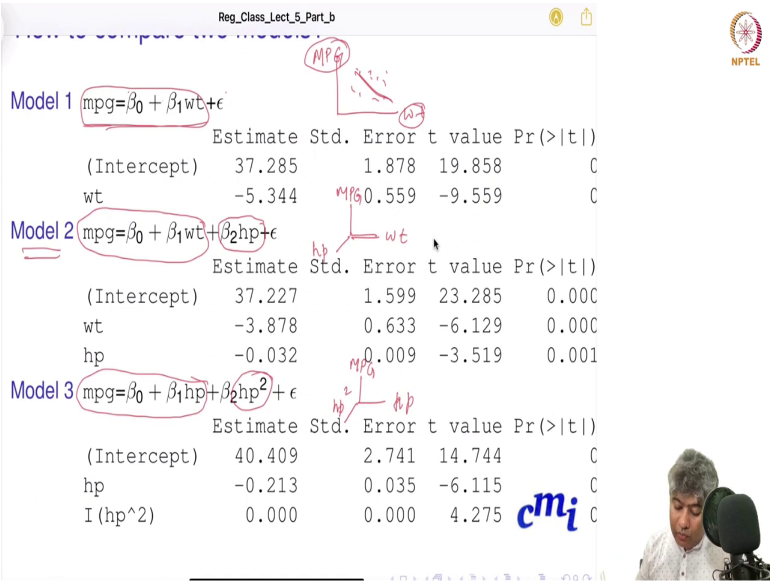
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 18
Comparing Models with Predictive Accuracy

Welcome back to the Part B of Lecture 5 and we are going to talk about that how we compare two models. In fact, we are going to talk about how we compare multiple models. Now, in this lecture we will start with three possible models.

(Refer Slide Time: 00:40)



Model 1 $mpg=\beta_0+\beta_1 wt+\epsilon$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.285      1.878  19.858        C
wt             -5.344      0.559  -9.559        C
```

Model 2 $mpg=\beta_0+\beta_1 wt+\beta_2 hp+\epsilon$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    37.227      1.599  23.285     0.00C
wt             -3.878      0.633  -6.129     0.00C
hp             -0.032      0.009  -3.519     0.001
```

Model 3 $mpg=\beta_0+\beta_1 hp+\beta_2 hp^2+\epsilon$

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    40.409      2.741  14.744        C
hp             -0.213      0.035  -6.115        C
I(hp^2)         0.000      0.000   4.275        C
```

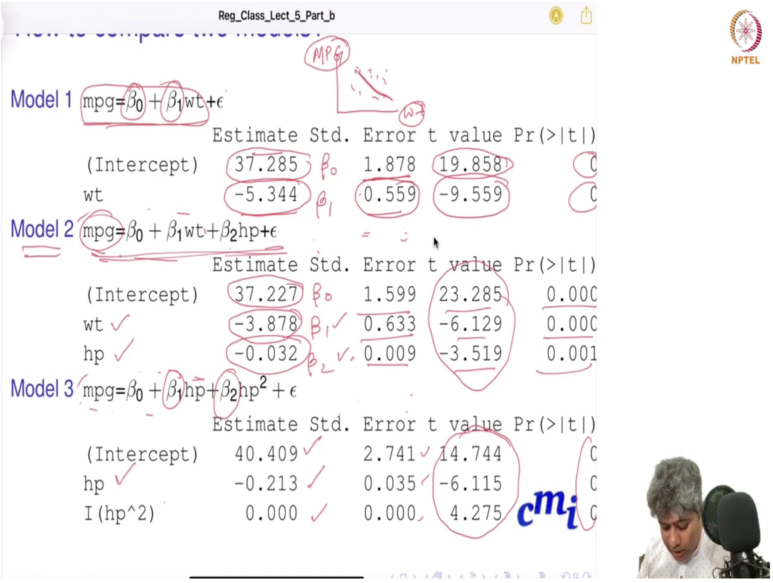
Now, one is model 1 which is miles per gallon as a function of beta naught plus beta 1 times weight plus epsilon. Now, this is one model with where you have essentially two variables. So, basically weight you are modelling miles per gallon as a function of weight, ok. So, you have weight on one axis and miles per gallon on the other axis and you are modelling as a function miles per gallon as a function of weight and that is what you are going to test.

Next model is function of weight. This is already in the part of the model 1, this part is same as model 1. In addition, now you are adding horsepower, ok. You are adding a second variable horsepower second independent variable or second feature horsepower in the model and this is your second model. So, it is going to be something like this like on one axis you are putting weight in one axis you putting horsepower and in one axis you are putting miles per gallon ok.

So, this is the second model and the third model is you are putting considering horsepower and the horsepower square. So, that means, it is a 2D model, but you are generating you are engineering a third axis horsepower square along with horsepower and putting miles per gallon on the third axis.

So, basically 2-dimensional data with the engineered feature you are putting it in the 3 dimension, ok. Now, what we are seeing is interesting stuff that we are seeing that if you look into you know carefully. Just a minute, let me just clean it up a little bit ok. So, now, if you see carefully that these are the estimates ok. These are the estimates of beta naught, this is the estimate of beta naught.

(Refer Slide Time: 03:21)



Model 1 $mpg=\beta_0+\beta_1 wt+\epsilon$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.285      1.878  19.858        C
wt            -5.344      0.559  -9.559        C
```

