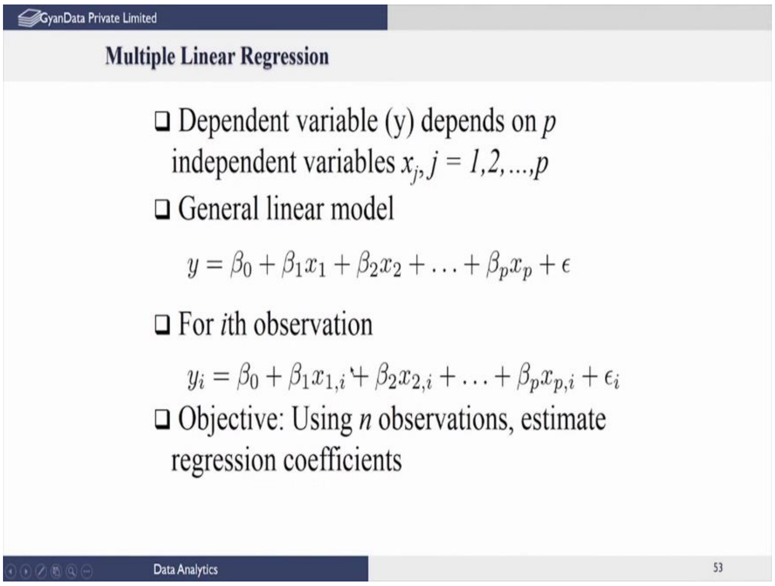
Python for Data Science
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 40
Multiple Linear Regression

Welcome, everyone to this lecture on Multiple Linear Regression. In the preceding lectures, we saw how to regress a single independent variable to a dependent variable. Particularly, we were developing a linear model between the independent and dependent variable. We also saw various measures by which we can assess the model that we built. In this lecture, we will extend all of these ideas to multiple linear regression which consists of one dependent variable, but several independent variables.

(Refer Slide Time: 00:49)



GyanData Private Limited

Multiple Linear Regression

❑ Dependent variable (y) depends on *p* independent variables $x_j, j = 1,2,\ldots,p$

❑ General linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

❑ For *i*th observation

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \epsilon_i$$

❑ Objective: Using *n* observations, estimate regression coefficients

Data Analytics 53

So, as I said that we have a dependent variable which we denote by y and several independent variables which we denote by the symbols x_j, where j equals 1 to p. There are p independent variables which we believe affect their dependent variable. We will try to develop a linear model between the dependent variable y and these independent p independent variables x_j; j equals 1 to p.

In general, we can write this linear model as before. We can say

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

where $\beta_1, \beta_{2,....,}\beta_p$ represents the slope parameters or the effect of the individual independent variables on the dependent variable.

In addition we also have an error. This error is due to error in the dependent variable, measurement of the dependent variable. In ordinary least squares we always assume that the independent variable measurements are perfectly measured and do not have any error whereas, the dependent variable may contain some error and that error is indicated as ϵ we do not know what this quantity is. We have assumed that it is a random quantity with zero mean and some variance.

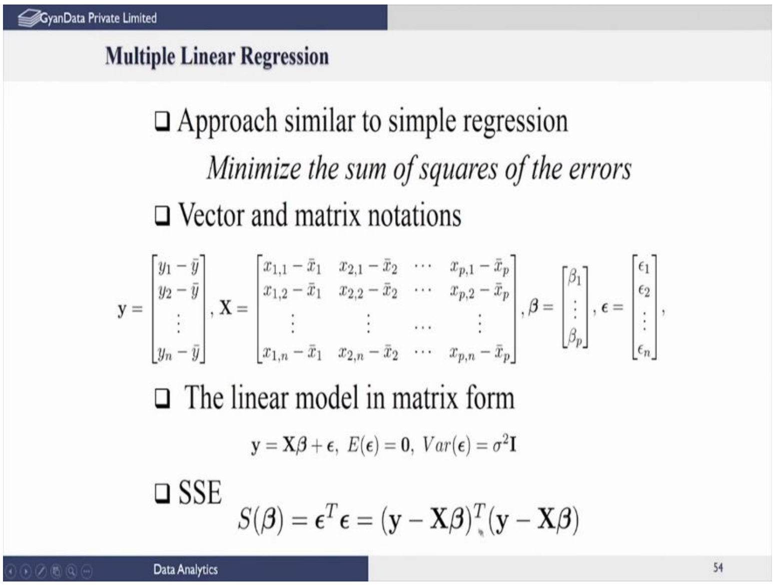
If we take the i th sample corresponding to this measurement of x_1 to x_p and y corresponding y we can say that i th sample dependent variable as

$$y = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \ldots + \beta_p x_{p,i} + \epsilon_i$$

and so on for i equals 1 to n.

We assume we have small n number of samples that we have obtained and our aim is to find the values best estimates of $\beta_0, \beta_1, \beta_{2,....,}\beta_p$ using these n sample measurements of x's and corresponding y. This is what we call multiple linear regression because we are fitting a linear model and there are many independent variables and we therefore, call the multiple linear regression problem.

(Refer Slide Time: 03:15)



GyanData Private Limited

Multiple Linear Regression

❑ Approach similar to simple regression

Minimize the sum of squares of the errors

❑ Vector and matrix notations

$$\mathbf{y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{1,1} - \bar{x}_1 & x_{2,1} - \bar{x}_2 & \cdots & x_{p,1} - \bar{x}_p \\ x_{1,2} - \bar{x}_1 & x_{2,2} - \bar{x}_2 & \cdots & x_{p,2} - \bar{x}_p \\ \vdots & \vdots & \cdots & \vdots \\ x_{1,n} - \bar{x}_1 & x_{2,n} - \bar{x}_2 & \cdots & x_{p,n} - \bar{x}_p \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

❑ The linear model in matrix form

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \; E(\epsilon) = \mathbf{0}, \; Var(\epsilon) = \sigma^2\mathbf{I}$$

❑ SSE

$$S(\boldsymbol{\beta}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Data Analytics

54

Again in order to find the best estimates of the parameters $\beta_0\ to\ \beta_p$, we actually set up the minimization of the sum squared of errors. In order to set it up in a compact manner using vectors and matrices we define the following notations. Let us define the vector y where which consists of all the n measurements of the dependent variable y_1 to y_n. We have also done one further things we have subtracted the mean value of all these measurements from each of the observations.

