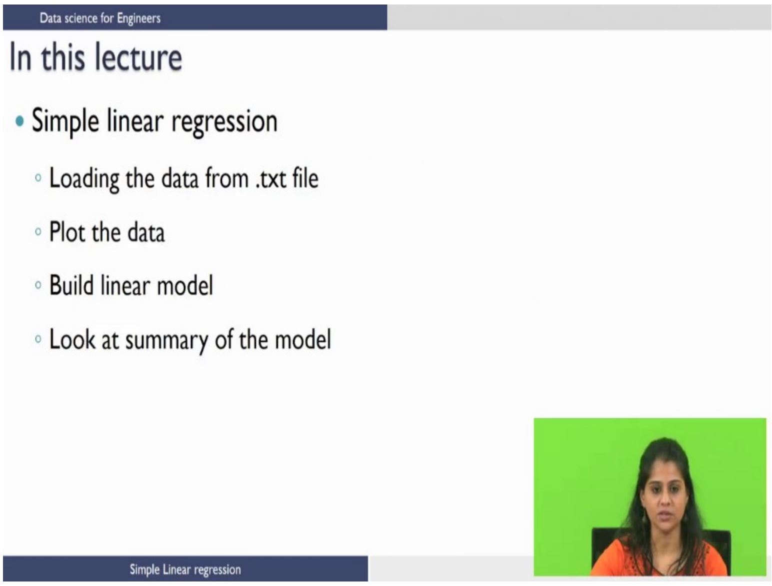
Data Science for Engineers
Prof. Shweta Sridhar
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 35
Simple Linear Regression Modelling

Welcome to the lecture on the implementation of simple linear regression using R

(Refer Slide Time: 00:19)



Data science for Engineers

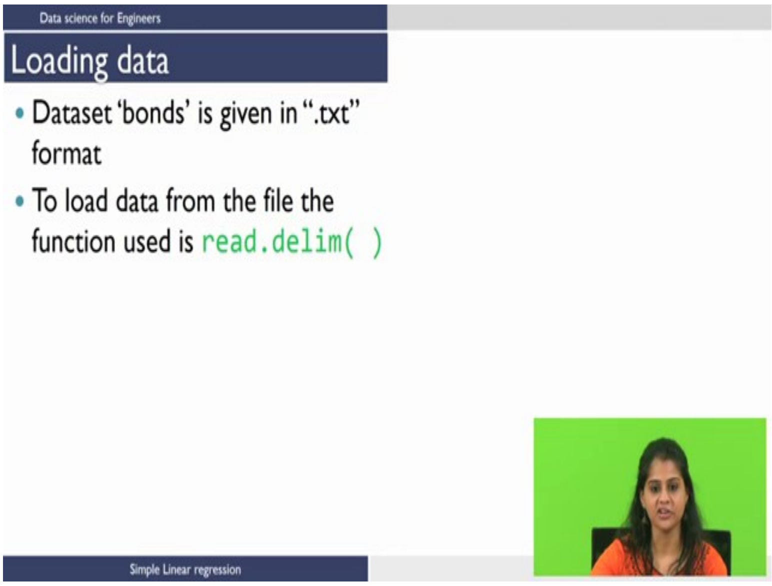
In this lecture

- Simple linear regression
  - Loading the data from .txt file
  - Plot the data
  - Build linear model
  - Look at summary of the model

Simple Linear regression

In this lecture, we are going to see how to implement simple linear regression in R. As a part of this lecture, we are also going to look at how to load the data from a text le, how to plot the data, how to build a linear regression model and how to interpret the summary of the model?

