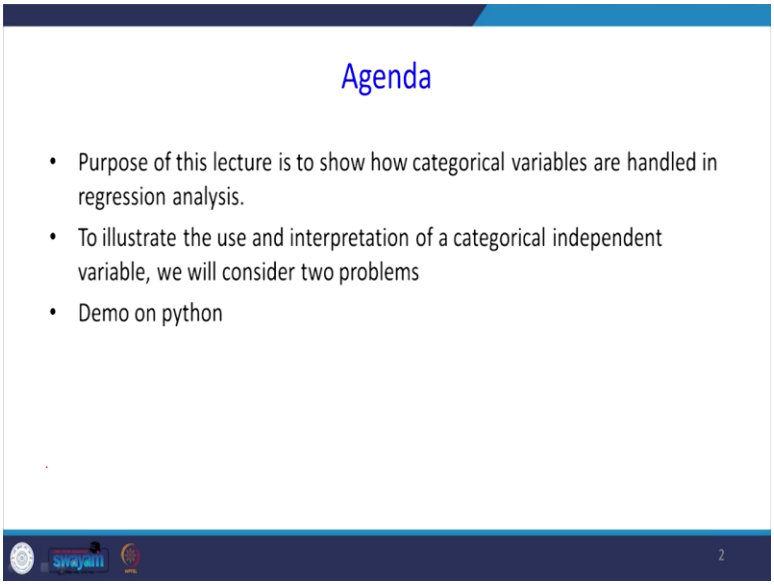
Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 35
Categorical Variable Regression

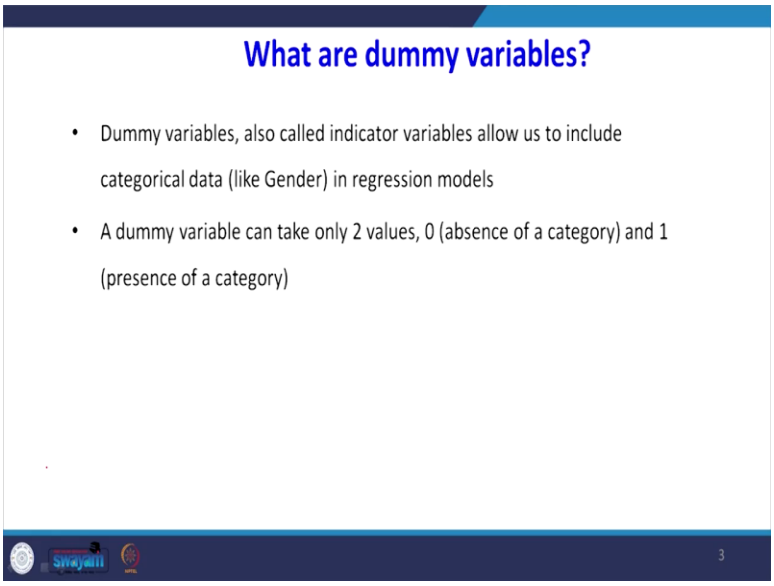
Dear students, in this lecture, we will see how to handle Categorical Variable linear regression analysis. Whenever we do a linear regression analysis, the assumption is the nature of independent and dependent variable has to be continuous variable. Sometimes what will happen we have to include the categorical variable into independent variable category? How to handle that kind of regression analysis that we will see in this class?

(Refer Slide Time: 00:57)



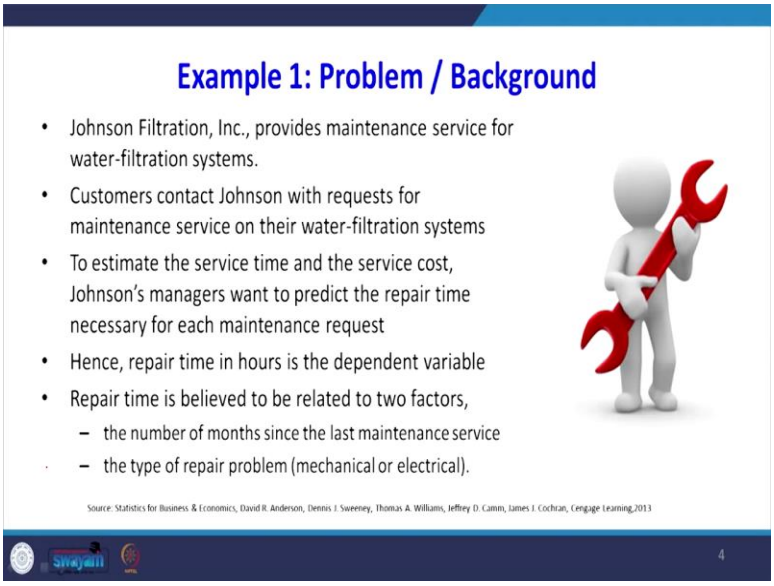
The Agenda of this lecturer is, to show how categorical variable are handled in regression analysis. Illustrate and will interpret how to do the categorical independent regression analysis. The same problem we will do in Python will explain how to code and how to do this categorical regression in Python programming.

(Refer Slide Time: 01:17)



Another name for categorical variable is called dummy variable dummy variable also called indicator variable. It allows us to include categorical nature in regression analysis. For example, gender is one of a categorical data where there is only two levels are possible male or female. If dummy variable can take only two values, when it is general category, for example, zero means absence of category and one means the presence of category. Here zero will be taken as their reference. With respect to zero, we will compare what will happen to another level of the categoricall variable.

(Refer Slide Time: 01:51)

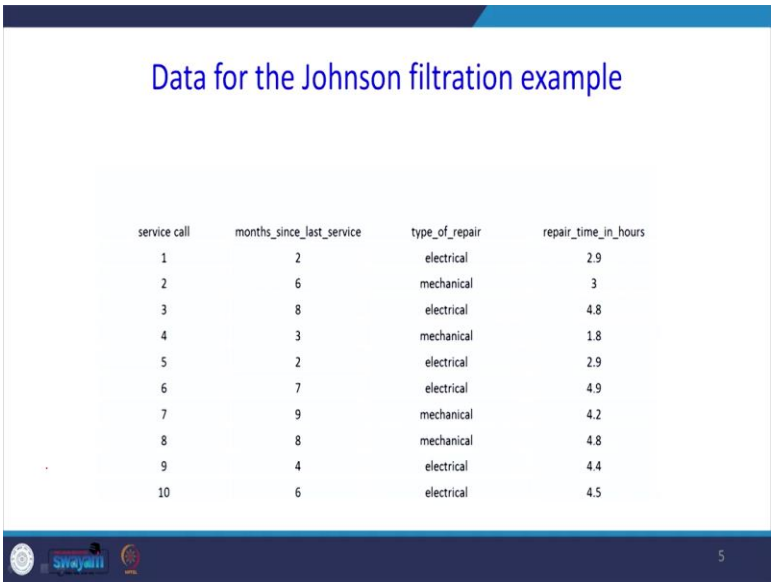


We will take a problem with the help of problem I will explain how to use categorical variable into the regression analysis and how to interpret it. This problem is taken from statistics for

Business and Economics from David Anderson Sweeney and Williams. It is Syncage Publication in 2003 to 2013 edition. Johnson filtration Incorporation provides maintenance service for water filtration systems.

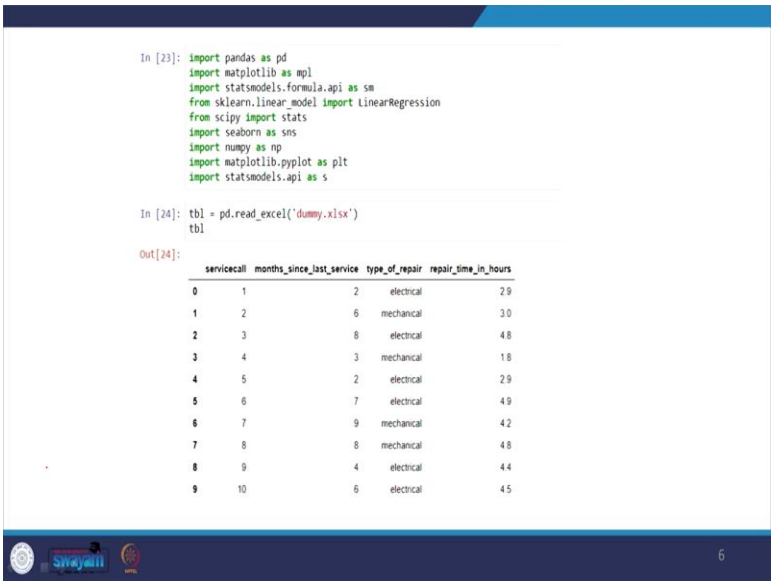
Customers contact Johnson's with a request for maintenance service on their water filtration system. To estimate the service time and the service cost Johnson's managers want to predict the repair time necessary for each maintenance request. Hence, the repair time in hours is the dependent variable. Repair time is believed to be related to two factors. One factor is number of months since the last maintenance service was done; second factor is the type of repair problem. Here the type of repair problem, mechanical or electrical is the categorical variable.

(Refer Slide Time: 02:53)



This is the given data. What is there is a Column 1 is the service call, the column 2 says months since the last service was done, in terms of month. Column 3 says the type of repair whether it is the repairs with respect to electrical system or mechanical system. The last column is repair time in hours. How much time it is taken for doing repairing?

(Refer Slide Time: 03:18)



```
In [23]: import pandas as pd
         import matplotlib as mpl
         import statsmodels.formula.api as sm
         from sklearn.linear_model import LinearRegression
         from scipy import stats
         import seaborn as sns
         import numpy as np
         import matplotlib.pyplot as plt
         import statsmodels.api as s

In [24]: tbl = pd.read_excel('dummy.xlsx')
         tbl
```

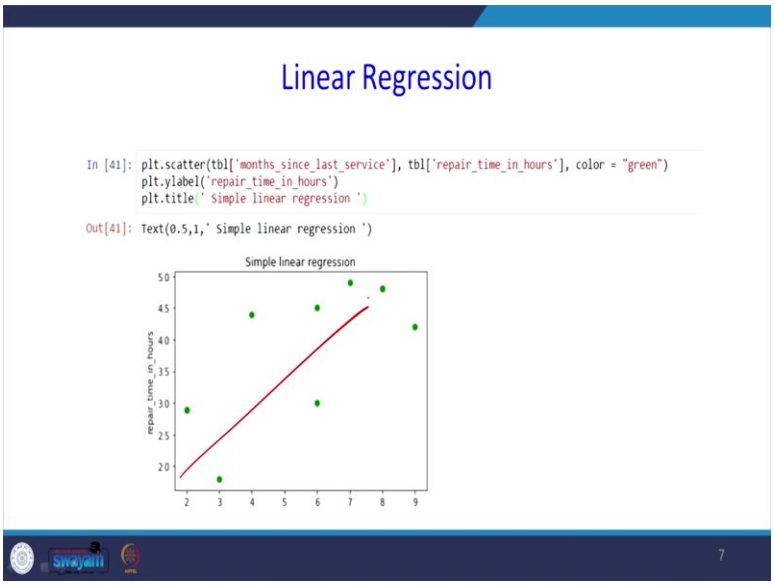
Out[24]:

	servicecall	months_since_last_service	type_of_repair	repair_time_in_hours
0	1	2	electrical	2.9
1	2	6	mechanical	3.0
2	3	8	electrical	4.8
3	4	3	mechanical	1.8
4	5	2	electrical	2.9
5	6	7	electrical	4.9
6	7	9	mechanical	4.2
7	8	8	mechanical	4.8
8	9	4	electrical	4.4
9	10	6	electrical	4.5

I have taken the screenshot of our python code get so I have to import necessary libraries like import Pandas as pd, import matplotlib as mpl, import statsmodel dot formula dot api as sm from sklearn linear underscore model import linear regression from Sci-Fi import stats import seaborn sns, import numpy as np, import matplotlib.pyplot as plt, import statsmodels dot api as s. First we will load this regression file it is a data file I have saved in the name of dummy dot xlsx that we are going to save any object called tv1.