So, the first one represents the first sample value of the dependent variable y_1 - the mean value of y over all the measurements $\bar{y}$. So, the first sample is mean shifted value of the first observation; the second coefficient or second value in this vector is the second sample value - the mean value of the dependent variable and so on for all the n observations we have. So, these are the mean shifted values of all the n samples for the dependent variable.

Similarly, we will construct a matrix X where the first column corresponds to variable independent variable 1; again, what we do is take the sample value of the first independent variable and subtract the mean value of the first independent variable. That means, we take the mean of all these n samples for the first variable and subtract it from each of the observations of the first independent variable.

So, the first coefficient here would be x_1, 1 represents the sample value of the first independent variable; first sample first independent variable - the mean value of the first independent variable, and we do this for all n measurements of the first independent variable. Similarly, we do this for the second independent variable and arrange it in the second column. So, this one represents the observation; the first observation of the second independent variable - the mean value of the second independent variable and we do this for all p variables independent variables.

So, this particular matrix X that we get will be a n x p matrix; n is the number of rows, p is the number of columns. You can view the first row as actually the sample first sample of all independent variables for the first sample of course, we have been shifted that value and the second row is the second sample and so on and each column represents a variable. So, first column represents the first independent variable and the last column represents the p th independent variable.

So, similarly we will represent all the coefficients β, except $\beta_{0,}$ in a vector form $\beta_1\ to\ \beta_p$ as a column vector. Here basically as I am sorry a row vector. So, β_1 is the first coefficient, βp is the coefficient corresponding the pth variable. So, we have βvector which is a p x 1 vector. We can also define the ϵ the noise vector as $\epsilon_1\ to\ \epsilon_n$corresponding to all the n observations.

Now, having defined this rotation we can write our linear model in the form y equals X times β+ ϵ. Notice that we have not included β_0, we have eliminated that indirectly by doing this mean subtraction. I will show you how that happens, but you can take in that right now we have only interested in the slope parameters. This linear model only involves the slope parameters β_1 to β_p does not involve the β_0 parameter because that has been effectively removed from the linear model using this mean subtraction idea.

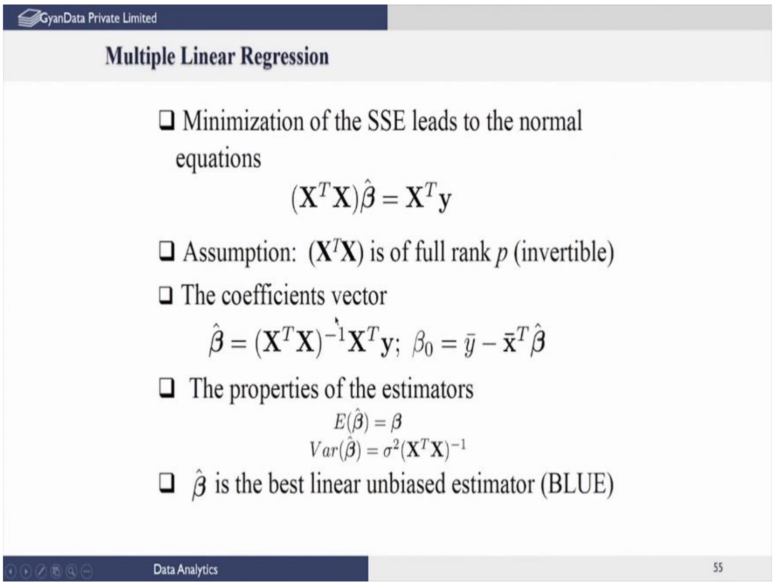
So, we can write our linear model compactly as $y = X\beta + \epsilon$ and we also make the user assumptions about the error that it is a zero mean vector in this case because it is a multivariate vector, 0 is a vector. So, expected value of ϵ_0 implies ϵ is a random vector with 0 mean and the variance; covariance matrix of ϵ is assumed to be $\sigma^2 I$.

Sigma squared identity in this form it means all the ϵ_s, ϵ_1 to ϵ_n have all have the same variance sigma squared homoscedastic assumption. And we also assume that ϵ_1 and ϵ_2 are uncorrelated or ϵ_i and ϵ_j are uncorrelated, if i is not equal to j. In which case we can write the covariance matrix of ϵ as $\sigma^2 I$.

Now, under this assumption we can go ahead and say we want to find the estimates of βso as to minimize the sum square of the errors. So, ϵ transpose ϵ is a compact way of saying the sum of all errors error squared of all the errors in all the n measurements. So, expanding this is nothing, but sum of ϵ_i^2, i equals 1 to n that is compactly written like this and this is what we want to minimize, but ϵ itself can be written as $(y - X\beta)$. So, we can write this whole thing as a $(y - X\beta)^T(y - X\ )$.

We want to minimize this which is a function of β by finding the best value of β.

(Refer Slide Time: 08:53)



GyanData Private Limited

Multiple Linear Regression

❑ Minimization of the SSE leads to the normal equations

$$(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T\mathbf{y}$$

❑ Assumption: $(\mathbf{X}^T\mathbf{X})$ is of full rank p (invertible)

❑ The coefficients vector

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y};\ \beta_0 = \bar{y} - \bar{\mathbf{x}}^T\hat{\beta}$$

❑ The properties of the estimators

$$E(\hat{\beta}) = \beta$$
$$Var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$$

❑ $\hat{\beta}$ is the best linear unbiased estimator (BLUE)

Data Analytics 55

So, if we set up this optimization problem to minimize the sum squared errors to find β, we will we can show we can by differentiating that objective function with respect to βand setting it equal to 0; we get what are called the first order conditions and these first order conditions will result in the following set of linear equations. We will get

$(X^TX)\hat{\beta} = X^Ty$.

Now, this is a p cross remember X is a n cross p matrix. So, X transpose is p cross n. So, this is a square matrix X transpose X is a square matrix of size p cross p and multiplied by the p cross 1 vector and similarly it is p cross 1 on the right hand side. So, these are p equations in p variables. Their linear equations in βX is all known, y is known. So, right hand side is like the, if you, what is that interpreted as AX= β.. It is a set of linear equations p equations in p variables which can be easily solved if a is invertible.

