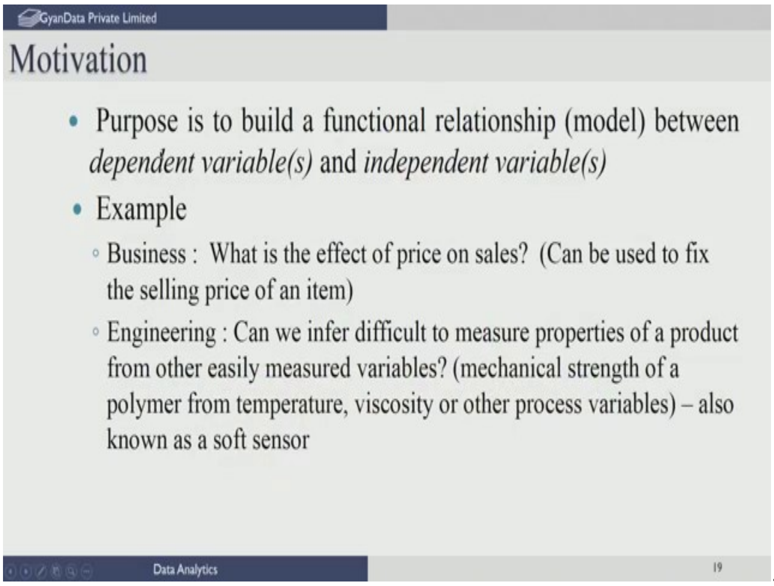
Data Science for Engineers
Prof. Shankar Narasimhan
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 32
Linear Regression

Welcome to this lecture on Regression Techniques. Today we are going to introduce to you the method of linear regression, which is very popular technique for analyzing data and building models. We will start with some motivating examples. What is it that regression does?

(Refer Slide Time: 00:37)



It is used to build a functional relationship or what we call model, between a dependent variable or variables there may be many of them and an independent variable again there might be more than one independent variable. We will de ne these variables and how you choose them for the intended purpose a little later, but essentially we are building a relationship between 2 variables you can take it in the simplest case and that relationship we also call it as a model.

So, in literature this is known as building a regression model. We can also call it as identification of a model. Sometimes this goes by the name of identification most popular term is regression.

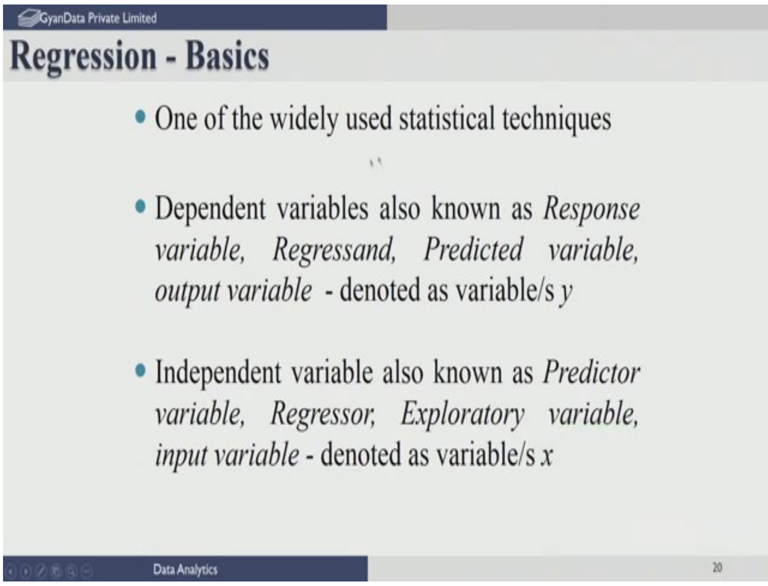
So, let us take some examples let us take a business case. So, suppose we are interested in finding the effect of price on the sales volume. Why do we want to actually find this effect, we may want to determine what kind of price? We want to set the selling price of an item in order to either boost sales or get a better market share.

So, that is why we are interested in finding what effect does price have on the sales. So, the purpose has to first define what why are we doing this in the first place in this case our ultimate aim is to x the selling price. So, as to increase our market share that is the reason we are trying to find this relationship.

So, similarly let us take an engineering example in this case I am looking at a problem where I am trying to measure or estimate the properties of a product, which cannot be measured directly by means of an instrument easily. However, by measuring other variables, we are trying to kind of infer or estimate this difficult to measure property. A case in point is the mechanical strength of a polymer for it. This is very difficult to measure on line continuously or the other hand process conditions such as temperature, viscosity of the medium can be measured. And from this it is possible to infer provided we have a model that relates the mechanical strength to these variables temperature viscosity and so on. First you develop a model then you can use the model to predict mechanical strength given temperature viscosity.

So, such a model is also known in the literature as a soft sensor or the software sensor and this model is very useful in practice. To continuously infer values of variables which are difficult to measure using an instrument, indirectly you are always inferring it through this model and other variables. So, these are cases where we have the purpose is very clear we are building the model for a given purpose and the purpose is defined depending on the area that you are working in.

