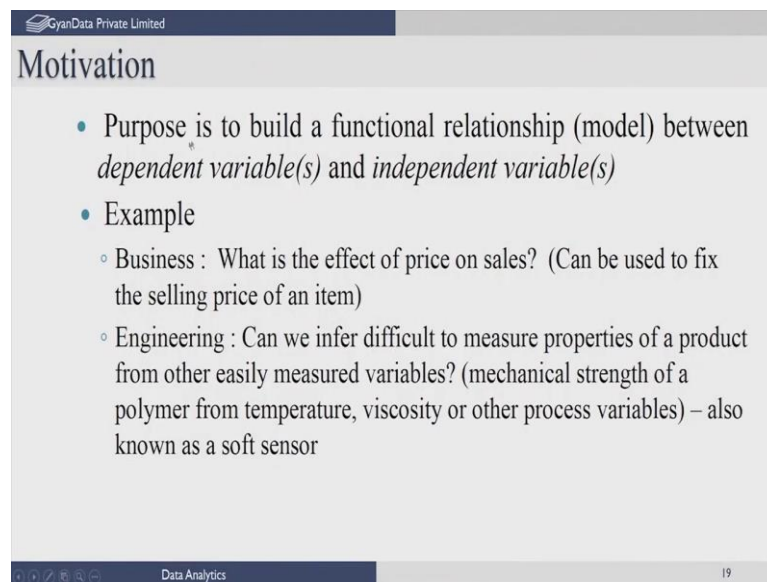


Python for Data Science
Prof. Raghunathan Rengasamy
Department of Computer Engineering
Indian Institute of Technology, Madras

Lecture - 27
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: 00:38)



GyanData Private Limited

Motivation

- Purpose is to build a functional relationship (model) between *dependent variable(s)* and *independent variable(s)*
- Example
 - Business : What is the effect of price on sales? (Can be used to fix the selling price of an item)
 - Engineering : Can we infer difficult to measure properties of a product from other easily measured variables? (mechanical strength of a polymer from temperature, viscosity or other process variables) – also known as a soft sensor

Data Analytics 19

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 define these variables and how you choose them for the intended purpose little later, but essentially we are building a relationship between two 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 fix 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 difficulty to measure property case in point is the mechanical strength of a polymer. This is very difficult to measure on line continuously or the other hand process conditions such as temperature viscosity of the 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)

GyanData Private Limited

Regression - Basics

- One of the widely used statistical techniques
- Dependent variables also known as *Response variable*, *Regressand*, *Predicted variable*, *output variable* - denoted as variable/s y
- Independent variable also known as *Predictor variable*, *Regressor*, *Exploratory variable*, *input variable* - denoted as variable/s

Data Analytics

So, regression happens to be one of the most widely used statistical techniques with data and typically there are two ah concepts here. The idea of a dependent variable which is known also in the literature as a response variable or a 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 the 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 variable. 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)

GyanData Private Limited

Regression types

- Classification of Regression Analysis
 - Univariate vs Multivariate
 - *Univariate*: One dependent and one independent variable
 - *Multivariate*: Multiple independent and multiple dependent variables
 - Linear vs Nonlinear
 - *Linear*: Relationship is linear between dependent and independent variables
 - *Nonlinear*: Relationship is nonlinear between dependent and independent variables
 - Simple vs Multiple
 - Simple: One dependent and one independent variable (SISO)
 - Multiple: One dependent and many independent variables (MISO)

Data Analytics 21

There are several different classifications and we are I was just going to give you a brief idea of that we can have what is called the univariate regression problem or a multivariate regression problem. The univariate is the simplest regression problem you can come up cross 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 between the dependent and independent variable can be arbitrary, can be quadratic, can be sinusoidal or can be any arbitrary non-linear function.

And we will wish to discover that non-linear function that best describes this relationship that 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 dependent 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: 06:51)

GyanData Private Limited

Regression analysis

- Is there a relationship between these variables?
- Is the relationship linear and how strong is the relationship?
- How accurately can we estimate the relationship?
- How good is the model for prediction purposes?