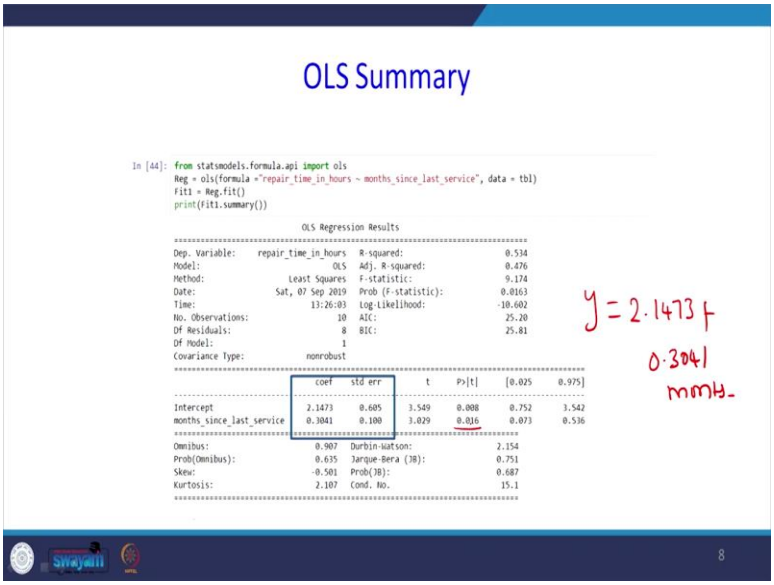
When you execute this one, we can see this is a data file. At the end of the class going to give the demo for this what are the codes which have done in it. There also we can understand the steps. Here this is the data, display the data.

(Refer Slide Time: 04:10)



But first will do the scatter plot between the months since last service and repair time in hours. When we look at this cat replied you see that there seems to be positive trend because when the month since last services more the repair time in hours also getting more. This is a simple linear regression considering only one independent variable. Here independent variable is continuous variable.

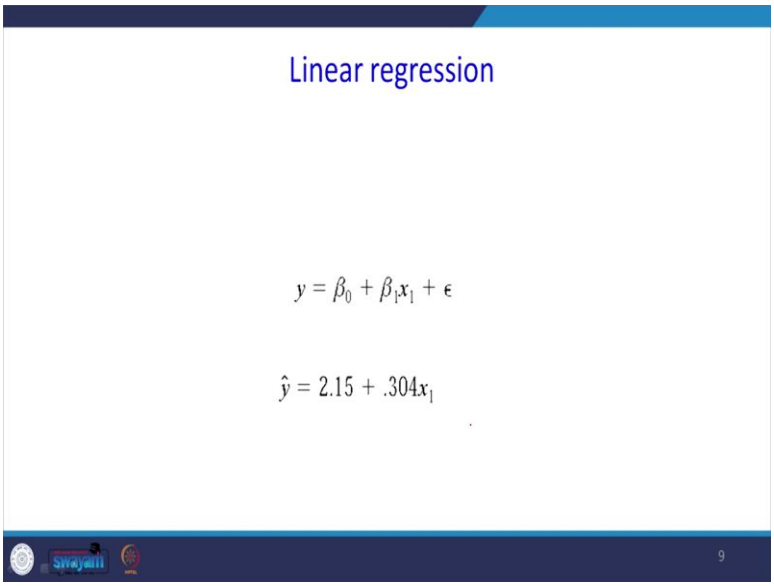
(Refer Slide Time: 04:41)



When you do the regression analysis, this is output of python. So, from stats model that formula dot AP import ols, ols is used for doing regression analysis. Here, the dependent variable is repair underscore time in hours tilde sign independent variables months since last service. When you look at this series, why intercept I can write Y equal to 2.1473 + 0.3041 because this is

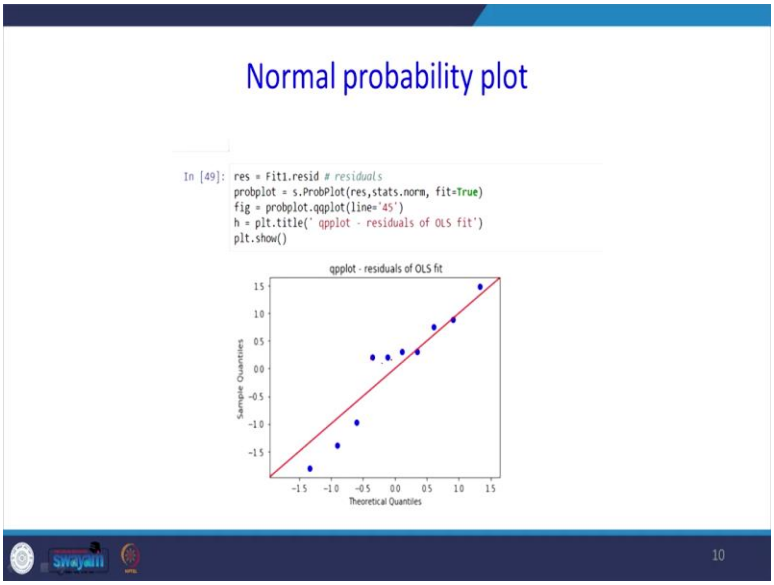
independent variables months since the last service was done. Look at the R square. R square is 53.4 % look at the P value of this independent variable here. Here, it is significant because it is less than 0.05. Now what we are going to do residual plots for this problem?

(Refer Slide Time: 05:38)



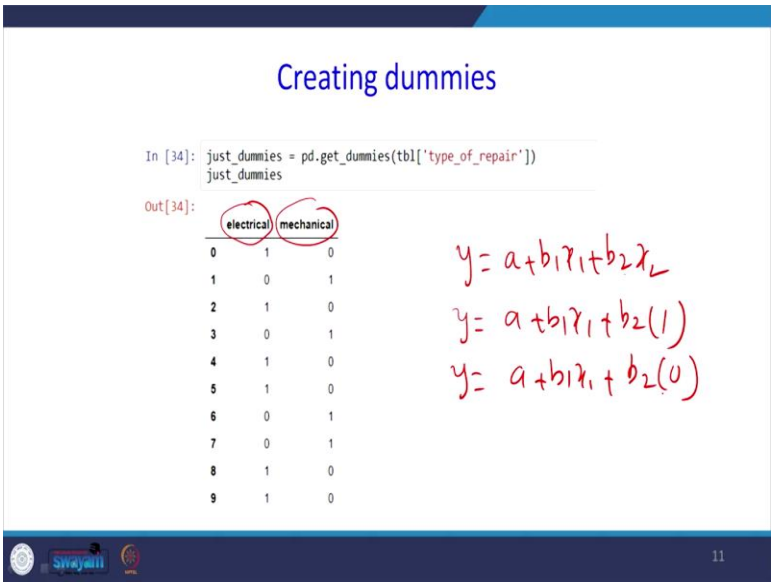
You look at this is 2.15. + 0.304 X1 is our regression model.

(Refer Slide Time: 05:47)



When we use this to do regression model, When you do the normal probability plot, look at this it has to all the probability points has to align with is red point. What is happening is there are so many points it is away from the red line. So, we can say that even if the, the residual plot is not appropriate, so the data, the error is not following normal distribution.

(Refer Slide Time: 06:10)



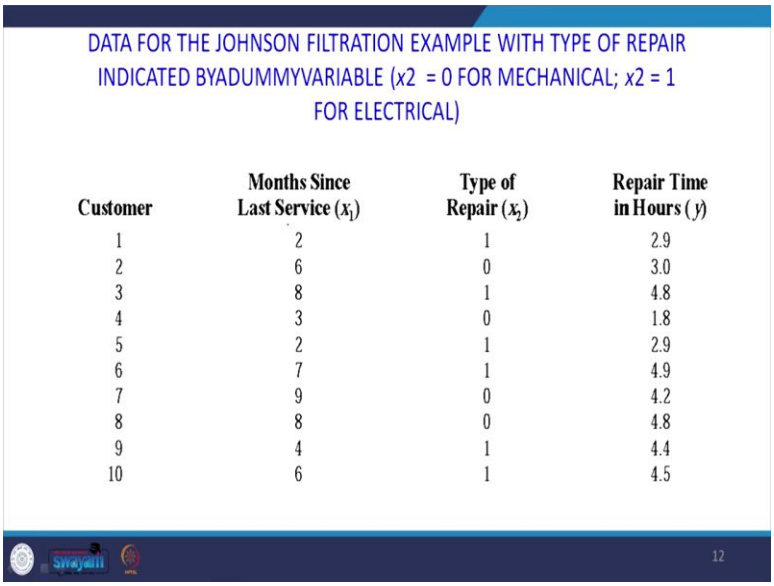
First, we will create a dummy variable for the categorical data. How to create a dummy variable for this categorical data? so that new dummy variable I going to call it is just underscore dummies equal to pd.get underscore dummies where the filename which column has to be converted into Dummies. So, the type of repair that is the value where we have written, whether the problem is related to mechanical or electrical.

So, when we display the just dummies see that that one variable is know, it is taken into 2 parts. One is for Electrical so, the presence of one says electrical; the absence of one says mechanical. There are 2 columns is there which is the dummy variable. So what happened both are same whether we can use this variable interval into our new regression model or this variable for Our new regression model, if you take electrical equal to 1.

So the equation be written as Y equal to a + b1 x1 + b 2 x 2. Here, X 1 is independent variable. The b2 value will be 1 if it is suppose we write if the problem the, this is the common regression equation. In this regression equation, when you substitute x2 equal to one that equation for Electrical problem related to electrical repair a + b 1 x 1 + b 21 this equation for repair due to electrical problem. Instead of this Y equal to a + b 1 x 1 + b 20 this is what problem related to mechanical. You can reverse also, no problem.

