Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-27
Classical Normal Linear Regression Model

Hello friends, welcome back to the lecture series on applied statistics, and econometrics. So, in this lecture, we are going to continue our discussion with multiple linear regression models. So, before we visit the models and analysis, let us have a look at today's agenda items.

(Refer Slide Time: 00:33)

Today's Agenda:

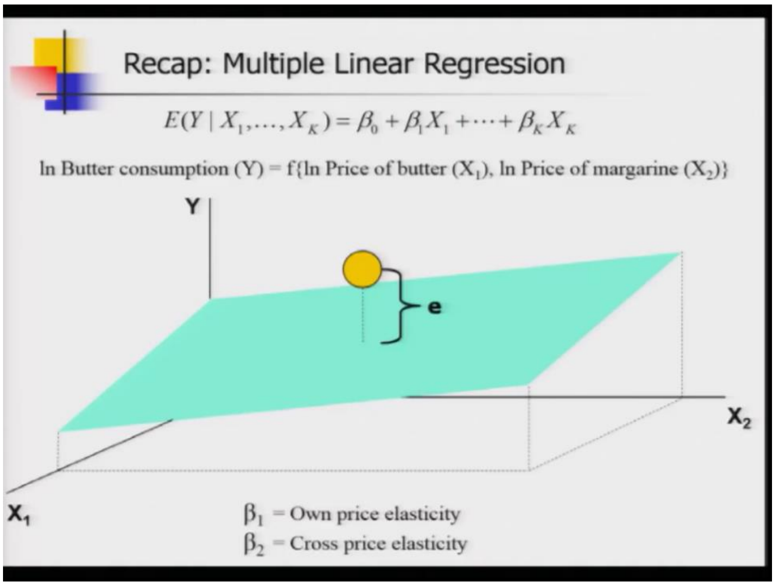
- CNLRM in matrix notation
- Gauss-Markov Theorem
- Goodness of fit measures
- Outlier detection

So, in this lecture, we are going to extend the classical linear regression model that we have seen in the previous lecture by adding one more assumption. And we are also now going to express our model in metrics notation, because we are going to involve more than 3 explanatory variables.

So, the next item in the list is a very important theoretical result known as Gauss-Markov Theorem, and then, we are going to revisit the goodness of fit measures. We have already done goodness of fit measures discussion, but we are going to extend that analysis here. And finally, we are going to talk about outliers and influential observation detection.

So, if you remember a linear regression model has one simple objective, and that is to find out the statistical dependence between one dependent variable or a response variable and a set of independent variables or regressors, and the relationship has to be linear in nature.

(Refer Slide Time: 01:41)



Recap: Multiple Linear Regression

$$E(Y \mid X_1, \ldots, X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$$

ln Butter consumption (Y) = f{ln Price of butter (X_1), ln Price of margarine (X_2)}

β_1 = Own price elasticity

β_2 = Cross price elasticity

So here in this slide, I am going to provide you a brief recap of multiple linear regression. So, we are going to start the slide with the concept of population regression function, and that is given by this expectation formula expected value of Y, given the variables X 1 to X k, and that as humans are linear functional form beta naught, the interceptor plus beta 1 times X 1 dot, dot, dot beta k times X k.

Now, it is very difficult to look at the multiple linear regression model because it is a k dimensional problem. So, if I want to now tell you the multiple regression story in a nutshell, let us reduce the dimension of the problem, and let us work with only to explanatory 20 variables.

So, here let us take an example from microeconomic theory. So, let us focus on butter consumption for the matter and that is our Y variable. Well, we take logarithm of butter consumption data, so that is basically my Y variable. And then I assume that my butter consumption depends on the log of price of butter, which is denoted by X 1 variable and log of price of margarine, and that is my X 2 variable. So, my linear regression model will be Y I equals two beta naught plus beta 1 X 1, I plus beta two X2 Y where all the variables Y X 1 and X 2 are in logarithm form.

So, now suppose, I got a data set, and I estimated the regression coefficients, and then I can now draw the fitted lines. Note that in this case, as we are dealing with more than one explanatory variable actually, we get a surface or a plane and that is drawn in this three-dimensional diagram where Y actually is the vertical axis, and that is measured in this log of butter consumption and then I have two X axis, one for the log of price of butter and the other one is for log of price of margarine.

Now, you see here, what I can infer from the diagram is the statistical dependence between these three variables. So, let us now first talk about the variable Y and X 1. So that is the log of butter consumption and the log of price of butter. So here, if you look at the slide, then you will see.

if I take a cross section of this diagram in the Y, X 1 plane then there is a downer slopping straight line. And that is because here I am observing the demand function for the butter. So as price of the commodity increases, quantity demanded, shall decrease, that is coming from the law of demand. And that is why I am observing downward slopping straight lane here.