Data Analytics 22

So, the first thing that the various questions that we want to first ask where 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 ahead with the assumption there exists a linear relationship, but you really want to know whether such a relation, linear relationship exists and how strong is this, how strongly the independent variable affects the response of the dependent variable. Also we are interest 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)

GyanData Private Limited

Regression methods

- Linear regression methods
 - Simple linear regression
 - Multiple linear regression
 - Ridge regression
 - Principal component regression
 - Lasso
 - Partial least squares
- Nonlinear regression methods
 - Polynomial regression
 - Spline regression
 - Neural networks

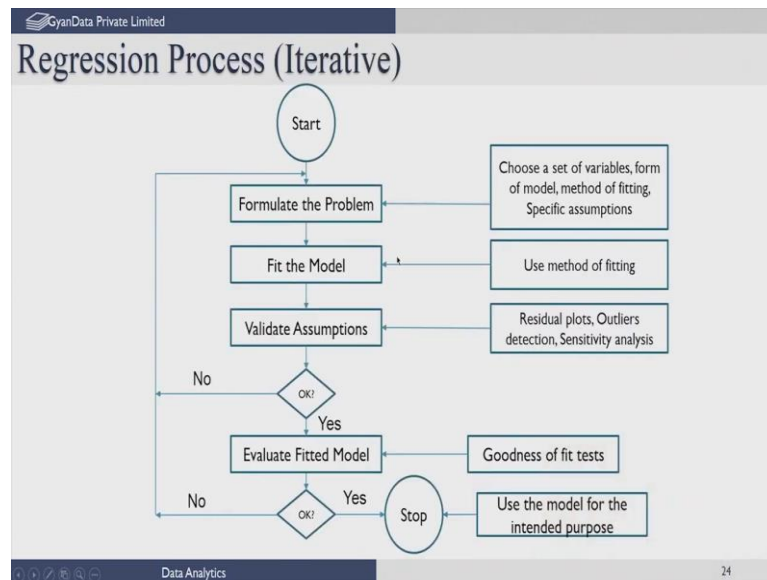
Data Analytics

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 tackled differently and there are techniques such as Ridge regression or Principal Component regression, Lasso regression, Partial Least Squares and so on and so forth which deals with these kind of difficulties that you might encounter in multi linear regression. Of course, in non-linear regression there is again a plethora of methods. I have only listed a 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 itself is not a once through process; it is iterative in nature. So, the first question that you should ask is 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 a 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 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 the 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 the 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 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 are 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 plots to kind of judge whether the assumptions we have made in developing the model are acceptable or not.

Sometimes you might have also a few data, experimental data. That data may be very bad, but you do not know this a priori you 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 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. So, this is sensitive analysis you do or if there are many variables you 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 used 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 the 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 a 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 have fewer number of observations, then there are other techniques. We will actually explain or how to evaluate fitted models with small samples that you have.

So, you first evaluate find out how well the model predicts on data that it has 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, linear model and so on and so forth that you have assumed it 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, ok.

It may also turn out that whatever you do you are not getting a good model, then maybe you should once look at the set of variables you have chosen and also the type of experiments that you have conducted. So, there could be problems with those that is probably affecting the model development phase. So when you all your atoms are 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 selected, did you miss out some important ones.

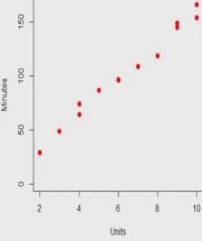
So, there was quite a lot to regression. What we are going to describe only a small part, we are not giving you the entire story. We are only providing a short story on how to formulate or fit a model for a linear case. How to validate assumptions and how to evaluate the fitted model, this is basically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

(Refer Slide Time: 15:12)

GyanData Private Limited

Ordinary Least Squares (OLS)

- Fourteen observations obtained on time taken in minutes for service calls and number of units repaired
- Objective is to find relationship between these variables (useful for judging service agent performance)



Units	Minutes
2	50
3	60
4	70
4	80
5	90
6	100
7	110
8	120
9	130
9	140
10	150

Data Analytics

25

So, let us take one small example which we will use throughout. This is a data, fourteen observations small sample which we have taken on a servicing problem service agents. These service agents let us say its 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 we 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 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 repaired. 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 fourteen such data points that 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 he is repairing every day and want to judge the performance of that agent, service agent in order to reward or appropriately kind of you know improves 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 he is doing nothing not repairing much, then there is some inefficiency in the in the maybe 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 thing or vice versa.

Now, for the sake of argument right now I have plotted as we said in the two variable cases 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 two variables, ok. Now, we want to build this linear relationship because from the trend of the data we believe a relationship exists. Let us go ahead with an assumption and just try to build this linear model ok.

