

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

Lecture - 38
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. 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 $j = 1$ to p . There are p independent variables which we believe affect the dependent variable.

(Refer Slide Time: 00:49)

Multiple Linear Regression

- Dependent variable (y) depends on p independent variables $x_j, j = 1, 2, \dots, p$
- General linear model
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$
- For i th observation
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \dots + \beta_p x_{p,i} + \epsilon_i$$
- Objective: Using n observations, estimate regression coefficients

We will try to develop a linear model between the dependent variable y and these independent p independent variables $x_j, j = 1$ to p . In general we can write this linear model as before we can say y the dependent variable $= \beta_0$, an intercept, $+ \beta_1$ times $x_1 + \beta_2$ times x_2 and so

on up to p β times $p \times p$, 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 assume that it is a random quantity with 0 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 the i th sample dependent variable $y_i = \beta_0 + \beta_1 x_{i1}$, the i th sample value of the independent variable 1. Similarly, x_2 y β_2 times x_2 y , where x_2 y represents the value of the second independent variable for the i th sample and so on + an error ϵ_i that corrupts the measurement of y_i and so on for $i = 1$ 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 β_0 β_1 β_2 up to β_p using these n sample measurements of x 's 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

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

□ The linear model in matrix form

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

□ SSE

$$S(\beta) = \epsilon^T \epsilon = (y - X\beta)^T (y - X\beta)$$

Data Analytics
54

Again in order to find the best estimates of the parameters β_0 to

β_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 will be x_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 cross 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 being 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 β_0 in a vector form β_1 to β_p as a column vector. Here basically as a I am sorry a row vector. So, β_1 is the first coefficient, β_p is the coefficient corresponding p th variable. So, we have β vector which is a p cross 1 vector we can also define ϵ , the noise vector, as ϵ_1 to ϵ_n corresponding to all the n observations. Now having defined this notation we can write our linear model in the form $y = x\beta + \epsilon$.

Notice that we have not included β_0 . We have eliminated that in directly by doing this mean subtraction I will show you how that happens. But you can take it 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 $y = x \beta + \varepsilon$ and we also make the usual assumptions about the error that it is a 0 mean vector in this case because it is a multivariate vector 0 is a vector. So, ε expected value $\varepsilon = 0$ implies ε is a random vector with 0 mean and the variance, covariance matrix of ε is assumed to be σ^2 identity.

σ^2 identity in this form it means all the epsilons, ε_1 to ε_n , have all have the same variance σ^2 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 β so as to minimize the sum square of the errors. So, $\varepsilon^T \varepsilon$ is a compact way of saying the sum of all errors squared of all the errors, in all the n measurements. So, expanding this is nothing, but $\sigma^2 \varepsilon_i^2 = 1$ 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 $y - x \beta^T y - x \beta$.

We want to minimize this which is a function of β by finding the best value of β . So, if we setup 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 β and setting it = 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 $X^T X$ into $\beta = X^T y$.

(Refer Slide Time: 08:54)

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
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}; \beta_0 = \bar{y} - \bar{x}^T \hat{\beta}$$
- The properties of the estimators
$$E(\hat{\beta}) = \beta$$
$$\text{Var}(\hat{\beta}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$
- $\hat{\beta}$ is the best linear unbiased estimator (BLUE)

Data Analytics

55

Now, this is a p cross, remember X is a n cross p matrix. So, X^T is p cross n . So, this is square matrix $X^T X$ is a square matrix of size p cross p and multiplied by the p cross n vector and similarly it is p cross 1 on the right hand side. So, these are p equations in p variables. The linear equations in β x is all known y is known. So, right hand side is like the if you, are what we, reinterpret this as a times some $x = b$. 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 $X^T X$ is a full rank matrix invertible. The meaning of this will become 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 at for the time being let us assume that X transpose X which is a square matrix is invertible it is a full rank matrix and then you can easily find the solution for β solve this linear set of equations by taking $a^{-1}b$ which is exactly $X^T X$ inverse $X^T y$. So, β 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 mean value of y - the mean vector X^T times the slope parameter. Notice that this is very similar to what we have in the univariate case where it says β_0 estimate is nothing but $\bar{y} - \bar{x}$ into β_1 . So, it is very similar to that you can see.

You can also compare the solution for the slope parameters, for the, with the univariate case which says β_1 is SXY divided by SXX . Notice that X transpose y represents SXY and $X^T X$ represents SXX in the univariate case you were diving SXY by SXX in the multivariate case

division is represented by an inverse. So, you get $X^T X$ inverse terms times $X^T y$. So, you can see that 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 as matrix times inverse times a vector that is a solution for β which is slope parameters.