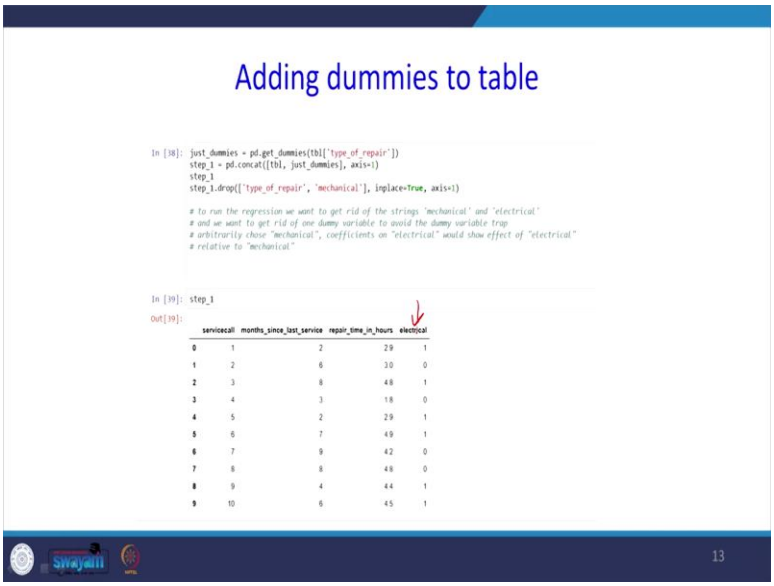
Mechanical can be taken as 1 and electrical can be taken as zero. There will not be problem in the interpretation.

(Refer Slide Time: 08:06)



This was the data which we have converted into dummy variable. Month since last service 1 represents problem related to electrical Zero represents problem related to mechanical. This was our Y is our dependent variable.

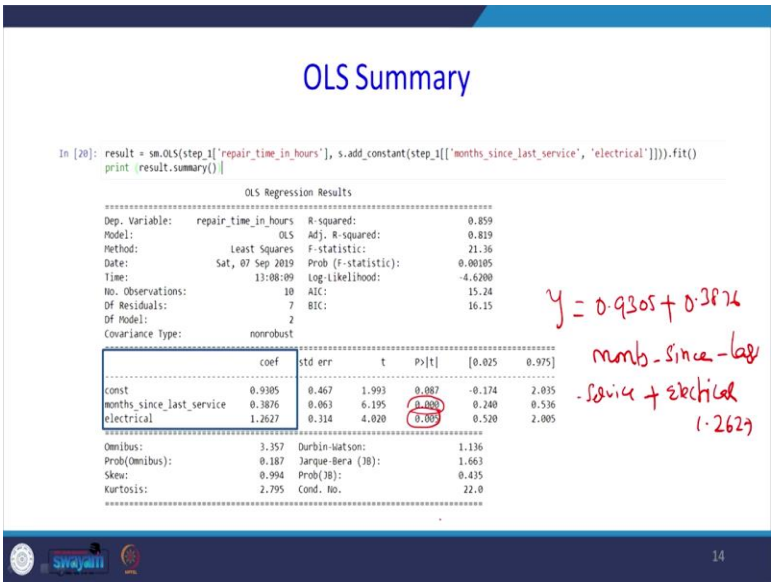
(Refer Slide Time: 08:22)



When you do the regression analysis, see that just underscore dummy is pd.get underscore dummies p1 is a type of repair. So here what I have done? I have displayed, I have dropped the certain columns what column I have dropped, I have dropped the column that is type of repair

then I have added only dummy variable with respect to the electrical repair. That is why this column has come. So, now this is going to this is the last column that is under electrical heading. It is going to be taken as independent variable that will do the regression analysis.

(Refer Slide Time: 08:54)

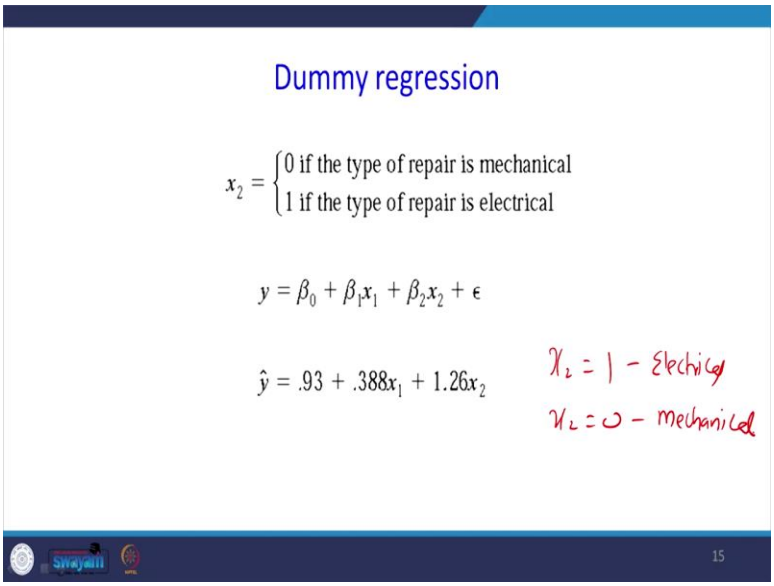


Result equal to sm.ols step underscore one repair underscore time underscore hours is taken as a dependent variable. Months underscore since underscore last underscore service taken as independent variable. So this electrical is taken as reference because that column where one means electric repair zero means mechanical repair. When you look at this, you see this equation can be written as Y equal to 0.9305 Plus months since last underscore service Seiko vision for this one is 0.3876 + electrical 1.2627.

So look at R square it is 0.85 previously, the R square was when there is only one independent variable I m going back previously asked for R is only for 0.534 when we introduce another variable what has happened, the R square is increased to 0.859. So, F statistics corresponding probability the p-value is very low 0.005 so as a whole this regression model is significant. When we look at the individual independent variable, for example, months underscore since there is independent variable 1, the P value is less than 0.01.

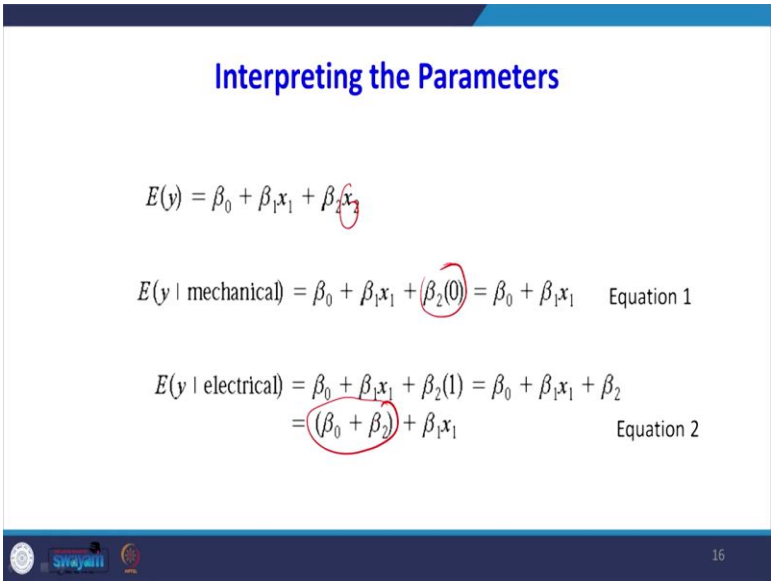
So we can say this variable is significant. Similarly, for the second one the type of repair, where electrical is taken as the reference this also less than 0.05, so, this is also a significant variable.

(Refer Slide Time: 10:42)



Now, this is the regression equation Y hat equal to 0.93 + 0.388 X 1 + 1.26 x2. If x2 equal to 1 means electrical if I say x2 equal to one it is related problem related to electrical if x2 be 0 it is related to mechanical.

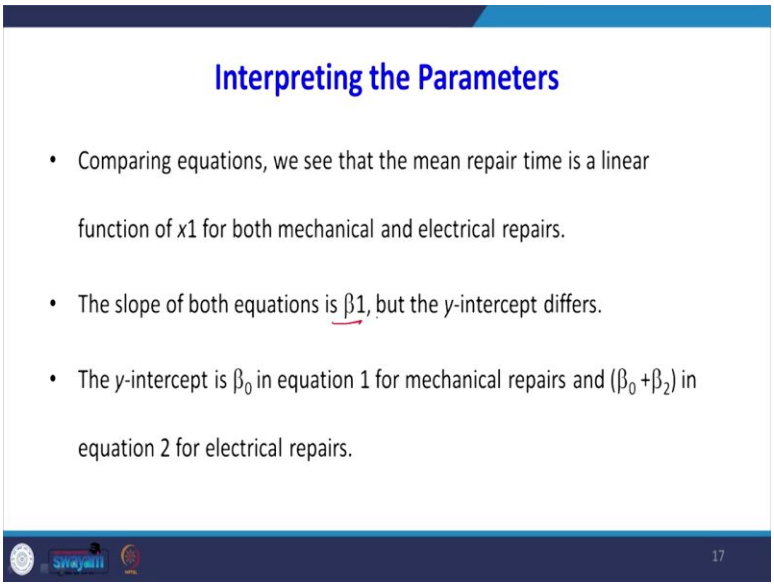
(Refer Slide Time: 11:15)



The most important part, that is, interpreting the parameters. We know the expected value of Y equal to beta 0 + beta 1 x 1 + beta 2 x2 when you substitute equal to 1, when you substitute this x 2 = 0 that equation for mechanical, problem related to mechanical. So, beta 0 + beta 1 x1 beta x2 0 so this term will become there Beta 0 + beta 1 X 1 will be there. When substitute this x 2 equal to one that equations for the problem related to electrical.

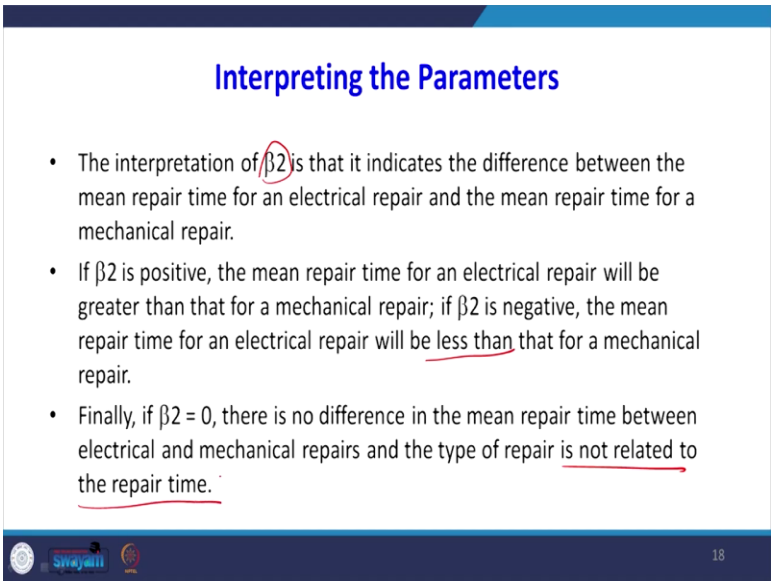
So, E expected value y electrical equal to beta 0 + beta 1 x1 so, beta 21 what is happening so, beta 0 beta 2 that can be grouped that this will be beta one x1. See, both equations are same, both equation having the same slope Beta 1 only it differs by this extra value in our Y intercept how much with Beta 2.