So, we assume that X^TX.is a full rank matrix invertible, the meaning of this will become a little clearer later and if it is not invertible we will have to do other things which we will again talk in another lecture. But, for the time being let us assume that X^TX which is a square matrix is invertible, it is a full rank matrix and then you can easily find the solution for $\hat{\beta}$ solve this linear set of equations by taking $A^{-1}B$ which is exactly $(X^TX)\ ^{-1}X^Ty$. So, $\hat{\beta}$ the coefficient vector can be found by this thing and this is the solution that minimizes the sum squared errors, this objective function that we have written.

So, once we get $β_1$ the slope parameters, $β_0$ can be estimated as the $\bar{y} - \bar{x}^T\hat{\beta}$. Notice that this is very similar to what we have in the univariate case, where it says $β_0$ estimate is nothing, but $\bar{y} - \bar{x}^T\beta_1$. So, it is very similar to that, you can see.

You can also compare the solution for the slope parameters with the univariate case which says $\widehat{\beta_1}$ = Xy divided by XX. Notice that X^Ty represents Xy and X^TX represents XX in the univariate case you are dividing X y by XX, in the multivariate case division is represented by inverse. So, you get X transpose X inverse times X transpose y.

So, you can see it is very very similar to the solution for the univariate case except that these are matrices and vectors and therefore, you have to be careful. You cannot simply divide, it is matrix times inverse times a vector that is a solution for β which is slope parameters. You can also estimate $β_0$ and β by doing what is called augmentation of the x vector with a constant value 1 1 1 in the final thing, but I did not use that approach.

Because, the main subtraction approach is a much better approach for estimating whether if for estimating $β_0$ and $\hat{\beta}$ slope parameters because this is applicable even to another case called the totally squares. The augmentation approach is valid only for ordinary least squares you cannot use it for total least squares which we will see again later. So, that is why I use the means subtraction route in order to obtain the estimates of the slope parameter first followed by the estimation of β0 using the estimate of the slope parameters in this value.

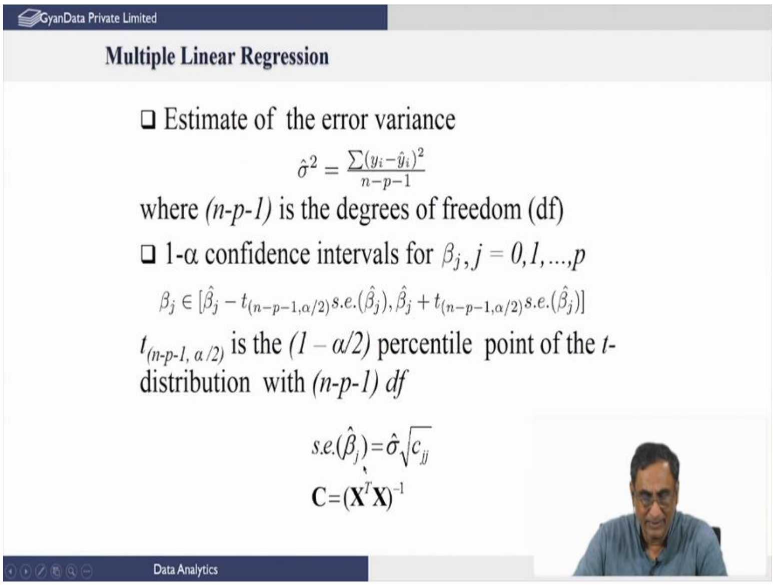
Now, you can also derive properties of these parameters β, we can show that the expected value of βhat is βwhich just means it is an unbiased estimate just as in the univariate case and we can also get the variance of this βhat. In this case it is a covariance matrix because it is a vector and we can show that the covariance matrix is sigma squared times this X transpose X inverse. Now, again you can go back and look at the univariate case, there the variance of $β_1$ the slope parameter will be σ^2 by S_{xx}.

In this case it is $\sigma^2(X^TX)^{-1}$. So, X^TXrepresents XX. σ^2 is the variance of the error corrupting the dependent variables. We may have a priori knowledge sometimes in most cases we may not be able to know this value of sigma square; we may not be given this so, we have to estimate the sigma squared from data and we will show how to get this.

These two parameters that actually we can show the first parameter says that the estimates of β1 the slope parameters are unbiased. So, $\hat{\beta}$ are unbiased estimator is an unbiased estimator of the true value β. Moreover it can show that among all linear estimators because βhat is a linear function of y.

Notice that $(X^T X)^{-1} X^T$ is nothing but a matrix which basically multiplies the measurements y; so, $\hat{\beta}$ can be interpreted as a linear combination of the measurements. Therefore, it is known as a linear estimator. Among all such linear estimators we can show that $\hat{\beta}$ has the least variance. Therefore, it is called a blue estimator or unbiased estimator with the best linear unbiased estimator that is what it blue represents; best in the sense of having the least variance.

(Refer Slide Time: 14:51)



GyanData Private Limited

Multiple Linear Regression

❑ Estimate of the error variance

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

where *(n-p-1)* is the degrees of freedom (df)

❑ 1-α confidence intervals for $\beta_j, j = 0,1,\ldots,p$

$$\beta_j \in [\hat{\beta}_j - t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_j), \hat{\beta}_j + t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_j)]$$

$t_{(n-p-1,\, \alpha/2)}$ is the *(1 – α/2)* percentile point of the *t*-distribution with *(n-p-1) df*

$$s.e.(\hat{\beta}_j) = \hat{\sigma}\sqrt{c_{jj}}$$

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$$

Data Analytics

Now, we can also estimate as I said sigma squared from the data and that sigma square estimate is nothing, but the after you fit the linear model you can take the predicted value for the ith sample from the linear model and compute this residual $y_i - \widehat{y\_i}$ which is the measured value - the predicted value for the ith sample, square it, take the sum over all possible samples n samples divided by n - p - 1. Again, if you go back to your linear case or univariate case you will find that the denominator is n - 2.