But now if I take a cross section of this surface or plane, and if I now look at a two-dimensional plot where I am measuring Y along the vertical axis and X 2 as the horizontal axis, then I see that there is a positive relationship between these two variables X 2 and Y. Why is this so? You do not have to assume microeconomic theory here, think from the logical perspective of a consumer. So here margarine and butter, these are substitute commodities. So, if price of one commodity increases, then of course the commodity demand for the other good will increase.

So, if price of margarine is going to increase then you can expect that demand for butter will go up, and that is why the variables, Y and X 2, they are showing a positive relationship in the Y, X 2 plane. Now, regression is not only about looking at the slope and noting down the nature of the relationship between two variables.

Regression is also about quantifying the strength of the relationship and also after getting the direction of the behavior of two variables. The measurement that, if there is a change in one variable by one unit then how the other variable is going to change by how many units it is going to change. So here, as we are dealing here with a log, log model here, the elasticities have very nice interpretation, and the slope coefficients beta 1 and beta 2 they can be called the elasticity.

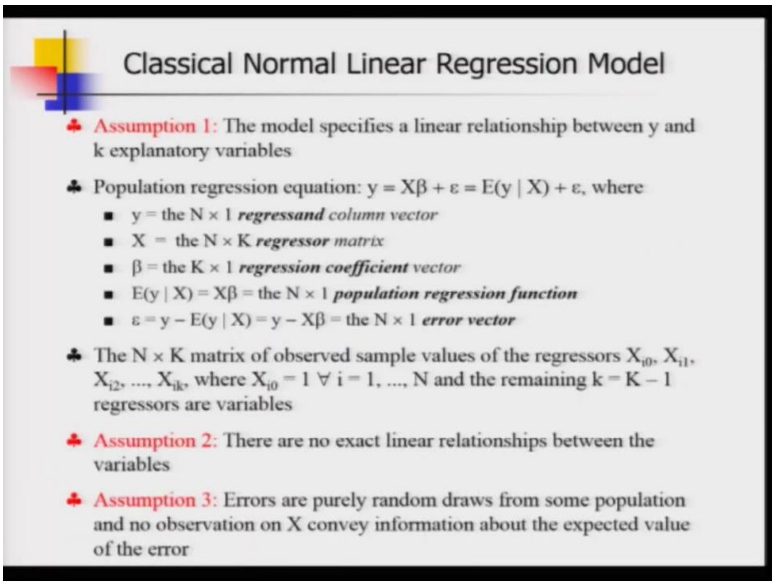
And here, if you look at the diagram again, the beta 1 actually gives me own price elasticity. So that means that if there is 1 percent change in the price of butter how many percentage of changing in butter consumption I am going to expect. And beta 2 gives me a measure of cross-price elasticity, so it tells me that if there is a 1 percent change in price of the margarine, then by what percent my butter consumption is expected to change.

And note that this cross-price elasticity theoretically can two types of values positive and negative. So here, as we are dealing with the substitute commodity like margarine, we are going to expect a positive number for beta 2. So, before we end the discussion on these brief recap on multiple linear regression models let me take you back to the slide once more, as I want to mention a very interesting point.

So, now let us consider this diagram again, and here you see that actually I do not have a straight line in two-dimensional plane, as my regression equation. Actually, I get what can be called a regression surface, and that is colored in green. So, all my fitted values from the regression equation, we lie on this regression surface. But note that the original observations on X 1, X 2, and Y are such that, that the observed bundle X 1, X 2 and Y that point may not lie on that surface.

So, that kind of a point is denoted by this yellow circle here. So, how do I measure the fitted residual value or the estimate of the error from my regression equation. So, I need to now calculate the distance of this white color ball or circle from the point that actually I observe for my values of X 1 and X 2 on this particular surface. So, this distance is given by this broken line.

(Refer Slide Time: 09:09)



Classical Normal Linear Regression Model

- Assumption 1: The model specifies a linear relationship between y and k explanatory variables
- Population regression equation: $y = X\beta + \varepsilon = E(y \mid X) + \varepsilon$, where
  - y = the $N \times 1$ ***regressand*** *column vector*
  - X = the $N \times K$ ***regressor*** *matrix*
  - β = the $K \times 1$ ***regression coefficient*** *vector*
  - $E(y \mid X) = X\beta$ = the $N \times 1$ ***population regression function***
  - $\varepsilon = y - E(y \mid X) = y - X\beta$ = the $N \times 1$ ***error vector***
- The $N \times K$ matrix of observed sample values of the regressors X_{i0}, X_{i1}, X_{i2}, ..., X_{ik}, where $X_{i0} = 1 \; \forall \; i = 1, \ldots, N$ and the remaining $k = K - 1$ regressors are variables
- Assumption 2: There are no exact linear relationships between the variables
- Assumption 3: Errors are purely random draws from some population and no observation on X convey information about the expected value of the error

So now, we are going to revisit that classical linear regression model the theory part of it, and we are going to augment that theoretical discussion with one additional and that is the assumption of normality. But before I introduce the assumption of normality in this theoretical setup, let me again, remind you that we start with the population regression function and then we get the stochastic version of it and then basically we have to estimate the unknown values of population parameter. So that is basically the crux of the classical linear regression model.