Model 2 $mpg=\beta_0+\beta_1 wt+\beta_2 hp+\epsilon$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   37.227      1.599  23.285    0.000
wt            -3.878      0.633  -6.129    0.000
hp            -0.032      0.009  -3.519    0.001
```

Model 3 $mpg=\beta_0+\beta_1 hp+\beta_2 hp^2+\epsilon$

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   40.409      2.741  14.744        C
hp            -0.213      0.035  -6.115        C
I(hp^2)        0.000      0.000   4.275        C
```

And, this is the estimate of beta one. So, this is my beta naught, this is my beta 1, this is the standard error. So, estimate divided by standard error will give you the t value. Similarly, minus 5.344 divided by 0.559 will give you negative 9.559, ok. Now, these values are in absolute sense quite large, definitely more than 3 sigma plus minus 3 sigma and you see the p value is very small, it is rounded of to 0 ok.

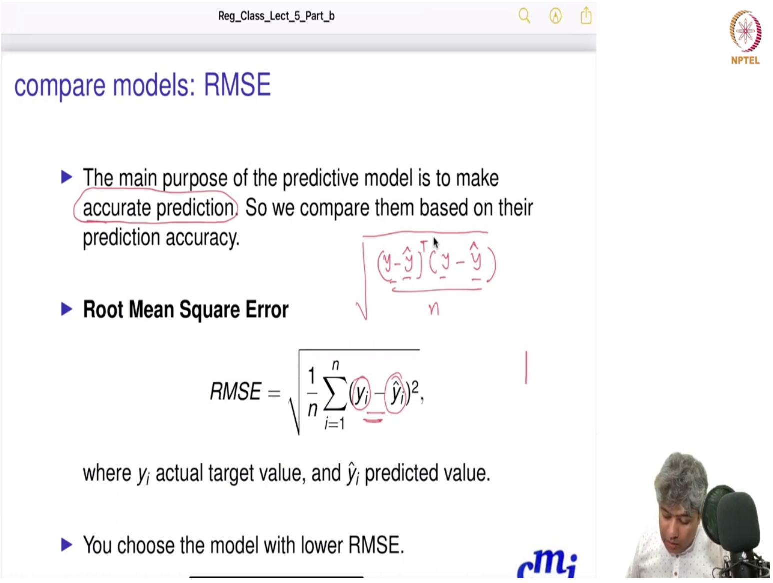
So, that means, this weight has a significant effect on the miles per gallon. Similarly, if you look into this model that it is beta naught, this is beta 1 and this is beta 2. Now, what we are seeing that the estimates these are the estimates, these are the standard deviation and these are the standard deviations and then these are the t values. And, here also we see the p values are extremely small the t values are very in absolute sense quite large plus minus 3 sigma far away from 3 sigma.

So, all these parameters are nonzero for sure we can say that you know these have significant effect. So, horsepower and weight does have a significant effect on miles per gallon and then final third model that what we are seeing is interesting is these are the estimates of beta naught, this is beta 1 and this is beta 2, this one beta 2, this is beta 1, these are the standard error.

And t values are very significant and the Pr values are so small they are on it up to 0; that means, horsepower and the quadratic effect of horsepower also have significant statistically significant effect on the miles per gallon. So, based on these model we can do all these inferences, but if I have to compare the performance of the model which model is a good model in terms of predictive accuracy how do we compare?

