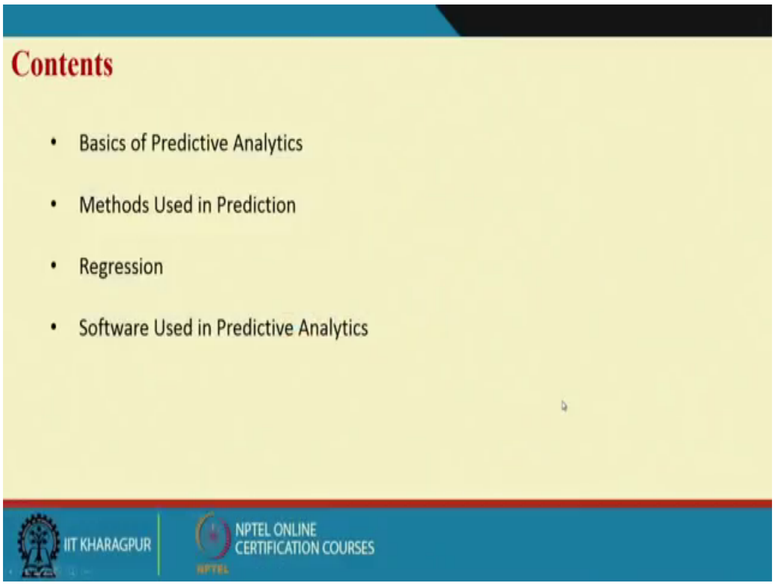
Industrial Safety Engineering
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture - 49
Accident Data Analysis: Regression

Hello everybody. Good day. Today we will discuss Accident Data Analysis Regression. In last class you have seen the accident data analysis using control charts. Before that, we are given we have seen the concepts of accident investigation, the data to be collected in during investigation as well as there are different recommendations.

So, today we will discuss regression, it is a very generic general technique or I can say generic technique for data analysis, need not be that limited to accident data analysis. So, as a result any regression lecture will be helpful to understand accident data analysis is which using regression.

(Refer Slide Time: 01:24)

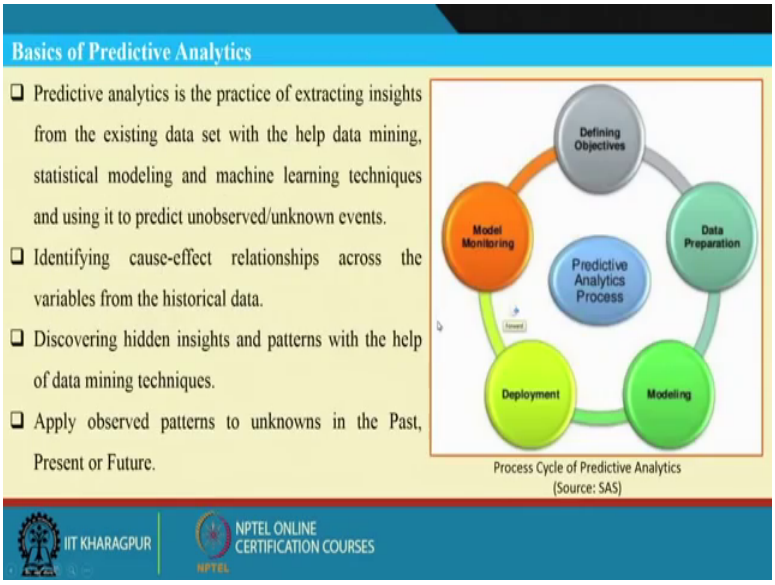


Contents

- Basics of Predictive Analytics
- Methods Used in Prediction
- Regression
- Software Used in Predictive Analytics

So, what we will discuss today? We just discuss today the concept of predictive analytics. And then what are the different methods used in prediction and then we will discuss on regression and I will show you some of the software's which are used in prediction or predictive analytics.

(Refer Slide Time: 01:40)



Basics of Predictive Analytics

- ❑ Predictive analytics is the practice of extracting insights from the existing data set with the help data mining, statistical modeling and machine learning techniques and using it to predict unobserved/unknown events.
- ❑ Identifying cause-effect relationships across the variables from the historical data.
- ❑ Discovering hidden insights and patterns with the help of data mining techniques.
- ❑ Apply observed patterns to unknowns in the Past, Present or Future.