(Refer Slide Time: 12:16)



Comparing equations 1 and 2 we see that the mean repair time is linear function of X1 for both mechanical and electrical repair. The slope of both equation is beta 1, but the y-intercept differs. The Y intercept is beta 0 in equation 1 for mechanical repairs and beta 0 + beta 2 in equation 2 for Electrical repairs.

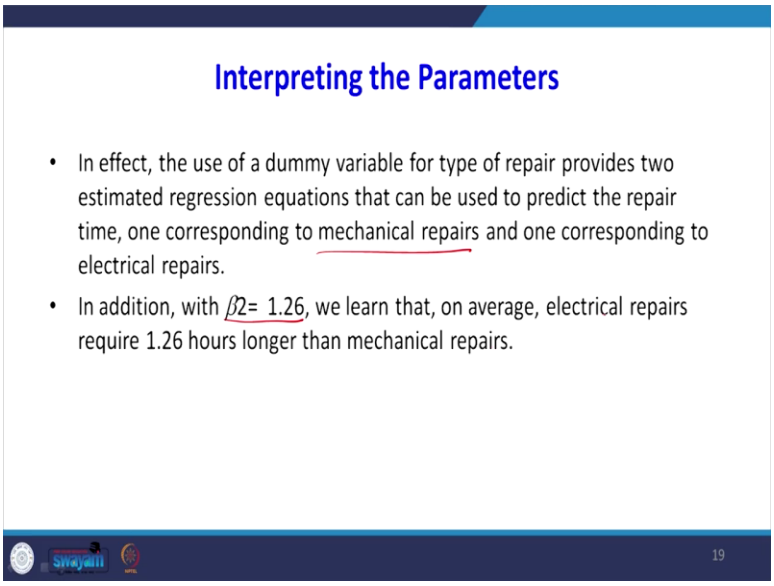
(Refer Slide Time: 12:45)



The interpretation of Beta 2 is that it indicates the difference between the mean repair time of electrical repair and the mean repair time of mechanical repair. So the time differs by with this unit of this Beta 2. Beta 2 is positive the mean repair time for electrical repair will be greater than that of the mechanical repair. In our problem, it is beta 2 is positive, if the beta 2 is negative the mean repair time for an electrical repair will be less than that of mechanical repair. If finally you Beta 2 equal to zero there is no difference in the mean repair time between electrical and mechanical repairs.

And the type of repair is not related to repair time. This is most important because after doing a dummy variable regression you have to interpret it. The interpretation is this way. The first thing is you have to look at what is the sign of this Beta 2. Beta 2 is positive or negative. Then in case the beta 2 is 0, we can save the type of the time taken to repair that filter is nothing to do with the type of problem it has occurred. Whether it is problems related to mechanical repair or problem related to electrical repair.

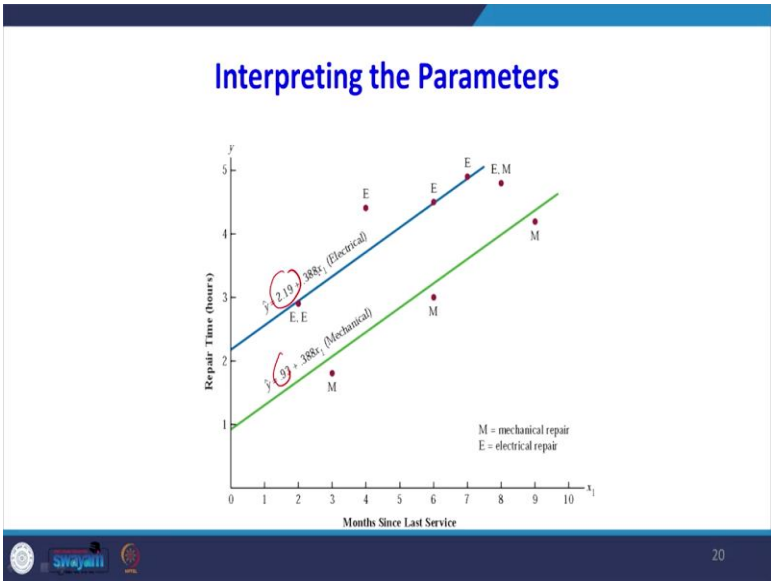
(Refer Slide Time: 14:01)



In effect, the use of dummy variable for type of repair provides to estimated regression equation that can be used to predict the repair time, one corresponding to mechanical repair and another corresponding to electrical repairs, in addition, beta 2, 1.26 we are getting this 1.26, going back, this 1.26. This 1.26 we learnt that the average electrical repairs required 1.26 longer than the mechanical repairs because for electrical repairs we have taken x1 = 1, for mechanical repair, we have taken x1 =0.

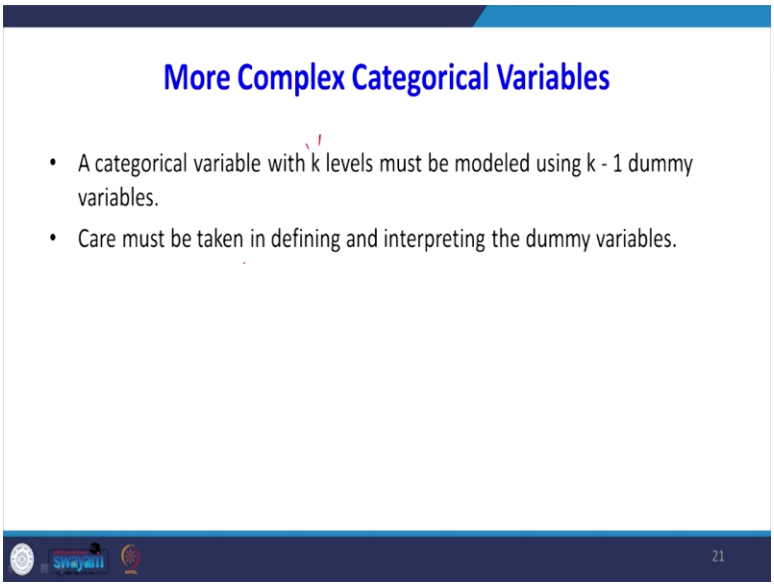
So, the electrical repair is taken as the reference. What is the meaning of that is that the 1.26 time units the electrical repair is taking longer time than mechanical repairs. Look at this picture.

(Refer Slide Time: 14:55)



The green one is for mechanical repair when substitute x2 = 0 here, the blue one is for electrical repair, very extreme cold one. Look at this one. 2.19, this is 0.19. Both the slopes are same. This slope has 0.388 for this equation and this equation. Only the intercept is differs.

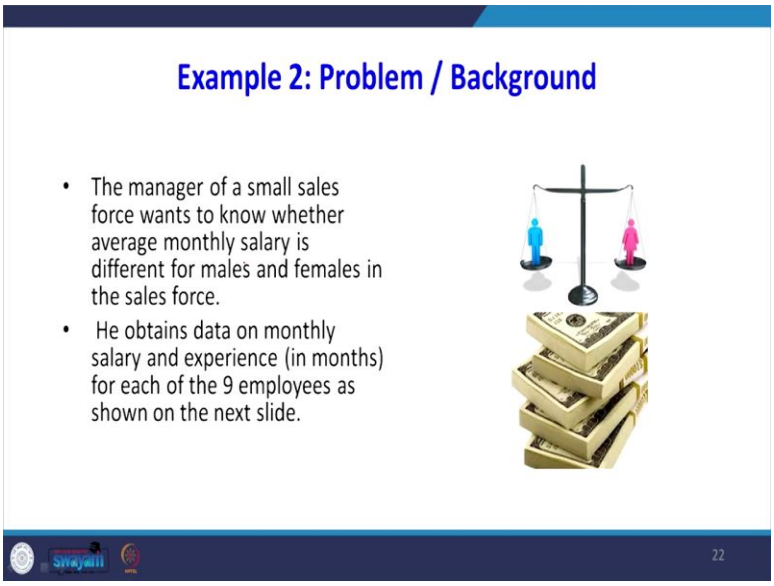
(Refer Slide Time: 15:13)



What is the logic is that here we have we have seen only two levels. Sometimes, there may be more than two levels. So, the number of a categorical variable with K levels must be modeled using k-1 dummy variable. What happened previously there was a 2 level, so we have taken only one dummy variable x2. So there are three levels you have to take 3 – 1 that is a 2 dummy variable. Care must be taken in defining and interpreting the dummy variable.

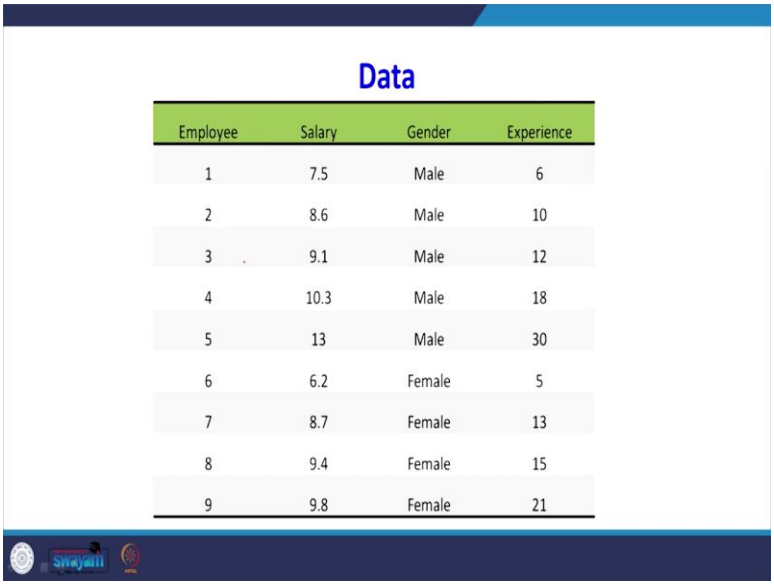
What is the care here is what is the value we have assigned is equal to 1. For example, electrical repair, you take an equal to one that equation is integrated with respect to x2 =1.

