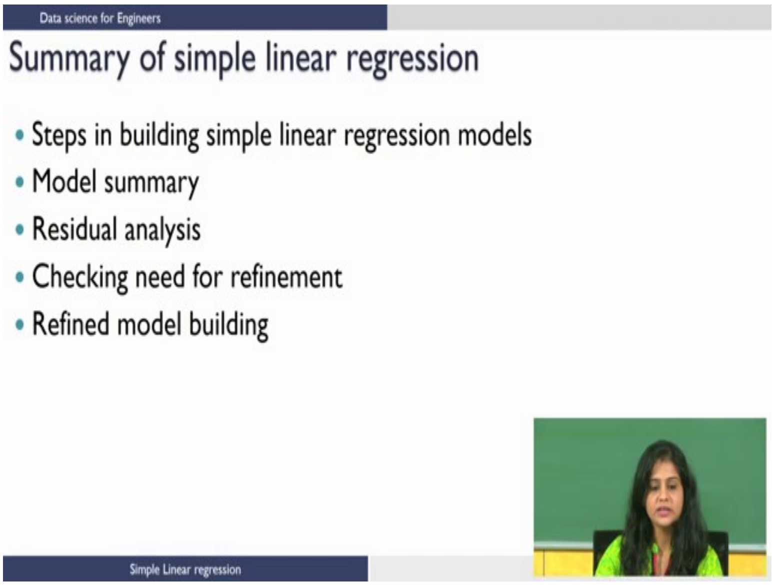
Data science for Engineers
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 40
Multiple Linear Regression Model Building and Selection

Welcome to the lecture on implementation of multiple linear regression to summarize from the previous lecture.

(Refer Slide Time: 00:23)



Data science for Engineers

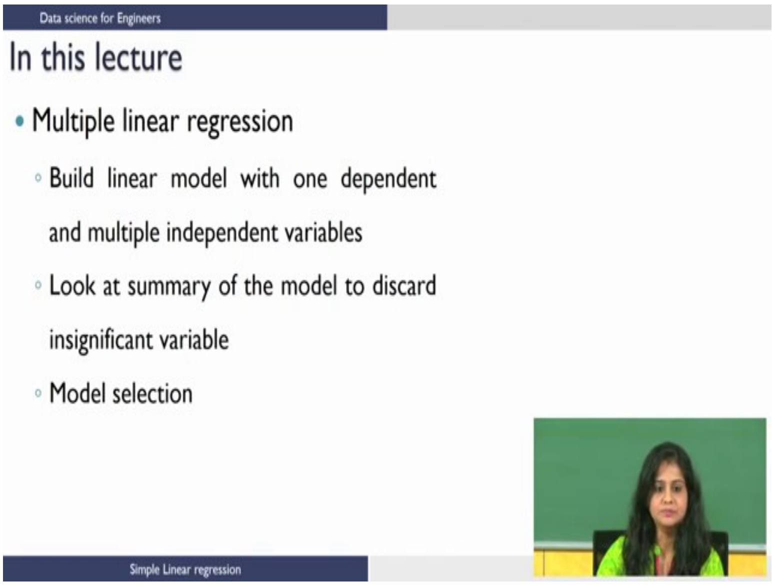
Summary of simple linear regression

- Steps in building simple linear regression models
- Model summary
- Residual analysis
- Checking need for refinement
- Refined model building

Simple Linear regression

We looked at steps in building a simple linear regression model where we looked at how to regress an independent variable with a dependent variable. As a part of this we also looked at how to assess the model that we have built and under that we looked at how to interpret the model summary and identify the significant variables. How to do residual analysis, how to check if the model needs refinement and we built a refined model.

(Refer Slide Time: 00:57)



