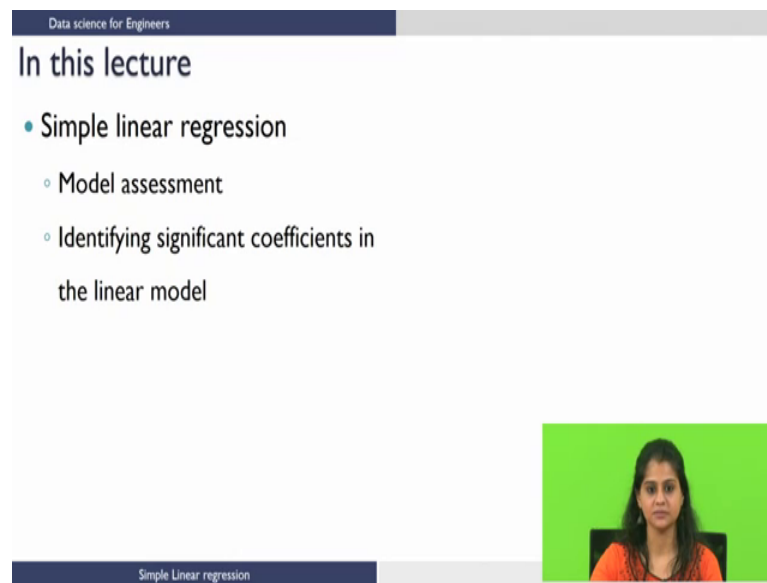


Data Science for Engineers
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 36
Simple Linear Regression Model Assessment

So, welcome to the second lecture on implementation of simple linear regression using R. In the last lecture, we saw how to read a data from a text file, how to visualize the data, how to build a linear model, and how to interpret it.

(Refer Slide Time: 00:31)



The slide is titled "Data science for Engineers" at the top left. The main heading is "In this lecture". Below it, there is a bulleted list:

- Simple linear regression
 - Model assessment
 - Identifying significant coefficients in the linear model

In the bottom right corner, there is a video inset showing a woman with dark hair wearing an orange top, speaking against a green background. At the bottom left of the slide, there is a footer that says "Simple Linear regression".

In this lecture, we are going to look at simple linear regression model assessment. As a part of this, we are also going to look at how to identifying, significant coefficients in the linear model.

(Refer Slide Time: 00:40)

Data science for Engineers

Model assessment

- How good is the linear model?
- Which coefficients of the linear model are significant (Identify important variables)
- Can we improve quality of linear model?
 - Are there bad measurements in the data (outliers)

Simple Linear regression



Now, let us start with model assessment. So, there are a few questions which we need to answer before we go into model assessment. After having built a model, we first need to check how good is a linear model. Now, we need to identify which coefficients in the linear model are significant.

Now, if you have multiple independent variables, then we also need to identify which of them are important. We also need to know can we improvise the model further. As a part of this, we are going to look at are there any bad measurements in the data. So, by bad measurements mean are there any other outliers in the data which could affect the model. This question alone will be handled in the next lecture.

So, let us look at how to answer the first two questions.

(Refer Slide Time: 01:23)

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
```



Simple Linear regression

Now, from the earlier lecture we saw how to look at the summary. We also know how to interpret it now. Now, this is the first just of summary that you get when you run the command.

So, I had regressed BidPrice with coupon rate from the data bonds and bondsmod was my linear model. I also have the estimates here which are nothing but the intercept and slope. So, let us look at the first level of model assessment.

(Refer Slide Time: 01:45)

Data science for Engineers

First level model assessment

- R^2 value=0.7516
- Hypothesis Testing
 - Coefficients
 - Full model and reduced model



Simple Linear regression

So, the first level of model assessment is done using the R squared value. Now if you go back and see, the R squared value for a model is 0.7516. Now this is pretty close to 1. Though not very close, but it is still closer to 1.

So, we can say that, yes, the model we have developed is reasonably good, but not really good. It also tells us the assumption that we made initially to begin with, that there is a linear relationship between x and y. We are also going to look at hypothesis testing, as a part of this; we are going to look at the hypothesis testing on coefficients and then on the full and reduced model.

Now, let us see what these full and reduced model are. So, first let us do the hypothesis test on coefficients.

(Refer Slide Time: 02:36)

Data science for Engineers

First level model assessment- Hypothesis test on coefficients

- In order to check if linear model is good we can check if the estimate $\hat{\beta}_1$ is significant
- Hypothesis Testing,
- Null Hypothesis $H_0: \hat{\beta}_1 = 0 \Rightarrow \hat{y}_i = \hat{\beta}_0 + \epsilon_i \rightarrow$ Reduced Model
- Alternate Hypothesis $H_1: \hat{\beta}_1 \neq 0 \Rightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i \rightarrow$ Full Model
- The confidence interval is computed to check if $\hat{\beta}_1$ is significant

Simple Linear regression

So, in order to check, if your linear model is good. We can check if the estimate beta 1 hat which is the slope, whether it is significant. So, my null hypothesis is, beta 1 hat which is the slope equal to 0.

So, this means that y i hat which is my predicted value is equal to beta naught hat, which is the intercept plus epsilon i. We also learnt in the earlier lecture, that this epsilon i is the residual. Now this becomes my reduced model. Since, my slope is equal to 0. So, what will the alternate hypothesis be in this case? So, my alternate hypothesis is that, beta 1

hat not equal to 0, and my $\hat{y}_i$ which is the predicted value is equal to $\beta_0 + \beta_1 \hat{x}_i + \epsilon_i$.

(Refer Slide Time: 03:36)

Data science for Engineers

First level model assessment- Hypothesis test on coefficients

- Test on $\hat{\beta}_1$ is a two sided test
- At $\alpha = 0.05$ i.e 95% confidence level