Defining Objectives

Data Preparation

Modeling

Deployment

Model Monitoring

Predictive Analytics Process

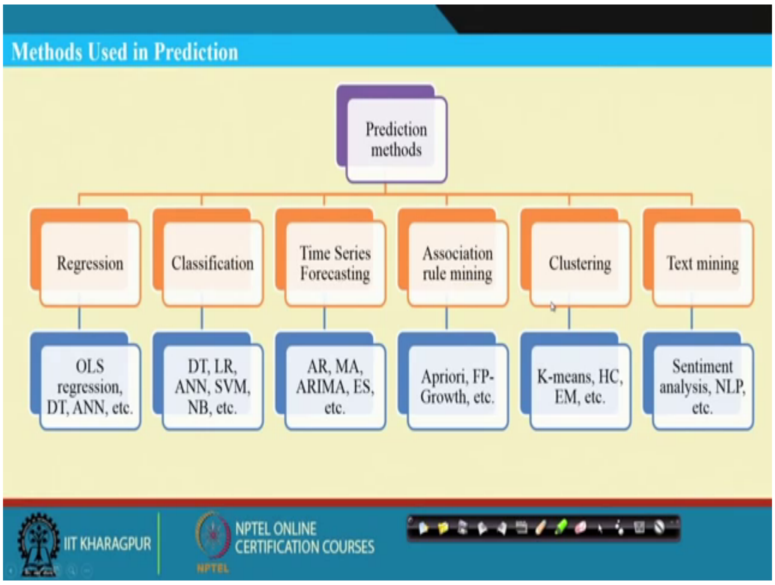
Process Cycle of Predictive Analytics
(Source: SAS)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, predictive analytics is basically the practice of extracting insights from the existing data set with the help of data mining, statistical modeling and machine learning techniques and using it to predict observe or unknown events. Identifying cause effect relationship across the variables from the historical data, discovering hidden insights and patterns with the help of data mining techniques, apply observed pattern to unknowns in the past, present or future. And this is the these are the steps what is given in SAS that process cycle of predictive analytics.

So, first you define the objectives, then data preparation, modeling, development, deployment and then your model monitoring and this is what is the predictive analytics process ok. So, just let me get the pointer here so that, we will be able to write something ok.

(Refer Slide Time: 03:15)

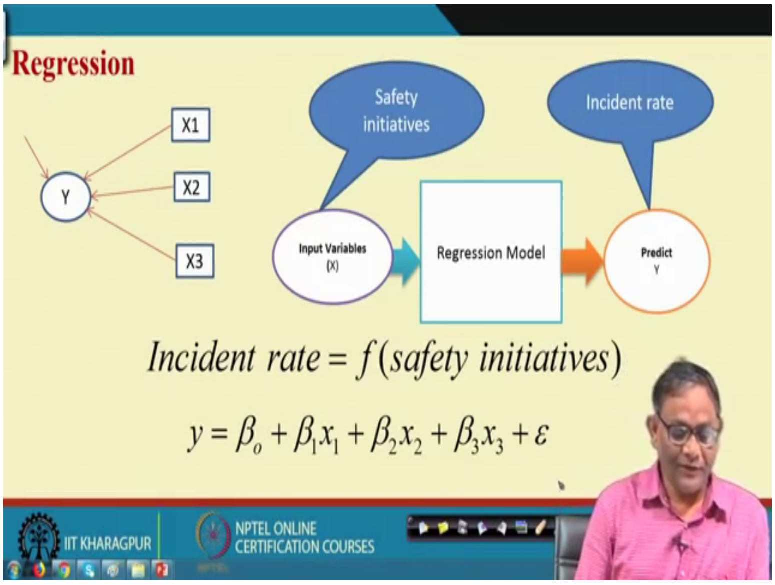


Methods Used in Prediction

So, fine so let us see next that what are the different methods. So, you will find out that there are large number of methods under different groups of prediction methods; one is regression, classification, time series forecasting, association rule mining, clustering, text mining and again under regression some of the techniques, under classification some of the techniques, under forecasting some of the techniques, under association rule mining techniques, clustering techniques and your text mining techniques ok.

So, we will not be able to complete all those things. The purpose of prediction is given to you, but in any one of the class of techniques will be useful here. So, today I will discuss regression and then we will discuss also classification ok.

(Refer Slide Time: 04:10)



Regression

$$Incident\ rate = f(safety\ initiatives)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

So, what is regression? So, pictorially if you say regression there will be two kinds of variables. One will be dependent variable another will be set of independent variable independent. Dependent variable is one that depends whose values or whose occurrence will depends on the independent variable which are basically affecting this. So, this can be taught in terms of also something like this; these are the causes and this is the effects ok, if sufficient information available from cause and effect relationship point of view.