Data science for Engineers

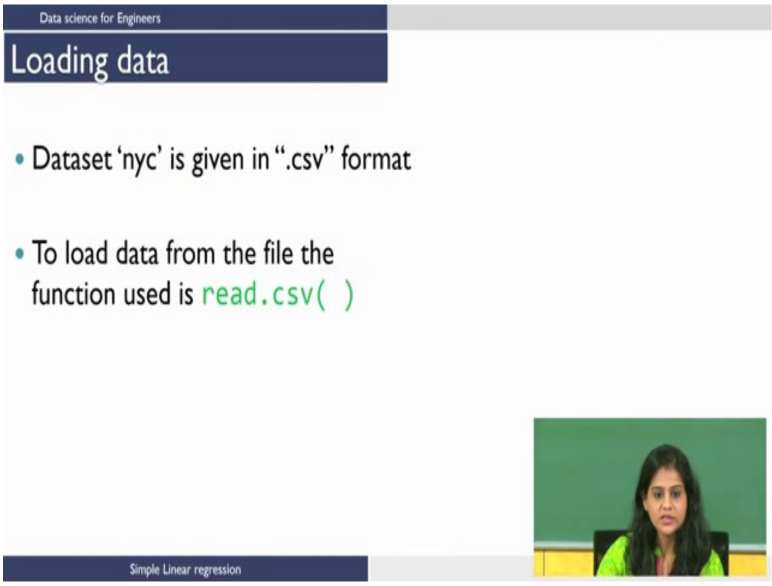
In this lecture

- Multiple linear regression
  - Build linear model with one dependent and multiple independent variables
  - Look at summary of the model to discard insignificant variable
  - Model selection

Simple Linear regression

In this lecture we are going to extend all of this to multiple independent variable so, it is called multiple linear regression and in this we are going to build linear model with one dependent and multiple independent variables. We are also going to look at the model summary and identify the insignificant variables and discard them and rebuild the model. We will also look at how to identify the subset of variables to build the model, this is called model selection.

(Refer Slide Time: 01:20)



Data science for Engineers

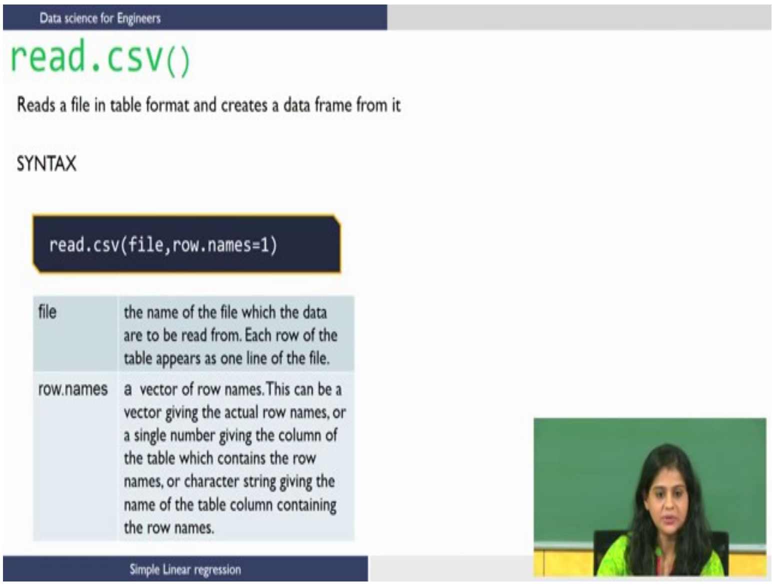
Loading data

- Dataset 'nyc' is given in ".csv" format
- To load data from the file the function used is `read.csv( )`

Simple Linear regression

So, let us start by loading the data so, the data set ‘nyc’ is given to you in a “csv" format and to load the dataset we are going to use the function read dot csv.

(Refer Slide Time: 01:30)



Data science for Engineers

read.csv()

Reads a file in table format and creates a data frame from it

SYNTAX

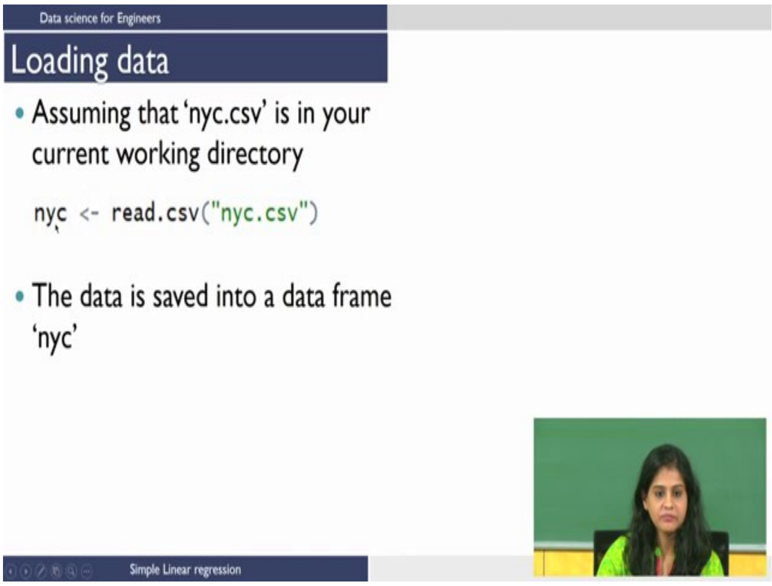
```
read.csv(file,row.names=1)
```

file	the name of the file which the data are to be read from. Each row of the table appears as one line of the file.
row.names	a vector of row names. This can be a vector giving the actual row names, or a single number giving the column of the table which contains the row names, or character string giving the name of the table column containing the row names.