So, we start with the formal assumption number 1 which says that a linear regression model specifies a linear relationship between Y and K number of explanatory variables. And then the population regression equation now can be written in matrix notation in much compact form and that is written as Y equals two X times beta plus epsilon and that can be further written as expectation of Y given X plus epsilon. And now, I am going to explain you what are these different components that you see in this equation.

So, let us look at the right-hand side variable, so that is basically my Y. Now, the last time when I wrote the linear regression equation, I was mentioning individual values. And that is why I had this sub script i after the dependent variables nomenclature Y, but now I have only Y, and that is basically giving me a column vector of the regression.

So, the dimension of this column vector is capital N times 1. So, there are in number of observations in this column vector so N is basically my sample size. Now, let us focus on capital X. So, capital X is N cross k regressor matrix. So, what do we mean by here? So, capital K is basically small k plus 1.

Why small k, of course, we are assuming there are k number of explanatory variables k is an arbitrary number and it will change from one example to the other one case to the other. And Y we have to add 1 to it because we have to have the intercept term. So, for the intercept, actually we have to now add one column in this data matrix, and this column will have all ones for these N number of observations. So that is the way capital K will be well to small k plus 1.

So, the next in the least ease the regression coefficient vector, and that is the beta vector. So that is capital K cross one column vector again, and then we have the expected value of Y given X and that is, capital X times beta. So that is capital in cross 1 population regression function.

And finally, epsilon is basically the difference between the observed value vector Y. So, epsilon is basically the difference between the observed value Y and the expected value of Y given X. And that is also again, in no in cross 1 vector that is also called error vector. So, now note that this in-cross k metric is also called a data metrics, and this is the metrics of observed sample values of the regressors X i naught, Xi 1 do, do, do Xi k. So Xi naught is basically that column, which will have values 1 for all observations, so this is inserted in the data matrix to take care of the intercept term.

And the remaining a small k number of regressors are basically the explanatory variables. Now, if you remember the last discussion you can have squared terms or cube terms of existing explanatory variables in the same data metrics.

So, next we focus on the assumption 2, and that is simple. It says, there are no exactly near relationships between the variables. So, it rules out the case of strong multicollinearity. So, if you remember I discussed the case of multicollinearity previously. And that says that two know two explanatory variables should be perfectly linearly correlated, so the product moment, Pearson correlation coefficient between two explanatory variables should not be exactly equal to 1.

Actually, it should not be greater than 0.9, because then you are going to see a very strong linear relationship between two explanatory variables later. Later part of the course, we are going to revisit this idea or concept of multicollinearity and we are also going to talk about some remedies.

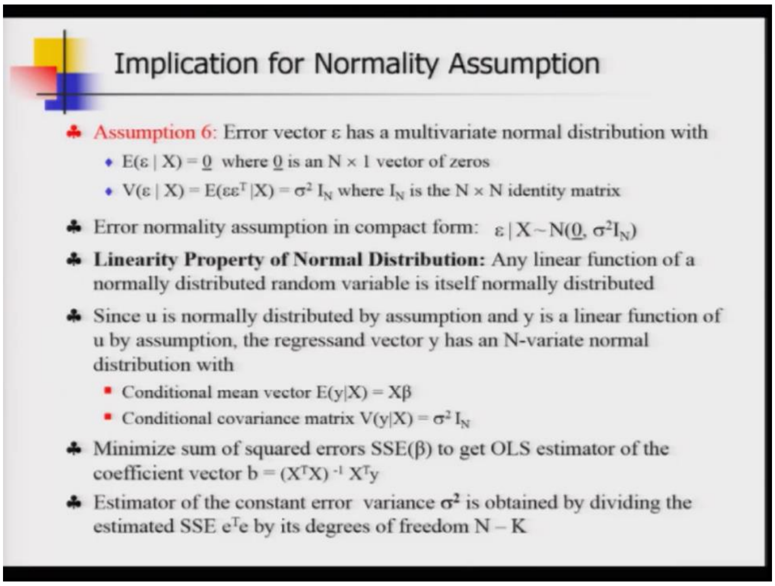
Now, we move on to assumption 3, which says that errors are purely random draws from some population and no observation on X convey information about the expected value of error. So, what does that mean? So, that says that your expected value of Epsilon given X is equal to 0, and this is popularly called the exogeneity assumption, so that means that there is no covariance or correlation between the error variable Epsilon, and the explanatory variable. Anyone of the explanatory variables that you have in the data metrics.