(Refer Slide Time: 03:43)

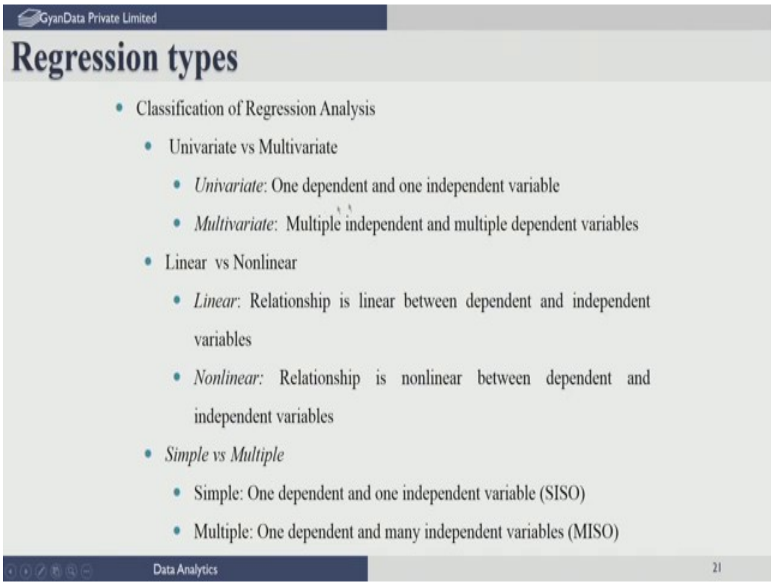


So, regression happens to be one of the most widely used statistical techniques with data and typically there are 2 curbing concepts here. The idea of a dependent variable which is known also in the literature as a response variable or regressand or a predicted variable or simply the output variable. The variable whose output we desire to predict based on the model. So, the symbolic way of denoting this output variable is by the symbol y.

On the other hand we have what is called the independent variable, this is also known in the literature sometimes as the predictor variable or the regressor variable as opposed to predicted and regressand or it is also known as the exploratory variable or very simply as the input variable. We will use the term independent variable for this and dependent variables for the response we will not use the other terms in this talk.

So, the independent variable is denoted by the symbol x typically. So, we have let us for the simple case assume that we have only one variable, which we denote by the variable x the independent variable and we have another variable called the dependent variable, which we wish to predict and we will denote it as y.

(Refer Slide Time: 05:04)



There are several different classifications and we are just going to give you a brief idea of that. We can have what is called a Univariate 2 regression problem or a multivariate regression problem. The univariate is the simplest regression problem you can come up across, which consists of only one dependent variable and one independent variable. On the other hand if you talk about a multivariate regression problem you have multiple independent variables and multiple dependent variables.

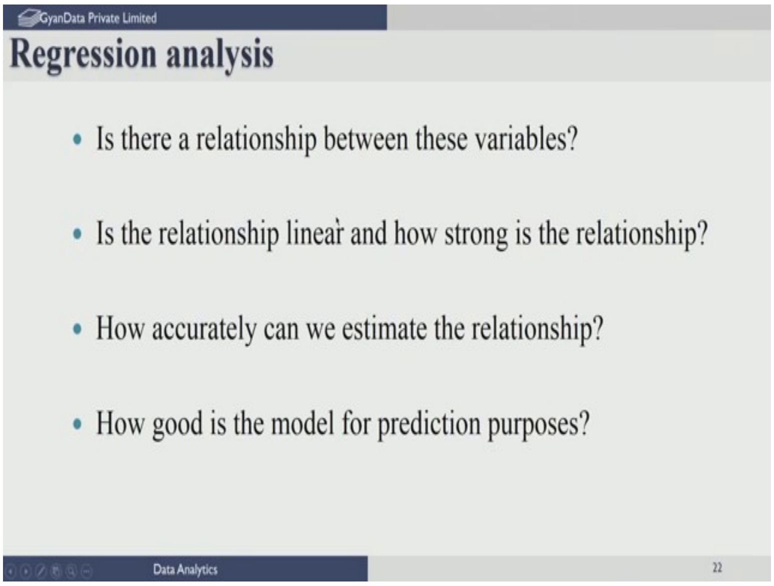
So, to understand the subject it is better to take the simplest problem understand it thoroughly and then you will see the extensions are fairly easy to follow. We can also have what is called linear versus. non-linear regression. Linear regression the relationship that we seek between the dependent and the independent variable is a linear function.

Whereas in a non-linear regression problem the functional relationship be-tween the dependent and independent variable can be arbitrary, can be quadratic, can be sinusoidal or can be any arbitrary non-linear function. And we wish to discover that non-linear function that best describes this relationship that is what forms part of non-linear regression.

We could also classify regression as simple versus multiple simple regression is the case of this single dependent and single depend independent variable also called the SISO system. And multiple regression linear regression is the case when we have one dependent variable and many independent variables or what is called the miso case multiple input single output.

So, these are various ways of denoting the regression problem, we will always look at the simplest problem to start with which is the simple linear regression, which consists of only one independent one dependent variable and analyze it thoroughly.