Simple Linear regression

So, the inputs for the function read dot csv it is similar to what we saw in the previous lecture for read dot delim. So, read dot csv reads the file in the table format and creates a data frame from it. So, the syntax is read dot csv and the inputs to the function are file and row names. So, file is the name of the file from which you want to read the data and row names is the vector giving the actual row names, could also be a single number.

(Refer Slide Time: 02:00)



Data science for Engineers

Loading data

- Assuming that ‘nyc.csv’ is in your current working directory

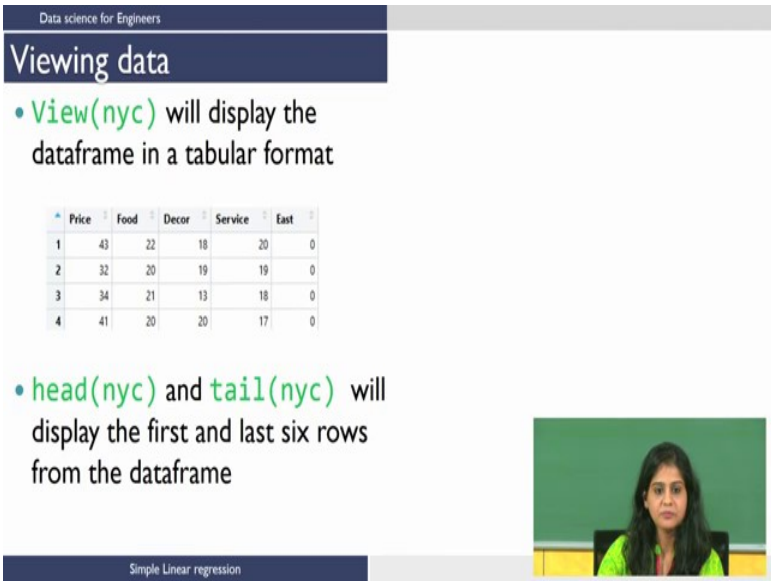
```
nyc <- read.csv("nyc.csv")
```

- The data is saved into a data frame ‘nyc’

Simple Linear regression

So, let us see how to load the data now so, assuming ‘nyc.csv’ is in your current working directory the command is read dot csv followed by the name of the le in double quotes. Now, once this command is executed it will create an object nyc which is a data frame. Now, let us see how to view the data.

(Refer Slide Time: 02:23)



Data science for Engineers

Viewing data

- View(nyc) will display the dataframe in a tabular format

	Price	Food	Decor	Service	East
1	43	22	18	20	0
2	32	20	19	19	0
3	34	21	13	18	0
4	41	20	20	17	0

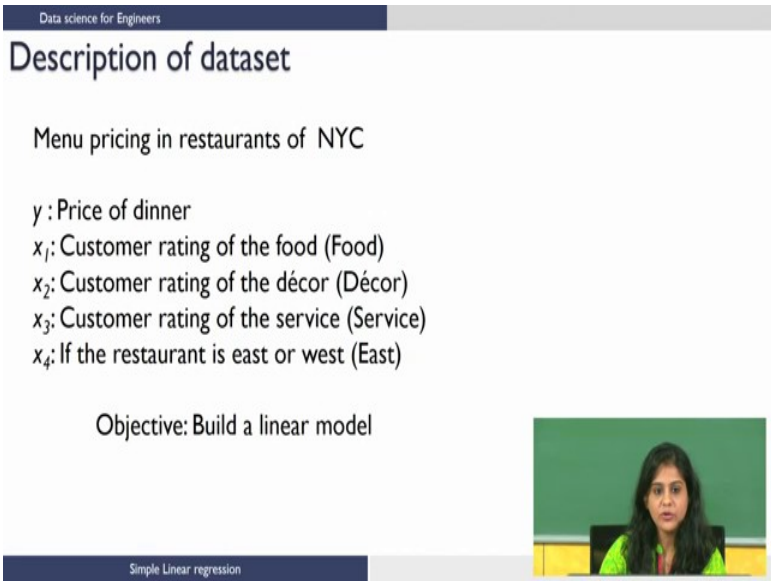
- head(nyc) and tail(nyc) will display the first and last six rows from the dataframe

