

Exploring Survey Data on Health Care
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 25
Measures of Variation – r^2 , R^2 , Adjusted R^2 , Pseudo R^2

Welcome friends, once again to the NPTEL MOOC module on exploring health data, we are on the verge of completing the 5th week. Here we are picking off one of the very pertinent topic, which is mostly discussed in research or it is required for writing research papers.

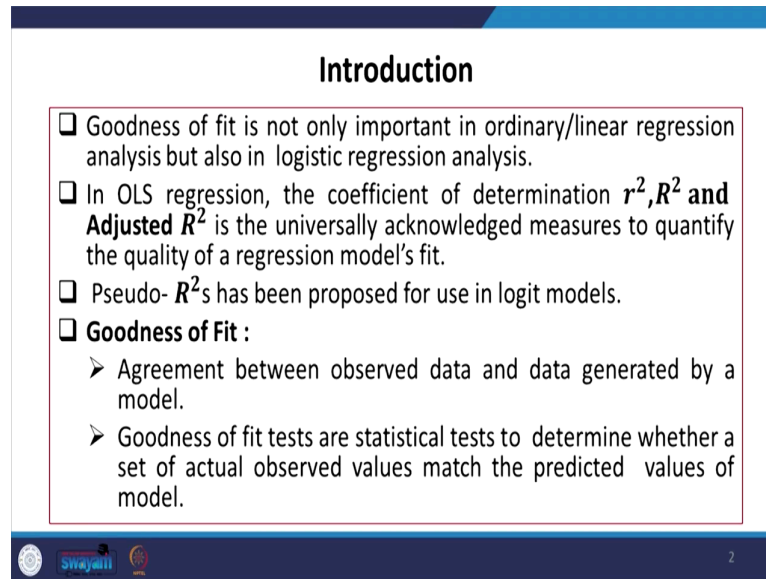
This is one kind post estimation check, we have kept the title of this particular lecture is on measures of variation maybe explained by r square, R square, adjusted r square or pseudo r square.

There are many doubts with researchers, and they usually caught off with this confusion how to deal with each of these issues, which is very essential while writing papers. Reviewers will check some of these indicators very seriously to understand whether your models are perfectly alright or not, in this regard we will clarify further details through our lecture.

So, this is measures of variation is also called goodness of fit. So, goodness of fit is not only important in ordinary or linear regression analysis, but also in logistic regression analysis as well. In the linear regression that is violates the coefficient of determination is basically called small r square capital R square and adjusted r square.

And is the universally acknowledged measures to quantify the quality of a regression models and its fit.

(Refer Slide Time: 02:12)



Introduction

- ❑ Goodness of fit is not only important in ordinary/linear regression analysis but also in logistic regression analysis.
- ❑ In OLS regression, the coefficient of determination r^2 , R^2 and **Adjusted R^2** is the universally acknowledged measures to quantify the quality of a regression model's fit.
- ❑ Pseudo- R^2 s has been proposed for use in logit models.
- ❑ **Goodness of Fit :**
 - Agreement between observed data and data generated by a model.
 - Goodness of fit tests are statistical tests to determine whether a set of actual observed values match the predicted values of model.

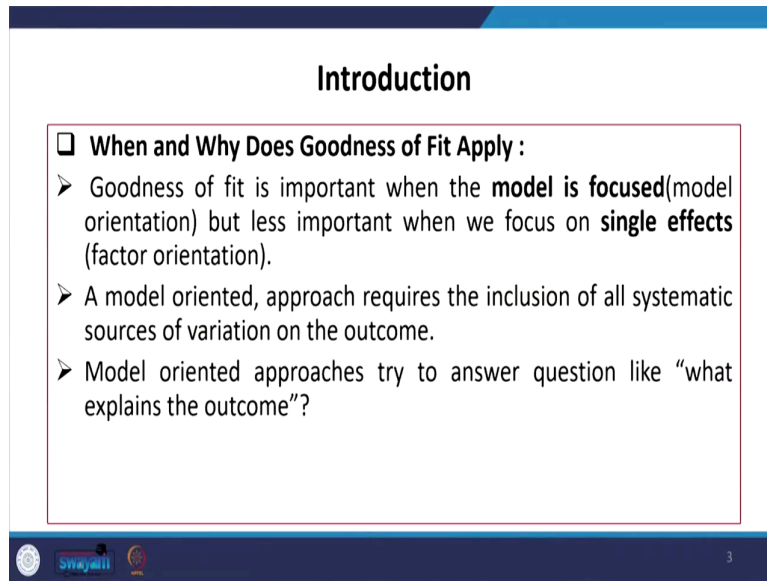
swayam 2

Pseudo r square has been proposed for using the, limited dependent variable model especially logit. So, what do you mean by goodness of fit? This is basically how your model is perfectly fit as per the variables and what is this average trend line.

This explains the agreement between observed data and the data generated by a model. So, the observed data and the data generated by the model any difference between this is actually explained with the help of goodness of fit. Goodness of fit test are statistical test to determine whether a set of actual observed values match the predicted values of the model.