(Refer Slide Time: 06:52)



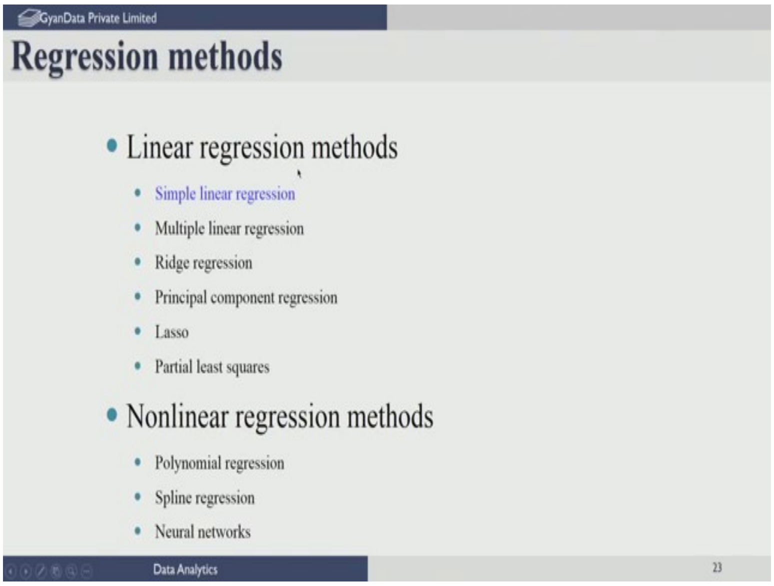
So, the first thing that the various questions that we want to first ask before we start the exercise is do we really think there is a relationship between these variables. And if we believe there is a relationship, then we would not want to find out whether such a relationship is linear or not.

Of course, in linear regression we are going at with the assumption there exists a linear relationship, but you really want to know whether such a relationship linear relationship exists. And how strong is this? How strongly the independent variable affects the response of the dependent variable?

Also we are interested since we are dealing with data that that has ran errors or stochastic in nature and we only have a small sample that, we can gather from there from the particular application. We want to ask this question, what is the accuracy of our model? In terms of how accurately we can estimate the relationship or the parameters in this model.

And if we use this model for prediction purposes subsequently how good it is? So, these are some of the questions that we would like to answer, even in the process of developing the regression model.

(Refer Slide Time: 08:06)

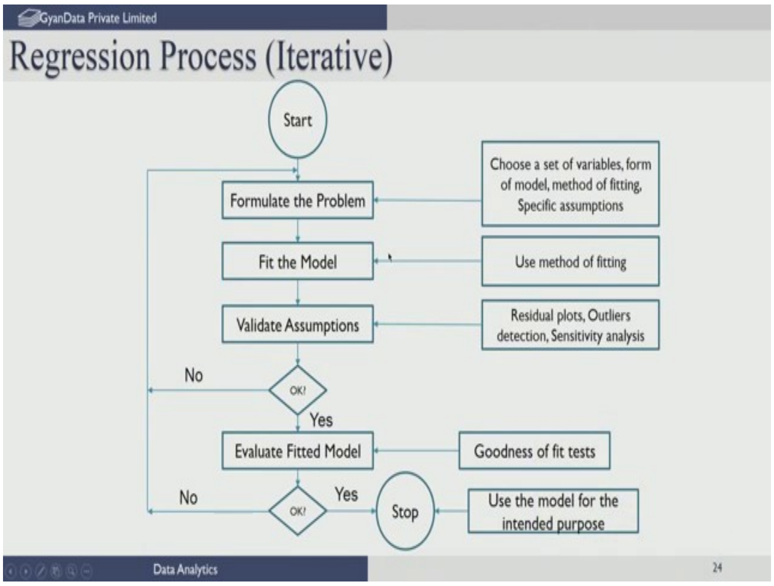


So, there are several methods also, that are available in the literature for performing the regression depending on the kind of assumptions you make and the kind of problems that may you may encounter. As I said the simple linear regression is the very very basic technique, which we will discuss thoroughly. Multiple linear regression is an extension of that for multiple independent variables, but there are other kinds of problems you may encounter, when you have several variables independent variables.

And those have to be attacked differently and there are techniques such as ridge regression or principal component regression, lasso regression, partial least squares and so on so forth, which deals with these kinds of difficulties that you might encounter in multi linear regression.

Of course, in non-linear regression there is again a plethora of methods, I am only listed just a few examples you could have polynomial or spline regression, where the type of equations or functional relationship you specify a priori, you can have neural networks or today a support vector regression. These are methods that are used to develop non linear models between the dependent and independent variable. Now let us take only the simple linear regression and go further.

(Refer Slide Time: 09:15)



So, you have to understand that the regression process it is itself is not a once through process, it is iterative in nature. So, the first question that you should ask us the purpose. Before we even start the regression you ask what is the purpose, what are you trying to develop the model for? Like I said in the business case we are developing the model in order to determine set the price selling price of the thing.

So, you are really interested in how this selling price affects sales. That is the purpose that you have actually got. In the case of a the engineering case we said the purpose is to replace a difficult to measure variable, by other easily measured variables and this model using a combination of the model and other easily measured variables we are predicting a variable, which is difficult to measure online. And then obviously, we can monitor the process using that that parameter.