(Refer Slide Time: 00:39)



Data science for Engineers

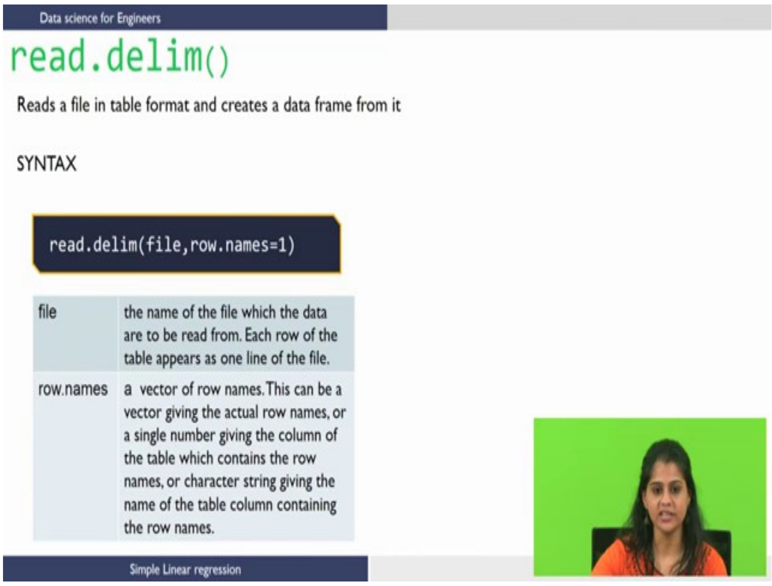
Loading data

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

Simple Linear regression

Now, let us see how to load the data. Now you have been given the data set bonds in the text format, the extension of the file is dot "txt". To load the data from the file, we use the function read dot delim.

(Refer Slide Time: 00:53)



Data science for Engineers

read.delim()

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

SYNTAX

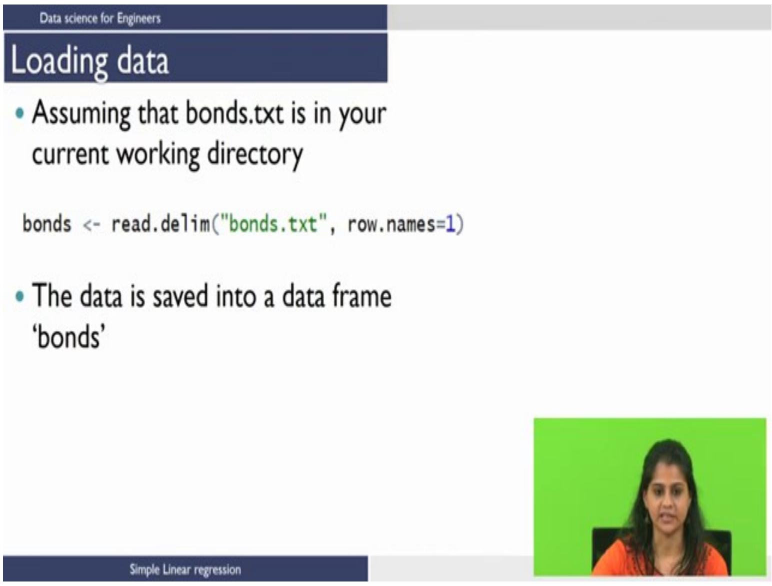
```
read.delim(file,row.names=1)
```

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, read dot delim reads a file in a table format and creates a data frame out of it. The input arguments to the function are file and row dot names. Now file is the name of the file from which you want to read the data and row dot names are essentially the row ids. It can be a vector of names or only a single number corresponding to the column name.

(Refer Slide Time: 01:16)



Data science for Engineers

Loading data

- Assuming that bonds.txt is in your current working directory

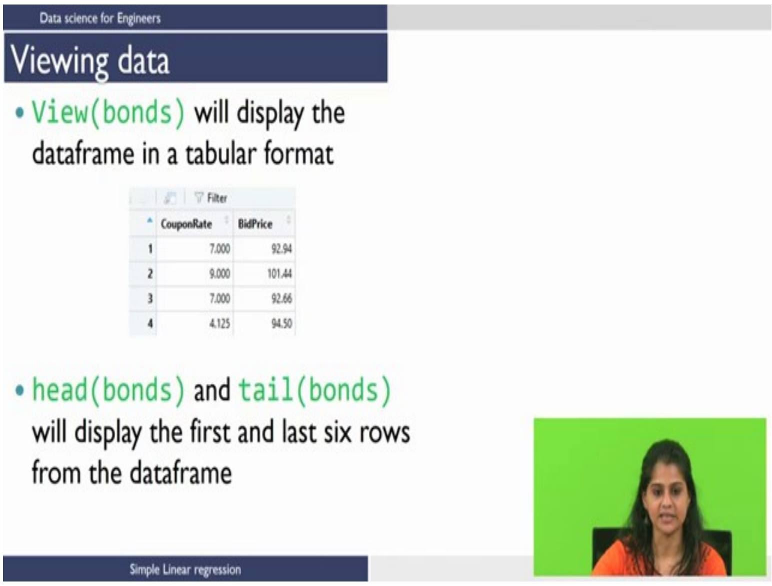
```
bonds <- read.delim("bonds.txt", row.names=1)
```