(Refer Slide Time: 15:54)



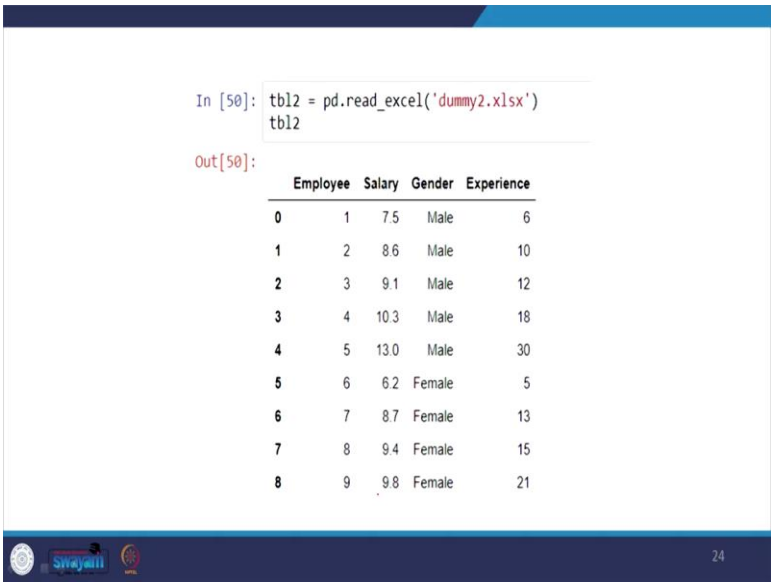
We will go for another problem. This problem is taken from statistics for management from Lemen N Ruby. The manager of a small sales force wants to know whether the average monthly salary is different for males and females in a sales force. He obtained a data on monthly salary and experience for each of 9 employees as shown in the next slide.

(Refer Slide Time: 16:20)



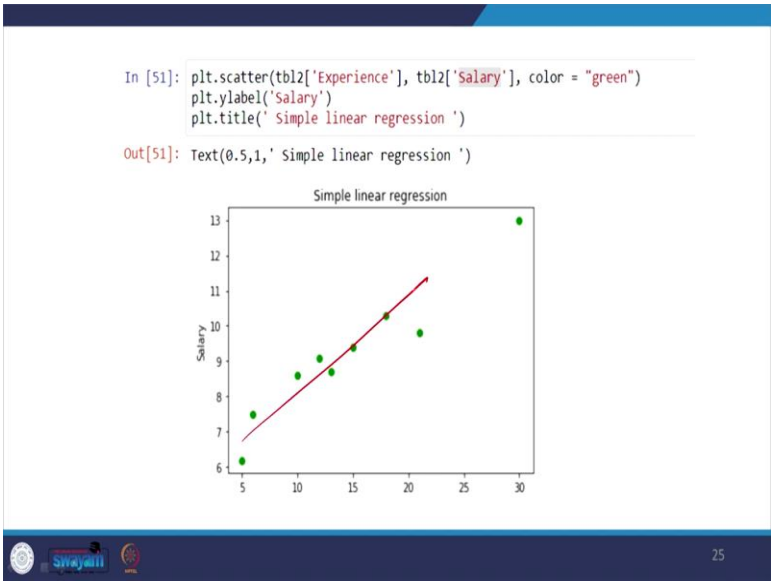
Look at this. This is there are nine employees their salary, there is gender, there is experience. Now what you are going to do in this example, what is the salary of the females even though they have equal experience with the male, whether females are discriminated or not when we can say that they are getting discriminated, even though they have equal experience with male, they are getting lesser salary that means the females are discriminated.

(Refer Slide Time: 16:49)



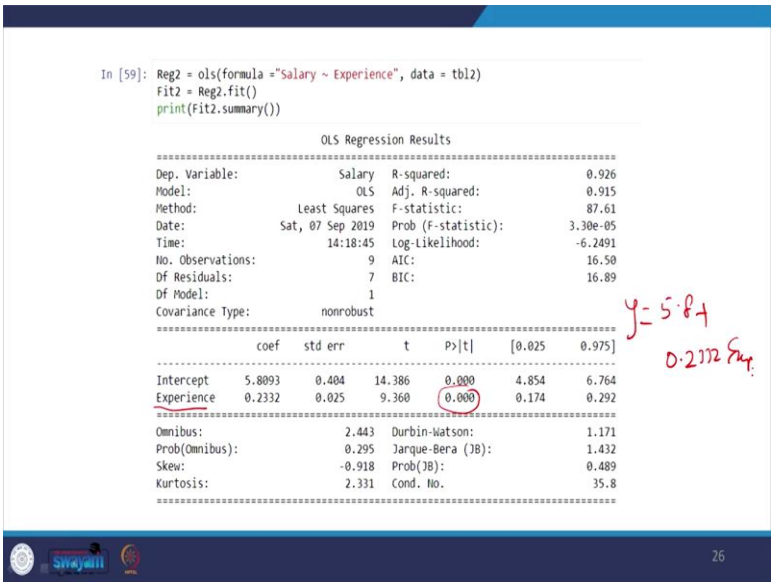
First, we will import the data. Here are imported in the object called tbl2 = pd.read underscore excel. The excel data where I have stored this problem is in the filename called dummy2. So, when I show this. Look at this, this is the employee salary, gender and experience. Next, what we are going to do?

(Refer Slide Time: 17:14)



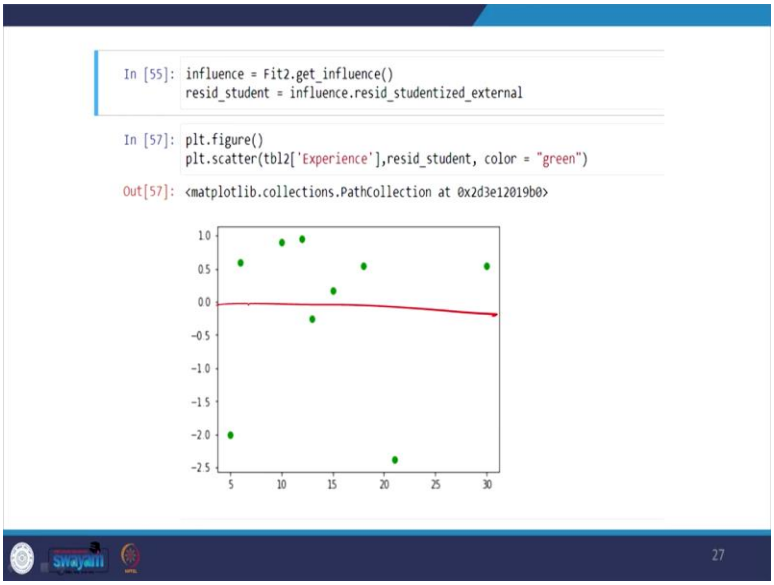
We are going to find out the scatter plot or is there any trend between the experience and the salary? It seems to be there is a positive trend. But look at the residual plot. What is this equation?

(Refer Slide Time: 17:29)



Y equal to see, R square is 0.926. See, the experience is the independent variable. Experience is the because p value is less than 0.05 we can say as the significant value. So we can write Y equal to 5.8 + 0.2332 experience. This is a regression equation. Ok now let us do the residual plot. For this we will do the error analysis.

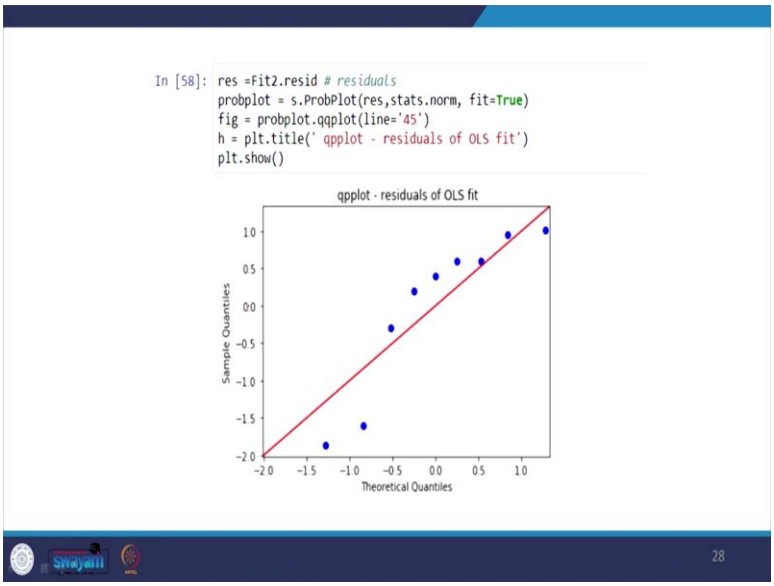
(Refer Slide Time: 17:58)



We will do the Residual analysis. You see that most of the points support, taken as the reference. This is a standardized residuals. Most of the points are you should be randomly it has to be distributed. Most of the points are above this way, there is a zero line. That means there is a

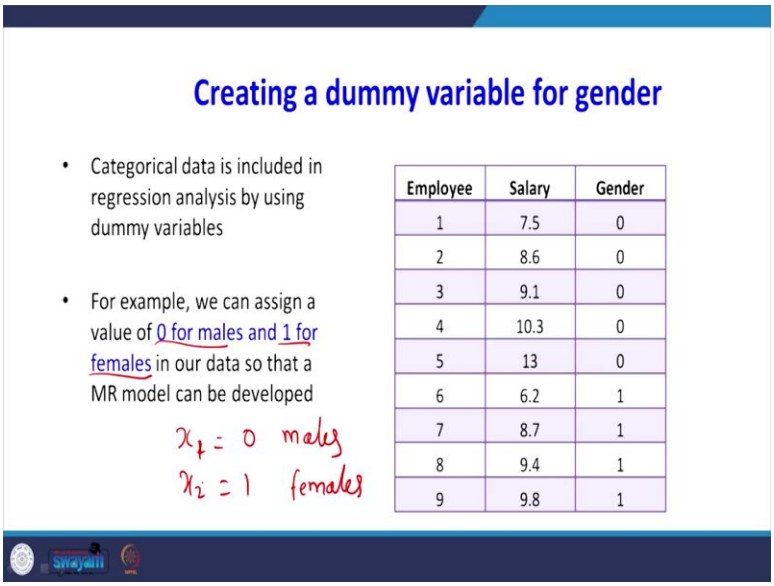
problem in assumption. Otherwise there might be some other variable that may affect the salary apart from experience.

(Refer Slide Time: 18:25)



Look at the see that quantile plot. You see that here also most of the points are above the pointers it has to sit on this red line, but it is not sitting on red Line then there is a problem in the assumption of that equal variance. That means error is not following equal variance.