So, the model which is predicted is actually matching with actual observed data or not, if it is not matching then there is certain problem with the goodness of fit. So, a model may be incorrect. So, like goodness of fit as we already said we need to understand when and why does goodness of fit apply.

(Refer Slide Time: 03:22)



The slide is titled "Introduction" and contains a list of points under the heading "When and Why Does Goodness of Fit Apply :". The points discuss the importance of model orientation versus single effects, the need for a model-oriented approach to include all systematic sources of variation, and the goal of model-oriented approaches to answer questions about the outcome.

Introduction

□ When and Why Does Goodness of Fit Apply :

- Goodness of fit is important when the **model is focused**(model orientation) but less important when we focus on **single effects** (factor orientation).
- A model oriented, approach requires the inclusion of all systematic sources of variation on the outcome.
- Model oriented approaches try to answer question like “what explains the outcome”?

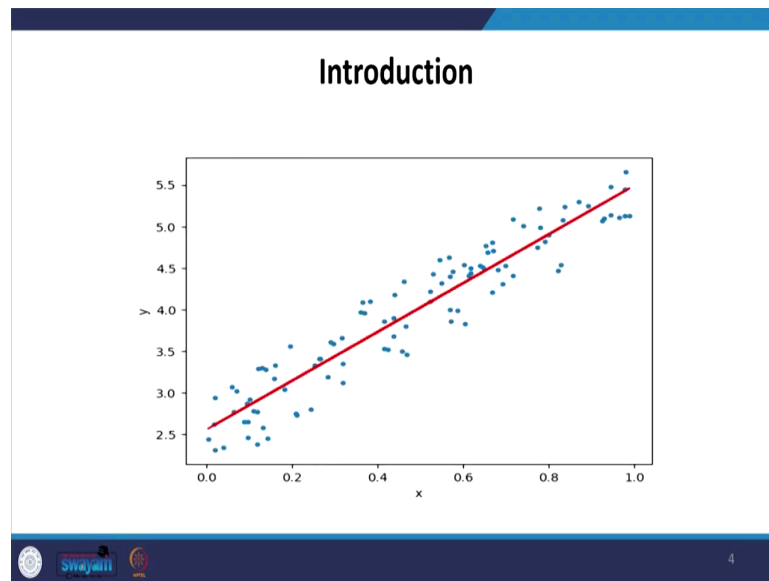
swayamii 3

Goodness of fit is important when the model is focused, or oriented, but less important when we focus on single effects that is factor orientation.

There are two way by which we should understand the orientation. One is called the entire model orientation, another is called single effect orientation. In case of single it is not that important, but in case of the entire model on a whole goodness of fit, is mostly discussed and has to be understood correctly.

A model oriented or approach requires the inclusion of all systematic sources of variation on the outcome. Model oriented approaches try to answer questions like what explains the outcome?

(Refer Slide Time: 04:13)



Now from this chart from this diagram you can easily understand, how our model is actually perfectly fitting to the trend line. So, trend line equation you can simply get it in excel as well as in all the packages.

Now, it seems that there exists a, strong correlation between these two variables. So, if you have different control variables, we wanted to map with its explained; explained variable we can see that it follows a perfect trend line. So, this gives rough idea about goodness of fit as well, but not exactly goodness of fit we are now clarifying how we can able to measure those variation.

(Refer Slide Time: 04:49)

Measures of Variation

- ❑ **Observed Y :**
 - $Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\varepsilon}_i$ ←
- ❑ **Explained portion of Y_i :**
 - $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$ }
- ❑ **Unexplained portion of Y_i**
 - $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$

5

If the variation is very huge; that means, it is not perfectly fit, when it is near about the trend line; that means, the variation is less. We need to understand the statistical significance of those variation and accordingly clarify whether your model is going to be fit or not.

Now, first concept in this case is called observed Y, then explained portion of the individual data that is Y_i then unexplained portion of Y_i . So, these three concepts mathematically should have been clarified. So, Y_i which we have mentioned as observed, that is based on the estimated β values of X and the control variables and it is for the model we have mentioned here. Then explained proportion of Y_i is finally based on the estimated value of Y_i hat that is basically considered with Y_i hat. And unexplained is basically the estimated error, the estimated error explains the unexplained portion of Y_i that is not explained through the model.


If that gap is going to be higher, then actually there are certain problems in the model. So, that is basically the difference between observed Y_i minus the explained portion of Y_i or estimated value of Y_i . So, this is basically the ε and this is what we are going to estimate in short.

(Refer Slide Time: 06:36)

Measure of Variation

☐ Total variation is made up of two parts
SST = SSR + SSE

- SST = total sum of squares $\sum (Y_i - \bar{Y})^2$ ✓
 - Measures the variation of the Y_i values around their mean \bar{Y} .
- SSR = regression sum of square $\sum (\hat{Y}_i - \bar{Y})^2$
 - Explained variation attributable to the linear relationship between x and y.
- SSE = Error sum of square $\sum (Y_i - \hat{Y}_i)^2$ ←
 - Variation attributable to factors other than the linear relationship between x and y.