So, the first thing is. So, naturally in a predictive model you have to say which model main goal is to come up with a model which will give you a very high predictive accuracy. Now, if you want to come up with a model which will give you a very high predictive accuracy then you have to compare between the models and hence we are saying that how to compare the two models or in fact, more than two we can go for 3, 4, 5 many models.

(Refer Slide Time: 06:23)



compare models: RMSE

- The main purpose of the predictive model is to make accurate prediction. So we compare them based on their prediction accuracy.

$$\sqrt{\frac{(y-\hat{y})^T(y-\hat{y})}{n}}$$

- **Root Mean Square Error**

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where y_i actual target value, and $\hat{y}_i$ predicted value.

- You choose the model with lower RMSE.

So, the next we will come with the concept called root mean square error. The main purpose of predictive model is to make accurate prediction. So, this is the main purpose. You have to make an accurate prediction that is the main purpose and the main purpose of this course is to develop a predictive model which will make accurate prediction.

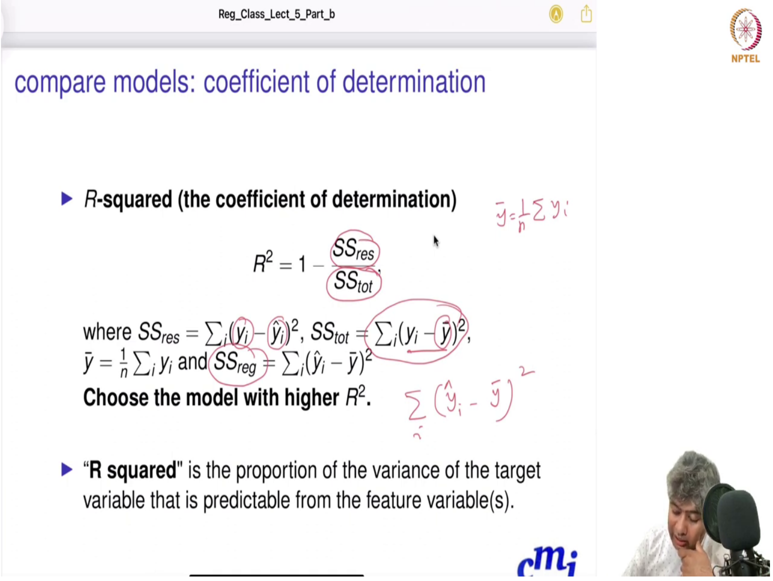
So, we can compare them based on their predictive accuracy. If the main goal is to make a accurate prediction so, we will just compare the two models based on their predictive accuracy. The model which will have a better predictive accuracy that will be better, ok. So, what we do the first thing natural thing that comes in our mind is you take the actual observed and y i hat is the predictive value of y.

What is the difference between them? Square them up take an average of that divide that. So, that is mean square error and then square root of that will give you the root mean square error.

Sometimes if you vectorize it, you can you get to write it in this way ok these are all in vector terms and then divided by n and of course, square root of that.

So, root mean square error is one of the extremely popular measure to you know calculate or estimate the predictive accuracy and then as if between the two model the one which will have a lower root mean square error of course, we will choose that with that error as the root mean square error, ok.

(Refer Slide Time: 08:40)



compare models: coefficient of determination

- R**-squared (the coefficient of determination)**

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}},$$

where $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$, $SS_{tot} = \sum_i (y_i - \bar{y})^2$, $\bar{y} = \frac{1}{n}\sum_i y_i$ and $SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2$