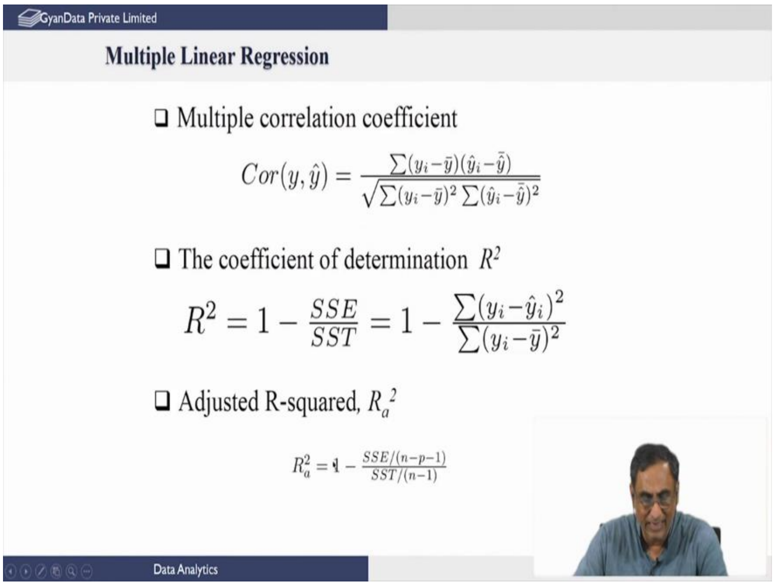
Here you have n - p - 1 because you are fitting p + 1 parameters; p is slope parameters + 1 offset parameter. Therefore, out of the n measurements p + 1 are taken away for the deriving the estimates only the remaining things are the degrees of freedom or the variability in the residuals is caused by the remaining n - p - 1 measurements and that is why you are dividing by n - p - 1; whereas, in the univariate case you would have divided by n - 2 because you are estimating only two parameters there.

So, you can see a one-to-one similarity between the univariate regression problem and the multiple linear regression problem in every derivation that we have given here. Now, once you have estimated $\hat{\sigma}$, the variance of the error used from the data you can go back and construct confidence intervals for each slope parameter. We can show that the true slope parameter lies in this confidence interval for any confidence interval you might choose; 1 - α; α represents like a level of significance.

So, if you say α is equal to 0.05. 1 - α would represents 0.95. So, that will be a 95 percent confidence interval, ok. Correspondingly, I will find the critical value from the t distribution n with n - p - 1 degrees of freedom and this represents α by 2 the lower value probability value from the t distribution and this upper critical value where the probability area under the curve beyond the value is α by 2.

So, n - p - 1 represents degrees of freedom. Notice that in the univariate case it would have been n - 2, very very similar. So, the confidence interval for βj for any given α can be computed using this particular formula and the term here s.e ($\widehat{\beta_J}$) represents the standard deviation of the estimate of $\widehat{\beta_J}$.

(Refer Slide Time: 17:40)



GyanData Private Limited

Multiple Linear Regression

- ❑ Multiple correlation coefficient

$$Cor(y,\hat{y}) = \frac{\sum(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum(y_i-\bar{y})^2\sum(\hat{y}_i-\bar{\hat{y}})^2}}$$

- ❑ The coefficient of determination R^2

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

- ❑ Adjusted R-squared, R_a^2

$$R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Data Analytics

And, that is given by the diagonal element diagonal element here of this quantity with sigma square replaced by the estimate here. So, we have computed the standard deviation of the parameter $\widehat{\beta_J}$ estimated parameter $\widehat{\beta_J}$ by using the estimated value of sigma multiplied by the diagonal element of $X^T X$. So, we are fitting the diagonal elements of the covariance matrix of β parameters, ok, that is all we have done.

So, this represents the diagonal element or the; square root of the diagonal element which represents the standard deviation of the estimated value of βwhich is what is used in order to construct this confidence interval. So, every one of this can be computed from the data as you can see and you can construct. Now, in the confidence level can later be used for testing whether the estimated parameter βis significant or insignificant as we will see later.

Now, we will can also compute the correlation between y and $\hat{y}$which tells you whether the predicted value from the linear model is resembles or closely related to the measured value. So, typically we will draw a line between the y the measured value and the predicted value and see whether it is these things fall on the 45 degree line and if it does then we think that the fit is good.

Another way of doing this is to find the correlation coefficient between y and y hat which is simply using the standard thing y_i - $\bar{y}$ multiplied by $\widehat{y_\iota}$- $\widehat{\overline{y_\iota}}$ summed over all quantities divided by the standard deviation of in y and the standard deviation in y hat, that is for normalization.

We could also use the coefficient of determination R squared; just as we did for the univariate case we can compute R squared as 1 – (sum squared error)/(sum square total) which is nothing, but $1 - \frac{\Sigma(y_i - \widehat{y_\iota})^2}{\Sigma(y_i - \overline{y_\iota})^2}$. So, if we take 1 - this we will actually we can show whether using the independent variables have we been able to get a better fit? If we have obtained a very good fit, then the numerator will be close to 0 and R squared will be close to one.

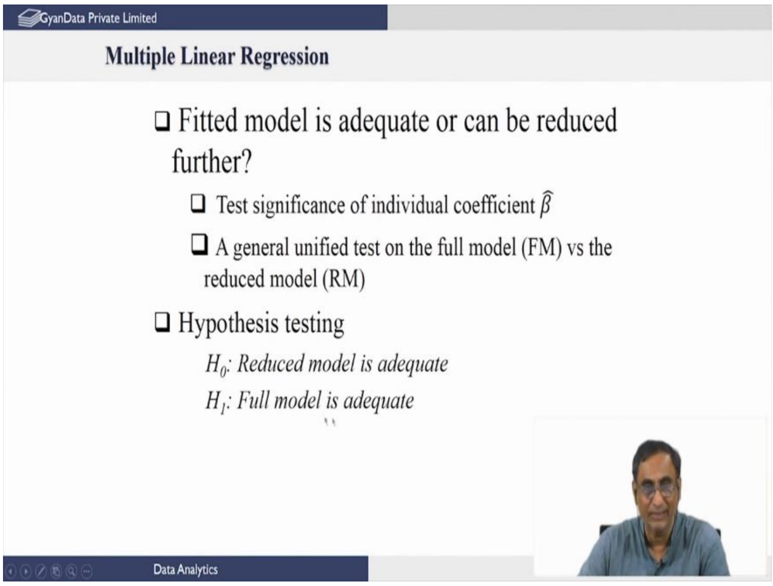
On the other hand, if we had not improved the fit because of X's any of the X's then the numerator will be almost equal to the denominator and therefore, this one will be close to 0. So, a value of R square close to 1 as before represents indication of a good linear fit whereas, a value close to 0 indicates the fit is not good. We can also compute the adjusted

R square to account for the degrees of freedom notice that the numerator has n - p - 1 degrees of freedom whereas, the denominator as n - 1 degrees of freedom.

