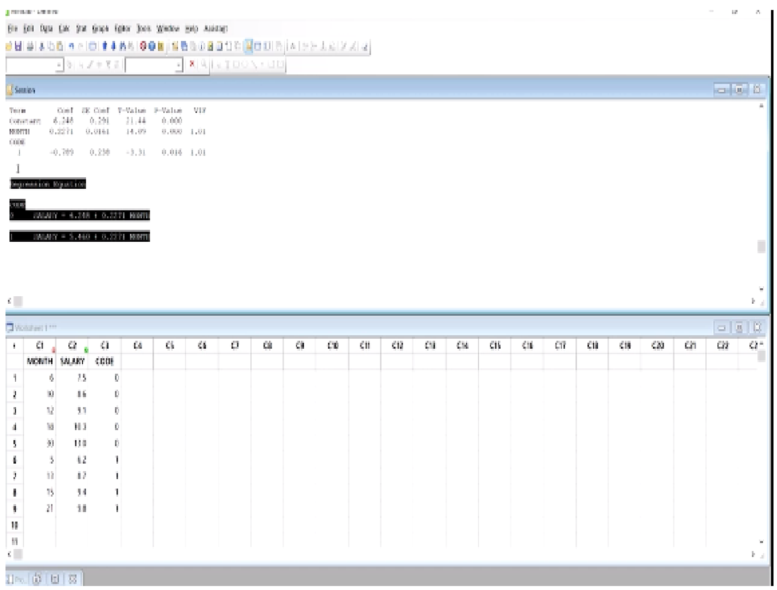
Base Salary($1000)Y = 6.25 + 0.227 Months employed(X1) - 0.789 Gender code(X2)

Predictor	Coef	SE Coef	t-ratio
Constant	6.2485	0.2915	21.44
Months employed(X1)	0.22707	0.01612	14.09
Gender code(X2)	-0.7890	0.2384	-3.31

S = 0.353037 R-Sq = 97.4% R-Sq(adj) = 96.5%

So, let us solve this equation and when we solve this equation, we will get this output, so let us solve this equation using Minitab.

(Refer Slide Time: 23:37)

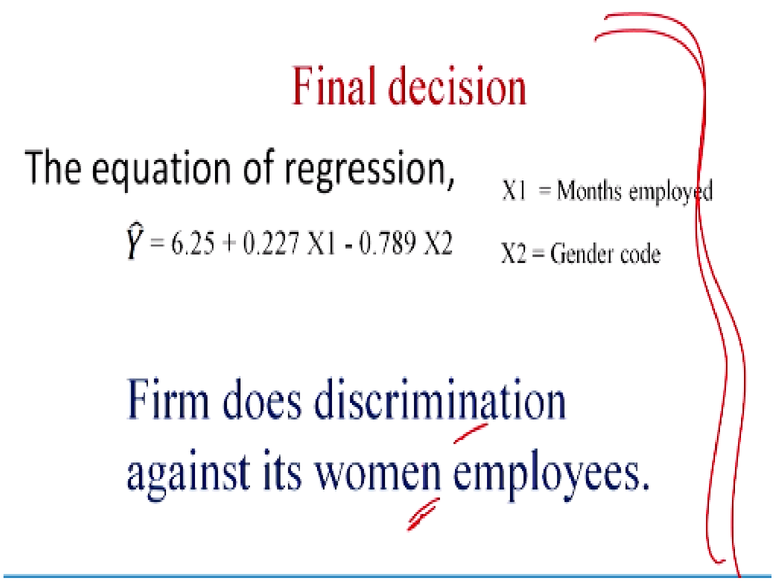


So, you can do coding over here. So this is for a men, this is for women, right, so regression fitted line, yeah. So initially, yeah so salary is of course month is this; in fact, we will take this code as year already. let us take code as continuous predictor for the time being okay, so this is the answer right, so you are getting this B value as negative one, right, so it means there is a discrimination taking place against women, right.

So, 6.24, .22 and .789, this is what the answer we were getting here, right, is not it but let us look at the other way wherein we will take this independent variable as a categorical independent variables. So we will go to stats fit regression model, so this code is now rather than this we will put it categorical predictor, right. So this is the code now, we have said that month since last employed if I take it for example, 10, right.

So, this becomes 2.27 and this is 6.24 so, it is total it is 9, right, salary is 9, but if it is women, right. Then if this is 10, then this is 2.27, then this 2.2768, so salary would be 8 and in this case it would be 9, so there is clear-cut case of discrimination against women as far as salaries are concerned, right. So this is the right way of solving question, if you have got coded independent variable, right.