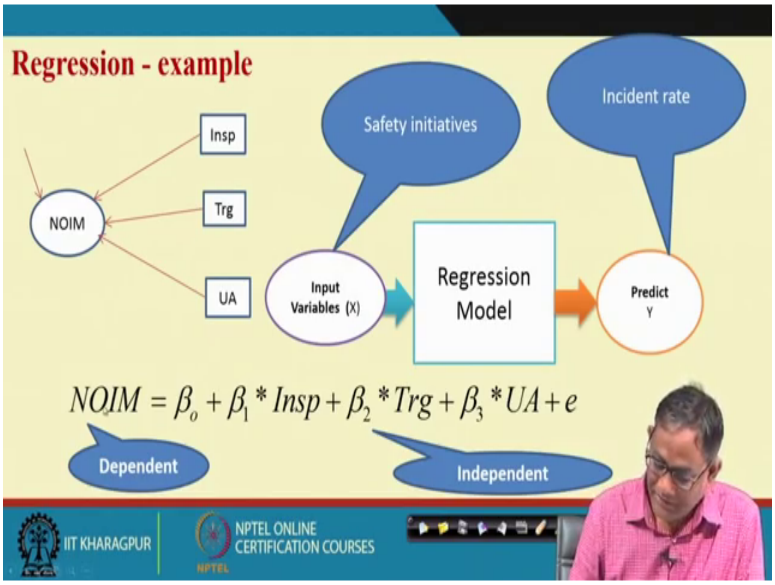
So, then from the safety or accident data analysis point of view, suppose, you want to predict accident incident rate and which is Y is incident rate, with the help of input variables what is basically the safety initiatives. So, different safety initiatives are taken in the industry and also every year or every month there is incident rate data.

So, you want to see the relationship between the incident rate and safety initiatives taken then you can go for regression analysis ok. And in that case, the regression will ultimately talks about the relationship between incident rate and safety initiative where incident rate is function of safety initiatives.

So, incident rate is the dependent variable and safety initiatives in terms of different initiatives like your inspection like, training other things, so, those are basically responsible for the value or incidence rate what you are observing. So in general, this is the equation. Suppose, there are 1 dependent variable, 3 independent variables; x 1, x 2 and x 3, so then the regression linear regression equation will be y equal to beta 0 beta 1 x 1 beta 2 x 2 beta 3 x 3 plus epsilon. Suppose, if there are p number of variables suppose x p, so then what will happen? This will be extended to x p beta p x p ok.

So, if you want the more detail about this regression analysis, so then I have a series of lectures also under applied multivariate statistical modeling that NPTEL lecture series on multiple regression or multiple linear regression. So, you may go through those videos also. Today we will not discussing of that depth, we will just show you the application of regression in safety data or accident data analysis.

(Refer Slide Time: 07:11)



Regression - example

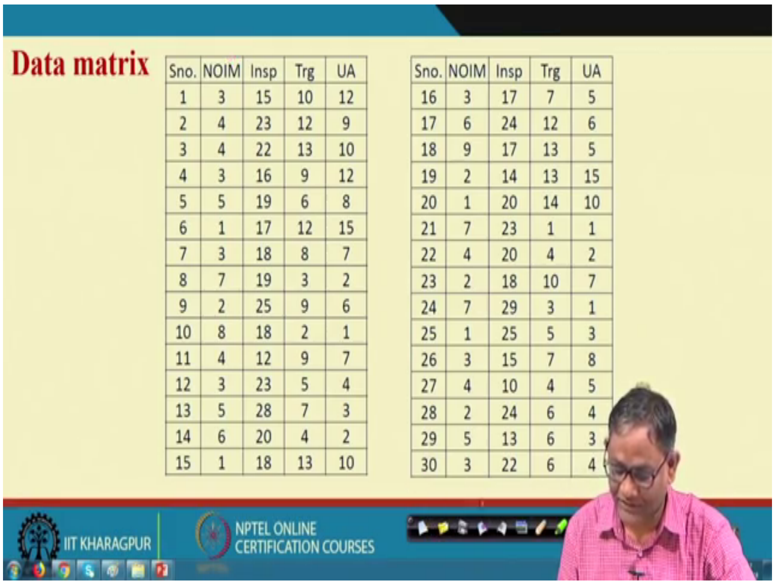
$$NOIM = \beta_0 + \beta_1 * Insp + \beta_2 * Trg + \beta_3 * UA + e$$

Now, we will see that suppose, we are interested to know that what is the dependent variable is number of injuries per month, number of injuries or incident per month and suppose inspection is 1 dependent independent variable, training is another and unsafe acts another.

So, hypothetically we are we will collect data, but if good inspection for hazard identification is practiced and hazard removal is practiced, it is quite possible that, this will ultimately lead to less number of incident per month. Similarly, if people are competent enough based on true training their competency level is improved then they will be less involved in accidents.

And also if you unsafe acts is more, then it is a chance that the NOIM will be more. If it is unsafe, acts are commutate unsafe acts are less then that will be fake. Maybe what happened we are assuming that this 3 are independent, but training and unsafe acts number also there may be some kind of correlation but for the time being we are assuming that they are independent. So, this is dependent and each of them are independent.