Simple Linear regression

Now view of nyc will display the data frame in a tabular format. There is a small snippet below which shows you how the output looks. So, I have price, food, decor, service and east as the 5 variables. So, say suppose if your data is really huge and you do not want to view the entire data then we can use head or tail function. So, head will give you the first 6 rows from a data frame and tail will give you the last 6 rows from the data frame.

So, now, let us look at the description of the data set we have already loaded it and we viewed it, but we do not know yet what the description is.

(Refer Slide Time: 02:57)



Data science for Engineers

Description of dataset

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 linear model

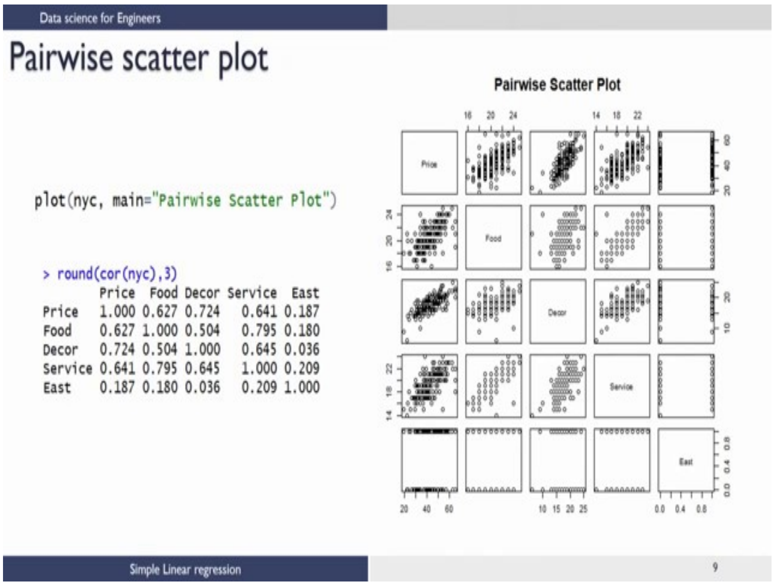
Simple Linear regression

So, the data is about menu pricing in restaurants of New York City. So, y which is my dependent variable is the price of the dinner, there are 4 other independent variables. So, I have food which is one of the independent variables it, is the customer rating of the food then I have decor which is the customer rating of decor, then I have service which is the customer rating of the service and east.

So, east is whether the restaurant is located on the east or west side of the city. So, now, our objective is to build a linear model with y which is price and with all the other 4 independent variables. Before we go on building a model let us say if our data exhibits some interdependency between the variables. So, for me to do that I am going to use a "pair wise scatter plot." So, I am going to use same function plot which we have earlier used.

Now, since I have multiple variable I am going to give the data frame as my input and I am just giving a heading as pair wise scatter plot.

(Refer Slide Time: 04:06)



Data science for Engineers

Pairwise scatter plot

```
plot(nyc, main="Pairwise Scatter Plot")
```

```
> round(cor(nyc),3)
        Price  Food Decor Service  East
Price   1.000 0.627 0.724   0.641 0.187
Food    0.627 1.000 0.504   0.795 0.180
Decor   0.724 0.504 1.000   0.645 0.036
Service 0.641 0.795 0.645   1.000 0.209
East    0.187 0.180 0.036   0.209 1.000
```

Simple Linear regression

9

On my right this is the output you will get so, we can see that all the variables are mentioned across the diagonals. So, when one moves from left to right the variables on my left will be in the y axis and the variables above or below will be on the x axis. So, let us take the first row for instance. So, I have price on the left. So, price is in the y and I have food below. So, food becomes the x axis.

Now, this is the plot for price versus food similarly I have price versus decor and price versus service and price versus east. I am going to the next row which is food on the y axis. So, if you take food versus decor the data is randomly scattered so, it does not show any correlation patterns, but whereas, if you see for food versus service you see strong patterns being exhibited here. So, let us see what the correlation is as such for all of these. So, correlation is a function and Professor Shankar has told you how it is computed.