 6

Then what is this variation in total, how we can clarify the variation? Variation is through three important concepts, one is called total sum of squares SST, this measures the variation of Y_i values around their total mean. You have to clarify these things i.e., \bar{Y} and then \hat{Y} .

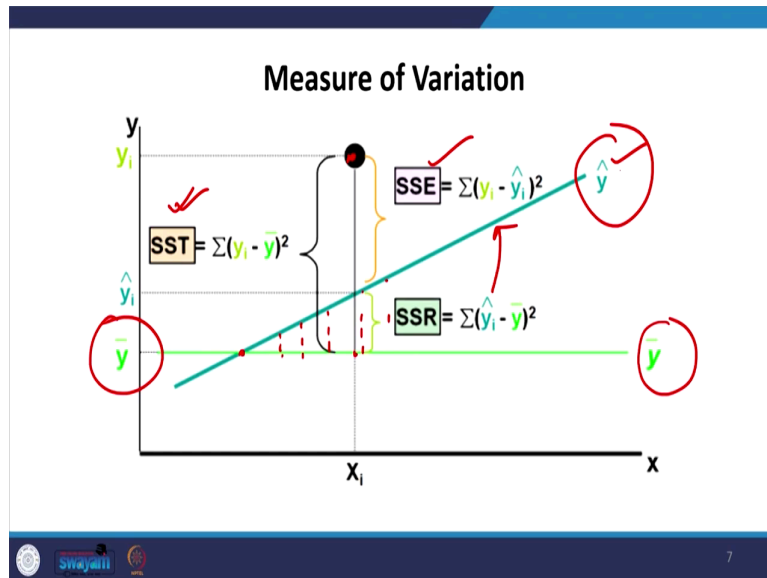
So, this is basically, each individual observed values minus the population average or the total average is basically capturing the difference, this difference is called total sum of squares. Then second one is called regression sum of square which is basically from the mean values for sure, but the estimated Y_i , not the individual Y_i , estimated \hat{Y}_i and its sum of squares. So, minus \bar{Y} if we take its square values. The difference between these two is basically called regression sum of square. Then that is basically called explained variation attributable to the linear relationship between x and y due to the linear relationship between x and y which we derived through the regression analysis if there are any variation that is basically captured through the SSR or regression sum of square.

Then what is SSE, error sum of square? Now each explained variation that is \hat{Y}_i and through the individual observed values, estimated values and observed values if there are certain differences that is basically called error sum of square which we have just explained couple of minutes back.

So, now every time we take the attempt to minimize the square in linear regression model, but here we are saying that if the square, since we are taking in positive values by taking a

square, those squares can be now comparable. So, variation attributable to factors other than the linear relationship between x and y is basically called these differences.

(Refer Slide Time: 09:04)



So, what is this SST, SSE and SSR? Now you can just see this is called total sum of square, what is this total sum of square? It is basically each y_i values minus the y bar, we have mapped it here. So, y bar each value that is from the point here till the y bar point the difference between these two and its square is called sum of squares total.

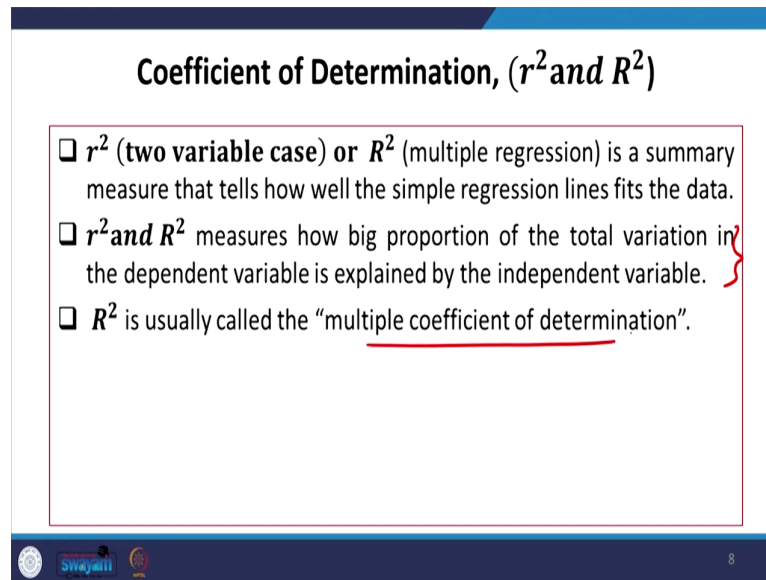
And that is called total sum of square in any model. Now sum of square due to regression is given here y_i hat minus y bar. So, y bar we have already explained as this. So, due to the regression trend line and its difference, whether the trend line is in fact, fitting to the average trend line of the total observed values.