```
> alpha=0.05
> n=35
> p=1
> qt(p = 1-(alpha/2),df = n-p-1)
[1] 2.034515
```

- $\hat{\beta}_1 = 3.0661$ and the standard deviation associated is $s_{\hat{\beta}_1} = 0.3068$
- Confidence interval for $\hat{\beta}_1$ is,

```
> 3.0661-(2.034515*0.3068)
[1] 2.441911
> 3.0661+(2.034515*0.3068)
[1] 3.690289
```

Simple Linear regression 24

Now, this becomes my full model. The confidence interval is computed to check if the slope is significant. Now, this test is a two sided test, since we see β_1 is equal to 0 or not equal to 0. At 95 percent confidence level; that is, at alpha equal to 5 percent. We get the critical value to be 2.0345. Now, let us see how to compute this critical value.

So, we know that alpha is equal to 0.05. And n here is the number of observations in your data. Now, in this case I have 35 observations. P becomes my number of independent variables. Here I have only my one independent variable. So now, we know from the statistics module how to compute the quantiles for a t from a t distribution.

Now, I give p equal to 1 minus alpha by 2 since it is a two sided test, and the number of degrees of freedom are given as n minus p minus 1. So, this command is in built in R you just need to give the inputs. You need to supply p and degrees of freedom.

After having done this we get the quantiles to be equal to 2.03. So, this is the critical value we are going to use this to compute the confidence interval. Now, we earlier saw from the summary that the slope, which is nothing but the β_1 is equal to 3.0661 and the standard deviation associated with it was 0.3068. So, the confidence interval is computed as the estimate plus or minus the critical value into the standard error.

So, by doing so we get the lower bound as 2.44, and the upper bound as 3.69. So now, we know that, this interval does not encompass 0; that is, anywhere between the interval I do not have 0. So, this itself is indicative of the fact that my beta 1 hat that is the slope is significant.

(Refer Slide Time: 05:36)

Data science for Engineers

First level model assessment- Hypothesis Test on models

- Computing F statistic

$$F_o = \frac{SST - SSE}{SSE / (n - 2)} = \frac{SSR}{SSE / (n - 2)}$$

$$SSE = \sum (y_i - \hat{y}_i)^2$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