So, cor is the function in R. I need to give the dataset with all the variables now round tells you to how many decimal points you want round off the number to. So, if I give round and I am giving the input as my correlation function and if I am saying 3 it means round of the number to 3 decimal places. So, let us see how to interpret the output. So, the correlation for price versus price will always be 1.

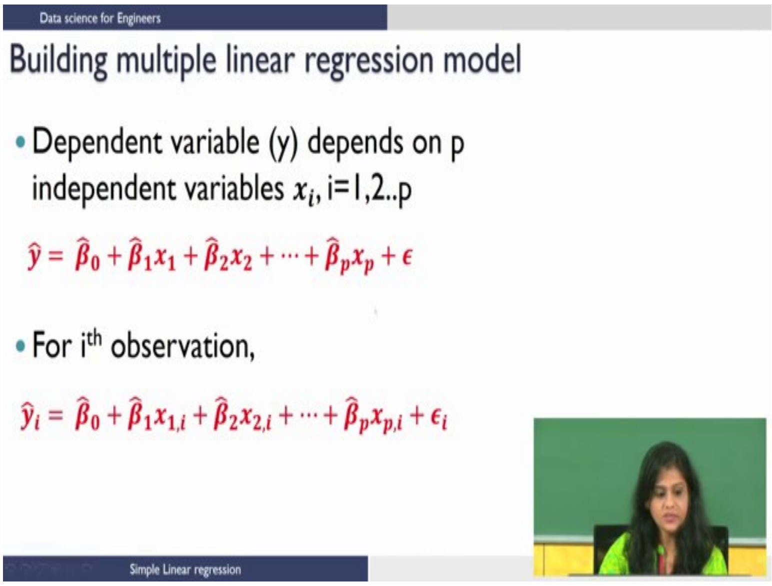
So, let us look at food and decor so, correlation between food and decor is 0.5 which is pretty low, but whereas, if you look at food and service it is almost equal to 0.8 which is quite high. So, we can see that food and service are correlated, but one of them can be dropped while building a final model. So, as we go along let us see which of the two we have to drop.

(Refer Slide Time: 06:25)

Model Building

Now, let us go on to model building.

(Refer Slide Time: 06:28)



Data science for Engineers

Building multiple linear regression model

- Dependent variable (y) depends on p independent variables x_i, i=1,2..p

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p + \epsilon$$

- For ith observation,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \cdots + \hat{\beta}_p x_{p,i} + \epsilon_i$$

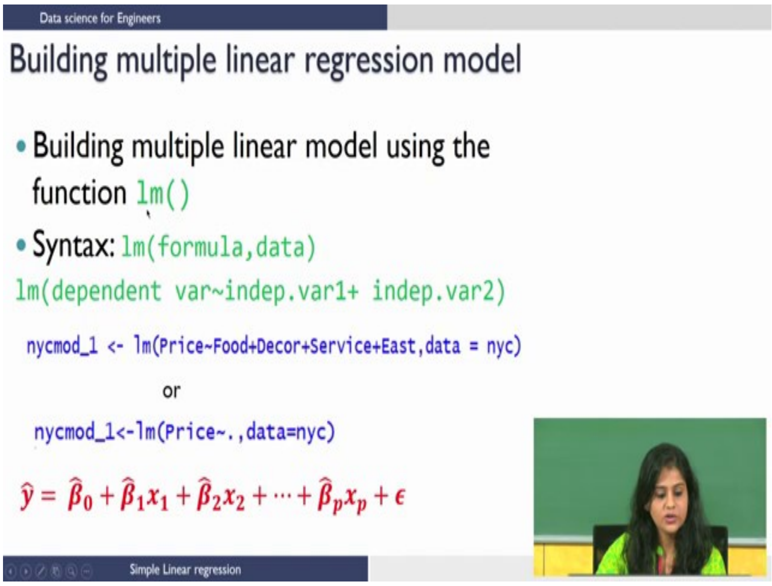
Simple Linear regression

So, like I earlier said, my dependent variable is only one here mean which is denoted by y. I have several independent variables which are denoted by x_i and i code ranges from 1 to p, where p is the total number of independent variables. Now let us see how to write this equation with multiple independent variables. Again I have ŷ which is the predicted value now I have $\hat{\beta}_0$ which is the intercept then I have $\hat{\beta}_1 x_1$ + $\hat{\beta}_2$ x_2 so on and so forth up to $\hat{\beta}_p$ x_p. So, $\hat{\beta}_0$ is the intercept and β_1

hat β_2 hat etcetera are the slopes.