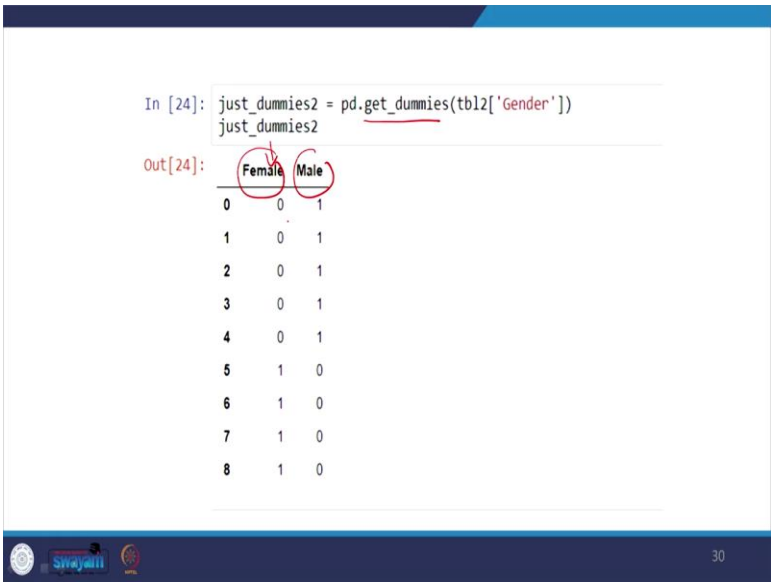
(Refer Slide Time: 18:42)



Now, what we have done in this data. Categorical data is included in the regression analysis by using dummy variable here what you have done? Zero for males, 1for females. What has taken

zero also as reference or one also reference? So, one for male, female is taken as a reference now data, so that a multiple regression model can be developed. We will do that one.

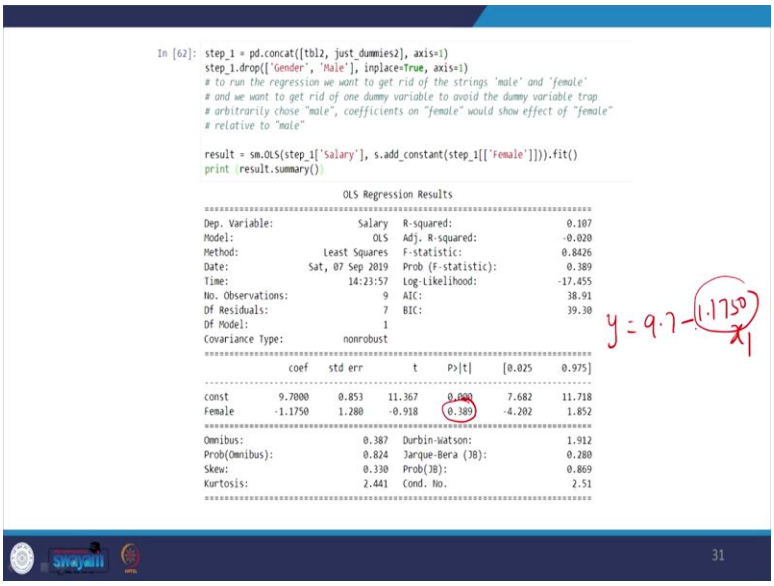
(Refer Slide Time: 19:08)



From the given data, I have converted into dummy variable, one dummy variable for fe female because there are two level female and male. So, male is taken as one female taken is zero. The coding is that zero is taken as male one is taken female. So in this we are going to take this column for our further analysis. So, how to interpret this 0 means female one means male.

In creating a dummy variable for gender, we are going to follow this notation x2 = 0 means male x2 equal to one is taken as a female. So, after creating dummy variable first how to create a dummy variable in Python just underscore dummies to that is a variable which I have given, pd.get underscore dummies. This was the command for making dummy variable. So we are going to take female column for further analysis. Zero means male one means female.

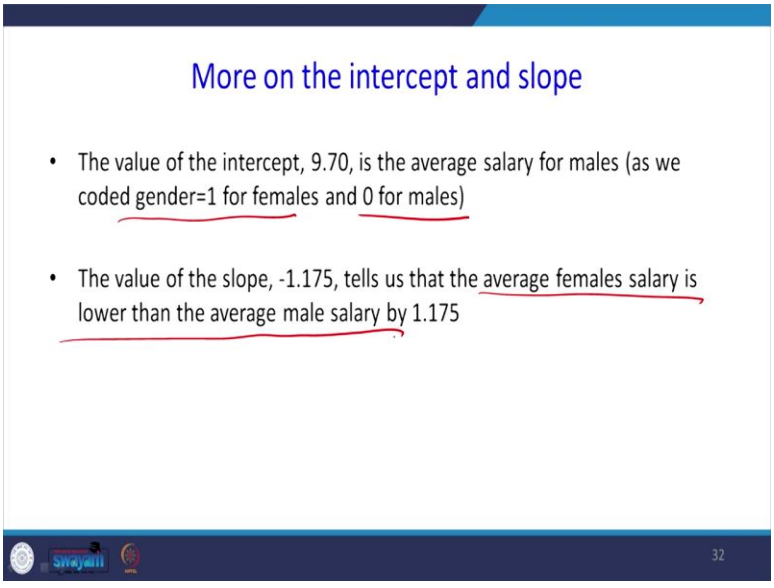
(Refer Slide Time: 20:10)



This was our Python output for that regression analysis. When you look at this, R square is 0.107 but look at here, first I will write the regression equation. Y = 9.7 -1.1750 x1. How to interpret this result you see that in the x1 is not the significant value here it is not significant. At the sample data level, what is the meaning of x1 R? Look at this. If you write x2=1 here x1 equal to 1 is not x2 it is x1.

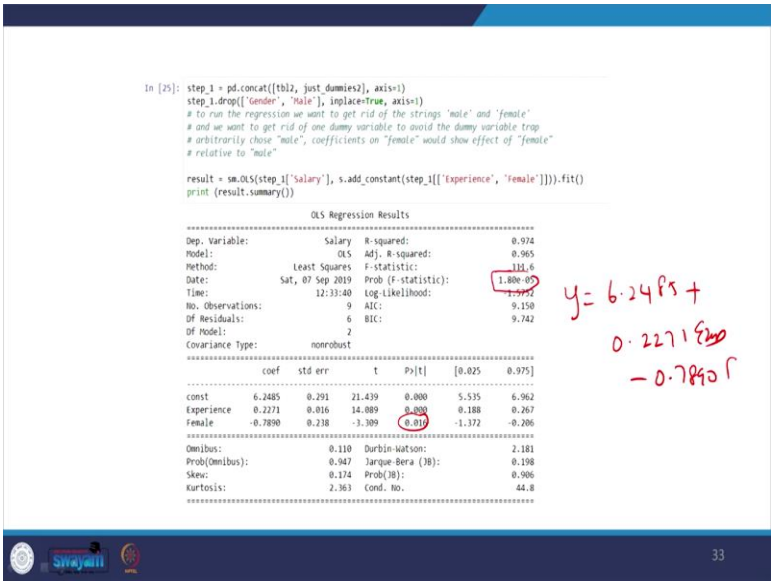
When you substitute x1 =1 this one, this coefficient says, it is negative. What is the meaning of these negative is this female is getting lesser salary when compared to male by this much unit because it is a negative sign, we go for interpretation.

(Refer Slide Time: 21:23)



The value of the intercept is 9.7 the average salary for males has been coded a gender 1 for female ok then, 1 for female and 0 for males. So, the value of the slope is - 1.175 tells us that the average salary is lower than the average male salary by 1.175. What is the meaning of this? Females are getting 1.175 units they are getting lesser salary when compared to male. If it is a positive then we can repeat that. When compared to male females are getting more salary because the negative you are saying that when compared to male, females are getting less salary.

(Refer Slide Time: 22:09)

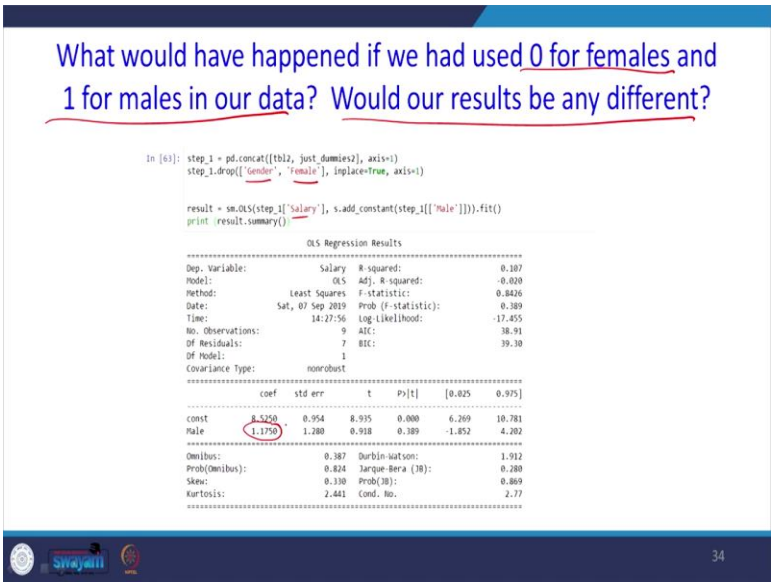


Now what we are going to do? We are going to introduce the previously we considered only gender. That is a female is taken as a reference. Now, we are going to introduce the experience also. When you introduce the experience also know the regression equation is Y equal to 6.2485 + 0.2 271, experience - 0.7890 female. Now look at the p-value these p-values now less that 0.05. Now, here the gender is significant variable.

In your previous slide, when you go back in this slide, you see that the p value is not significant. So we cannot say there is a gender discrimination. We can write a regression equation with the help of sample data, but at the population level, there is no connection between Gender and their salary because the relation between x1 that is there gender and the salary there is no relationship. That means here the both female and male are getting same salary.

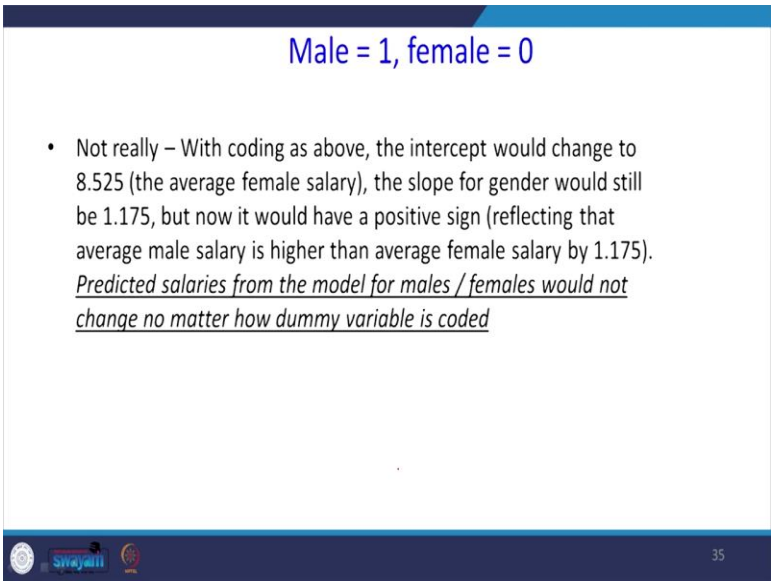
But when we introduce our experience was one of the variable now the general also is significant, so by considering experience and the gender, now gender also one of their significant because you are the P value is less than 0.05. We look at the f value the f value is very low. The probability value also very low as a whole model, this model is significant individually also all the variables are significant.