Therefore, we can do an adjusted R squared which divides the SSE by the appropriate degrees of freedom. We can say this is the error due to per degree of freedom that is there in the fit whereas, the denominator represents the error because we have fitted only the offset parameter there are n - 1 degrees of freedom, this is the error per degree of freedom. So, this kind of a thing is a much also a good indicator instead of using R squared we can use adjusted value of R square.

So, these are all very very similar again to the univariate linear regression problem.

(Refer Slide Time: 21:16)



GyanData Private Limited

Multiple Linear Regression

- ❑ Fitted model is adequate or can be reduced further?
  - ❑ Test significance of individual coefficient $\hat{\beta}$
  - ❑ A general unified test on the full model (FM) vs the reduced model (RM)
- ❑ Hypothesis testing

  H_0: *Reduced model is adequate*

  H_1: *Full model is adequate*

Data Analytics

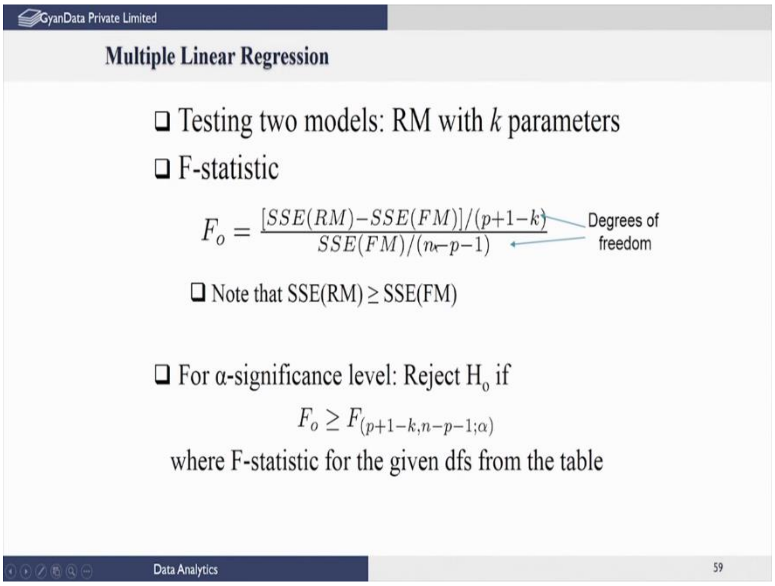
So, we can use; what we can check R squared and see whether the value is close to 1 and if it is we can say maybe the linear model is good to fit the data, but that is not a confirmatory test. We have to do the residual plot as we did in linear regression univariately in linear regression. And, that is what we are going to do further.

So, we are going to find whether the fitted model is adequate or it can be reduced further; what this reduced further means we will explain. In the univariate case there is only one independent variable, but here there are several independent variables. Maybe not all independent variables have an effect on y; some of the independent variables may be irrelevant. So, one way of trying to find whether a particular independent variable has an effect is to test the corresponding coefficient.

Notice we have already defined the confidence interval for each coefficient and we can see whether the confidence interval contains 0 in which case we can say the corresponding independent variable does not have a significant effect on the dependent variable and we can perhaps drop it, ok.

Or we can also do what we call the test F test just as we did in the univariate regression problem we can test whether the full model is better than the reduced model. The reduced model contains no independent variables whereas, the full model can contain all or some of the independent variables you can do many kinds of tests and we will do this. So, we can test whether the reduced model which contains only the constant intercept parameter is a good fit as opposed to including all the independent variable some or all the independent variables that is what we call the full model.

(Refer Slide Time: 23:02)



GyanData Private Limited

Multiple Linear Regression

- ❑ Testing two models: RM with *k* parameters
- ❑ F-statistic

$$F_o = \frac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$$

Degrees of freedom

- ❑ Note that SSE(RM) ≥ SSE(FM)

- ❑ For α-significance level: Reject H_0 if

$$F_o \geq F_{(p+1-k, n-p-1; \alpha)}$$

where F-statistic for the given dfs from the table

Data Analytics

59

We will consider a specific case here where we do the F test statistic for the case when we have a reduced model and compare it with the full model. The reduced model we will consider with k parameters. Specifically, let us consider the reduced model with only one parameter which means that we have only the constant intercept parameter we would not include any of the independent variables and compare it with the full model which contains all of the independent variables including the intercept.

So, the reduced model is one which contains only the offset parameter and no independent variables, the full model is the case where we consider all the independent variables and the intercept parameter. So, the number of parameters we are estimating in the reduced model is only one. So, k equals 1 and the full model is the case where we have all the independent variables p independent variables. So, we are estimating p + 1 parameters in the full model, ok.

So, what we do is perform a fit and compute the sum squared errors which is nothing, but the difference between y, the measured value and the predicted value. So, we will first take the model containing only the offset or the intercept parameter and estimate. In this case of course, y bar will be the best estimate and we will compute sum squared errors which is nothing, but the variance of the measurements for the dependent variable.

Then, we would also perform a linear regression containing all the parameters independent variables. And, in this case if we compute the difference between y and y predicted and take the sum squared errors that is the SSE of the full model. So, when we want to compare whether we want to accept the full model as compared to the Reeves model what we do is take the difference in the sum squared errors. Remember, the sum squared errors for the reduced model will be greater than the sum squared errors for the full model because the full model contains more number of parameters and therefore, you will get a better fit.

So, the difference in the fit, which is difference in the sum squared errors between the reduced model fit and the full model fit that is the numerator divided by what we call the degrees of freedom. Notice, the full model has p + 1 parameters. p independent variable + the offset and the reduced model in this particular case contains only one parameter. So, k equals 1. So, the degrees of freedom will be p.

So, you divide this difference in the sum squared errors by p. Denominator is the sum squared errors of the full model which contains n - p - 1 degrees of freedom because p + 1 parameters have been fitted. Therefore, the degrees of freedom is to total number of measurements - p - 1. And, so, we divide that sum squared errors for the denominator by the number of degrees of freedom and then take this ratio as defined and that is your F statistic.