The trend line value or the estimated values is explained through this. This is basically called y hat and the total mean of the average value, this is basically mean of your regression line or mean of your or estimated value through the regression minus the total mean which is basically constant since we are mapping, it is here as y bar.

So, each difference value of y_i , y hat every time the difference is basically called regression some of square. Now comes your error sum of square which we have just mentioned, which is basically the error. Error means what each value that is provided minus the value on the average trend line or the estimated value we have derived.

How each value is different than that of the estimated value? If that difference we can plot, it is basically called sum of square errors. Now how we are going to define the measures of variation?

(Refer Slide Time: 11:38)



Coefficient of Determination, (r^2 and R^2)

- ❑ r^2 (two variable case) or R^2 (multiple regression) is a summary measure that tells how well the simple regression lines fits the data.
- ❑ r^2 and R^2 measures how big proportion of the total variation in the dependent variable is explained by the independent variable.
- ❑ R^2 is usually called the “multiple coefficient of determination”.

We are going to take certain ratios. So, those ratio will explain us about coefficient of variation.

So, now let us clarify one by one, r square and R square. r square is basically two variable case its binary relationship, that is called two variable relationship not multiple relationship.

So, in that case two variables if you are taking, which is simply called r square usually r is taken the correlation coefficient of two variables, but R if you are taking that is called multiple correlation, R is for multiple correlation, r is called correlation between two variables i.e., bi-variate correlation. Similarly, its goodness of fit or its coefficient of variation in this context is written as r square and R square.

So, R square and r square is in fact, a summary measure that tells how well the simple regression line fits the data. r square and R square measures how big proportion of the total variation in the dependent variable is explained by the independent variable. How much its variation is explained by the independent variables. Then R square is usually called the multiple coefficient of determination.

(Refer Slide Time: 13:14)

Coefficient of Determination, r^2

r^2 generally used in single equation model. This model contains only one independent variable.

(a) $r^2 = 0$, [(b),(c),(d),(e) $\in (0,1)$],
(f) $r^2 = 1$

So, next we are clarifying the differences r square generally use in single equation model. This model explains only one independent variable. Now here like in this example you can see some overlapping and there is x and y there is complete differences, there are certain portion of overlapping like these.

Now, r square value in case of the diagram b then b c d and e. So, like within the range of 0 and 1 if there is no correlation at all, I mean explanation by the independent variable in that case it is going to be 0, if it is 100 percent explained by the independent variable then in that case it is called one.

So; that means, the variation is accordingly major, let us explain it further. What do you mean by coefficient of determination as R square? This is usually called multiple coefficient of determination.

(Refer Slide Time: 14:37)

Coefficient of Determination, R^2

- R^2 is usually called the “multiple coefficient of determination”.
- R^2 measure what proportion of the total variation in the dependent variable is explained by the entire model, in other words all independent variables.
- $R^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2} = \frac{SSE}{SST} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum \hat{\epsilon}_i^2}{\sum(Y_i - \bar{Y})^2}$
- $1 \geq R^2 \geq 0$
- Minimizing RSS implies that maximizing R^2 .

10

R square measures what proportion of the total variation in the dependent variable explained by the entire model, in other words all independent variables how they are explaining the total variation.

So, R square is in fact, is measured as the error sum of square, the error that that has been created is due to the total sum of square. How much is the error component out of the total sum of square is basically called R square. And what do you mean by error? Basically error is how much it explains from the model and how much was the residual.

So, TSS minus the RSS is nothing, but called sum of square errors. So, 1 minus RSS divided by TSS is called R square. Now this is RSS, residual sum of square or this is regression sum of square is noted as epsilon estimated divided by the total sum of square. So, we have explained.


Now, question arises here, after estimating these what is the range of this formula or this value? We know that we have taken R square it's not only R could be minus 1, but when we take square of it the simple term converted to positive.

So, in that case its maximum ranges are from 0 to 1. So, minimizing RSS regression sum of square implies that, maximizing the R square; that means, when this component is lesser, 1 minus this one so; that means, R square is going to be higher.

(Refer Slide Time: 16:51)

Coefficient of Determination, R^2

- R^2 therefore lies between **0 and 1**. the closer it is to 1, the better is the fit.
The closer it is to 0, the worse is the fit.
- R^2 is a no decreasing function of the number of explanatory variables or regressors present in the model.
- Number of regressors increases, R^2 almost invariably increases and never decreases.
- Stated differently, an additional X variable will not decrease R^2 .

 11

So, accordingly you can do it. So, R square therefore, is between 0 and 1. The closer it is to 1 the better is the fit. The closer it is to 0 the worse is the fit. So, this is one of the important interpretation. R square is no decreasing function of the number of explanatory variables. Or regressors present in the model, when your regressors are rising it is not decreasing rather R square is fitting to the model'.

It's no decreasing function of the number of explanatory variable when your explanatory variables are rising it's not an decreasing function rather it might be an increasing function.