- The data is saved into a data frame 'bonds'

Simple Linear regression

Now, assuming that the data is in your current working directory, the command reads as read dot delim. So, within quotes I have bonds dot txt and I am giving row dot names = 1. Now, once the command is executed, an object of bonds is created, which is a data frame. Now let us see how to view the data. View of bonds will display the data in a tabular format, the snippet below shows the table.

(Refer Slide Time: 01:40)



Data science for Engineers

Viewing data

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

Filter

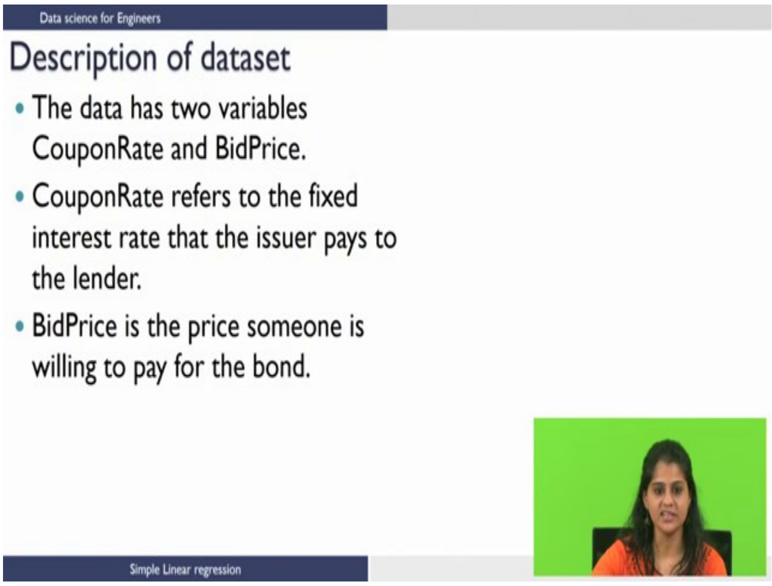
	CouponRate	BidPrice
1	7.000	92.94
2	9.000	101.44
3	7.000	92.66
4	4.125	94.50

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

Simple Linear regression

We can also view the first few rows of any data set, head and tail functions will help us to do that. Now, head of bonds will give us the first 6 rows from the data and tail of bonds will give us the last 6 rows from the data.

(Refer Slide Time: 02:02)



Data science for Engineers

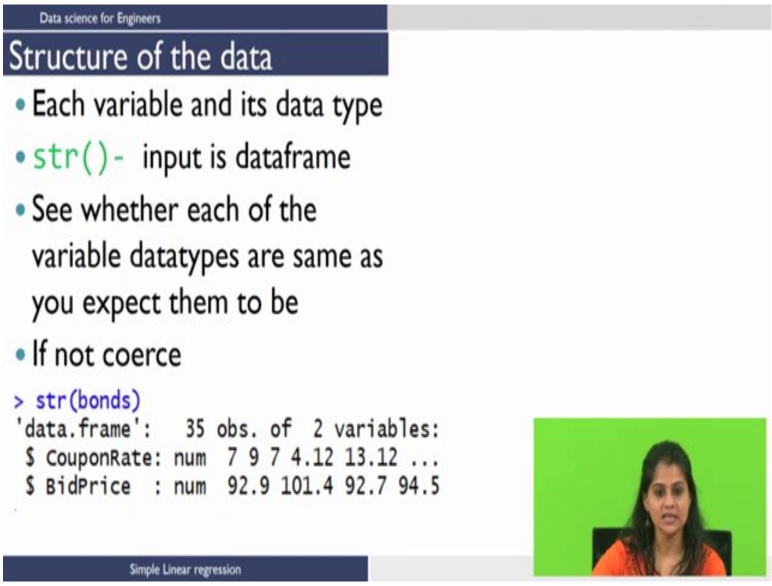
Description of dataset

- The data has two variables CouponRate and BidPrice.
- CouponRate refers to the fixed interest rate that the issuer pays to the lender.
- BidPrice is the price someone is willing to pay for the bond.