You can also estimate β_0 and β_1 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 mean subtraction approach is a much better approach for estimating whether if for estimating β_0 and β , β slope parameters because this is applicable even to another case called the total least 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 mean subtraction route in order to obtain the estimates of the slope parameter first followed by the estimation of β_0 using the estimates of the slope parameters in this manner.

Now, you can also derive properties of these parameters β . We can show that the X vector value of $\hat{\beta}$ is β which just means it is an unbiased estimate just as in the univariate case and we can also get the variance of this $\hat{\beta}$ in this case it is a covariance matrix because it is a vector and we can show that the covariance matrix is σ^2 times this X transpose X inverse.

Now, again you can go back and look at the univariate case. There the variance of β_1 slope parameter will be σ^2 by SXX in this case it is σ^2 into $X^T X$ inverse. So, $X^T X$ represents SXX . σ^2 is the variance of the error corrupting the dependent variables. We may have a priori (Refer Time: 13:35) knowledge sometimes in most cases we may not be able to know this value of σ^2 we may not be given this. So, we have to estimate the σ^2 from data and we will show how to get this.

These two parameters that actual we can show the first parameter says that the estimates of β_1 the slope parameters are unbiased. So, $\hat{\beta}$ are unbiased estimator it is an unbiased estimator of the of the true value β . Moreover you can show that among all linear estimators because $\hat{\beta}$ is a linear function of y . Notice that $(X^T X)^{-1} X^T$ is nothing but 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 a 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: 14:52)

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)

□ $1-\alpha$ confidence intervals for $\beta_j, j = 0, 1, \dots, p$

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

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

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

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

Data Analytics
56

Now, we can also estimate as I said σ^2 from the data and that σ^2 estimate is nothing but the after you fit the linear model you can take the predicted value for the i th sample from the linear model and compute this residual $y_i - \hat{y}_i$ which is the measured value - the predicted value for the i th sample, square it take the sum of all possible samples, n samples, divided by $n - p - 1$. Again if you go back to your linear case 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 α set 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 cost by the remaining $n - p - 1$ measurements and that is why you are diving 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 multi multiple linear regression problem in every derivation that we have given here. Now, once we 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 may choose $1 - \alpha$, α represents like a level of significance.

So, if you say $\alpha = 0.05$, $1 - \alpha$ would represents 0.95. So, that will be a 95 percent confidence interval. 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 upper lower value probability value from the t distribution and this is the upper critical value where the probability area under the curve beyond the value is α by 2. So, $n - p - 1$ represents the degrees of freedom notice that in the univariate case it would have been $n - 2$, very very similar.

So, the confidence interval for β_j for any given α can be computed using this particular formula and the term here se of β_j represents the standard deviation of the estimate of β_j and that is given by the diagonal element, diagonal element here of this quantity with σ square replaced by the estimate here.

So, we have computed the standard deviation of the, of the parameter β hat j estimated parameter β_j by using the estimated value of σ multiplied by the diagonal element of $X^T X$. So, we are fitting the diagonal elements of the covariance matrix of β parameters that is all we have done. So, this represents the diagonal element or the square root of the diagonal element which represents standard deviation of the estimated value of β which 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, the confidence level can later be used for testing whether the estimated parameter β is significant or insignificant as we will see later.

(Refer Slide Time: 18:39)

□ 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

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



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 t is good. Another way of doing this is to find the correlation coefficient between y and \hat{y} which is simply using

the standard thing $y_i - \bar{y}$ multiplied by $\hat{y}_i - \bar{y}$. Summed over all quantities divided by the standard deviation of y and the standard deviation in \hat{y} 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 $1 - \frac{\text{sum squared error}}{\text{sum squared total}}$. Which is nothing, but numerator is $y_i - \hat{y}_i$, the residual squared divided by $y_i - \bar{y}$ squared which is the variance in y basically. 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 t . If we have obtained a very good t then the numerator will be close to 0 and R squared will be close to 1.