So, there might be positive latency between rise in the explanatory variables to that of the fit in the model. So, basically, suppose you wanted to estimate any context let it be on health care, you wanted to estimate the impact of food quality on quality of health care on medical expenditure. Food quality food consumption and on medical expenditure. So, it might happen that you have taken very specific variables or very limited variables two or three variables and wanted to feel that your R square values would be very higher. But one suggestion we usually give to our researchers that while your R square is low, actually first attempt you need to fit your model through including the most important variables.

When you increase your variables it is expected that model is going to be better fit. So, there are other couple of suggestions also like R square if it is very low then which kind of data you are looking at whether it is cross sectional or not, whether it is cross sectional panel or not, whether it is time series or not, as we know that when your data is more or less repetitive

in future or almost the same trend line is being followed; that means, your line that is fitted is going to be most closer to the one which you have dealt in the previous round of data.

So, when your data is time series, R square value is expected to be much higher. When it is cross sectionals there are huge variations in the cross sectionals observations therefore, the R square value might be lesser and in between panel data comparative is better value of R square, but even if the data is having represented with less goodness of fit that does not mean your model is bad this is one of the indicators where you may cross check.

But we need to take enough question enough attempt to increase the value of R square this is what we are guiding. So, that your regression model is going to be robust. Number of regresses increases R square almost invariably increases and never decreases this is what we have just guided.

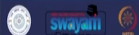
So, number of regresses has to be higher. Because if one of the variable is expanding around 20 percent of the total model if you are including other important variable those may actually cover up the gaps.

Stated differently that any additional X variable will not decrease R square. If you are adding another variable the R square is not going to be decreased.

(Refer Slide Time: 20:43)

Adjusted R^2

- A disadvantage attached to the R^2 measure is, it is an **increasing function of the number of regressors**. It means more the number of regressors, higher the value of R^2 .
- To avoid this problem, another measure is used to assess the goodness of fit of a model, called **Adjusted R^2** .
- Adjusted R^2** explicitly takes into account the number of regressors included in the model.
- Denoted as $\overline{R^2}$ (R-bar square).
- Computed from the R^2 :
$$\overline{R^2} = 1 - (1 - R^2) \frac{n-1}{n-k}$$

12

So, R square we have discussed. Now another important dimension in our model is called adjusted R square. Adjusted R square has to be also discussed, that is basically taken with R square bar.

And this is. In fact, adjusted with the number of parameters in the model, number of regresses in the model. We simply said R square without adjusting its parameter or degrees of freedom, when you adjust or you conditioning upon your constraints in the model and then your R square value is actually better resulted.

So, R square value adjusted could be negative as well we are just coming to it. So, like the disadvantage attached to the R square measure is that, it is an increasing function of the number of regressors. It means more the number of regressor higher the value of R square this is what we have said, but in case of adjusted one when your degrees of freedom are very less of course, you may not able to fit the data to the model.

In that case adjusted R square give the right picture, you are not just simply including more independent variables. Independent variables might be redundant as well those may not explain the model rightly, you are for simply including it your adjusted R square is going to be very less. So, it might be negative also its not necessarily be increasing every time.

To avoid this problem another measure is used to assist this number of regressor issues regarding model fit, is called adjusted R square. Adjusted R square explicitly takes into account the number of regressors included in the model, this is what is in fact, the answer which we have discussed. this is denoted with R bar square and computed as $1 - \frac{R^2}{n - k}$ and its degrees of freedom.

So, basically what is this we are trying to get the value of R square this is nothing, but R square it adjusted with $n - 1$ in the numerator divided by number of parameters that is k . So, the word adjusted here means adjusted for the degrees of freedom which depends on the number of regressions that is called k in the model.

(Refer Slide Time: 23:34)

Adjusted R^2

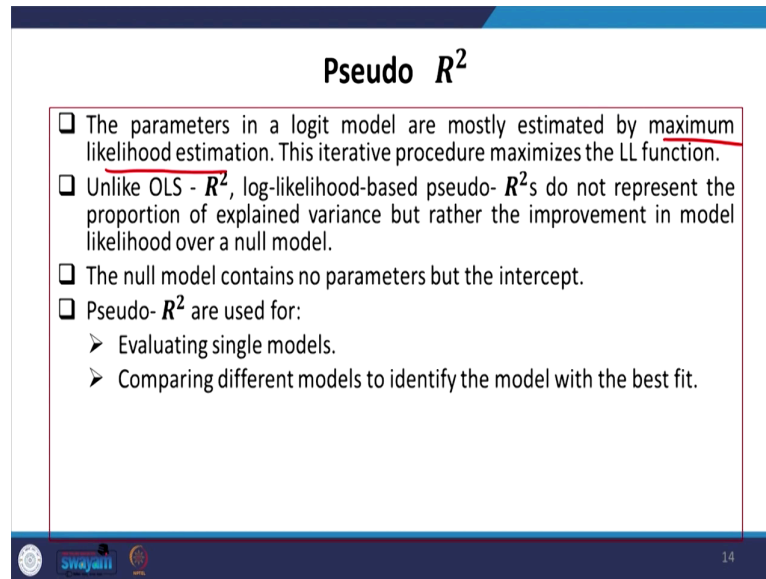
- The word adjusted here means **adjusted for the degrees of freedom** which depends on the number of regressors (k) in the model.
- If $k > 1$, $\bar{R}^2 < R^2$, the number of regressor in the model increases, the \bar{R}^2 becomes increasingly smaller than the R^2 .
- The R^2 is always positive, but the \bar{R}^2 can sometimes be **negative.**
- \bar{R}^2 often used to compare two models that have the same dependent variable.
- Adjusted R^2 is adjusted for the number of independent variables, so it will increase only if the **new variable increases the explanatory power of the model more than would be expected** by the chance.