Simple Linear regression

Now, let us look at the description of the dataset, now the data has 2 variables coupon Rate and Bid Price. Now, coupon rate refers to the fixed interest rate that the issuer pays to lender. Bid price is the price someone is willing to pay for the bond.

(Refer Slide Time: 02:19)



Data science for Engineers

Structure of the data

- Each variable and its data type
- str()- input is dataframe
- See whether each of the variable datatypes are same as you expect them to be
- If not coerce

```
> str(bonds)
'data.frame':   35 obs. of  2 variables:
 $ CouponRate: num  7 9 7 4.12 13.12 ...
 $ BidPrice  : num  92.9 101.4 92.7 94.5
```

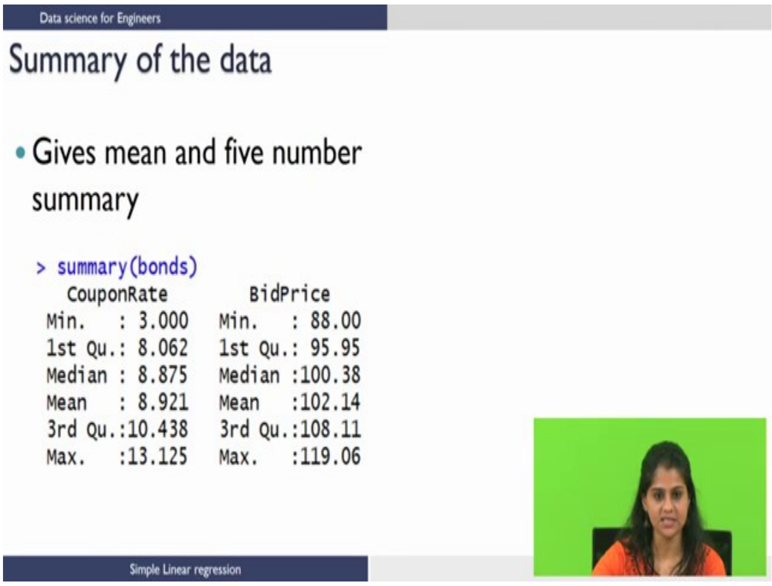
Simple Linear regression

Now, we have seen how to load the data and how to view the data. Let us now see what the structure of the data is. By structure I mean that each variable and it is data type. We use the function str and the input to the function is a dataframe. Now we exactly want to see whether the variable data type are same as what we expected them to be, if not we need to coerce them to the respective data types.

So, now this should ring a bell, because we have learned the function as dot followed by the name of the data type and we will use this function to coerce it if the variable is not of the desired type. Now for this dataset, I run the function I say str of bonds now bonds is the name of my data frame. So, the output reads as data frame bonds is of the type data frame it has 35 observations of 2 variables. The first column being coupon rate, which is of the type numeric and I have the first few values being displayed.

The next column Bid Price is also of the type numeric and the first few values of the same column are being displayed.