Choose the model with higher R^2**.**

$\bar{y} = \frac{1}{n}\sum y_i$

$\sum_i (\hat{y}_i - \bar{y})^2$

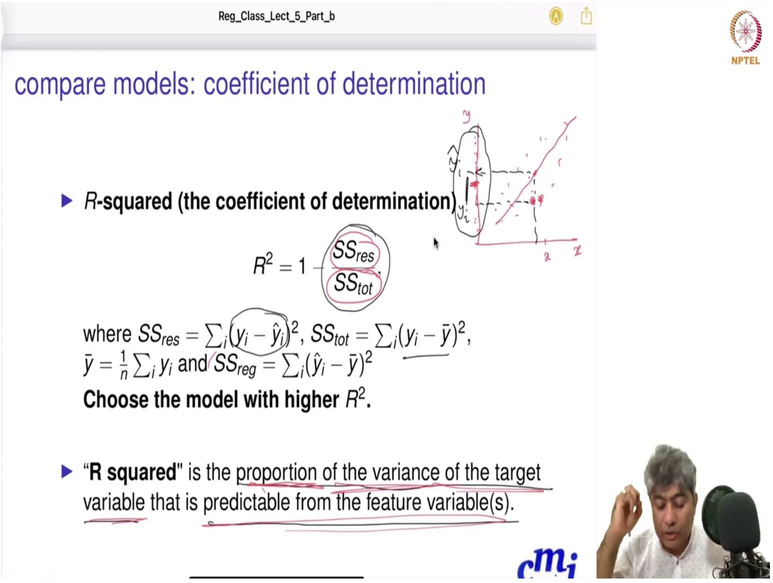
- "**R squared**" is the proportion of the variance of the target variable that is predictable from the feature variable(s).

Next there is another popular measure is called coefficient of determination or R-square. This measure is also extremely popular let me explain you what this measure is. This is basically one sum of squares of residual divided by sum of squares of total what is sum of squares of residual.

So, you have the predicted value and you have the original value. So, original value minus predicted value square and total sum. So, that is sum of squares of residuals and then you have y bar; y bar is just average over all the y i's in your data overall y i. So, if you just take the subtract that from all y's and then square it and take the sum that is sum of total sum of squares. So, this will give you total sum of squares.

And, what is sum of squares of regression? So, basically predicted value minus y bar whole square and then if you just take the sum over that that will give you the regression sum of squares. Now, carefully you see I will try to draw a picture and so that you get a slightly better sense of what is happening here. Let me just give you a sense of what is happening here.

(Refer Slide Time: 10:22)



compare models: coefficient of determination

- *R*-squared (the coefficient of determination)

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where $SS_{res} = \sum_i (y_i - \hat{y}_i)^2$, $SS_{tot} = \sum_i (y_i - \bar{y})^2$,
$\bar{y} = \frac{1}{n}\sum_i y_i$ and $SS_{reg} = \sum_i (\hat{y}_i - \bar{y})^2$
Choose the model with higher R^2.

- "**R squared**" is the proportion of the variance of the target variable that is predictable from the feature variable(s).

So, suppose you have 1 x and 1 y 1 x and 1 y. So, it is a simple you know simple straight line model that you are trying to fit through the points. So, effectively all values on y-axis that is where you see the values. I mean if I just project all these value on the y-axis these are the values, then y bar will be somewhere in the middle, the average of all the all these values.

Now, for a particular x you did a prediction suppose this y was here that is where you have observed the y, but your prediction is here, right. So, naturally let me take a little bit different colorm ok. So, this is my suppose this is the predicted value. So, the predicted value is y i hat and this is the original value ok here you have the original value. So, original value is here y i. So, lexically y i minus y i hat is the residual y i minus y i hat is my residual.

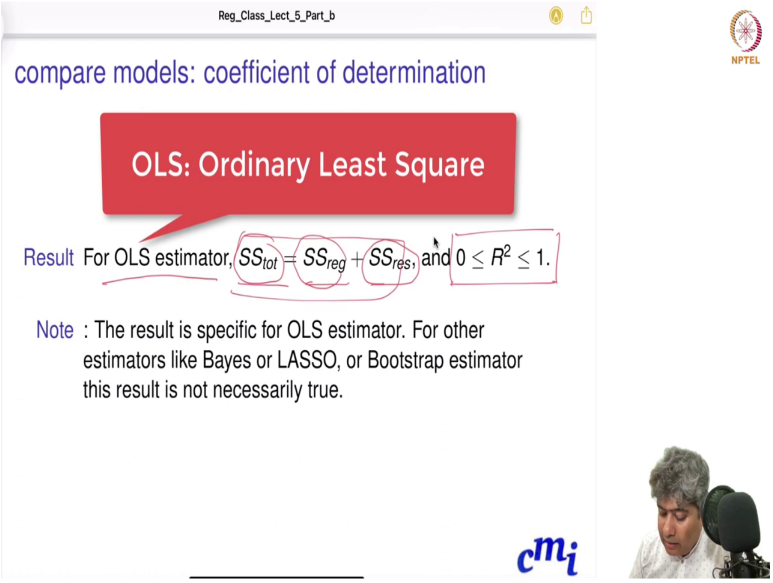
And, how much that means, this residual is the residual that is not being explained by the model right that is the variability that my model cannot explain that is the part of the variability in the entire data that my model cannot explain; whereas, y i minus y bar whole square is the total variability that you have in the response y ok the total variability that you have in the response y.