On the other hand if we are not improved the t 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, value of R square close to 1 as before represents indication of a good linear t whereas, a value close to 0 indicates the t is not good. We can also compute adjusted R square to account for the degrees of freedom notice that the numerator has $n - p - 1$ degrees of freedom whereas, the denominator has $n - 1$ 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 t whereas, the denominator represents the error because we have fitted only the p set parameter there are $n - 1$ degrees of freedom this is the error per degree of freedom. So, this kind of a thing is 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: 21:17)

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, we can check R squared and see whether the values close to one and if it is we can say maybe linear model is good to the data, but that is not a confirmatory test. We have to do the residual plot as we did in linear regression, univariate 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. Or, we can also do what we call the test, F test, just as we did univariate regression problem we 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 test 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 variables, 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 $SSE(RM) \geq SSE(FM)$
- ❑ 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 will 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 α set parameter and no independent variables the full model is a 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 1, so $k = 1$ 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.

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 α set or the intercept parameter and estimate. In this case of course, \bar{y} will be the best estimate. And we will compute sum squared errors which is nothing but the variance of the measured measurements for the dependent variable. Then we will also perform a linear regression containing all the parameters, independent variables, and in this case we will 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 reduced model what we do is take the difference in the sum squared errors remember the sum squared errors for the reduced model will become greater than the sum squared errors for the full model because the full model contains more number of parameters and therefore, you 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 as $p + 1$ parameters p independent variable + the α set and the reduced model in this particular case contains only 1 parameter, so, $k = 1$

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 the total number of measurements - $p - 1$. So, we divide the 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 we want to test the null hypothesis against this alternative we find the test criteria for the α level of significance. We will 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 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 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 using in what we call the sequential method for subset selection that will be discussed in the 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. Let us go through simple example in order to what you call revisit these ideas.

(Refer Slide Time: 27:51)

The slide is titled "Multiple Linear Regression" and is part of a presentation by GyanData Private Limited. It discusses "Menu pricing in Restaurants of NYC". It defines the dependent variable y as the "Price of dinner" and the independent variables x_1, x_2, x_3, x_4 as "Customer rating of the food (Food)", "Customer rating of the décor (Décor)", "Customer rating of the service (Service)", and "If the restaurant is east or west (East)" respectively. The objective is to "Build a model". The slide footer includes navigation icons, the text "Data Analytics", and the page number "60".

So, in this case we have what is called the price data where customers are being asked to rate the food and the other aesthetics of a particular restaurant and we also and cost of the particular dinner also data(Refer Time: 28:15) 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 New York Westside is probably a little poorer whereas, the east side probably is a little richer neighborhood.

So, location of the restaurant also would indicate, would have a 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 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 .

Multiple Linear Regression

The figure displays a 5x5 matrix of scatter plots illustrating the relationships between five variables: Price, Food, Decor, Service, and East. The diagonal cells contain the variable names. The upper triangle shows scatter plots of the variables against each other. The lower triangle shows the same scatter plots with regression lines. The 'Price' vs 'East' plot is highlighted with a red border.

The variables and their approximate ranges are:

- Price: 14 to 24
- Food: 18 to 24
- Decor: 10 to 20
- Service: 0.0 to 0.8
- East: 0.0 to 0.6

The scatter plots show that Price is positively correlated with Food, Decor, and Service, and negatively correlated with East. Food is positively correlated with Decor and Service. Decor is positively correlated with Service. Service is positively correlated with East.

The third one is the scatter plot between price and service and the last one is price versus location. And similarly you can actually develop a scatter plot between food and decor which is here or food and service and so on.

So, for example, if we look at the scatter plot between food and decor it 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 that it is 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 and decor is, seems to be pointed out or, indicated by this scatter plots let us go ahead and build one.

(Refer Slide Time: 31:17)

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

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

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 it 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 σ 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. 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, they are significant, but service has a high p value therefore, it seems to indicate that this coefficient is insignificant is equal almost = 0 that is what this indicates. If you look at the east which is this independent location parameter that as 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 statistics 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 t or explanation of the price and therefore, you should actually include this. Whether you should include all of them or only some of them we can do different kinds of test to find that.

What we have done in this particular case is only compare the model with-out 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 have made. The reduced model is one containing only the intercept and the full model is one which contains intercept and all four independent variables and thus 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. But it may turn out that all of them is not necessary and that we will we will examine further. So, the corresponding t that we obtained is this. As I said that from the, what you call, 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 t. For the time being let us actually remove this and try the t.