(Refer Slide Time: 03:32)



Data science for Engineers

Summary of the data

- Gives mean and five number summary

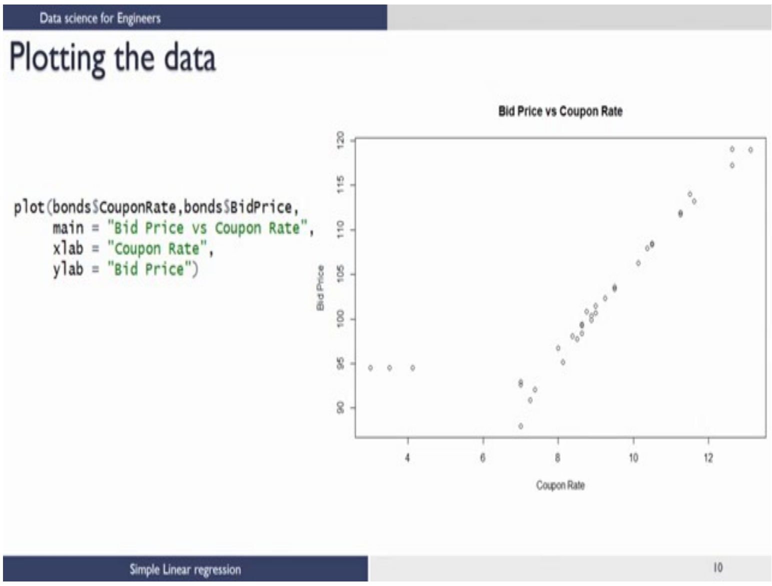
```
> summary(bonds)
   CouponRate        BidPrice
 Min.   : 3.000   Min.   : 88.00
 1st Qu.: 8.062   1st Qu.: 95.95
 Median : 8.875   Median :100.38
 Mean   : 8.921   Mean   :102.14
 3rd Qu.:10.438   3rd Qu.:108.11
 Max.   :13.125   Max.   :119.06
```

Simple Linear regression

Now, let us look at the summary of the data. So, the summary function followed by the name of the data frame in this case bonds will give us 5 number summary and mean from the data.

Now, the first column which is coupon rate, I have the 5 number summary and the mean and I also have the 5 number summary and the mean for the second column. So, till now we have seen how to load the data, how to view the data, We have also looked at the structure and the summary.

(Refer Slide Time: 04:02)



Data science for Engineers

Plotting the data

```
plot(bonds$CouponRate,bonds$BidPrice,
     main = "Bid Price vs Coupon Rate",
     xlab = "Coupon Rate",
     ylab = "Bid Price")
```

Simple Linear regression

10

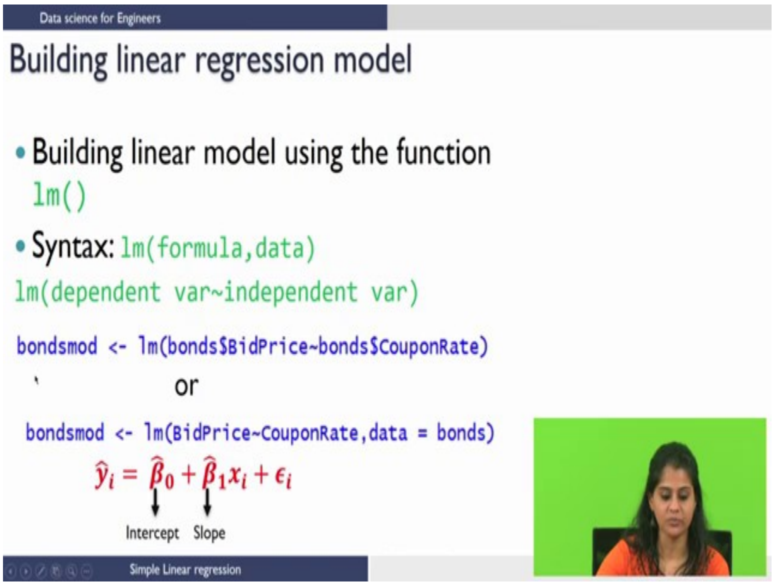
Now, let us see how to visualize the data. So, to visualize the data I use the plot function. We have covered the plot function earlier in the visualization in r section, now the input to the plot function are basically x and y. In this case x refers to my coupon rate and y refers to my bid price. So, in order to access the variables, I need to give the name of the data frame followed by a dollar symbol. So, I say access coupon rate from the bonds data and access bid price from the bonds data.

So, I can also give a title to my plot. So, inside the parameter main you can specify the title of your plot, xlab is nothing, but x label. So, I am assigning it as x “Coupon Rate" and y label I am assigning it as “Bid Price". So, the plot is on right hand side.