Now, you are taking the ratio of the two. So, this gives you a sense that how much variability my model cannot explain with respect to the total variability, ok and then if you just subtract that from then that this must be value between 0 and 1 and if you just subtract that from the one that gives you that how much variability of my model can be my model can explain how much variability of the response.

Of course, you higher the R-square better the model and predictive accuracy is. So, of course, the model with the better higher R-square we will choose that model. Now, R-square is a proportion of the variance of the target variable that is predictable from the features feature variables. So, that is extremely important concept.

So, R-square is a proportion of the variance of the target variable that is predictable from the feature variable that is the concept that is that R-square is trying to capture.

(Refer Slide Time: 13:49)



Reg_Class_Lect_5_Part_b

compare models: coefficient of determination

OLS: Ordinary Least Square

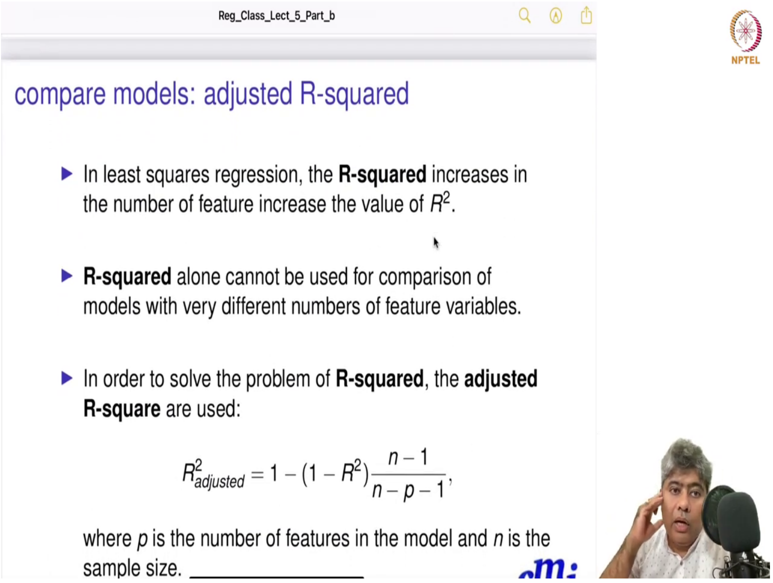
Result For OLS estimator, $SS_{tot} = SS_{reg} + SS_{res}$, and $0 \leq R^2 \leq 1$.

Note : The result is specific for OLS estimator. For other estimators like Bayes or LASSO, or Bootstrap estimator this result is not necessarily true.

Now, I am going to tell you a result without explaining the proof and all that for ordinary least square estimator for ordinary least square estimator, sum of squares of total is equal to sum of squares of regression plus sum of squares of this equals. So, this result is only true for the OLS estimator. This result is specific to only this estimator; for other estimator like Bayes estimator LASSO estimator or bootstrap estimator this result is not necessarily true. We have to be bit careful about that.

Now, naturally if you see this result kind of gives you that z R-square to be between 0 and 1, this result that typically we know in the popular context of analytics that R-square has to be always 0 and 1. Yes, R-square always has to be 0 and 1, but it is true only for OLS estimator, ok and it is not necessarily you may this result may valid for other kind of estimators.