(Refer Slide Time: 23:47)



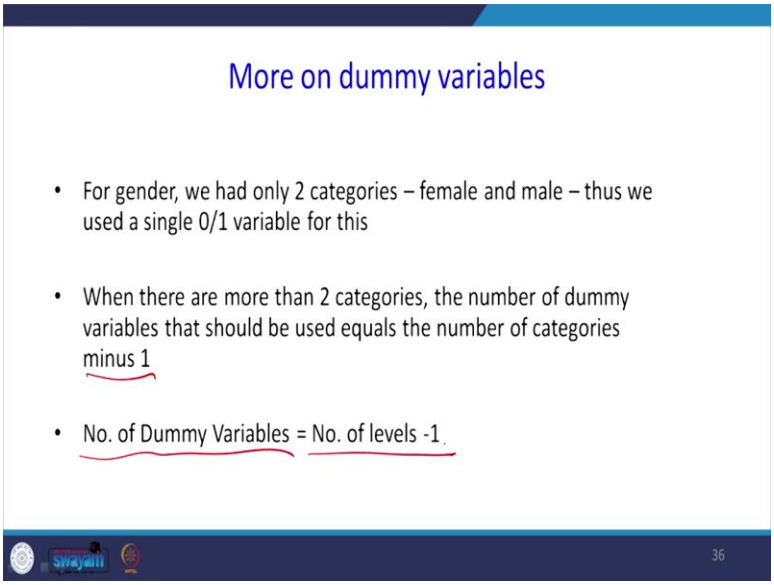
What would happen if we used zero for females and one for males in our data. Would our results be any different right? So for that purpose we have done some modification here. For example, gender female it is just reversed. You see that there is a difference in intercept but the slope is same but the slope sign is different. So what is the meaning here? The male right, because the male is 1,the males are getting 1.175 unit of higher salary when compared to females.

(Refer Slide Time: 24:27)



So what happened is there any difference in the result not really. With the coding as above, the interested change to 8.525 see that 8.525 the slope of the gender would still 1.175, but it would have a positive sign reflecting that the average male salary is higher than average female salary by 1.175. So predicted salaries from the model for males and females would not change no matter how the dummy variable is coded.

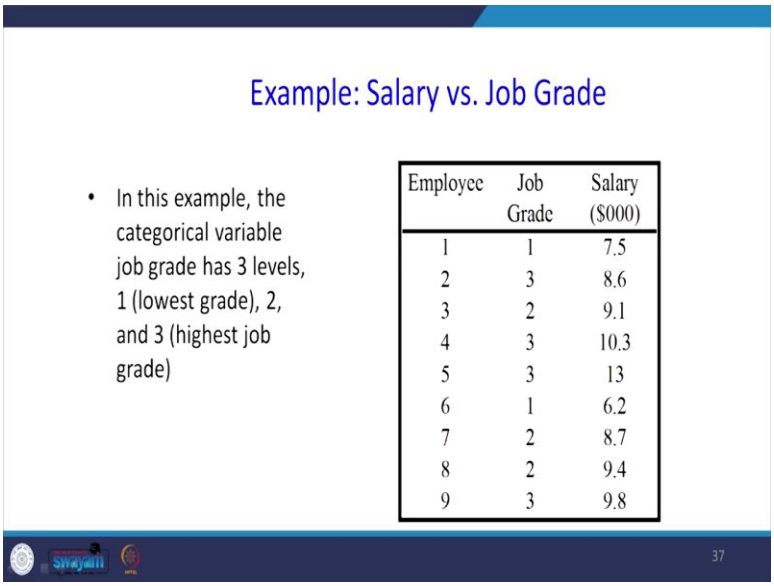
(Refer Slide Time: 25:00)



Sometimes, what will happen, they may be more than one dummy variable, how that in our problem. We have only two levels, sometimes there are three levels. We should have to Dummy variable will see that example. For gender, we had only two categories female and male does we used a single dummy variable 0,1 variable for this. When there are more than two categories the

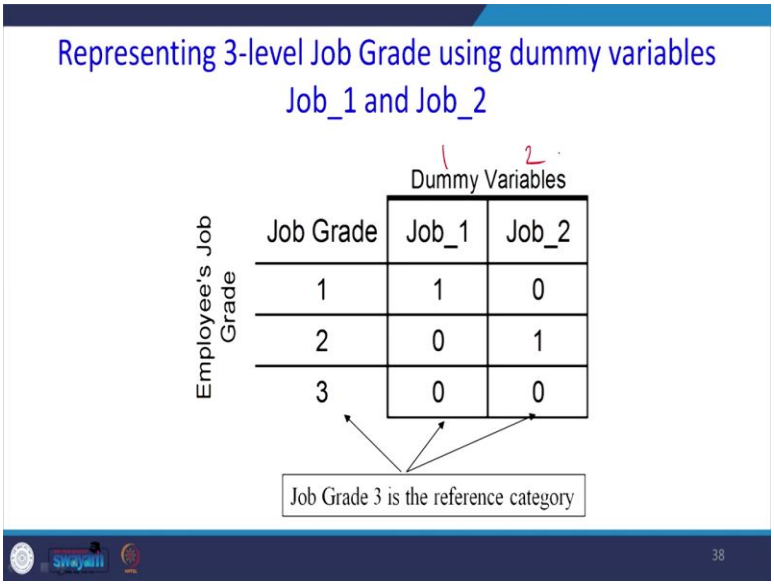
number of dummy variables that should be used = the number of categories -1. So, the number of dummy variable is number of levels -1.

(Refer Slide Time: 25:33)



You see that there is one example where the job grade is there are three levels. 1, 2, 3, in this example, the categorical variable job grade as three level so, 1, 2, 3, 1 means lowest Grade, 2 means medium and 3 means highest grade. We are going to have three levels in our categorical data three levels are level 1, level 2, level 3.

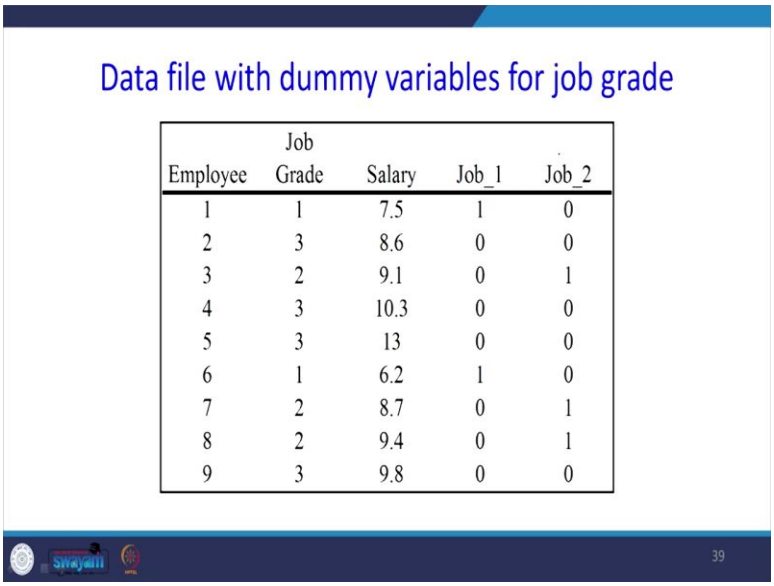
(Refer Slide Time: 25:54)



There are 3 levels and we are going to have only 2 Dummy variables. Job 1 say taken as 1, 0 job2 taken as 0,1 job3 is 0,0. So now, we can say this 0, 0 is taken as a reference, ok. So, the

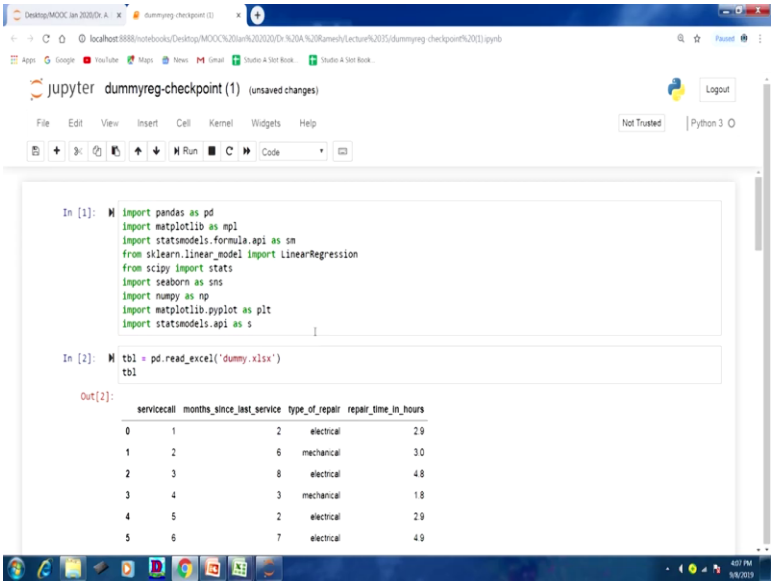
presence of 1,0 will explain category 1; 0,1 will explain category 2; 0,0 will explain category 3. So here what is happening is there are 3 levels. But we are going to have only two dummy variable dummy variables. Dummy 1 and dummy variable 2.

(Refer Slide Time: 26:33)



Now, this is a new data set how this data set can be used for doing dummy variable regression. The interpretation is already I have explained to you now will go for demo of this code which I have shown in our, this presentation.