So, ε is the error. So, if you could recall from your earlier lectures in OLS, the assumption is that, so, error is present only in the measurement of dependent variable and not on the independent variable. So, independent variables are free of errors whereas, there is always some error present in the measurement of y. So, this ε is an unknown quantity which has 0 mean and some variance, now for any i th observation this is how my equation is written.

(Refer Slide Time: 07:42)



Data science for Engineers

Building multiple linear regression model

- Building multiple linear model using the function `lm()`
- Syntax: `lm(formula,data)`

```
lm(dependent var~indep.var1+ indep.var2)
```

```
nycmod_1 <- lm(Price~Food+Decor+Service+East,data = nyc)
```

or

```
nycmod_1<-lm(Price~.,data=nyc)
```

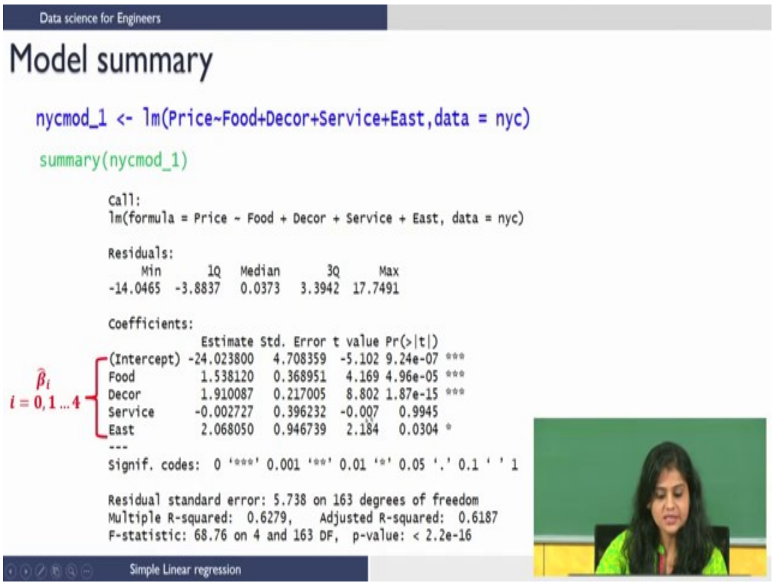
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p + \epsilon$$

Simple Linear regression

So, now, let us go and build a model. So, the function to build a multiple linear model is same as what we used in the univariate case. Here also I am going to use lm now again the syntax is l m and there are 2 input parameters formula and data. Now the syntax is slightly different compared to the univariate case. So, I have my dependent variable here then I have a tilde sign and how many ever independent variables I have I am going to separate them with a + sign. Say for instance I have 2 independent variables in my data. So, I am regressing the dependent variable with 2 independent variables so, the 2 independent variables have to be separated by a + sign.

So, now, let us see how to do it for our data nyc. So, again I have l m so, I am regressing price with all the 4 input variables which is food, decor, service and east and I am taking these variables from the data nyc. So, you can either separate the independent variables by a + sign. So now, if you want to say regress price with all the 4 inputs, there is another way you can write the same command. So, I say regress price and then I give a tilde sign and then I say a dot. So, this means regress price with all the input variables from the data nyc So, if you are going to give all the input variables for regression then you can go with this, but if you have a subset of variables that you want to build a model with, then you can specify the variables separated by a + sign. So, just to reiterate this is the form of my equation. So, now, let us go and see how to interpret the summary. So, after having built this model I am going to look at the summary of it.

(Refer Slide Time: 09:30)



Model summary

```
nycmod_1 <- lm(Price~Food+Decor+Service+East,data = nyc)

summary(nycmod_1)
```

```
Call:
lm(formula = Price ~ Food + Decor + Service + East, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.0465  -3.8837   0.0373   3.3942  17.7491 

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
```

$\hat{\beta}_i$ $i = 0,1 \dots 4$

So, this snippet gives you a just at the summary. So, if you could recall in the first lecture of simple linear regression, we looked at what each of this line here means in depth. So, we have the formula in the first line we have the residuals and the 5 point summary of them ,then we look in at the coefficients. So, here we say that intercept, food, decor, service and east and these are the coefficients for these variables.