(Refer Slide Time: 26:11)



So, let us look at one more question, so in this case, the final decision which we will take is that firm does discriminate against its women employees, right.

(Refer Slide Time: 26:25)

Transforming Variables and fitting curves

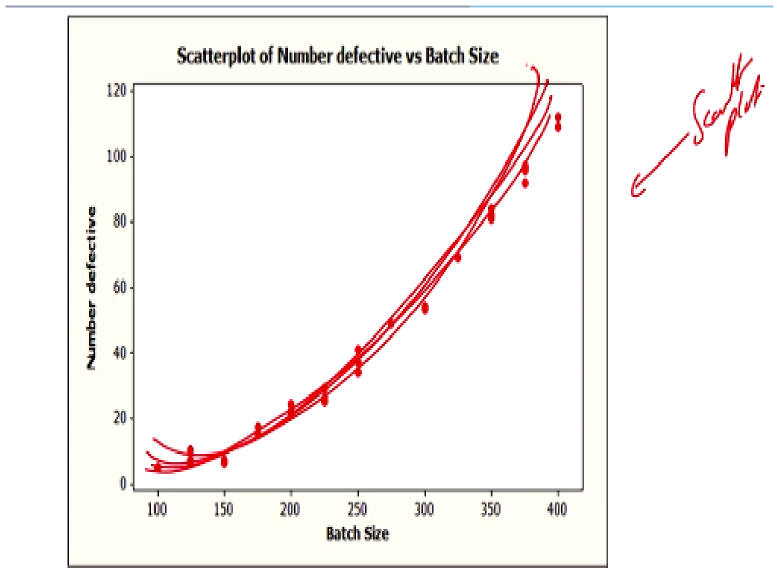
Now, let us look at one more case of modelling of regression, let us say is called transforming variables in fitting curves, so let us say are you are receiving raw materials from your vendors, you are manufacturing something and you receive raw material in terms of batches, so in different batch sizes, right.

(Refer Slide Time: 26:46)

Batch Size	Number defective
100	5
125	10
125	6
125	7
150	6
150	7
175	17
175	15
200	24
200	21
200	22
225	26
225	29
225	25
250	34
250	37
250	41
250	34
275	49
300	53
300	54
325	69
350	82
350	81
350	84
375	92
375	96
375	97
400	109
400	112

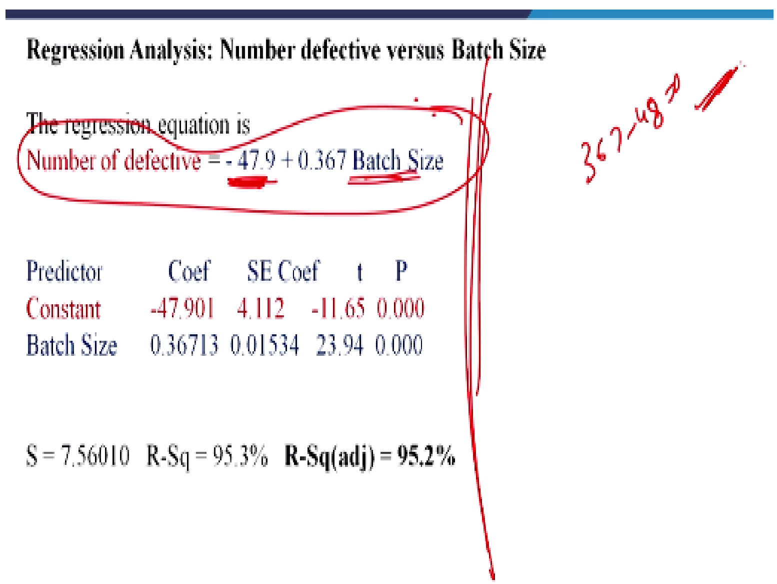
So, when batch size is 100, you will always get 5 defectives, right, so you have collected data over above batch size and number of defectives, so you want to know is there any relationship between the batch size and number of defectives.

(Refer Slide Time: 27:04)



So, first of all you should draw scatterplot, this is your scatter plot, is not it, so you can see over here, there is a non-linear relationship right.

(Refer Slide Time: 27:24)



So, let us find out this relationship, when you solve this equation, this is how your line up estimation or regression equation looks, it is -47.9, .36* batch size, right, so if batch size is let us say 1000, right, so this becomes 367, right – 47 or let us say 48, so this becomes whatever it is, right, so that would be number of defectives.