(Refer Slide Time: 08:31)



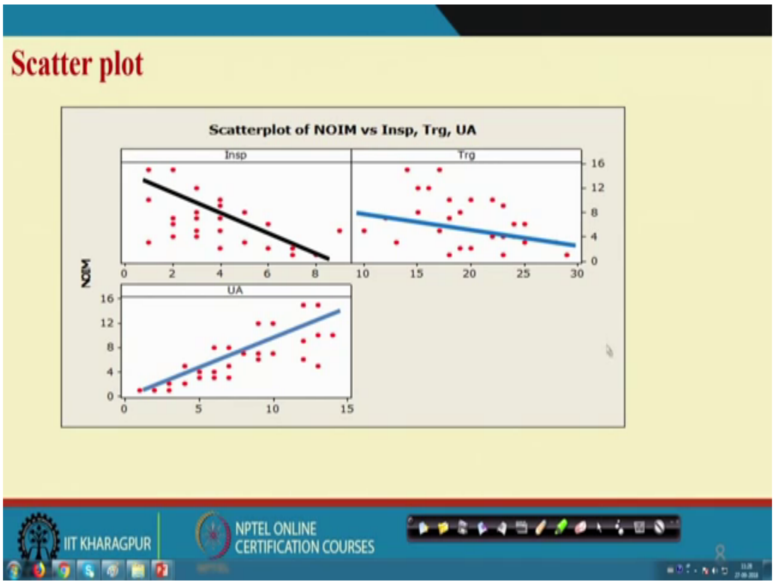
Data matrix

Sno.	NOIM	Insp	Trg	UA
1	3	15	10	12
2	4	23	12	9
3	4	22	13	10
4	3	16	9	12
5	5	19	6	8
6	1	17	12	15
7	3	18	8	7
8	7	19	3	2
9	2	25	9	6
10	8	18	2	1
11	4	12	9	7
12	3	23	5	4
13	5	28	7	3
14	6	20	4	2
15	1	18	13	10

Sno.	NOIM	Insp	Trg	UA
16	3	17	7	5
17	6	24	12	6
18	9	17	13	5
19	2	14	13	15
20	1	20	14	10
21	7	23	1	1
22	4	20	4	2
23	2	18	10	7
24	7	29	3	1
25	1	25	5	3
26	3	15	7	8
27	4	10	4	5
28	2	24	6	4
29	5	13	6	3
30	3	22	6	4

Then let us assume this data. Suppose, we have collected 30 data points, 30 data and the number of incident per month, then inspection in some in maybe number of inspection, training maybe number percentage of people train or number of people trained and unsafe acts number of unsafe acts identified; means what I mean to say, these are the variables which will be measured using appropriate measurement scale. For the time being you please understand that this kind of data are available and you want to know what is NO relationship between NOIM and inspection training and unsafe acts. These data are not real data, they are basically hypothetical data. So, you want to feed the equation.

(Refer Slide Time: 09:30)

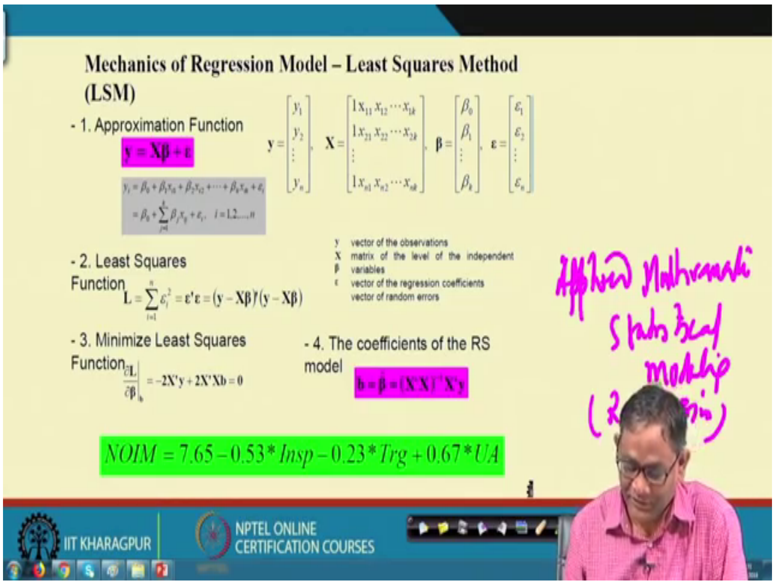


Scatter plot

Then first we everybody will be interested to see that as NOIM is the dependent variable, whether the inspection, unsafe acts and training they have any linear relationship or not. The best way to start with the use of scatter plot, scatter plot basically having 2 axis, this axis and this axis. So, here x axis is inspection and y axis is NOIM, which is number of incident per month. So, if you see the observation plot, so, there is a decline trained and that decline trained is represented by the straight line.

Similarly, for training also there is a decline train and for unsafe acts it is increasing trained ok. But if you see the data from training and NOIM that see these are the scatters, so scattered may be the best straight line, what is printed here that may not be a significant fit or may not be able to explain much of the variability in NOIM. Never the less this plot ultimately tells you yes there are some linear relations and that can be captured through linear regression. So, first if is scatter plot between dependent and all independent variables.