(Refer Slide Time: 15:16)



compare models: adjusted R-squared

- In least squares regression, the **R-squared** increases in the number of feature increase the value of R^2.
- **R-squared** alone cannot be used for comparison of models with very different numbers of feature variables.
- In order to solve the problem of **R-squared**, the **adjusted R-square** are used:

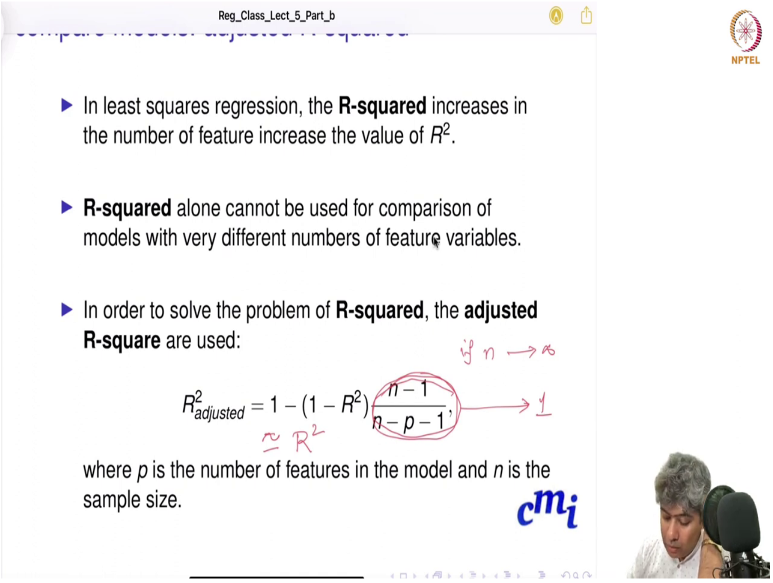
$$R^2_{adjusted} = 1 - (1 - R^2)\frac{n-1}{n-p-1},$$

where p is the number of features in the model and n is the sample size.

Next what we will talk about is something called adjusted R-square. So, the in the least square regression what happens that if the R-square the R-square increases with the number of feature increases in the model. So, in the model if you keep adding the number of independent variables or if you keep adding the feature variables, then what happens the R-square keep increasing. So, R-square cannot alone cannot be used for comparison of the models in with very different number of features.

So, in order to solve this problem adjusted R square was proposed and what we found that adjusted R-square is basically 1 minus 1 minus R-square times this constant that this coefficient kind of thing.

(Refer Slide Time: 16:12)



- In least squares regression, the **R-squared** increases in the number of feature increase the value of R^2.
- **R-squared** alone cannot be used for comparison of models with very different numbers of feature variables.
- In order to solve the problem of **R-squared**, the **adjusted R-square** are used:

$$R^2_{adjusted} = 1 - (1 - R^2)\frac{n-1}{n-p-1}, \xrightarrow{\text{if } n \to \infty} 1$$

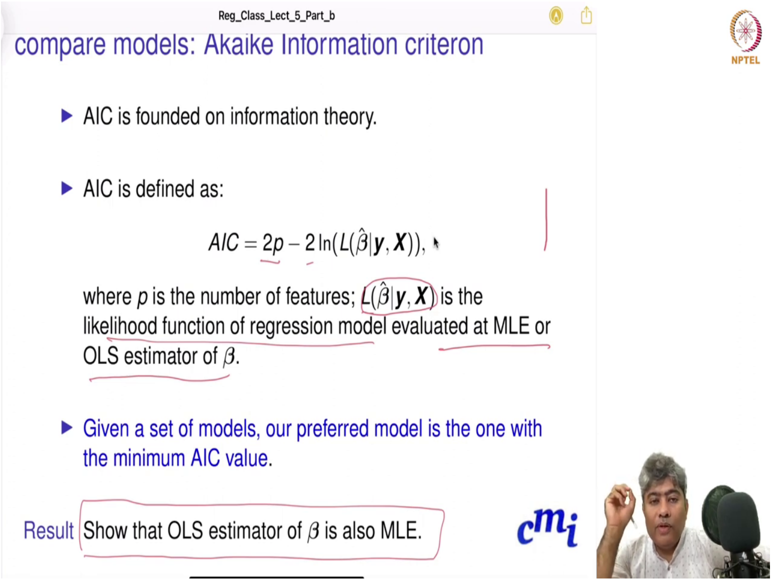
$$\approx R^2$$

where p is the number of features in the model and n is the sample size.

But, if you notice this if n goes to infinity; that means, if you keep increasing the number of sample in yours data in the for very large number for very large samples for a very very big data. This will go to 1 this coefficient this part n minus 1 divided by n minus p minus 1 will go to 1. If this go to 1, then effectively adjusted R-square will be approximately equal to R-square.