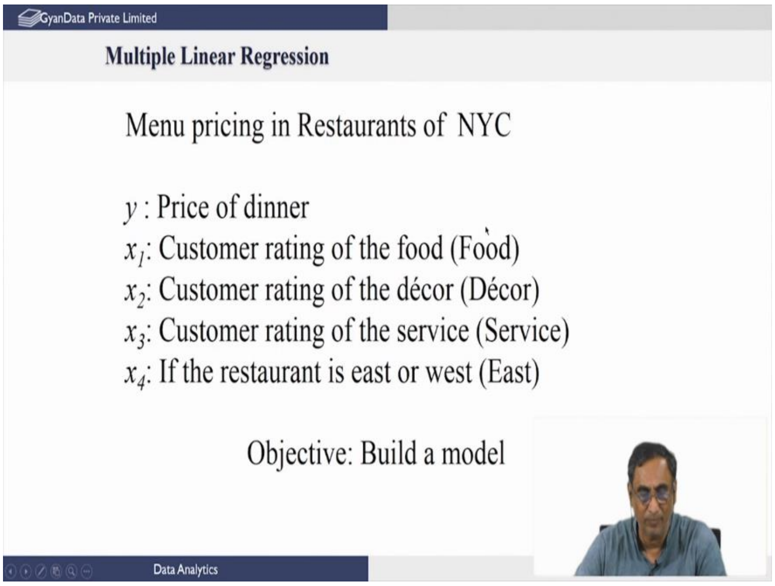
Now, in order to reject if we want to reject the null hypothesis or if you want to test the null hypothesis against its alternative, we find the test criteria for the α level of significance. We would take it from the F distribution where the numerator degrees of freedom is p + 1 - k for this particular case it is exactly p. And the denominator degrees of freedom is n - p - 1 and α level of significance we use and we compute the test criteria critical value from the F distribution.

Then, we compare the test statistic with the critical value and if the test statistic exceeds the critical value at this level of significance then we reject the null hypothesis that is we will say the full model is a better choice and the independent variables do make a difference. And, this is a standard thing that R function will provide. This particular comparison between the reduced model which has no independent variables and the full model which contains all the independent variables in multi linear regression. Of course, you can choose different reduced models and compare with the full model.

For example, you can take the reduced model by leaving out only one of the independent variables so, that we will have p parameters. We can compare it with the full model and again perform a test to decide whether the inclusion of that independent variable makes a difference or not. So, this kind of combination can be done depending on what stage you are and that will be you using in what we call the sequential method for subset selection that will be discussed in a later lecture.

But, essentially the R functions only provide a comparison between the reduced model which contains no independent variable and the full model which contains all of the independent variables.

(Refer Slide Time: 27:50)



GyanData Private Limited

Multiple Linear Regression

Menu pricing in Restaurants of NYC

y : Price of dinner
x_1: Customer rating of the food (Food)
x_2: Customer rating of the décor (Décor)
x_3: Customer rating of the service (Service)
x_4: If the restaurant is east or west (East)

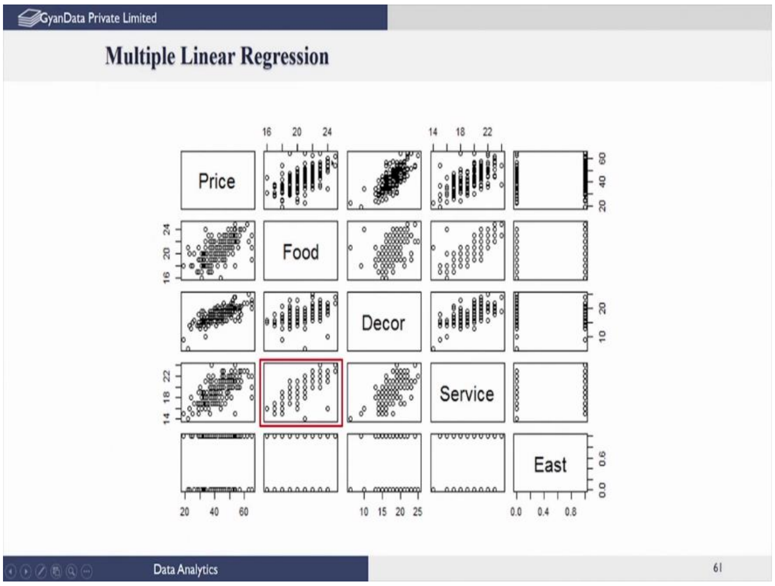
Objective: Build a model

Data Analytics

Let us go through a simple example in order to what are called revisit these ideas. So, in this case we have what is called the a price data where customers have been asked to rate the food than the other aesthetics of a particular restaurant. And, we also and the cost of the particular dinner also a data were obtained for these restaurants and the location of these restaurants whether on they are on the east side of a particular street in New York or the west side. Typically, in New York west side is probably a little poorer whereas, the east side probably is a little richer neighbourhood. So, location of the restaurant also would indicate would have an effect on the price.

So, these are the four independent variables people data was obtained on the quality of the food, the decor and service all this was rated by the customers and at the location of this restaurant and the price of dinner in that served in that restaurant was also taken. So, you would expect that the quality of the food the service level all of this would have a very direct influence on the price in the restaurant. And, a linear model was built between y and the independent variables x 1 to x 4.

(Refer Slide Time: 29:16)



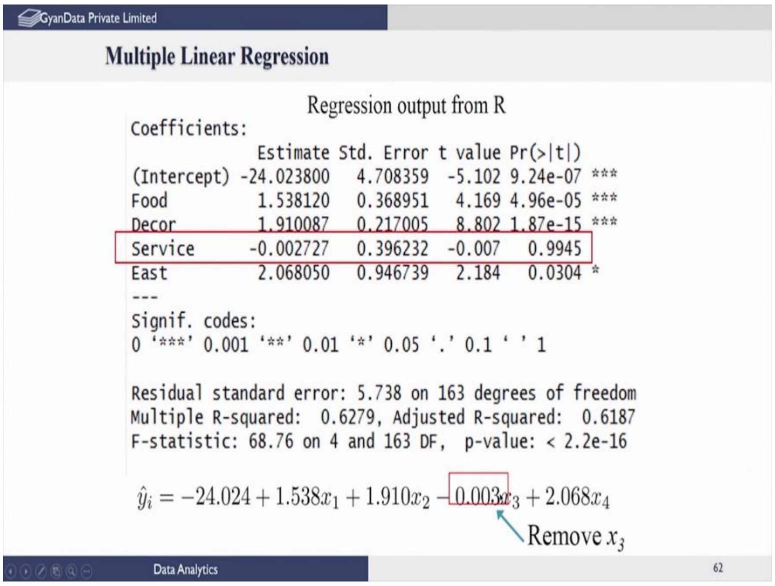
So, before we build a model we do a scatter plot as usual any visualization and here you because there are several independent variables we have not just one plot scatter plot between y and x 1. For example, in this case remember price is y, y versus x 1 this particular plot shows the correlation or the scatter plot for y versus x 1 or y versus food, price versus food. The second one is the scatter plot price and decor the third one is the scatter plot between price and service and the last one is price versus location, ok.