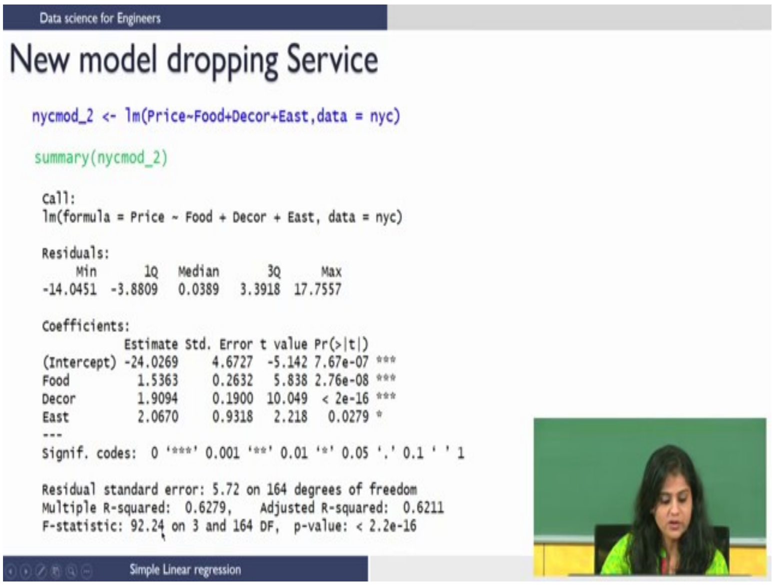
So, for each of these coefficients, I am given an estimate value some variance associated with it a t value which is the ratio of estimate by the standard error and some probability value. So, if you look at the p value. So, the p value for intercept is very low compared to our significance level which is 0.05. So, this tells you that intercept is one of the important terms that have to be included in the model. The same goes with p value for food and decor they are also less than 0.05 and so we need to retain them and the stars tell you how significantly different are they from 0.

Whereas, if you look at the p value of service, it is 0.9945 which is really high compared to our significance level. So, this tells you that this term, service, is not important and if you look at the estimated value it is very very close to 0. So, this tells you that service is not an important term in explaining the price.

So, now, if you see the p value for east, though it is not very very low compared to food and decor it is though it does not have a p value which is very low as that compared to food and decor, it is still OK and the significance star is only one which tells you that look if I have a significance level of say 0.025 or 0.01 then I can reject this term, but till then I can always keep it.

So, let us look at the r squared value. The r squared value is 0.628 and the adjusted r square is 0.619 and the f statistic value is really high which is 68.76. So, this tells you that compared to the reduced models which are the only intercept my full model is performing better and I should retain it. So, now, that we know service is not significant, let us build a new model dropping service.

(Refer Slide Time: 12:11)



Data science for Engineers

New model dropping Service

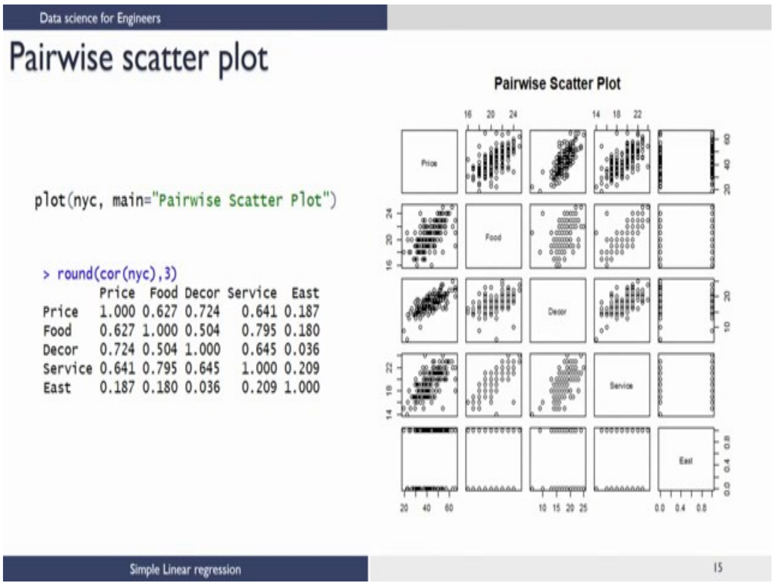
```
nycmod_2 <- lm(Price~Food+Decor+East,data = nyc)

summary(nycmod_2)

Call:
lm(formula = Price ~ Food + Decor + East, data = nyc)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.0451  -3.8809   0.0389   3.3918  17.7557 

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
```

Simple Linear regression

So, I have dropped service and I have built a new model and I am calling it nycmod_2. So, let us jump on to the coefficient section. So, the estimates are not drastically different before and after removing the service variable. So, this tells you that service is not very important. So, again if you look at the p value it tells you that these variables are very significant and if you look at the r squared value here down.