(Refer Slide Time: 10:54)



Mechanics of Regression Model – Least Squares Method (LSM)

- 1. Approximation Function

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$
$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1,2,\ldots,n$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1\, x_{11}\, x_{12} \cdots x_{1k} \\ 1\, x_{21}\, x_{22} \cdots x_{2k} \\ \vdots \\ 1\, x_{n1}\, x_{n2} \cdots x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

y vector of the observations
X matrix of the level of the independent
β variables
ε vector of the regression coefficients
vector of random errors

- 2. Least Squares Function $\mathbf{L} = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

- 3. Minimize Least Squares Function $\left.\frac{\partial \mathbf{L}}{\partial \boldsymbol{\beta}}\right|_{b} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0$

- 4. The coefficients of the RS model $\mathbf{b} = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$NOIM = 7.65 - 0.53 * Insp - 0.23 * Trg + 0.67 * UA$

Applied Mathematic
Statistical
modelling
(3 ...)

So, then there is the mathematics behind it. So, I will not discuss the mathematics here, so you do not required to know also the mathematical details of it, but for the time being only you please keep in mind that, this is what is the general form, i basically stands for observation. So, if you have we have 30 observations, so, n equal to 30 and this is the equation algebraic form and this is the matrix form of the regression.

And then when you are writing in matrix form, so that time basically the y n number of observation, x first is the intercept for n number 1's, then the variable x 1, variable x 2, like variable x k; k number variables are there. If it is p number of variable you will be ended with by p number of variables.

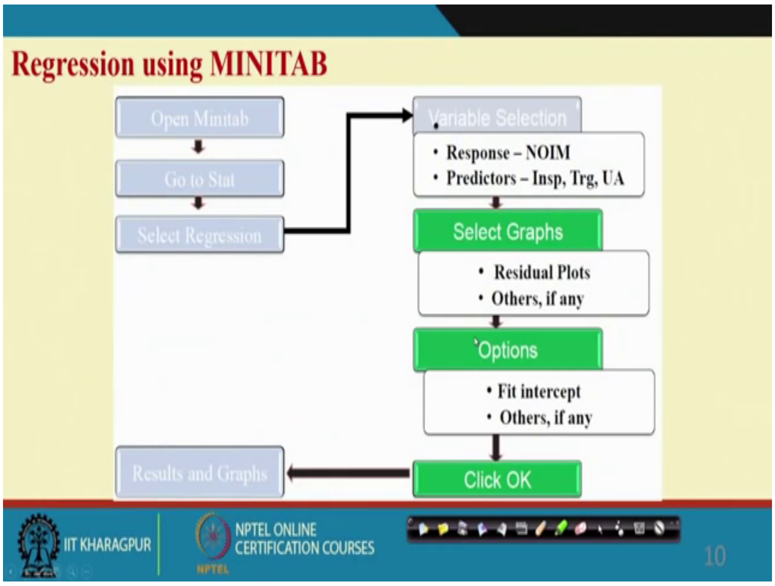
So, as there are k number of independent variables, so and with intercept, so there are k plus 1 number of that coefficient between dependent and each of the independent variable, so that is to be estimated which is basically the talking about the relation between a dependent and independent, so that is represented by beta and this is error term. So, then the ultimate aim will be to minimize the sum total of errors.

So, when you feed the model and then the predicted or predicted value or fitted values and the actual value, the difference will be considered that is the error and then all sum of all differences, the square differences need to be minimized and that is what is the least square function. If you minimize these, so ultimately you will get a matrix equation like this where this equation gives you the estimate of the parameters.

So, using software you can get like Minitab we have used, when you will get that NOIM is 7.65, that is beta 0 which is basically the intercept then minus 0.53 into inspection minus 0.23 into training and plus 0.67 into unsafe acts. So, that mean if inspection increases, NOIM means decreases. Training is better NOIM decreases unsafe act is more, NOIM will be more. So, that is quite logical and although the data is hypothetical, but it is giving you conceptually correct measure.

So, from mathematics point of view everything is given here. If you want more mathematics, see my that NPTEL lecture on multivariate applied multivariate statistical modeling, applied multivariate statistical modeling regression series regression ok. So, we are talking about only linear regression, non-linear regression we are not talking about here.