So, if for very large number of for very large data set where p is reasonably reasonable size that adjusted R square will be same as the R square. p is the number of feature in the model and n is the sample size.

(Refer Slide Time: 17:03)



compare models: Akaike Information criteron

- AIC is founded on information theory.
- AIC is defined as:

$$AIC = 2p - 2\ln(L(\hat{\beta}|\mathbf{y}, \mathbf{X})),$$

where p is the number of features; $L(\hat{\beta}|\mathbf{y}, \mathbf{X})$ is the likelihood function of regression model evaluated at MLE or OLS estimator of β.

- Given a set of models, our preferred model is the one with the minimum AIC value.

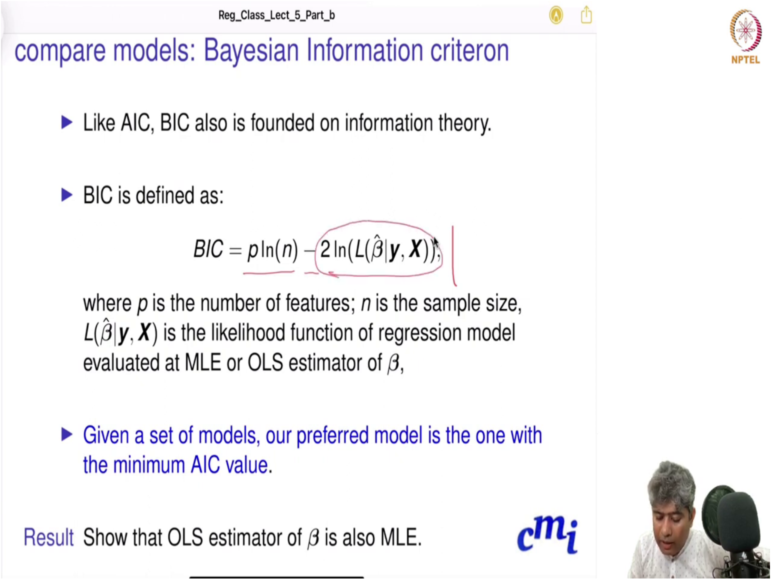
Result Show that OLS estimator of β is also MLE.

Next I will talk about Akaike information criteria. The idea about Akaike information criteria is being founded in the information theory. Now, AIC is defined as 2p minus 2 times log likelihood evaluated at MLE, where p is the number of features and L beta hat y given y and x is the likelihood function of the regression model evaluated at MLE or OLS estimator of beta.

Now, given a set of models our preferred model is the one with the minimum AIC value because we want to minimize the remember that for the set of parameters where the best choice will be when the AIC is minimum even the definition. One result it sort of you know I am giving you as a you know trying by yourself that show that OLS estimator of beta is also maximum likelihood estimator for standard regression model that we have discussed so far.

So, you take if you can take that model and you know OLS estimator of beta and that maximum likelihood estimator of beta turns out to be exactly same. So, can you show that? So, it is a; it is a small theory problem to solve. It is a theoretical exercise.

(Refer Slide Time: 18:51)



compare models: Bayesian Information criteron

- Like AIC, BIC also is founded on information theory.
- BIC is defined as:

$$BIC = p\ln(n) - 2\ln(L(\hat{\beta}|\mathbf{y}, \mathbf{X})),$$

where p is the number of features; n is the sample size, $L(\hat{\beta}|\mathbf{y}, \mathbf{X})$ is the likelihood function of regression model evaluated at MLE or OLS estimator of β,

- Given a set of models, our preferred model is the one with the minimum AIC value.

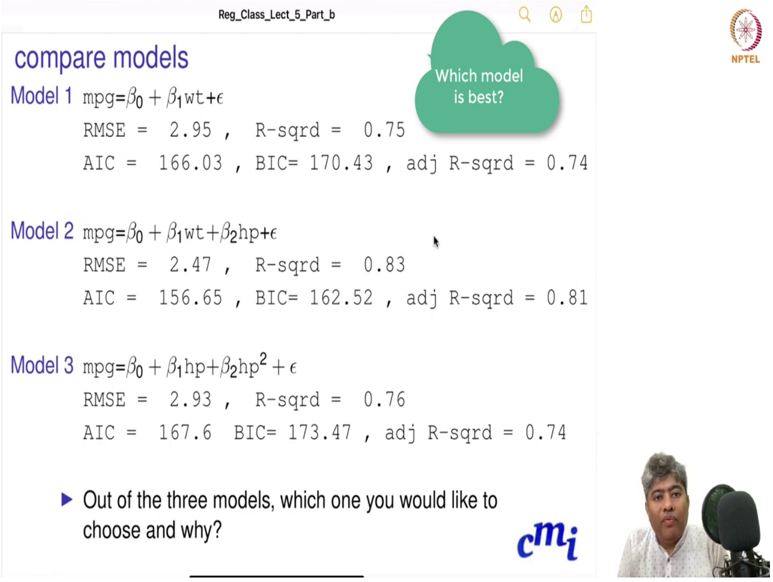
Result Show that OLS estimator of β is also MLE.