So, the title is bid price versus coupon rate like how we have assigned it on the y axis I have bid price and I have labeled it as bid price and on the x axis, I have coupon rate and I have labeled it as coupon rate.

Now, we see a linear trend. Now there are some points which are completely outside the range of coupon rate. Now let us see if our linear model will help us to identify these points.

(Refer Slide Time: 05:19)



Data science for Engineers

Building linear regression model

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

```
lm(dependent var~independent var)
bondsmod <- lm(bonds$BidPrice~bonds$CouponRate)
```

or

```
bondsmod <- lm(BidPrice~CouponRate,data = bonds)
```

$$\underset{\text{Intercept}}{\hat{y}_i = \hat{\beta}_0} + \underset{\text{Slope}}{\hat{\beta}_1} x_i + \epsilon_i$$

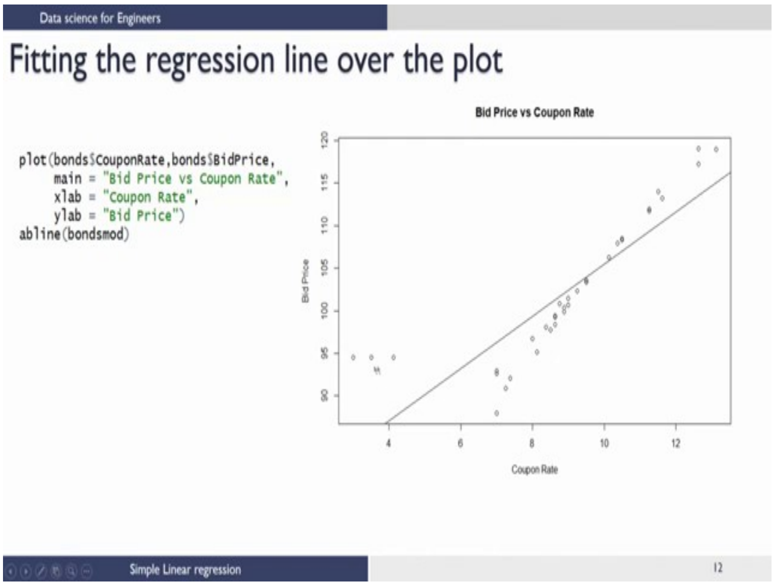
Simple Linear regression

So, to start with let us build a linear regression model. So, building a linear model is done using the function lm. So, the inputs for the function are formula and data, by formula I mean I am regressing dependent variable versus the independent variable.

So, how is it translated to a formula? So, I have dependent variable I have a tillet sign followed by the independent variable. So, the tillet sign tells us regress the dependent variable with the independent variable So, there are 2 ways to build the linear model, let us see how to do that. The first way tells us linear model which is a function. So, I am accessing the individual variables which is bid price and coupon rate using the data frame followed by the dollar symbol. So, take bid price from the bonds data and take coupon rate from the bonds data and regress them.

There is also another way. So, instead of mentioning the name of the data frame to access the variables, we can directly mention the name of the variable and give data = bonds. So, access these variables from the bonds data. So, assuming our equation is of the form ŷi = $\hat{\beta}_0$+ β1 xi + ε i. So, ε i is the error which will be called as residuals. So, $\hat{\beta}_0$is the intercept and β1 is the slope So, hereafter these estimates will be referred to as intercept and slope. So, now that we have built a linear model and have saved it as an object bondsmod let us see how to fit the regression line over the plot.

(Refer Slide Time: 07:02)



Data science for Engineers

Fitting the regression line over the plot

```
plot(bonds$CouponRate,bonds$BidPrice,
     main = "Bid Price vs Coupon Rate",
     xlab = "Coupon Rate",
     ylab = "Bid Price")
abline(bondsmod)
```