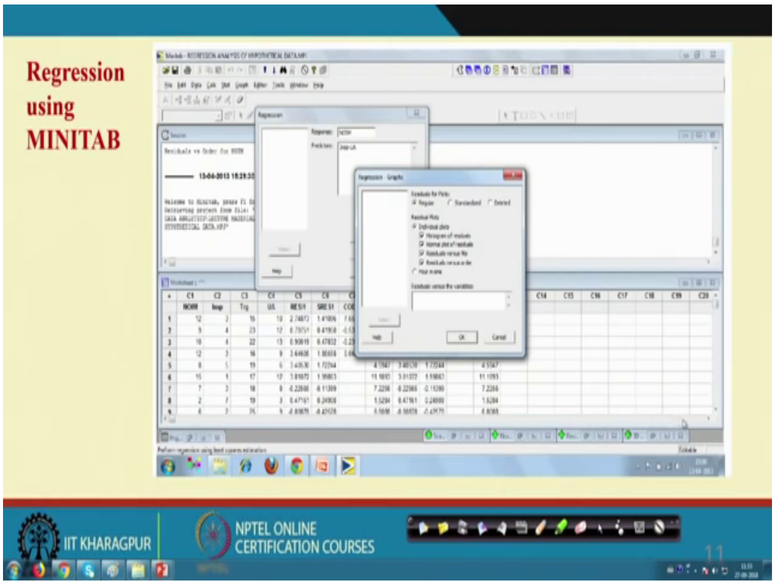
(Refer Slide Time: 14:15)



Regression using MINITAB

Then you see that, the regression can be done using Minitab or any of the any software there are many software. So, if you want to use Minitab, so open Minitab, go to statistics, select regression then, there will be variable selection; one is response another one is predictors. Under response, you select NOIM under predictors you select these 3. Then you can select graphs residual plot and others you can say the fit intercept others if any then click and results and graphs will be displayed ok. So, this is basically the steps using Minitab.

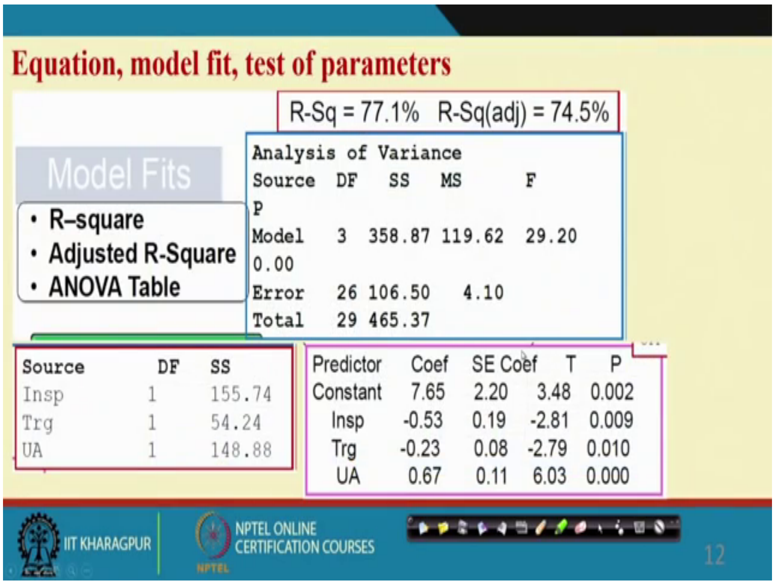
(Refer Slide Time: 14:59)



Now, if you want the snapshot, you see that Minitab snapshot given here. So, now, ultimately when you are going for that stat then regression then, ultimately you have to put the response values here. Then what are the predictors you are putting here, then what are the plots you want and there are other options also you just click click click and you will get and this is what is the first part is the data, second here the coefficients and here the predicted or fitted values are given and also some errors are giving there ok.

So, if you collect data, use Minitab and you will get the solution to regression equation ok.

(Refer Slide Time: 15:49)



Equation, model fit, test of parameters

R-Sq = 77.1%   R-Sq(adj) = 74.5%

Model Fits

- R-square
- Adjusted R-Square
- ANOVA Table

```
Analysis of Variance
Source  DF   SS     MS       F
P
Model    3  358.87 119.62  29.20
0.00
Error   26 106.50   4.10
Total   29 465.37
```

```
Source        DF    SS
Insp           1    155.74
Trg            1     54.24
UA             1    148.88
```

Predictor	Coef	SE Coef	T	P
Constant	7.65	2.20	3.48	0.002
Insp	-0.53	0.19	-2.81	0.009
Trg	-0.23	0.08	-2.79	0.010
UA	0.67	0.11	6.03	0.000

IIT KHARAGPUR   NPTEL ONLINE CERTIFICATION COURSES   12

Exactly that one we have done here. So for our data, so what are the issue here? Issue is that, our dependent variable you know NOIM, independent variable; inspection, training and unsafe acts. So, their contribution in terms of sum square given here and their degrees of freedom are also giving there.

So, we will not discuss degree of freedom, we will be you just follow some book. And so then what happened? First is that whether the model is fit to the data or not linear regression. So, we use R square adjusted R square and ANOVA table these 3 measures we will use.