```

> SSE<-sum((bonds$BidPrice-bondsmod$fitted.values)^2)
> SSE
[1] 575.3418
> SSR<-sum((bondsmod$fitted.values-mean(bonds$BidPrice))^2)
> SSR
[1] 1741.263
> n=35
> (SSR/SSE)*(n-2)
[1] 99.87401
```
- This F statistic is returned by the summary command

Simple Linear regression

1

Now, let us do a hypothesis test on the models. So, to do so we use the F statistic. So, let us go back and revisit what the F statistic is. So, F statistic is nothing but my sum squared residual divided by the sum square error by the degrees of freedom for the denominator. So, for the sum square residual, we know it is of the form of summation of $y_i \text{ hat} - \bar{y}$ the whole square. So, I have only one degrees of freedom, since, I am using only one parameter to compute it.

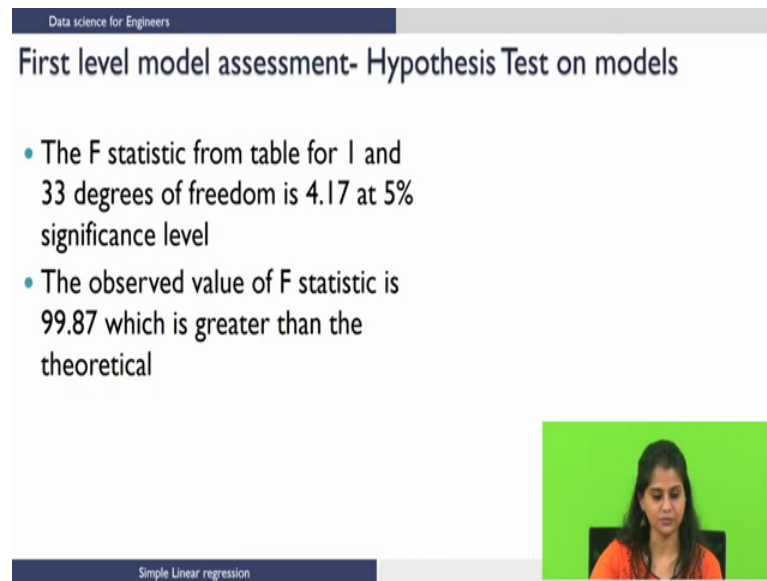
Whereas for the sum square error it is the summation of $y_i - y_i \text{ hat}$ whole square; Now, I am using two parameters to compute it. So, the degrees of freedom reduced by 2. So, hence I have the denominator as n minus 2. So, this is how you would compute the sum squared error. So, I am summing my y_i which is nothing but from bid price bond staller BidPrice. I have the fitted values, and I know the mean which is $\bar{y}$ of the bid price. I am squaring the term, and I am summing it.

Now, we know that my num the number of observations we have a 35 from the data. So, from the formulae we know our F statistic is computed as SSR by SSE into n minus 2.

Degrees of freedom $n - 2$ go to the numerator, and we get the F statistic to be equal to 99.87.

Now, this F statistic is what is returned by the summary, which is given in the last line of the summary.

(Refer Slide Time: 07:16)



The slide is titled "First level model assessment- Hypothesis Test on models" and is part of a "Data science for Engineers" presentation. It contains two bullet points:

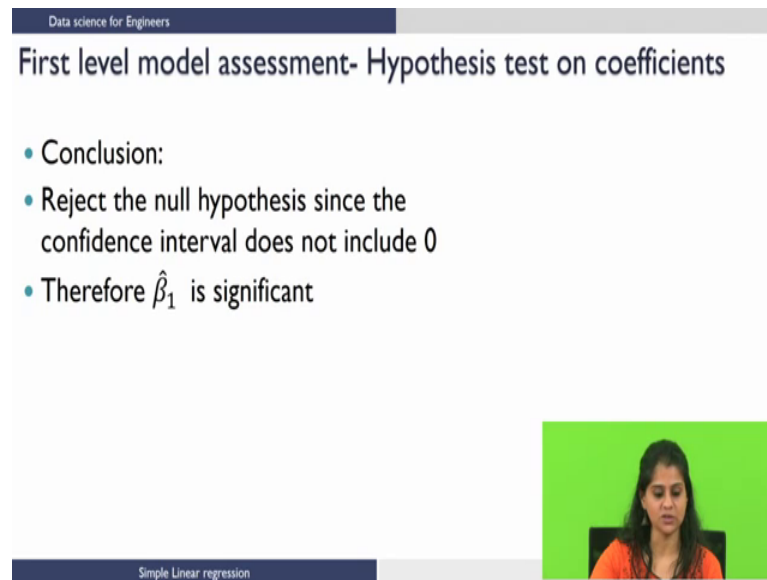
- The F statistic from table for 1 and 33 degrees of freedom is 4.17 at 5% significance level
- The observed value of F statistic is 99.87 which is greater than the theoretical

A small video inset in the bottom right corner shows a woman with dark hair wearing an orange top. The slide footer includes "Simple Linear regression".

So, let us see what conclusions can we draw from these two tests. We know that the F statistic from the table. 1 and 3 degrees of freedom is 4.17 at 5 percent significance level.

What we observe is 99.87, at 1 and 33 degrees of freedom. Now, this is greater than the theoretical value that we get from the distribution.

(Refer Slide Time: 07:40)



The slide is titled "Data science for Engineers" and "First level model assessment- Hypothesis test on coefficients". It contains a bulleted list with the following text:

- Conclusion:
- Reject the null hypothesis since the confidence interval does not include 0
- Therefore $\hat{\beta}_1$ is significant

In the bottom right corner, there is a small video inset showing a woman with dark hair wearing an orange top, speaking against a green background. The bottom of the slide has a dark blue bar with the text "Simple Linear regression".

So, what conclusion can we draw now? So, we know that, we can reject the null hypothesis, since the confidence interval does not include 0. And hence beta 1 hat which is the slope is also significant.

So, in this lecture, we saw how to assess the model. We also looked at how to answer some of the important question that gets associated while assessing a model. We also saw how to identify the significant coefficients. In the next lecture we will look at how to identify outliers and how to improvise a model.

Thank you.