13

If k is greater than 1 then adjusted R square and its bar that is adjusted R square (\bar{R} square) is less than that of R square this is going to be an important question in almost all analytical context. So, if k is greater than 1 then adjusted R square is going to be lesser. And it is usually the case of course, in multiple regression your k is greater than 1, then the number of regression in the model increases. So, adjusted R square becomes increasingly smaller than that of just R square.

The R square is always positive, but the R adjusted square can become sometimes be negative. So, this is to be highlighted and it is very important in your preparation. So, adjusted R square often used to compare two models that have the same dependent variable. Adjusted r square is useful comparison. So, adjusted R square is adjusted for the number of independent variables. So, it will increase only if the new variable increases the explanatory power of the model more than would be expected by the chance.

(Refer Slide Time: 24:48)



Pseudo R^2

- ❑ The parameters in a logit model are mostly estimated by maximum likelihood estimation. This iterative procedure maximizes the LL function.
- ❑ Unlike OLS - R^2 , log-likelihood-based pseudo- R^2 s do not represent the proportion of explained variance but rather the improvement in model likelihood over a null model.
- ❑ The null model contains no parameters but the intercept.
- ❑ Pseudo- R^2 are used for:
 - Evaluating single models.
 - Comparing different models to identify the model with the best fit.

14

So, next one is called pseudo R square. Pseudo R square is though less often used it is specifically discussed in some context, some of the particular model, it's not called R square it's called scheduled by certain degrees of freedom scheduled by certain goodness of fit.

The parameters in a logit model are mostly estimated by maximum likelihood estimation not by the regression coefficient. It's not a coefficient rather it's through the maximum likelihood estimation. This iterative procedure maximizing the likelihood function unlike OLS base R square the log likelihood-based estimation gives the pseudo R square value. That do not represent the proportion of explained variance, but rather the improvement in the model likelihood over a null model.

So, basically when we are actually discussing certain likelihood of being included in a variable in a model. So, actually each fractional points are not discussed in that particular model, they are not discussing the marginal effect as such. So, each of the represented variance are not actually captured in those kinds of limited dependent models.

So, the log likelihood functions give certain values about R square, but those are called pseudo. The null model contains no parameters, but the intercept. Pseudo R square are used for evaluating single models. Comparing different models to identify the model with the best fit.

(Refer Slide Time: 26:42)

Pseudo R^2

- ❑ Pseudo- R^2 based comparisons of different models using the same sample are straightforward.
- ❑ Comparisons of different models resting on several samples can be reliably processed only as long as the pseudo-R2 used is free of explicit sample effects.
- ❑ More than 80 percent of publications that report a pseudo-R2 do not specify which measure they used (Hoetker 2007).

Pseudo R square based comparisons of different models using the same sample are straightforward. Comparisons of different models resting on several samples can be reliably processed only as long as the pseudo R square used is free of explicit sample effects.

More than 80 percent of publications that report a pseudo R square do not specify which measure they used as per the author we have mentioned.

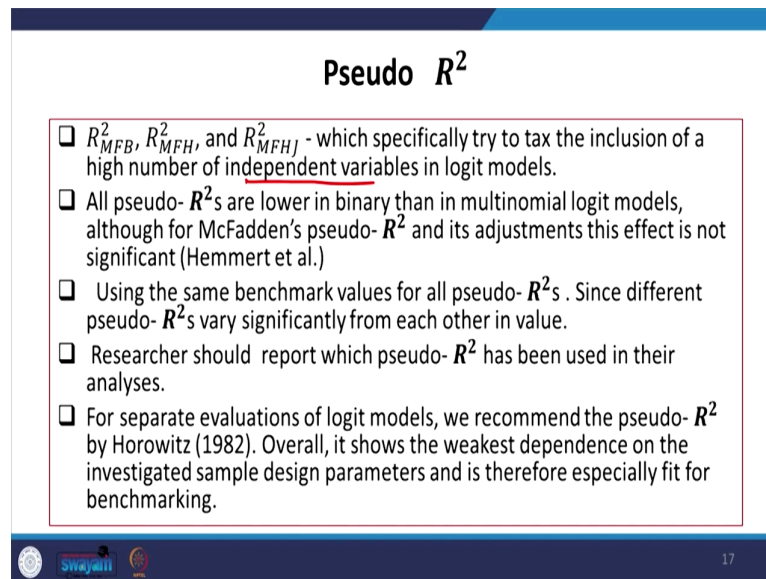
(Refer Slide Time: 27:15)