So, R square is here 77.1 percent, so R square talks about that what is the proportion of variance of the dependent variable is explained by the independent variables or by the model, the regression model. So, what are the proportion of dependent variable is proportion of variance of dependent variable is explained by the regression model. Now, this R square is sensitive to the sample size, so adjusted R square is computed which is basically insensitive to sample size; sample size means, number of observations considered for the model fitting.

So, adjusted R square also 74.5 percent. So, it is considering the accident data it is a good figure. So, we can say the model is a good model because it is able to explain 77 percent of variance of the dependent variable. Another one is the analysis ANOVA table, ANOVA table having the source of variation. So, and then source is model then error these 2.

Then the model your F value is 29.20, it is basically hypothesis test that whether the in whether the all the independent variables are insignificant, they are not contributing to the prediction or to the explanation of the NOIM y variable or not ok. And then it is 29.20, it is a large value in comparison to the theoretical value.

So, this value to be compared with theoretical value and the theoretical value ultimately here model degree of freedom is 3, error degrees of freedom is 26, so then, the theoretical value will be 3 F distribution has 2 degrees of freedom, one is numerator that is 3 and in denominator degree of freedom.

Then you consider a significance level which is alpha, alpha usually alpha is usually 0.05 which is talking about the error in accepting the hypothesis ok. So, what happened? This one this value, if I put this equal to F 3 26 0.05, so you will get a value which will be much less than this value. So, that means, it talks about that the model is fit to the data or that mean linear model is a fit. So, this is the first thing.

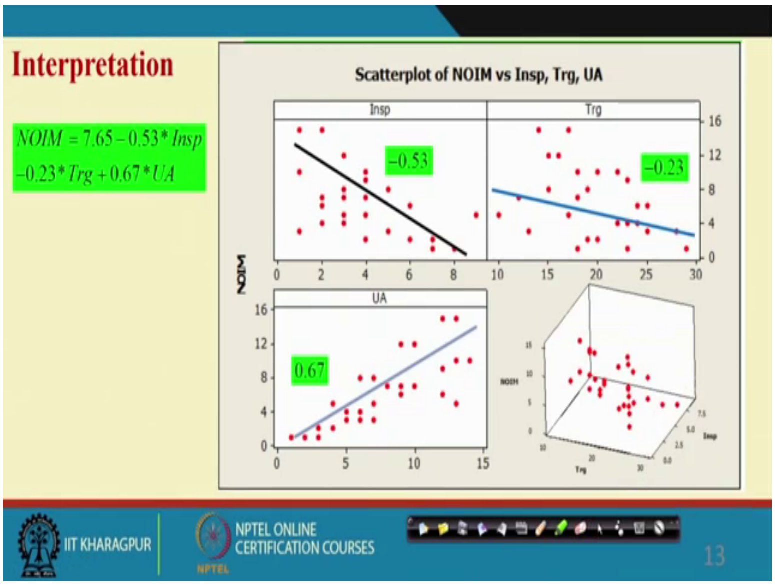
Then once the data is fit to the model then you have to check whether the parameters, the coefficient values, the coefficient values they are significant or not. So, how many coefficients are there? There are 4 coefficients; one is constant intercept then the first independent variable, second independent variable, third independent variable or s s k independent variable.

Then if you see the coefficient the 7.65 then, this is this we all know these. Then their standard error is also computed and then the T value, that is the T statistics is used here and corresponding P values are also given. Means as T value is for 3.48, so it says that the intercept 7.65 is significant with probability value 0.002, means considering this, this is the influence of the intercept. So, then you are committing only 0.002, that is error probability ok.

Similarly, inspection the value the inter that contribute other way I can say the relationship value is minus 0.53, so that is having only 0.009 probability of that not having this value. So, similarly for these and this; so, from this parameter test also what we are finding out? We are finding out of the all the parameter including intercept they are significant. So that means, the regression model is accepted from adequacy point of view, from parameter test point of view also ok.

So, you can use this model for prediction purposes. Please keep in mind the data what I have used here, these are hypothetical data. When you collect the real data, you may find problem that R square is not adequate, some of the coefficient values are positive, but where in from concept point of view, it should be negative and so many of the coefficient will not be significant. So, many issues will be there. So, that requires your in-depth knowledge about regression analysis ok.

(Refer Slide Time: 21:46)