So, the purpose for each thing has to be well defined, then that leads you to the selection of variables. Which is the output variable that you want to predict and what are the input variables that you think are going to affect the output variable. And so, you choose to set of variables and take measurements, get a sample, do design of experiments, which is not talked about in this whole what we called lecture. So, we will do proper design of experiments in to get what we call meaningful data and once you have the data, we have to decide the type of model. When we say type of model it is a linear model or non-linear model.

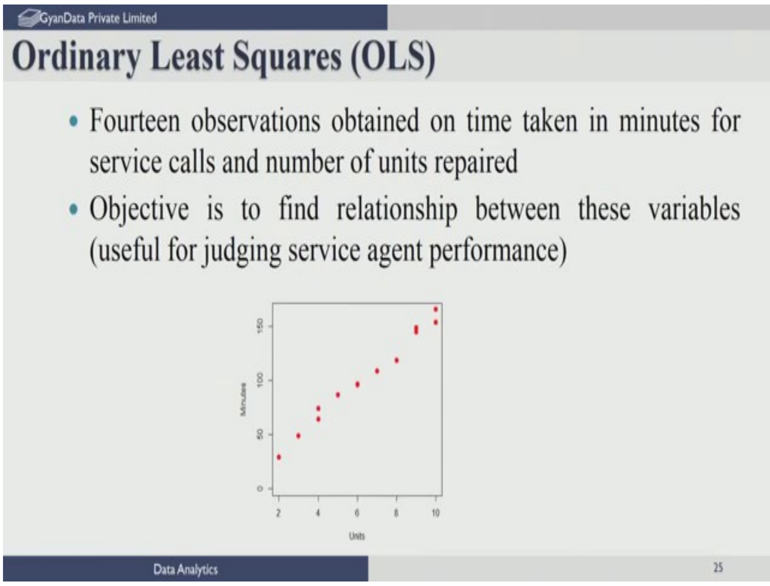
So, let us say we have chosen one type of model, then you have to actually choose the type of method that you are you going to use in order to derive the parameters of the model or identify the model as we call it.

Once you have actually done that unfortunately when we use a method it comes with a bunch of assumptions associated you would like to validate or verify, whether the assumptions we have made, in deriving this model are correct or perhaps they are wrong. What this is done by using, what we call residual analysis or residual plots. So, we will examine the residual to kind of judge, whether the assumptions were made in developing the model are acceptable or not.

Sometimes you might have also a few data you may experiment and data may be very bad, but you do not know this a priori, you would like to throw them out they might affect the quality of your model. And therefore, you would like to get rid of these bad data points and only use those good experimental observations for building the model. How to identify such outliers or bad data is also part of the regression. You remove them and then you actually have to redo this exercise.

Finally once you develop the model you want to actually do sensitivity analysis. Is there a if we have a small error in the data how well how much it affects the response variable and so on.

(Refer Slide Time: 15:12)



So, this is sensitive analysis you do or if there are many variables we would like to ask this question are all variables equally important or should I discard one of the input variables and so on. So, these are the things that you would do and once you have built the best model that you can from the given data and the set of variables you have chosen then you proceed further.

So, the data that you use in building this model or regression model is also called the training data. You have used the model to train you to use the data to train the model or estimate the parameters of the model. Such that a data set is also denoted as the training data set. Now once you have built the model you would like to see how well does it generalize can it predict the output variable or the dependent variable for other values of the independent variable which you have not seen before.

So, that comes to the testing phase of the model. So, you are evaluating the fitted model using data which is called test data. This test data is different from the training data. So, when you do experimental observations, if you have a lot of data you set apart the sum for training and remaining for testing typically 70 or 80 percent of the data experimental data is used for training or fitting the parameters. And the remaining 20 are used to test the model. This is typically done, if you have a large number of data points.

If you fewer number of observations then you there are other techniques we will actually explain or how to evaluate written models with small samples that you have. So, you first evaluate find out how well the model predicts on data that it does not seen before and once you have satisfied with it, then you can stop. Otherwise if this model that you have developed even the best model that you have developed under whatever assumptions linear model and so on so forth that you assumed, is not adequate for your purpose you go ahead and change the model type, you may want to actually now consider a non-linear model maybe introduce a quadratic term or you might want to more look at a more general form and redo this entire thing.

It may also turn out that whatever you do you are not getting a good model then maybe you should even look at the set of variables you have chosen and also the type of experiments that we have conducted. So, there could be problems with those that is probably affecting the model development phase. So, when all your attempts have failed you may want to even look at your experimental data that you have gathered. What how did the how did you conduct the experiments, whether there was any problem with that or the variables when you select and did you miss out some important ones.

So, there was quite a lot to regression. What we are going to describe only a small part we are not giving you the entire story, we are only providing a short story on how to formulate or t a model for a linear case. How to validate assumptions and how to evaluate the fitted model. This is basically going to be the focus of the lectures.

So, let us take one small example, which we will use throughout. This is a data of 14 observations small sample, which we have taken on a servicing problem service agents. These service agents, let us say it is like Forbes aquaguard service agent that comes to your house. They go visit several houses and they take a certain amount of time to kind of service the unit or repair it if it is down.