(Refer Slide Time: 28:00)

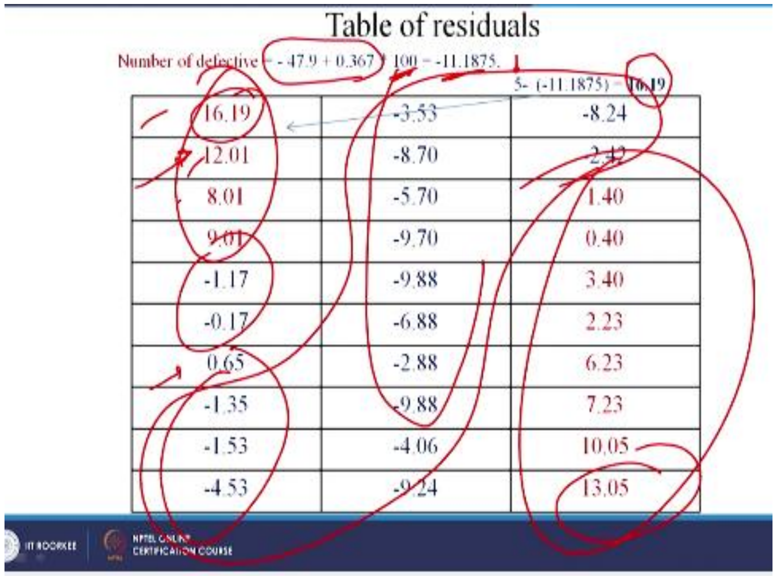


Table of residuals

Number of defective = - 47.9 + 0.367 * 100 = -11.1875.

5- (-11.1875) = 16.19

16.19	-3.53	-8.24
12.01	-8.70	2.42
8.01	-5.70	1.40
9.01	-9.70	0.40
-1.17	-9.88	3.40
-0.17	-6.88	2.23
0.65	-2.88	6.23
-1.35	-9.88	7.23
-1.53	-4.06	10.05
-4.53	-9.24	13.05

Now, when you look at the residuals for this equation, so this is your regression equation what you calculated, if batch size is 100, initially in equation itself. When batch size is 100, let us find out what is its fitted value, so fitted values is this one, okay. So and this was your number of defectives 5, so 5 -; - this becomes 16.9, so this is for residual, similarly if you put the second independent variable which is what; 125.

So, when you put 125, you will get some value over here and then number of defective minus that value, this would be the second residual and so on, so this first, second, third so, this is the last residual, right. So if you look at these residuals carefully, you will find that all these are positive, negative, one positive, then all these are negative, just see, yeah in fact up to this point, these many negative and these positives right.

So, there is a pattern of residuals, some of them are positive, some of them are negative, they are not randomly distributed it means there is some problem with the model, right. So what we try to do is; just by looking at this plot, we will say that this is not a linear relationship, this is a non-linear relationship. So what we do; we will take the square of batch size which is our independent variable.

So, 100 square = 10,000, this square is equal to this and so on, right, so let us now run a regression between number of defective, batch size and batch sizes square.

(Refer Slide Time: 30:10)

The regression equation is
Defects = - 7.02 + 0.000721 Batch size square

Predictor	Coef	SE Coef	T	P
Constant	-7.0203	0.9883	-7.10	0.000
Batch size square	0.00072119	0.00001153	62.56	0.000

S = 2.95196 R-Sq = 99.3% R-Sq(adj) = 99.3%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	34101	34101	3913.31	0.000
Residual Error	28	244	9		
Total	29	34345			

So, this is the batch sized square, so this is R square or let us look at adjusted R square, so 99.3 is the model fit earlier what it was; when we did not take square, it was 95% right, now it is 99, just see how much improvement has taken place in the model, right. so you if you take let us say batch size square and batch size both then what happens?

(Refer Slide Time: 30:43)

The regression equation is
Number defective = 6.90 - 0.120 Batch Size + 0.000950 (Batch Size)2

Predictor	Coef	SE Coef	T	P
Constant	6.898	3.737	1.85	0.076
Batch Size	-0.12010	0.03148	-3.82	0.001
(Batch Size)2	0.00094954	0.00006059	15.67	0.000

S = 2.42306 R-Sq = 99.5% R-Sq(adj) = 99.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	34186	17093	2911.35	0.000
Residual Error	27	159	6		
Total	29	34345			

This is 99.5, earlier it was 99.3, so this is even a better model. So you can fit a model depending upon the values of independent and dependent variables, with this we come to an end of this particular course whatever syllabus which was designed for this course with this lecture I am completing the last topic, if you have got any questions or queries, you can always approach me, thank you.