Different Pseudo R^2

Source	Pseudo- R^2	Range of values
Aldrich and Nelson (1984)	$R_{AN}^2 = \frac{2 \cdot (LL_V - LL_0)}{2 \cdot (LL_V - LL_0) + n}$	$0 \leq R_{AN}^2 < \frac{-2 \cdot LL_0}{n - 2 \cdot LL_0}$
Veall and Zimmermann (1992)	$R_{VZ}^2 = \frac{2 \cdot (LL_V - LL_0)}{2 \cdot (LL_V - LL_0) + n}$	$0 \leq R_{VZ}^2 < 1$
Maddala (1983) and Cox and Snell (1989)	$R_{MS}^2 = 1 - \exp\left(-\frac{2 \cdot (LL_V - LL_0)}{n}\right)$	$0 \leq R_{MS}^2 < 1 - \exp\left(-\frac{2 \cdot LL_0}{n}\right)$
Cragg and Uhler (1970) and Nagelkerke (1991)	$R_{CU}^2 = \frac{1 - \exp\left(-\frac{2 \cdot (LL_V - LL_0)}{n}\right)}{1 - \exp\left(-\frac{2 \cdot LL_0}{n}\right)}$	$0 \leq R_{CU}^2 < 1$
McFadden (1974)	$R_{MF}^2 = 1 - \frac{LL_V}{LL_0}$	$0 \leq R_{MF}^2 < 1$
Ben-Akiva and Lerman (1985)	$R_{MFB}^2 = 1 - \frac{LL_V - m}{LL_0}$	$\frac{m}{LL_0} \leq R_{MFB}^2 < \frac{LL_0 + m}{LL_0}$
Horowitz (1982)	$R_{MFH}^2 = 1 - \frac{LL_V - \frac{m}{2}}{LL_0}$	$\frac{m}{2 \cdot LL_0} \leq R_{MFH}^2 < \frac{2 \cdot LL_0 + m}{2 \cdot LL_0}$
Hensher and Johnson (1981)	$R_{MFHJ}^2 = 1 - \frac{\frac{n \cdot (j-1) - m}{LL_0}}{n \cdot (j-1)}$	$\frac{m}{m - n \cdot (j-1)} \leq R_{MFHJ}^2 < 1$
Estrella (1998)	$R_E^2 = 1 - \left(\frac{LL_V}{LL_0}\right)^{-2 \cdot LL_0}$	$0 \leq R_E^2 < 1$

There are various pseudo R square reported in different article, some of are by default available in Stata, like McFadden this one is by default available. There are other frequently used methods as well we are not going to the depth of it, those who have interest to do further research they can read between the lines and read correctly there are different ranges also defined for their pseudo R square values.

(Refer Slide Time: 27:48)



Pseudo R^2

- ❑ R^2_{MFB} , R^2_{MFH} , and R^2_{MFHJ} - which specifically try to tax the inclusion of a high number of independent variables in logit models.
- ❑ All pseudo- R^2 s are lower in binary than in multinomial logit models, although for McFadden's pseudo- R^2 and its adjustments this effect is not significant (Hemmert et al.)
- ❑ Using the same benchmark values for all pseudo- R^2 s . Since different pseudo- R^2 s vary significantly from each other in value.
- ❑ Researcher should report which pseudo- R^2 has been used in their analyses.
- ❑ For separate evaluations of logit models, we recommend the pseudo- R^2 by Horowitz (1982). Overall, it shows the weakest dependence on the investigated sample design parameters and is therefore especially fit for benchmarking.

17

So, McFadden and others like R square of these MFB, MFH as we have mentioned in this table, and MFHJ which specifically try to tax the inclusion of a high number of independent variables in large model. Also like we discuss about adjusted R square and R square it includes degrees of freedom the number of variables included in the model are also adjusted with the model.

Similarly, here also there are different suggestions given. All pseudo R square are lower in binary than in multinomial logit models. So, if it is binary; that means, we are not actually explaining the other variation due to more variables. So, when there are more multinomial logit models; that means, explanations are actually given by third category in the dependent variable in that case accordingly the value of R square are different.

Although for McFadden pseudo R square and its adjustment the effect is not significant. Using the same base mark values for all pseudo R square since different pseudo R square vary significantly from each other in value same base mark value of pseudo R square also used. Researchers should report which pseudo R square has been used in their analysis.

That is more important for separate evaluations of logit model we recommend the pseudo R square by which overall it shows the weakest dependence on the investigated sample design parameters and is therefore, specifically fit for benchmarking.

So, McFadden pseudo R square we have already said that this is mostly used by the Stata software. McFadden R square and it is the default pseudo R square value reported by Stata package.