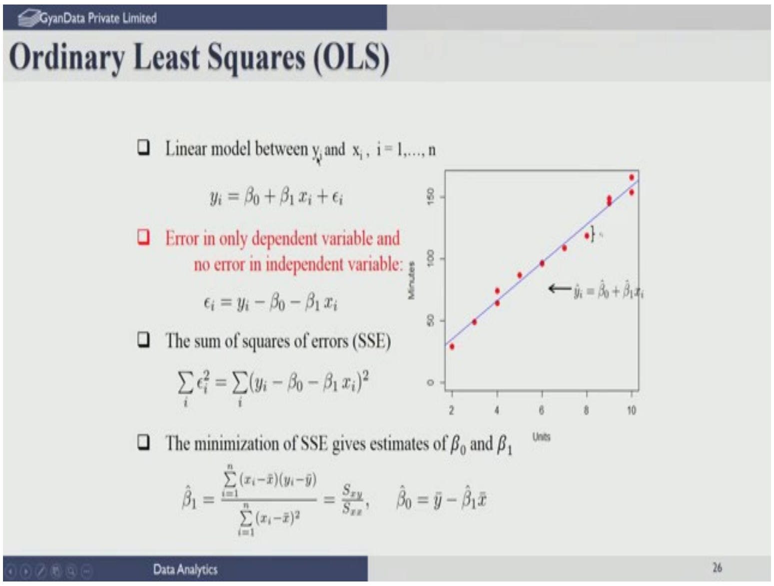
So, they will report the total amount of time taken in minutes let us say for through that they have spent on servicing different customers and the number of units that they have serviced in a given day. So, let us assume that every day the service agent goes out on his rounds and notes the total amount of time he has actually spent and tells at the end of the day reports to his boss the number of units that he has repaired he or she has replied. Let us say that there are several such agents roaming around the city and so on and each of them come back and report.

Let us say there are 14 such data points of the same person or multiple persons that you have actually gathered and from this the question that we want to actually answer let us say is given this data suppose as an agent gives you data, you monitor him for week or month on how many how much time they spending, and how many units is repairing every day and want to judge the performance of the agent service agent. In order to reward or appropriately kind of you know improve his productivity.

So, if you know a relationship between the time taken and number of units repaired which you believe should happen if somebody takes more time and is doing nothing not repairing much, then there is some inefficiency in them maybe he is wasting too much time in between travel or whatever. So, we need to find out right. So, the purpose is to actually judge the service agent performance and do performance incentives in order to improve productivity of these agents. So, we are interested in developing a relationship between number of units and the time taken by something or vice versa now.

For the sake of argument right now I have plotted as we said in A 2 variable case you can visually plot the data scatter plot. So, you plot the data first I have taken units on the x axis and the minutes on the y axis, I will discuss shortly whether we should choose units as the independent variable or minutes as the dependent variable or vice versa, but for the time being I have just plotted it on the x and y axis and look at the spread of data and it looks like there is a linear relationship between the 2 variables. Now we want to build this linear relationship, because from the trend of the data we believe a linear relationship exists let us go ahead with an assumption and just try to build this linear model.

(Refer Slide Time: 18:08)



Now, comes the exact mathematical form in which we state this problem. We have data points x I of the independent variable we have n data points in the example n is 14 and y is the dependent variable which we want to use for prediction or whatever purpose that we want for this model.

In this case as I said, given both x and y I would like to rate if some service agent comes and tells this is the time, I have spent and this is a number of units I have I have repaired and based on it is performance for a month or a week you would like to rate the service agent that is the purpose for building this model.

So, what we do is we have these 14 observations we have taken and we are going to set up the model form. So, as I said we are decided to go with a linear model and the linear model in general can be written like this between y and

x. And this is said that every data point whatever b y b y, whether it is time or units in the previous case, which you have taken as the dependent variable. Can be written in terms of the independent variable as β_0 naught a constant called the intercept term + β_1 times x i, where β_1 represents what we call the slope.

So, β_1 tells you how a change in x_i affects the response variable y. So, β_0 is a constant it is an o set term, but given xi the independent variable, it is not possible to get whatever observations we have, observations always contain some error and that error is denoted by ε i. So, ε i represents an error. Now we have to ask this question what is this error due to. There could be several reasons for this error. This could be because we have not measured xi precisely or yi precisely or it could be that the model form we have chosen is perhaps inadequate to explain the relationship, between x and y.

And therefore, whatever we are unable to explain is denoted as ε i and it is called a modeling error. So, in this particular ordinary least squares regression that we are going to deal with we will assume whenever we set up the problem xi the independent variable is assumed to be completely error free. In the sense, we have measured xi exactly. There is no error in reporting of xi. On the other hand the dependent variable yi could contain error. We allow for errors in the reporting of yi, but the error is not what we call a systematic error it is a random error that is how we are modeling ε i or you could also look at ε i as a modeling error.

In this particular case where you can say this linear model is only an approximation of the truth and anything that we are not able to explain perhaps can be treated like a random error modeling error. Whatever be the reason the most important thing to note is that this particular model form re ordinary least squares methodology a formulation does not allow error in the independent variable.

So, when you choose the independent variable one of the things you to do carefully is that you should ensure that this thing is the most accurate among the 2 variables. If you have A_2 variable case you should choose the independent variable as the one which is the most accurate one. In fact, it should be probably error free.