Simple Linear regression

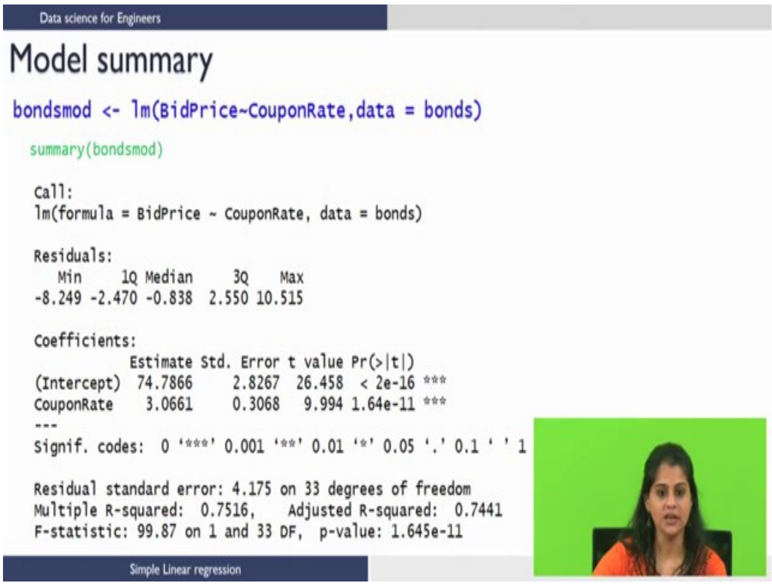
12

We will use a function called ab line and the input for the function is bondsmod which is my linear model.

Now, we have already gone back and seen how to plot coupon rate and bid price. Now in addition to the plot you need to mention this command. So, ab here refers to the intercept and slope. If your equation is of the form y = a + bx, then a is my intercept and b is my slope.

In this case a is β_0 and b is β1. So, let us see how the plot looks. So, on my right I have the plot we are now able to see how the regression line fits. It fits pretty badly and it is also not identify the outliers. So, we can say that regression line is indeed getting affected by these outliers.

(Refer Slide Time: 07:59)



Data science for Engineers

Model summary

```
bondsmod <- lm(BidPrice~CouponRate,data = bonds)

  summary(bondsmod)

 Call:
 lm(formula = BidPrice ~ CouponRate, data = bonds)

 Residuals:
    Min     1Q Median     3Q    Max
 -8.249 -2.470 -0.838  2.550 10.515

 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
 CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 4.175 on 33 degrees of freedom
 Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
 F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

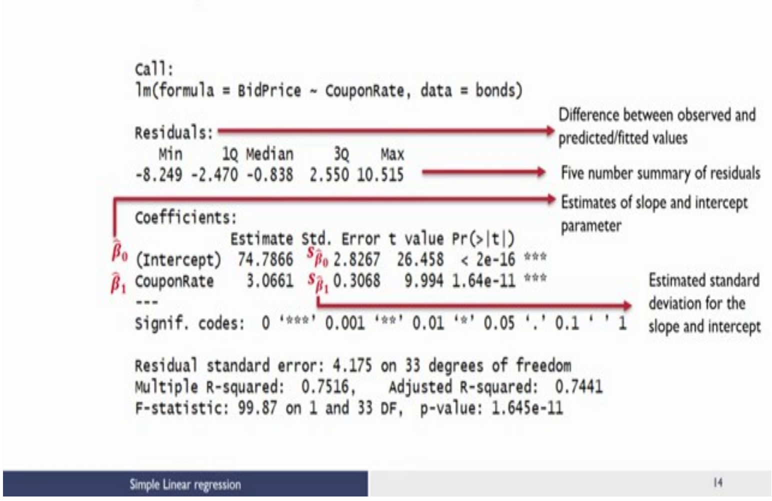
Simple Linear regression

So, now let us take a look at the model summary. So, I have regress bid price versus coupon rate from the data bonds, you can also use the other command.

So, summary is a function the input to the summary function becomes a linear model. So, we have bondsmod as the linear model, this is the first look at the summary. So, this is how it looks when you run the command and this is how it would look in the callzone.

So, we have 4 sections of output we have call, we have residual, we have coefficients, and we have some few heuristics at the bottom. So, now, let us look at each of these and what they mean in depth.

(Refer Slide Time: 08:37)



```
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
   Min     1Q Median     3Q    Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Difference between observed and predicted/fitted values