This measure is defined as McFadden R square is equal to one minus log of lc divided log null, lc denotes the maximize likelihood value from the current fitted model divided by the null that is basically denoting the corresponding values, but the, but for the null model. The null model with only an intercept and no covariates are discussed.

So, some of the important points were finally, going to comprehend and explain. In time series data it is really easy to get a high R square value because of the trend components in the data, but in cross section data and panel data we usually get low values of R square and this is quite common. Low R square is perfectly acceptable it's not going to be making much problematic.

(Refer Slide Time: 30:55)

Important Point

- In time series data, it's really easy to get a "high R^2 value, because of the trend components in the data.
- In cross-section data and Panel data, low R^2 values are common.
- Low R-square is perfectly acceptable.
- R square just measures the proportion of variations explained by included explanatory variables out of total variations in dependent variables.
- Other model fitness and adequacy measures should be focus such as individual significance of coefficients and overall fitness of the model using t and F statistics, respectively.
- Additionally, look at multicollinearity and heteroscedasticity.

19

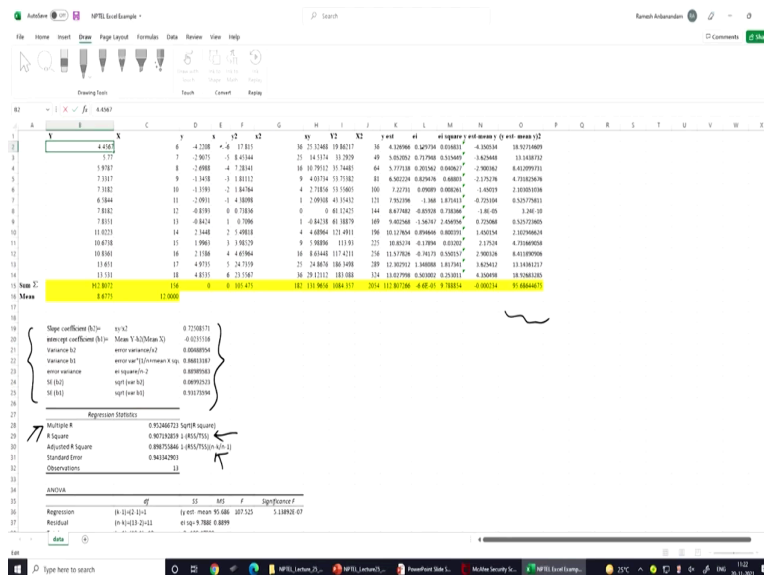
R square just measures the proportion of variation explained by included explanatory variables out of the total variation independent variables.

Other model fitness and adequacy measures should be focused such as individual significance of coefficients and overall fitness of the model using t and f statistics respectively as well. Additionally look at multilocularity and heteroscedasticity to find out the right approach.

So, we can explain you with the help of excel sheet just for your reference I am not going to explain in detail since we have already covered the time. I am just going to show it through our excel sheet example if you just wanted to calculate in your survey data as well you can follow this approach.

The formula should be entered very correctly and you can able to get the result rightly. And here in your screen the works which we wanted to show it to you.

(Refer Slide Time: 31:59)



We have y values and x value. Now what we will do, we will find out the estimated value y bar we have taken some of the these values are derived.

So, like here you can just have a check, you can just click and you will find out the y values and x values accordingly. Then we have taken their x square then y square. Likewise, it is required in our estimation then between that these coefficient x and y, then Y and X that those are the observed variables.

Estimated variables we will also require. So, Y square and X square is also calculated y estimated and the error we can also estimate and once with the error is estimated basically y

minus \hat{y} is not it. So, once that is done, you click on each of the component you will find the commands. And accordingly its square will be taken that is basically called e_i square.

So, y estimated value minus in the n column y estimated value minus the mean value of y this is what we have already discussed in the diagram and we will take the square term at the last square here. So, finally, the slope coefficient is defined as $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$ this is what at the bottom we discuss everything, intercept coefficient that is mean of that y minus β value times its mean value of x .

And variance of these coefficient can be also derived, error variance then variance divided by its $1 + n(\bar{x} - \bar{x})^2$ and its error variance adjusted with the degrees of freedom. Then its square root of each of those variance is derived after doing these minimum estimations we require for our calculation.

So, mean multiple R square we can calculate like this, and then the R square value that is basically $1 - \frac{RSS}{TSS}$, this is calculated here total sum of square, then adjusted R square is also calculated with the help of its degrees of freedom, we have already all defined its degrees of freedom correctly and number of observation in this case is 13.

So, accordingly you can able to calculate. Rest of the details once we click on this particular table all its commands we have said and that will be useful for your calculation correctly.

So, I am not going to much detail of it since we have already crossed the time for this particular lecture. So, that next time it will be easier for you to clarify. So, this is what we have done it. So, finally, we have shown one excel sheet to you and that excel sheet is going to be very useful those who wanted to work for a small scale of data, but in large scale of data you have to go give these commands and there are readily available commands as well to calculate the R square values.

So, with this I think I need not explain much we will look forward for your participation in the next class.

Thank you.