So, let us take the case of the units and minutes. Typically the number of units repaired by a service agent will be reported exactly, because he will have a receipt from each customer and saying that the unit was serviced, you give this back and the total number of receipts that the service agent has gathered precisely represents the number of units service.

So, there cannot be an error unless somebody transcribes this thing, the error in transcription, this can be exactly counted and you would have a precise idea it is an integral number that cannot be an error in this. On the other hand the amount of time taken could vary because of several reasons; one because this guy reported the total time he actually started out on the day and when he returned to the office end of the day. And this could involve not just the service time, but also travel time and depending on the location, the travel time could vary, it could vary from time of day, depending on the traffic, it could also vary because of congestion or a particular even that has happened.

So, the time that has been reported contains other factors that we may not have precisely considered, unless the service agent goes with that stopwatch and measures exactly the time for repair. Typically you will report the total time spent in in servicing all of these units including travel time and so on that is the kind of data that you might get.

So, you should regard the minutes as only an approximation, you can not say that is only due to servicing, but also could have other factors which you treat as random disturbance as random error. So, it is better in this case to choose units as the x variable, because that is precise that has no error and y where minutes as the dependent variable.

Notice there might be an argument saying that you know you should always choose the variable which you wish to predict as the dependent variable, need not once you build a model you can always decide to use this model for predicting x given y or y given x.

So, it does not matter how you cast this equation, how you build this model, it is more important that when you apply ordinary least squares, you should actually ensure that the independent variable is an extremely accurate measurement or it represents the truth as closely as possible. Whereas, y could contain other factors or errors and so on and it is this method is tolerant to it.

So, this goes if on the other hand if you believe both x and y contain significant error, then perhaps you should consider other methods called total least squares or principal component regression that we will talk about later. If positive not in this lecture, but if we have the time we will do it later.

So, essentially what I am saying is that once you have decided based on purpose based on the kind of quality of the of the measurements, what is the independent dependent variable, then you can go ahead and say given all the observations n observations, what is the best estimate of β_0 and β_1. As I said that β_0 is the intercept parameter and β_1 is the slope actually geometrically interpret β_0, β_0 represents the value of y when x is 0. So, when you put x = 0 and you look at where this line intercepts the y axis this vertical distance is β_0 and the slope, which represents the slope of this regression line that is β_1 so, your estimating the intercept and slope.

So, now what is the methodology for estimating this β_0 and β_1. So, what we will do is we will do a kind of a thought experiment you give values of β_0 and β_1 and then you can draw this line. So, we will ask different people let us say, values of β_0 and β_1 and draw appropriate lines. Again the line shape that the slope and the intercept will be different depending on what value you propose for β_0 and β_1. Then once you have done this we will actually go back and find out how much deviation is there between the observed value and the line. In this particular case we will say the observed value let us take this observed value is yi corresponding to this xi, which is 8.

Now, the line if this particular equation is correct then this is the predicted value of y, which means for this given value of x i according to this equation you believe y predicted should be here. And then this deviation between the observed value and the predicted value, which is on this line, the vertical distance is what we call the estimated error.

So, you do not know what the actual error is, but if you propose values for β_0, β_1, immediately I am able to derive an estimate for this error which is the vertical distance of the point from that line. We estimate this error for all data points.

So, we compute ei for every data point yi using the proposed parameters β_0 and β_1 and the value of the independent variables we have for all the observations. Now what we do is we can say as a metric, what is the best line? We propose that the best line is 1, which minimizes the deviations some square deviations or the distances .

So, overall the data points, we will compute this distance which is geometric distance is nothing, but square of this value we will compute this and sum over all the data points n data points. And we try to find

β_0 and β_1, which minimizes this sum squared value or minimizes the sum of the vertical distances or the point from that line.

So, the notion of a best t line in the least square sense or the ordinary least square sense is one that minimizes the vertical distance of the points from the proposed line. Now you can, once you set up this formulation, then we can say then who over gives the best β_0 and β_1 will have the minimum vertical distance of the points from that line. And this can be done now analytically instead of asking you now for this β_0 and β_1 I try to solve this optimization problem, which means minimize this, find out β_0 β_1, which minimizes this and this what is called the unconstrained optimization problem with 2 parameters you differentiate this with respect to β_0 set it = 0 for those called the first order conditions.

Those of you have done a little bit of optimization will know that our calculus, will know all I have to do is differentiate this function with respect to β_0 set it = 0 differentiate this function with respect to β_1 set it = 0 and solve the resulting set of equations. And finally, I will get the solution for β_0 and β_1, which minimizes this sum squared error. So, the least squares technique uses this as a criterion in order to derive the best values of β_0 and β_1.

Of course, you can counter by saying I will use some other metric maybe I should have used absolute value. That will make the problem difficult this method was proposed in the late 1700s by Gauss or another person called Legendre and it has become popular as a methodology although in recent years other methods have taken over.

So, the method of least squares is a very popular technique and it gives you parameters analytically for the simplest cases. So, you get β_0 estimated. So, the estimate that you derive is not what you it is not that you should you should treat this estimate as actually the truth it is an estimate from data. Had you given me a different sample maybe I would have got a different estimate remember that. The estimate is always a function of the sample that you are given.