Similar to AIC there is another information criteria called Bayesian information criteria like AIC, BIC is also founded in the information theory, ok. So, BIC is defined as p times log n minus 2 times log n of L beta hat given y, X. So, this is same like AIC and in AIC what you had 2 times p, this is typically known as the penalty of AIC and this is p times log n.

So, some people do like BIC over you know over the AIC because Bayesian information criteria sort of make a balance between the number of samples in data set and the number of features in your data set. So, BIC is slightly preferred over the AIC, but you know honestly that does not make much of a difference what my experience is they are technically in a all practical purpose even when you will do data analysis generally they do not violate each other.

So, where is the p is the number features and you know n is the sample size. So, given a set of models our preferred a model will be the one with which with the BIC value minimum BIC value, alright.

(Refer Slide Time: 20:32)



compare models

Which model is best?

Model 1 mpg=$\beta_0 + \beta_1$wt+ϵ

RMSE = 2.95 , R-sqrd = 0.75

AIC = 166.03 , BIC= 170.43 , adj R-sqrd = 0.74

Model 2 mpg=$\beta_0 + \beta_1$wt+$\beta_2$hp+ϵ

RMSE = 2.47 , R-sqrd = 0.83

AIC = 156.65 , BIC= 162.52 , adj R-sqrd = 0.81

Model 3 mpg=$\beta_0 + \beta_1$hp+$\beta_2$hp$^2 + \epsilon$

RMSE = 2.93 , R-sqrd = 0.76

AIC = 167.6 BIC= 173.47 , adj R-sqrd = 0.74

- Out of the three models, which one you would like to choose and why?

Now, I am asking you this question you know pause your video for 5 minutes and think about that which out of these three models which model you would like to choose and why so. Pause your video and maybe you take 5 minutes and think about it that which model would be better and which model will be we would like to choose out of these three models that we have discussed.

I believe now you have found the correct answer. So, let us try to figure out the answer. So, based on RMSE, so, RMSE out of three models if the ones we will choose which has the minimum RMSE. So, based on RMSE if you if the first one has 2.95, the second one has 2.47 and the third one is 2.93. So, out of three models we will prefer the 2nd model ok. So, based on RMSE we would prefer 2nd model.

Now, let us look at the R square. We would prefer the model which will have a R square bit higher than the other models. Now, for the 1st model has 0.75, 2nd model has 0.83. So, 2nd model is better than the 1st model because R square is higher than the 1st model and the 3rd model is 0.76. So, out of these three numbers again R square for the 2nd model is higher, ok.

Now, let us look into the AIC. AIC, minimum the AIC better the model is that is the criteria. So, what we are seeing that the 1st model has AIC value 166.03, the 3rd 2nd model has 156.65 and the 3rd model is 167.6. So, we would prefer the 2nd model again based on the AIC value.

Let us try with the BIC value again if the BIC value is minimum we will choose the model which has the minimum BIC value. So, based on the for three BIC value 170, 162 and 173; obviously, 162 is the minimum. So, we will choose the 2nd model over the 1st and 3rd model.

Now, adjusted R square and R square is similar, so, higher the adjusted R square you will prefer that model and out of these three models 0.74, 0.81 and 0.74, 0.81 is higher. Therefore, we will choose adjusted R square compared to the adjusted R square 0.81 for model 2. So, based on all five metric what we found that model 2 is better compared to model 1 and model 3. So, that is how we generally choose the best model out of the all models that we are considering in our basket of models.

So, in the next lecture, we will try to understand the complexity of the models. So, we will stop here now and see you in the next video.