And, similarly you can actually develop a scatter plot between food and decor which is here or food and service and so on. Even though we consider all these variables that we have obtained like food decor service location as independent, it is possible when we select these variables they are not truly independent. There might be inter dependencies between the what we call the so called independent variables, that can give rise to problem in a regression which we will see later that what we call the effect of co-linearity.

But, a scatter plot may reveal some interdependencies between the independent so called independent variables. So, for example, if we look at the scatter plot between food and decor it is seems to be completely randomly distributed this does not seem to be any quite correlation. However, food and service seems to be very strongly correlated. There seems to be a linear relationship between food and service. So, perhaps you do not need to include both these variables. We will see later that is this true, but in this just a scatter plot itself reveals some interesting features.

And, so, we will now go ahead and say perhaps a linear model between price and food then decorous is seems to be pointed out or indicated by the scatter plots let us go ahead and build one.

(Refer Slide Time: 31:16)



GyanData Private Limited

Multiple Linear Regression

Regression output from R

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.023800   4.708359  -5.102 9.24e-07 ***
Food          1.538120   0.368951   4.169 4.96e-05 ***
Decor         1.910087   0.217005   8.802 1.87e-15 ***
Service      -0.002727   0.396232  -0.007   0.9945
East          2.068050   0.946739   2.184   0.0304 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.738 on 163 degrees of freedom
Multiple R-squared:  0.6279, Adjusted R-squared:  0.6187
F-statistic: 68.76 on 4 and 163 DF,  p-value: < 2.2e-16
```

$$\hat{y}_i = -24.024 + 1.538x_1 + 1.910x_2 - \boxed{0.003}x_3 + 2.068x_4$$

Remove x_3

Data Analytics

62

And, if we apply the R function lm to this data set and we examine the output, we will get this output from R and tells that the intercept term is - 24.02 and the slope parameters the coefficient multiplying food is 1.5, the coefficient multiplying decor is 1.9 and so on so forth.

It also gives you the standard error for each coefficient as well as the offset parameter which is nothing, but the sigma value for the estimated quantities and it also gives you the probability values p-values as we call them. And, notice if the p-value is very low, it means that this coefficient is significant. We cannot take it that this value is 0, ok. Any low value of this indicates that the corresponding coefficient is significantly different from 0.

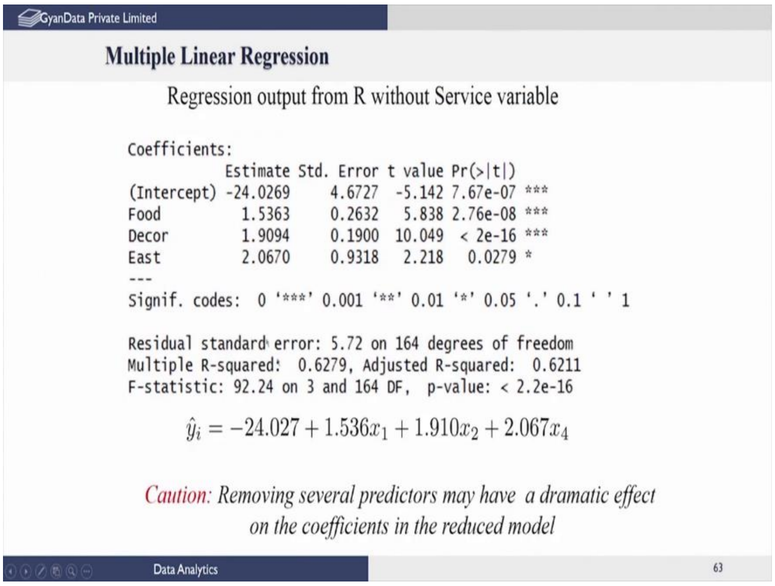
So, in this case the first three has very low p-values and therefore, the significant, but service has a high p-value. Therefore, it seems to indicate that this coefficient is insignificant is equal almost equal to 0 that is what this indicates. If you look at the east which is this independent location parameter that has does not have a very low p-value, but it is still not bad 0.03 and therefore, it is significant only is insignificant only if you take a level of significance of 0.025 or something like that. If you take 0.1 or 0.05 and so on you will still consider this east this coefficient to be significant and that is what this is basically pointing out, this star indicates that.

So, now we will go ahead and try to actually look at the F value also, the F statistic says that the full model as compared to the reduced model of using only the intercept is actually significant; which means, the constant model is not good and including these variables results in a better fit or explanation of the price and therefore, you should actually include this.

Whether you should include all of them are only some of them we can do different kinds of tests to find out what we have done in this particular case is only compare the model without any of these independent variables which is called the constant model with all of these variables included. That is the only two model comparisons we are made the reduced model is one containing only the interceptor and the full model is one which contains intercept and all four independent variables and that is the p-value it has given the corresponding F statistic.

So, we are saying that including these independent variables is important in explaining the price, ok. So, but it may turn out that all of them is not necessary and that we will we will examine further. So, the corresponding fit that we obtain is this. As I said that from the, what you call the as the confidence interval for the slope parameter for service we can say that we can remove this it is insignificant and perhaps we can remove this and try the fit. For the time being let us actually remove this and try the fit.

(Refer Slide Time: 34:42)



GyanData Private Limited

Multiple Linear Regression

Regression output from R without Service variable

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.0269     4.6727  -5.142 7.67e-07 ***
Food          1.5363     0.2632   5.838 2.76e-08 ***
Decor         1.9094     0.1900  10.049  < 2e-16 ***
East          2.0670     0.9318   2.218   0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.72 on 164 degrees of freedom
Multiple R-squared:  0.6279, Adjusted R-squared:  0.6211
F-statistic: 92.24 on 3 and 164 DF,  p-value: < 2.2e-16
```

$$\hat{y}_i = -24.027 + 1.536x_1 + 1.910x_2 + 2.067x_4$$

Caution: Removing several predictors may have a dramatic effect on the coefficients in the reduced model