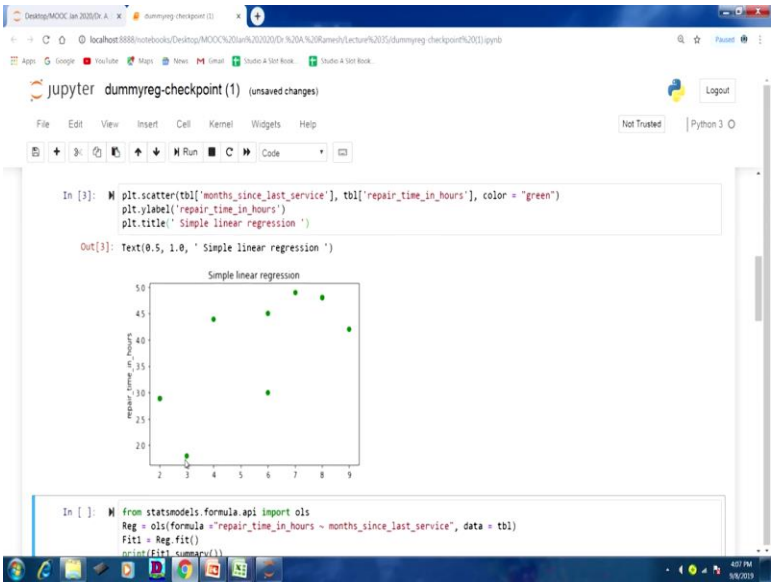
(Refer Slide Time: 26:49)



I have prepared already code for that person. First I am going to remove this output by clicking kernel restart and clear output. I have cleared the output now I am going to run this one. So as

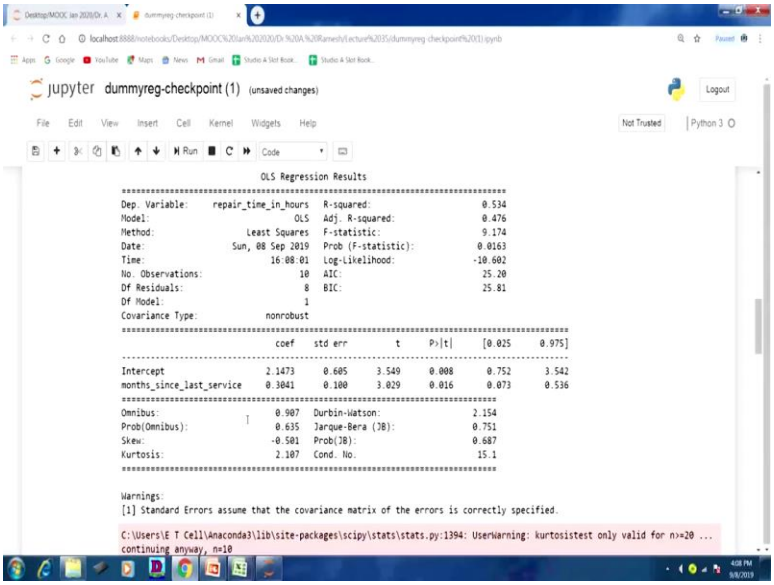
you know, this is Shift Enter so again shift enter this is the data. This data shows service call months since last service type of repair. Next we will go for scatter plot.

(Refer Slide Time: 27:17)



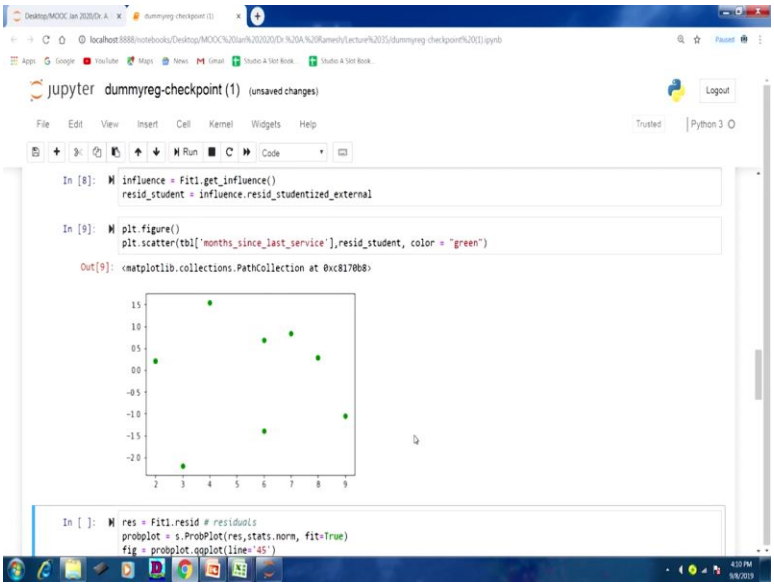
Scatter plot shows that there is a correlation between month since last service and repair time in hours next will go for simple linear regression where were taken only one independent variable.

(Refer Slide Time: 27:35)



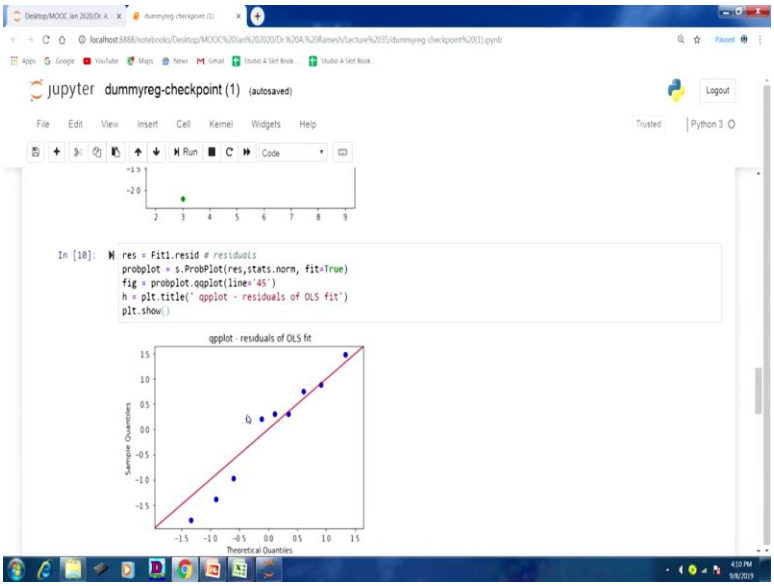
When you look at this here, this is equal to 2.14 + 0.3 041 x1, suppose these variables x1. Look at the p-value this p-value is less than 0.05. So this variable is significant value variable, R square also it is good above than 0.5. See, when there is a f statistic this is also less than 0.05. So, as a whole model it is valid.

(Refer Slide Time: 28:04)



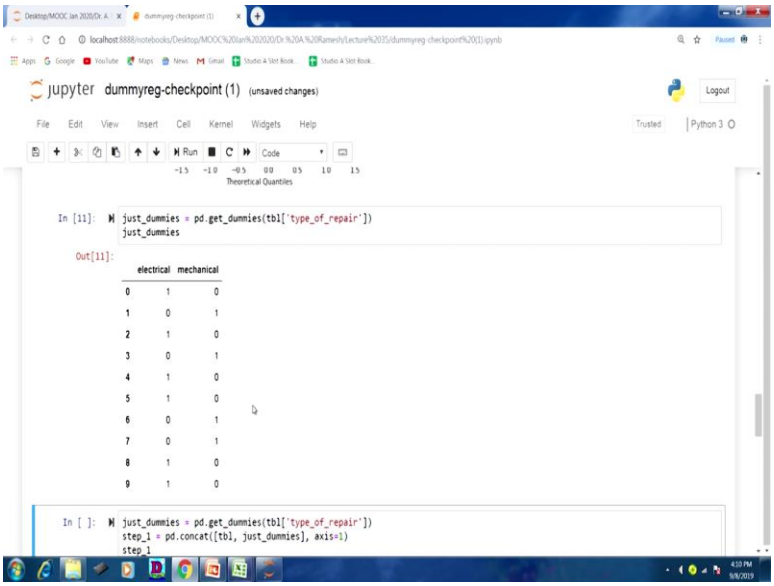
Now, we will plot standardized residual plot. When you look at the standardized residual plot this is the pattern. See there are that some points which are going above -2 how to interpret the standardized residual plot all the points should be between – 2 to + 2. But it seems that there are some variable which goes beyond -2. So it is violating our model assumptions. Now we will go for this.

(Refer Slide Time: 28:38)



These also see that these into some pattern continuously three lines are below this line. There are so many points are above this line. There are also problems in variance of the error variable also.

(Refer Slide Time: 28:51)



Now we will convert the data into dummy variable. This is dummy variable electrical is taken as one mechanical is taken 20 know after converting into dummy variable to drop the column dummy variable belongs to mechanical. After Dropping we can see this route for duplicate this now. There is no mechanical column only electrical column is there.

I will do for this data set will go for regression analysis two independent variable one is months underscore since last service another one is type of repair that is electrical is taken as reference. When we look at the p-value the p-value are both independent variables less than 0.05, so the significant model in this equation when you substitute x 2 equal to one will get a regression equation for problem related to electrical. When you substitute x 2 equal to zero will get a regression equation for problem related to mechanical repairing system.

Now we will be going for another problem. This is our second problem, where the salary is the dependent variable. Experience is independent variable gender also independent variable. When we plot that between experience and salary there is a positive relationship. Now will take salary and experience, experience is an independent variable. You see that experiences a significant because less than 0.005. R square is 0.26.

There is no problem in this. Now, we look at the standardized residual plot that most of the points there is not equally plotted, most of the point above zero there is no randomness in the distribution. There seems to be some pattern in the residuals. We will go for checking the normality of the variance error now. See that it is this also following some kind of a pattern and then also not sitting on the exactly the diagonal line.

Now will go for create a dummy variable for the gender. There is one is for female another one is male now will drop this one. So, female is taken as one male is taken as zero when you do there is regression analysis where Gender is taken as a female now, you see that Y equal to 9.7 plus minus 1.175 female. So, females are getting less salary than the male but look at the P value, when you consider only the gender, the P value is more than 0.05.So this gender variable is not significant.

When you, when you bring another variable is an experience when you look at our previous code, it is only gender is taken gender also it is not significant because the p-value 0.389. Now will take Gender and experience together, let us see what is happening. When you take Gender and experience together, you see that the P value for female is less than 0.05. The experience also, listen 0.05. Both the variables are significant, but the female is getting less salary, when compared to male even though they have equal experience.

Now, what will happen when you reverse the code? Suppose, we have taken female equal to 1 male equal to zero now, what will happen? When you reverse that code send male equal to 1 and female equal to zero what will happen if there will not be any change in the result. Only the sign of usually the male is taken that was - 1.17. Now female is taken as reference. So we are getting only the positive value of 1.17.

Only the difference in the Y intercept, otherwise all interpretations are same. In this lecture by using dummy variable regression I have taken 2 problems with the help of python code I have explained how to do a dummy variable regression and I have also interpreted the result. We know what is the dummy variable regression is sometime the gender is one example for dummy variable regression because there are two possibilities, male and female.

Similarly the job category, Category 1, category 2, category 3, these are dummy variable. For this purpose we have learnt how to do a regression analysis, the next class very important topic that is logistic regression are going to see that one before seeing Logistic regression. There is a one principle called maximum likelihood principle. I will explain what is the maximum likelihood principle? With the help of some examples, then we will go per Logistic regression in the next class. Thank you very much.