(Refer Slide Time: 18:09)

GyanData Private Limited

Ordinary Least Squares (OLS)

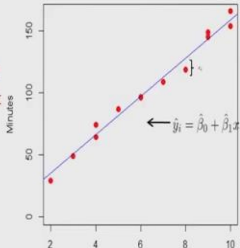

- Linear model between y_i and x_i , $i = 1, \dots, n$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$
- Error in only dependent variable and no error in independent variable:

$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$
- The sum of squares of errors (SSE)

$$\sum_i \epsilon_i^2 = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$
- The minimization of SSE gives estimates of β_0 and β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Data Analytics

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 spent and this is a number of units I have ai have repaired and based on its 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 fourteen 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 be 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 a constant called the intercept term plus $\beta_1 x_i$ where β_1 represents what we call the slope.

So, β_1 tells you how change in x_i effects the response variable y , ok. So, β_0 is a constant. It is an offset term, but given x_i 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 x_i precisely or y_i precisely or it could be that the model form we have chosen is perhaps inadequate to explain the relationship between x_i and y_i 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 x_i the independent variable is assumed to be completely error free in the sense we have measured x_i exactly. There is no error in reporting of x_i .

On the other hand the dependent variable y_i could contain error. We allow for errors in the reporting of y_i , 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 or for 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 should do carefully is that you should ensure that this thing is the most accurate among the two variables. If you have a two 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 ok. 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 serviced.

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. They 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 event 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 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 cannot say that is only due to servicing, but also could have other factors which you treat as random disturbance or 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 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 ok. If 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 read that β_0 is the intercept parameter and β_1 is the slope to 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, you are estimating the intercept and slope. So, now what is the methodology for estimating this β_0 , ok? 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. The line shape that the slope and the intercept will be different depending on what value you proposed for β_0 and β_1 , then once you have done this you will actually go back and find out how much deviation is there between the observed value and the line ok.

In this particular case, we will say the observed value let us take this observed value is y_i corresponding to this x_i 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, ok. 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, ok. So, we compute e_i for every data point y_i using the proposed parameters β_0 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 one which minimizes the deviations sum squared 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 fit 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 whoever 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 two parameters. You differentiate this with respect to β_0 , set it equal to 0 for those called the first order conditions.

Those of you have done a little bit of optimization will know that or calculus will know all I have to do is differentiate this function with respect to β_0 , set it equal to 0, differentiate this function with respect to β_1 , set it equal to 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 square 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 we 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, our and another person called Legendre and they 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 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 had given me.

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 Pearson's correlation coefficient was close to 1 or -1, you said that there is a you could interpret that there may exist a linear relationship.

Similarly you can see $\widehat{\beta}_0$ is 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 apriori in this particular case for example if you do not service any units which means you have not travelled you are not let us say you are on holiday, then clearly you would have taken zero time for servicing. So, I know in this particular case perhaps that that if you process zero units, you should not have taken any time. and 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 $y=\beta_1x$ and in which case the solution for β_1 will turn out to be again S_{xy} by S_{xx} . Except that this S_{xy} is a cross covariance not about the mean, but about 0 which means you set \overline{xy} equals 0 in this expression and you will get σ_{xy} in the numerator over all data points divided by $\sigma_{(x_i-\bar{x})^2}$.

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 fit, ok.

If you know it and you want demand it makes physical sense, then you are well within your rights to ah force β_0 equal to 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: 32:26)

OLS: Testing Goodness of Fit

- Prediction using the regression equation: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Coefficient of determination - R^2 is a measure of variability in output variable explained by input variable

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Variability explained by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Total variability in y
- R^2 values: Between 0 and 1
 - Values close to 0 indicates poor fit
 - Values close to 1 indicates a good fit (However, should not be used as sole criterion to judge that a linear model is adequate)
- Adjusted \bar{R}^2

$$\bar{R}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2 / (n - p - 1)}{\sum (y_i - \bar{y})^2 / (n - 1)}$$

So, this is as far as getting the solution is concerned. So, now once you get the solution you can ask for every given x_i what is the corresponding predicted value of y_i using the model. So, you plug in your value of x_i 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 x_i , I can estimate the corresponding y_i using the model and geometrically if you actually try to estimate y_i given this point, it will fall on this line ok. You draw the vertical line which intersects this particular regression line and that particular point on that line will represent \hat{y}_i 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 fit line is the blue line represents the best fit 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,

$$\bar{R}^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2 / (n - p - 1)}{\Sigma(y_i - \bar{y})^2 / (n - 1)}$$

. So, essentially this particular quantity called R squared will be between 0 and 1 we can show, ok.