So, the r squared value before and after removing service is not changed much this itself is an indicator that service is not helping us in explaining the variation in price. The adjusted r square has changed a bit and that is only because we have removed one variable and the degrees of freedom change. The f statistic again is really really high telling you that full model with food, decor and east is performing better compared to your reduced model with only the intercept.

(Refer Slide Time: 13:15)



Data science for Engineers

Pairwise scatter plot

```
plot(nyc, main="Pairwise Scatter Plot")
```

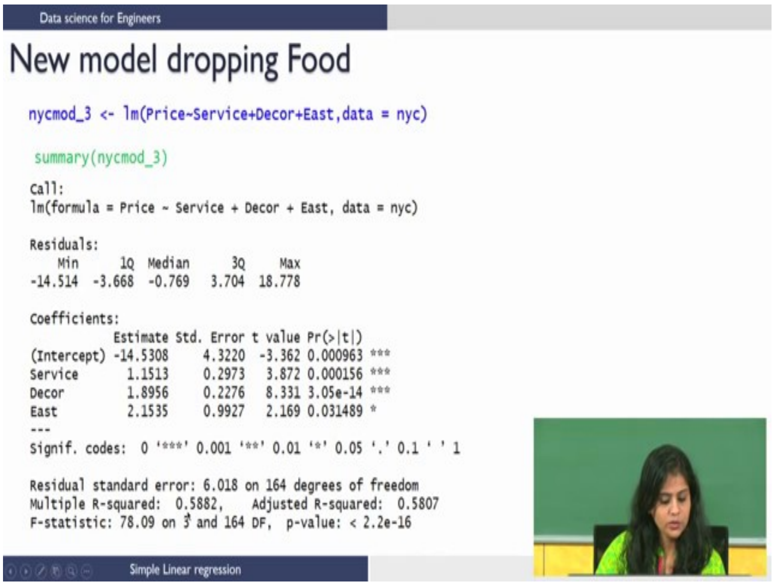
```
> round(cor(nyc),3)
        Price  Food Decor Service  East
Price   1.000 0.627 0.724   0.641 0.187
Food    0.627 1.000 0.504   0.795 0.180
Decor   0.724 0.504 1.000   0.645 0.036
Service 0.641 0.795 0.645   1.000 0.209
East    0.187 0.180 0.036   0.209 1.000
```

Simple Linear regression

15

If you recall from the scatter plot, we saw there was a high correlation between food and service. So, now, we built a model dropping service, let us now retain service and build a model dropping food. So, I have dropped food from here. So, let us take a look at this summary.

(Refer Slide Time: 13:28)



Data science for Engineers

New model dropping Food

```
nycmod_3 <- lm(Price~Service+Decor+East,data = nyc)

summary(nycmod_3)
Call:
lm(formula = Price ~ Service + Decor + East, data = nyc)

Residuals:
    Min      1Q  Median      3Q     Max
-14.514  -3.668  -0.769   3.704  18.778

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.5308     4.3220  -3.362 0.000963 ***
Service       1.1513     0.2973   3.872 0.000156 ***
Decor         1.8956     0.2276   8.331 3.05e-14 ***
East          2.1535     0.9927   2.169 0.031489 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.018 on 164 degrees of freedom
Multiple R-squared:  0.5882,    Adjusted R-squared:  0.5807
F-statistic: 78.09 on 3 and 164 DF,  p-value: < 2.2e-16
```

Simple Linear regression

If you take a look at this summary though the p value tells you that all the variables are significant, if you look at the r squared value it has dropped from 0.628 to 0.588 which is a huge decrease and even the adjusted r square has decreased. So, this tells you that service is less important and food is explaining the price in a much better sense than service.

So, the r squared value and the scatter plots tell us to go ahead with the linear model where we still need to verify the assumptions we make on the errors using residual analysis. So, this task we are going to leave it to you as an exercise you can do it and verify these assumptions.

Thank you.