So, we denote such estimates by this hat always implies it is an estimated quantity and the estimated value of β_1 turns out to be the cross covariance between x and y divided by the variance of x. You can prove this. So, remember you this cross covariance is essentially like a Pearson's coefficient. So, the Pearson's coefficient said if the coefficient, Persons correlation coefficient, was close to 1 or - 1, you said that there is you could interpret that there may exist a linear relationship.

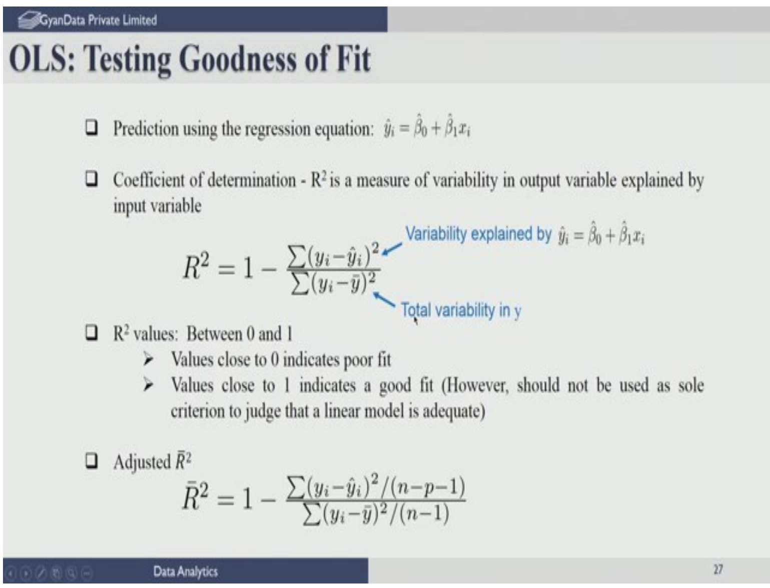
Similarly you can see $\beta 1$ is a function of that coefficient. It depends on the cross covariance between x and y and β_0 the intercept turns out to be nothing, but the mean of y - the estimated value of β_1 slope parameter multiplied by the mean of x. This is your intercept parameter. Of course, one could also ask suppose I know that if x is 0 y is 0 I know that a priori.

In this particular case for example, if you do not service any units which means you have not traveled you are not let us say you are on holiday then clearly you would have taken 0 time for servicing. So, I know in this particular case perhaps that that if you process 0 units you should not have taken any time.

So, therefore, the intercept should pass through 0. If you know it and you want you want to force this line to pass through 0 0, the origin, then you should not estimate β_0 you should simply remove this parameter and simply write $y = \beta_1 x$. And in which case the solution for β_1 will turn out to be again Sxy by Sxx except that this Sxy is a cross covariance not about the mean, but about 0. Which means you set $\overline{xy} = 0$ in this expression and you will get σxy in the numerator over all data points divided by σxi square not σ xi - $\overline{x}$ square.

So, essentially you are taking the variance around 0 and the cross covariance around 0 0 0 and then you will get the estimated value of β_1. Of course, β_0 in that case is assumed to be 0. So, the line will pass through 0 0 and you will get another slope. You are forcing the line to pass through 0 0. Remember you have to be careful when you do this, because it will, unless you are sure that should pass through the origin, you should not force this thing you will get a bad t. If you know it and you want demand it it makes physical sense then you are well within your rights to force $\beta_0 = 0$ do not estimate it. That can be done by simply taking the cross covariance and variance around 0 instead of around their respective means .

(Refer Slide Time: 32:25)



So, this is as far as getting the solution is concerned. So, now, once you get the solution you can ask for every given xi, what is the corresponding predicted value of y i using the model. So, you plug in your value of xi in the estimated model which is using the estimated parameter $\hat{\beta}_0$ and $\hat{\beta}_1$ and I will call this prediction $\hat{y}_1$ it is also an estimated quantity for any given xi,I can estimate the corresponding yi using the model.

And geometrically if you actually try to estimate yi given this point it will fall on this line. You draw the vertical line which intersects this particular regression line and that particular point on that line will represent y i hat for every point. So, for this point it is actually the corresponding predicted value will lie on this line here if this is the best t line. The blue line represents the best t line in the least square sense.

So, you can do this for any new point which you have not seen before in the test set also. Let us look at some couple of other measures which you can derive from this. We can talk about what is called the coefficient of determination r squared, which is defined in this manner. It is just 1 - difference between the observed value and the predicted value squared difference summed over all data points, divided by the variance of y, which is $(yi - \overline{y})^2$.

So, essentially this particular quantity called r squared will be between 0 and 1 we can show. So, how to interpret this. The denominator is the variability in yi what do you mean by that? That is if you are given just yi and trying to find out how much variance there it is there in the data this particular thing divided by n of course, gives you the variance of y.

So, you can say this much variability exists in the data suppose I build the model and try to predict yi if xi had a influence on y, then I should be able to reduce it is variability I should be able to do a better prediction and the difference between yi and ŷi should be lower. If xi had a strong influence in determining y .