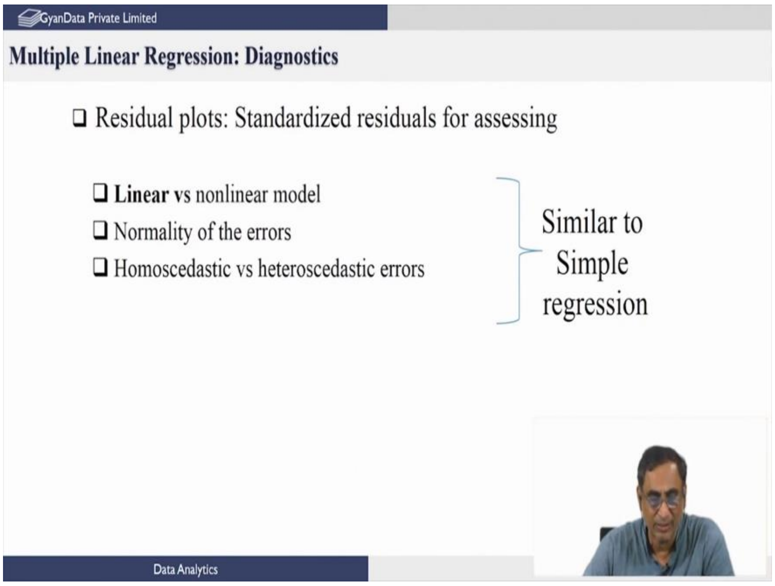
Data Analytics

63

We have done that. We have only included now food, decor in east and done the regression again and it turns out that the regression thing is still what you call the R squared value is improved, not improved significantly, but not reduced. And, F value is significant and we get the more or less the same coefficients for the other parameters also the intercept term and the slope parameters.

It indicates that x 3 is not adding any value to the prediction of y. The reason for this as we said if you look at the scatter plot service and food are very strongly correlated. Therefore, only either food or service needs to be included in order to explain price and not both right. And, in this case service is being removed, but you can try removing food as the variable and try to fit between price, decor and decor service and east and you will find that the regression is as good as retaining food and eliminating service.

(Refer Slide Time: 35:44)



GyanData Private Limited

Multiple Linear Regression: Diagnostics

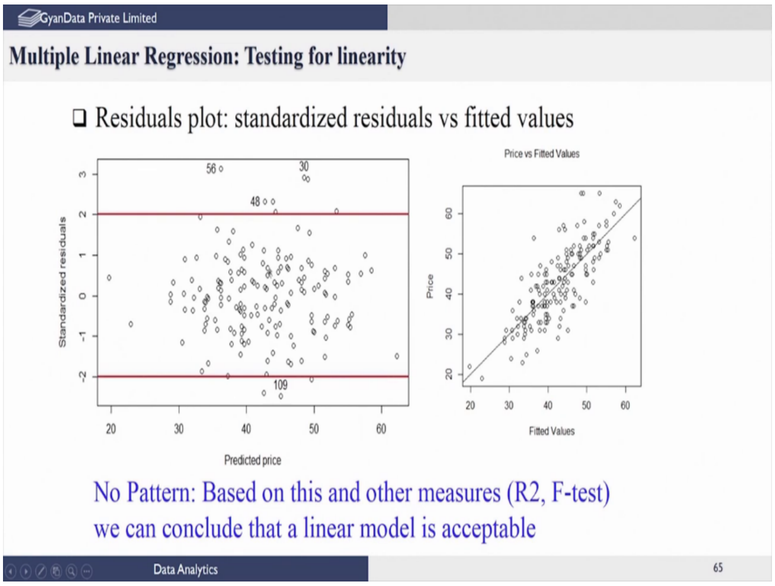
- ❑ Residual plots: Standardized residuals for assessing
  - ❑ **Linear** vs nonlinear model
  - ❑ Normality of the errors
  - ❑ Homoscedastic vs heteroscedastic errors

Similar to Simple regression

Data Analytics

R squared value and the F statistics seems to indicate that we can go ahead with the linear model, but we should further examine the standardized residual plot for concluding whether the linear model is or not. There should be no pattern in the residuals.

(Refer Slide Time: 36:00)



GyanData Private Limited

Multiple Linear Regression: Testing for linearity

- ❑ Residuals plot: standardized residuals vs fitted values

No Pattern: Based on this and other measures (R2, F-test) we can conclude that a linear model is acceptable

Data Analytics

65

So, let us actually do the residual plot. Here we have taken the standardized residuals and plotted it against what is called the predicted price value or the fitted value. Remember, this is y i hat; y i hat is only one variable so, you need to generate only one plot and we have also shown here the in red lines the confidence interval for the standardized residuals and anything above this outside of this interval indicates out layers. So, for example, 56, sample number 48, sample number 30 and 109 and so on so forth may be possible outliers. And, but there is no pattern in the standardized residuals it spread randomly within this boundary and therefore, we can say since there is no pattern a linear fit is acceptable.

So, here the quality of the fit is shown here. So, the actual price, the measured value versus the y i hat predicted value is shown and a linear model seems to explain the data reasonably well. The last thing is, we have these out layers; if you want to improve the fit you may want to remove let us say the out layer which is farthest away from the boundary. For example, you may want to remove 56 and redo the linear regression multi multiple linear regression and again repeat it until there are no out layers that will improve the R squared value and the fit quality of the fit a little more, ok.

So, we have not done that we leave this as an exercise for you. So, what we have done is we have seen that whatever was valid for the univariate regression can be extended to the multiple linear regression except that scalars there will get replaced by vectors and matrices corresponding. What was the variance there will become a variance covariance matrix here, what was a vector the scalar there might mean scalar might become a mean vector here.

So, you will see a one to one correspondence, but the residuals plot and interpretation of confidence interval for βall of this the F statistic are more or less similar, except that understand in the multiple linear regression there are several independent variables. All of them may not be relevant. We may be able to take only a subset and I will actually handle subset selection as a separate lecture. For the time being we are just an a significance test on the coefficient in order to identify the irrelevant independent variables, but there are better approaches and we will take that up in the following lectures.