So, how to interpret this? The denominator is the variability in y_i . What do you mean by that? That is if you have given just y_i and try to find out how much variance there 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 y_i . If x_i had a influence on y , then I should be able to reduce its variability, I should be able to do a better prediction and the difference between y_i and \hat{y}_i should be lower if x_i had a strong influence in determining y , ok. So, the numerator represents the variability which is explained by the explanatory variable x_i or the independent variable x_i , ok.

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 x_i has a very little impact on explaining y and probably there is no relationship between y and x . On the other hand if the numerator is close to 0 and then you get r square close to 1, it implies that the x_i can explain the variation in y_i which means there is a strong relationship between x_i and y_i . So, values close to 0 indicates a poor fit, values close to 1 indicates a good fit, 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 year last class if you try to find the Anscombe data for the four datasets, you will get all R squared close to 1 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.

Value close to 1 is a good starting point. Yes you can now be a little bit assurance you get that the linear model perhaps can explain relationship between y_i and x_i ok. There is also something called the adjusted R square which we 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 x_i has no influence and drop that i , β_i 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 \bar{y} .

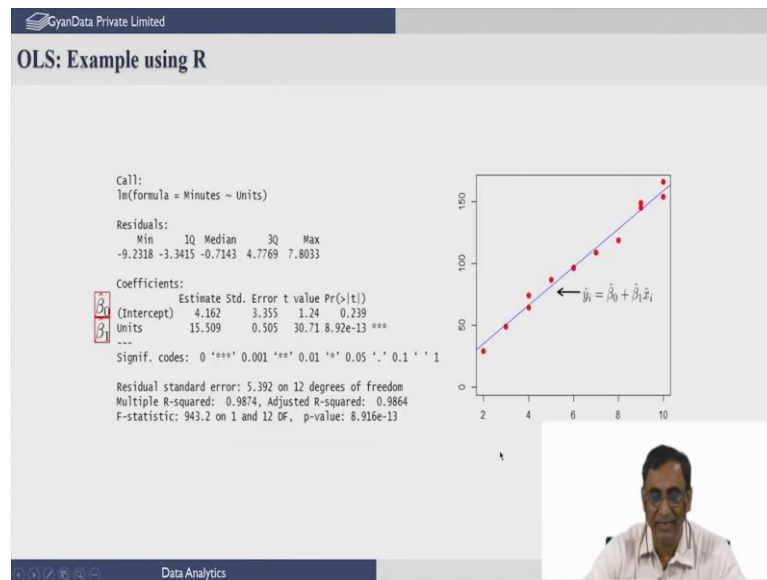
So, you can regard the denominator as fitting a model with just the parameter β_0 . On the other hand the numerator you have used two parameters to fit the model. Whenever you use more parameters typically you should get a better fit. So, generally the numerator value is obtained because you have used two parameters whereas, the denominator you used only one parameter.

So, you have to account for the fact that you have used more parameters to obtain a better fit and not because there is a linear relationship between y_i and x_i . So, you should go back and account for this what we call the number of parameters you have used and 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, $n-1$ am sorry p equals 1 which happens to be only β_1 . So, $n-2$ would represent the number of degrees of freedom used to in 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 squared essentially is makes this adjustment and give the it is different from R squared, but it is a more accurate way of what we call judging whether there is a good linear good model between the dependent and the independent variable. And in this case p equals 1 because I have only one explanatory variable, but this it can extend into a many independent variable case where p is the number of independent variables you have chosen for fitting the model.

(Refer Slide Time: 38:32)



Ok 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 a variable is the independent variable and what is the dependent variable you indicate. In this case we have indicated minutes has the independent variable and units has 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 the 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 fourteen 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 two parameters; the β_0 which is the first. The intercept is called the β_0 . Estimated value is here and slope parameter the estimated value is 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 explain 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 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 ah what you call plots in order to judge

and conclude that a linear model is adequate. We will do this and the other things that outputs that are given as in the subsequent lectures I will explain them and we will see you in the next lecture.s

Thank you.