Five number summary of residuals

Estimates of slope and intercept parameter

Estimated standard deviation for the slope and intercept

Simple Linear regression

14

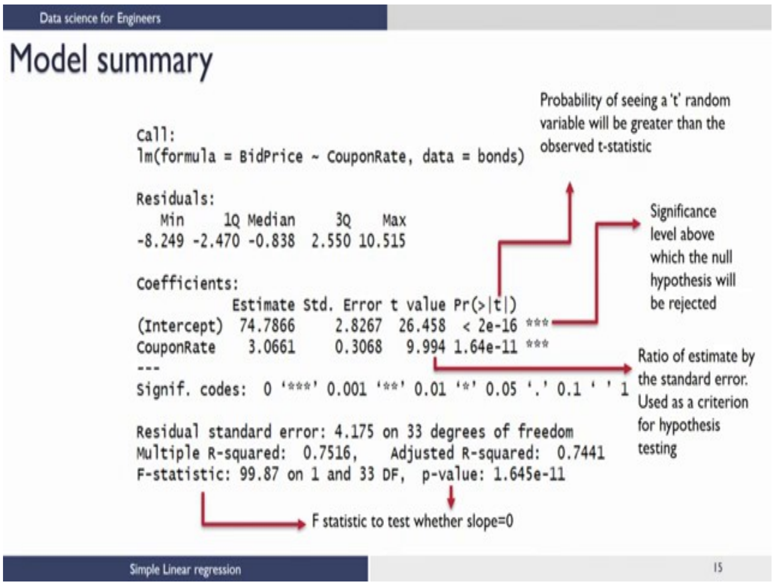
Now, call displays the formula which we have used. So, in this case I have used the formula bidprice versus coupon rate so regress bidprice, which is my dependent variable with my independent variable which is coupon rate and from the data bonds.

Now, this is the way for you to check if you have given the right dependent and independent variable. The next section is residual. So, what are residuals are nothing, but difference between the observed and predicted values. So, in our equation earlier we saw we had a parameter called εi. So, that εi corresponds to residuals below the residuals is the 5 number summary for the residual.

So, the next section is coefficients. We see 2 rows which is intercept and coupon rate and certain set of values associated with them. Now these intercept and coupon rate are nothing, but $\hat{\beta}_0$ and $\hat{\beta}_1$ hat, we earlier saw for our equation y = $\hat{\beta}_0$+ $\hat{\beta}_1$ into xi + ε I, $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ is the slope.

Now, let us see what other 4 parameters in the column have to say. Now I have the first column which is estimate. Now this is nothing, but the estimate for the slope and intercept parameter. The next column is standard error. So, standard error is the estimated standard deviation associated for the slope and intercept.

(Refer Slide Time: 10:08)



Data science for Engineers

Model summary

```
Call:
lm(formula = BidPrice ~ CouponRate, data = bonds)

Residuals:
   Min      1Q Median      3Q     Max
-8.249 -2.470 -0.838  2.550 10.515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.175 on 33 degrees of freedom
Multiple R-squared:  0.7516,    Adjusted R-squared:  0.7441
F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

Probability of seeing a ‘t’ random variable will be greater than the observed t-statistic

Significance level above which the null hypothesis will be rejected

Ratio of estimate by the standard error. Used as a criterion for hypothesis testing

F statistic to test whether slope=0

Simple Linear regression

15

I have the next column as t value. So, what is t value? It is the ratio of estimate by the standard error and it is also an important criterion for the hypothesis testing. The column after that is the probability.

So, it is the probability of seeing a t random variable, which will be greater than the observed t statistic. So, we can see few stars being indicated at the end. So, what are these stars. These stars tell us the significance level above, which the null hypothesis will be rejected.

So, what is the null hypothesis? The null hypothesis is that the estimates will be = 0. At the last line I have an F statistic and a corresponding p-value associated with it. Now the F statistic is again used to test the null hypothesis which is nothing, but slope = 0.

So, in this lecture we saw how to load a data, how to plot and how to visualize, how to build a linear model and how to interpret the results from the linear model? So, in the next lecture we will see how to assess our model and we will see if we can improvise our model.

Thank you.