So, the numerator represents the variability, which is explained by the explanatory variable x or the independent variable x i. So, if the numerator is approximately equal to the denominator then you basically get 1 and R squared will be close to 0, the implication of this is xi has a very little impact on explaining y and probably there is no relationship between y and x. If on the other hand if the numerator is close to 0 and then you get R square close to 1 it implies that the xi can explain the variation in yi, which means there is a strong relationship between xi and yi.

So, values close to 0 indicates a poor t, values close to one indicates a good t, but the problem does not end there. If you get R squared close to 1 you should not conclude your job is done and the linear relationship is good and so on. And the Anscombe data for example, when we saw last class, if you try to find the Anscombe data for the 4 datasets you will get all r squared close to one and that does not mean that the linear model is good.

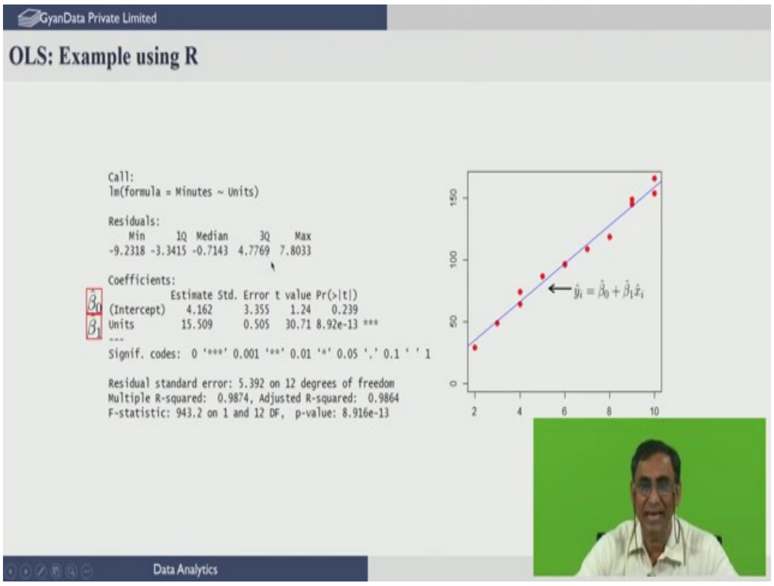
You should look at other measures before you conclude, conclusively determine, that a linear model is good a. A value close to one is a good starting point. Yes you can now be a little bit assurance, you get that the linear model perhaps can explain the relationship between yi and xi. There is also something called adjusted r squared, which will come across which is essentially this. If you look at the denominator, you can say that if you try to estimate a constant value. Suppose you say xi has no influence and dropped that yβi and try to estimate β_0 in the least square sense. You will find that the best value of β_0 is actually best estimate is just $\overline{y}$.

So, you can regard the denominator as fitting a model with just the parameter β_0. On the other hand the numerator you have used 2 parameters to t the model. Whenever you use more parameters typically you should get a better t. So, generally the numerator value is obtained, because you have used 2 parameters. Whereas, the denominator you used only 1 parameter so, you have to account for the fact that, you have used more parameters to obtain a better t and not because there is a linear relationship between yi and xi.

So, you should go back and account for this what we call the number of parameters you have used or the number of degrees of freedom that is used in estimating the numerator. For example, you have n data points and in this case the p = 2 parameters. So, n - I am sorry p = 1, which happens to be only β_1. So, n - 2 would represents the number of degrees of freedom used to estimate this numerator variability whereas, n - 1 is used to estimate the denominator variability, because you have used only the parameter β 0 for denominator whereas, you used 2 parameters to estimate the numerator.

So, you should adjust this by dividing the number of degrees of freedom and the adjusted R square essentially makes this adjustment and give it is different from R squared, but it is a more accurate way of what I call judging whether there is a good linear good model between the dependent and independent variable. And in this case p = 1, because I have only 1 explanatory variable, but this can extend to many independent variable case where p is the number of independent variables you have chosen for fitting the model.

(Refer Slide Time: 38:31)



Finally we will end with the R command. The R command for fitting a linear model is just called lm, if you have loaded the data set and then you say that what kind of what you call variable is the independent variable and what is the dependent variable, you indicate. In this case we have indicated the minutes as the dependent variable and units as the dependent variable and these are variables that forms part of the data set, they are defined as these variables and therefore, you are using them.

So, loading of the data set you would have already seen, lm is the one that you used to build the model, you indicate what is the dependent and independent variable. And then you will get an output that is given here first you will get the range of residuals, which I said is the estimated value of ε i for all the data points in this case all the 14 residuals are not given the max value min value the first quartile third quartile in the median are given here.

And I will only now look at 2 parameters the β_0, which is the first the intercept is called the β_0 estimated values here and the slope parameter the estimated values 15.5 for this particular data set. Now I will also now only focus on this particular line, which talks about the R squared value, which we explained to judge the quality of the model it is a very high R squared you get or the adjusted R squared.

So, from this we can conclude maybe a linear model is explains their relationship between x and y very precisely, but we are not done yet we have to do residual analysis we have to do further what you call plots in order to judge and conclude that linear model is adequate. We will do this and the other things that outputs that are gives as in the subsequent lectures I will explain them. And, we will see you in the next lecture.

Thank you.