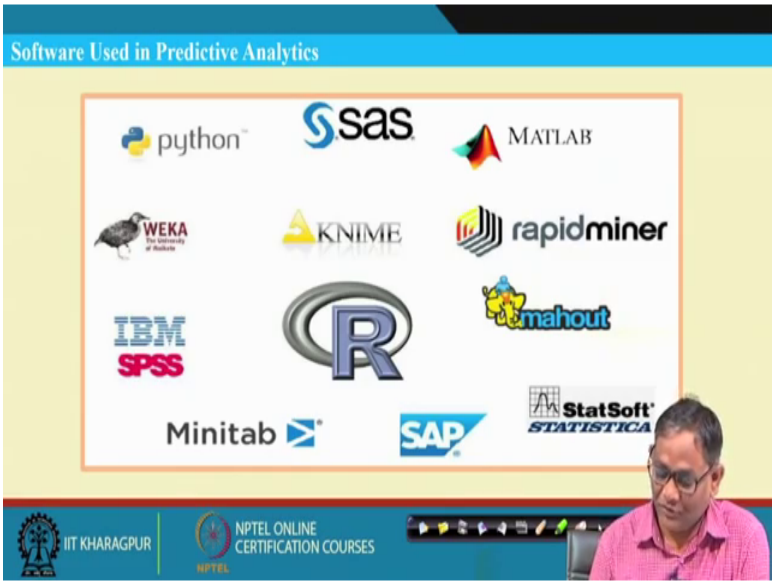
Interpretation

$NOIM = 7.65 - 0.53 * Insp - 0.23 * Trg + 0.67 * UA$

Then I just finish this with the interpretation here. So, what is the interpretation? So your; we have accepted this model. So then, if you it this negative relationship is justified, negative relationship justified this is negative, trained is negative this is positive justified also.

Now, what happen? This is basically talking about training and inspection together within NOIM, some met some q plot we have shown you. So, the interpretation is that, if you improve the training; obviously, the number of incident per month will be reduced, improve your inspection that also reduce and if you reduce the unsafe acts then your number of incident per month will also reduce ok. So, that is the way you have to understand the regression things.

(Refer Slide Time: 22:48)

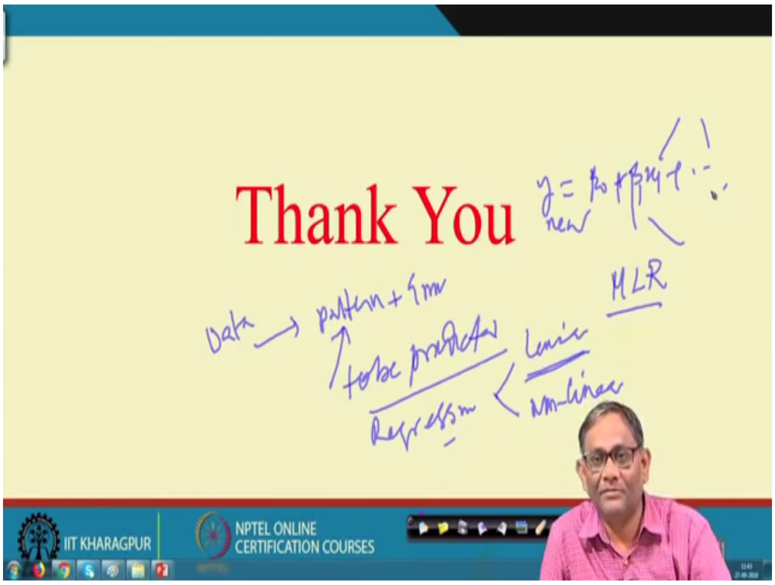


Software Used in Predictive Analytics

So, there are many software which can be used like python, sas, Minitab, WEKA, rapidminer, SPSS, Minitab. So, what or MATLAB, StatSoft software there are many more things are there. So, open source software like R is very famous, python is also very famous and if you see that commercial software like sas, that SPSS, Minitab, MATLAB, so, all those.

So, these are the software StatSoft etcetera, these are the software which will help your analysis easier because, you do not required to understand in the mathematical treatment of regression ok. So, that is a good thing to happen.

(Refer Slide Time: 23:46).



Thank You

So, thank you very much, I hope that you understand. What I tried to tell you here is that, you if you have data that data consist of pattern plus error, this pattern to be can be predicted to be predicted, there are large number of prediction models, one of them is regression. So, if you use regression, then there are the different steps. So, the regression can be linear, can be non-linear, linear and non-linear regression.

So, we have seen the multiple linear regression only multiple linear regression. So, that the steps is there, you first conceptualize the dependent and independent variables, then collect the relevant data, then understand the new software, understand the mathematical basis or basic mathematics, put the data into the software and then you will get regression equation.

Then you have to fit the find the fitness of the model through R square adjusted R square or F test you can do that. Then you do the test of parameters; that mean each of the independent variable coefficient values to be tested using T state. You all those things permit you that the model is fit and parameters are significant then use this model for prediction purposes.

So, as if you have a model like y equal to beta 0 plus beta 1 x 1 like this, so then a new y new given x new independent new, independent object observation related then you will be able to find out that what will be the new y.

So, that is nothing but suppose I have you have develop a model considering last 5 years data, so there may be in monthly data maybe next month what will be the number of in incident in the plant what we which we have considered then fit the values like training value, inspection value and unsafe act value in the equation and then the equation will give you what is the what will be the NOIM or number of incident per month for the next month or subsequently in other months also. So, that is the way you have to use the model.

Thank you very much.