(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

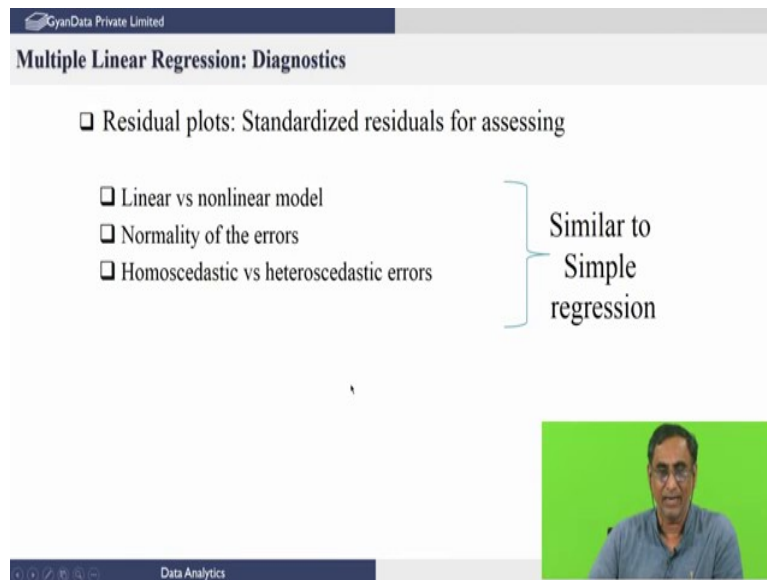
Data Analytics
61

We have done that. We have only included now food, decor and east and done the regression again and it turns out that 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 and the slope parameters.

It indicates that x_3 is not adding any value to the prediction of y . The rea-son 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 t 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: 35:45)



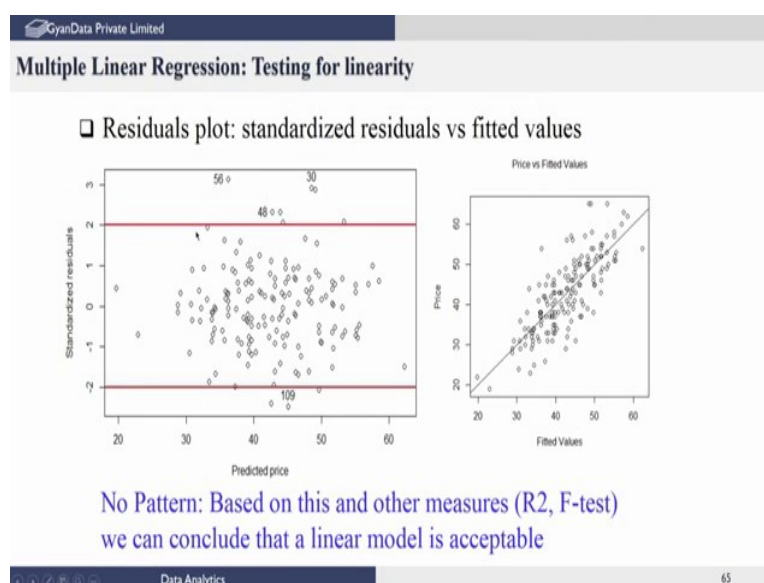
The screenshot shows a presentation slide from 'GyanData Private Limited' titled 'Multiple Linear Regression: Diagnostics'. The slide lists three diagnostic checks for residual plots: 'Linear vs nonlinear model', 'Normality of the errors', and 'Homoscedastic vs heteroscedastic errors'. A blue bracket on the right groups these three items under the text 'Similar to Simple regression'. In the bottom right corner, there is a small video inset of a man with glasses speaking. The bottom of the slide features a navigation bar with icons and the text 'Data Analytics'.

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

Similar to Simple regression

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 no pattern in the residuals. So, let us actually do the residual plot.

(Refer Slide Time: 36:00)



Here we have taken the standardized residuals and plotted it against the, what is called the predicted price value or the fitted value. Remember this is \hat{y}_i , \hat{y}_i has 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 outliers. 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 is spread randomly within this boundary and therefore, we can say since there is no pattern a linear model is acceptable. So, here the quality of the fit is shown here. So, the actual price measured value versus the \hat{y}_i that the predicted value is shown and a linear model seems to explain the data reasonably well.

The last thing is we have these outliers if you want to improve the fit you may want to remove let us say the outlier which is farthest away from the boundary.

For example, you may want to remove 56 and redo the linear regression multiple multiple linear regression and again repeat it until there are no outliers. That will improve the R^2 value and the fit quality of the fit little more.

So, we have not done that we leave this as 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 a variance there it will become a variance covariance matrix here, what was a vector there 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 β all of this, the F statistic 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 done a significance test on the coefficient in order to identify the irrelevant independent variables, but there are better approaches and we will take it up in the following lectures.