So, now before we start talking about the normality assumption let me remind you about assumption 4 and 5. I have not written assumptions 4 and 5, in this slide because, I believe that you remember them from the previous discussion.

Can you tell me what are they? So, they are actually the assumptions of homoscedasticity, and zero serial correlation. So, what is homoscedasticity, let me remind you briefly. So here, we assume that the variance of error term will not change with respect to the X values. So it is constant over the data set that you have. So, the variance of Epsilon is equal to sigma squared. You can say in the language term.

So, in assumption 5, we know what do we mean. In mathematical statistics term we say that covariance between Epsilon i and Epsilon j is equal to zero when i is not equal to j. So that means that if you will take any two values of the error terms they are not correlated with each other.

(Refer Slide Time: 16:33)



Implication for Normality Assumption

- Assumption 6: Error vector ε has a multivariate normal distribution with
  - $E(\varepsilon \mid X) = \underline{0}$ where $\underline{0}$ is an $N \times 1$ vector of zeros
  - $V(\varepsilon \mid X) = E(\varepsilon\varepsilon^T \mid X) = \sigma^2 I_N$ where I_N is the $N \times N$ identity matrix
- Error normality assumption in compact form: $\varepsilon \mid X \sim N(\underline{0}, \sigma^2 I_N)$
- **Linearity Property of Normal Distribution:** Any linear function of a normally distributed random variable is itself normally distributed
- Since u is normally distributed by assumption and y is a linear function of u by assumption, the regressand vector y has an N-variate normal distribution with
  - Conditional mean vector $E(y|X) = X\beta$
  - Conditional covariance matrix $V(y|X) = \sigma^2 I_N$
- Minimize sum of squared errors $SSE(\beta)$ to get OLS estimator of the coefficient vector $b = (X^TX)^{-1} X^Ty$
- Estimator of the constant error variance σ^2 is obtained by dividing the estimated SSE e^Te by its degrees of freedom $N - K$

Finally, we start with the assumption of normality. And let me tell you one thing, this assumption, number 6, which is talking about the normality is not required to conduct a classical linear regression model.

So, if you remember two lectures back only, we have derived the OLS estimators and for that matter we did not assume normality anywhere. And if you also remember in the first module of the course when we did car fitting by least squares principle there also we did not assume normality.

So, why do we have to now introduce another assumption on a normality? You will see later on, we are going to conduct hypothesis testing and for hypothesis testing we need to assume certain distributional assumptions, and that is why we are bringing normally in the picture, there is also another reason. If you remember we also spoke about another estimation method, which is called the maximum likelihood estimation method.

And for that, you have to make some distributional assumption as well. So, we are here seeing the OLS principle how it is applied to find the estimates of the regression coefficients of a classical linear regression model, but one can also adopt the maximum likelihood estimation method MLE.

And for that you have to assume some distributional form for the errors. And that is why normality assumptions comes handy and people make normality assumptions in the context of classical linear regression model.

So, here we say that my error vector Epsilon has a multivariate normal distribution. Why it is a multivariate normal distribution because here you are dealing with N number of observations. So basically, in that error vector you have N items, Epsilon 1 to Epsilon N. Now, I am measuring that each one of these vector components like Epsilon 1, Epsilon 2, Epsilon 3 do, do, dot N they are all random variables. So basically, there are N number of random variables. So, if you bring in the concept of normality then you have to bring in the concept of multivariate normality.

So here I, as you, that Epsilon follows a multivariate normal distribution with mean expected value of Epsilon given X the data matrix is equal to a vector of 0. Why it is a vector of 0? Because here I am not talking about one particular random variable, I am talking about N number of random variables together. So, this 0 underscored is the N cross 1 vector of zeroes.

And this multivariate normal distribution will have a variance which is given by expected value of Epsilon, Epsilon, transpose or Epsilon, Epsilon prime given the data metrics X, and that will finally take the form of Sigma square times. IN, where sigma squared is the constant variance, and IN is basically N cross N identity matrix.

So, if you remember your metrics algebra lessons from school or college, then you can actually, have a picture in front of your eyes now. So IN is basically identity matrix where the diagonal elements are all ones and the off diagonal remains are all 0. So basically, you just have to multiply a constant sigma square in front of that IN matrix and that will give you the variance metrics for the Epsilon.

So, in a compact form we can see that error normality assumption results in this expression, Epsilon given data matrix X follows normal with mean vector of 0, and then sigma squared IN covariance.

So, this sigma square IN is basically a variance covariance matrix where the covariance items are all 0 because of the assumption of zero serial correlation. So, only the variance stamps are going to stay in that variance covariance metric. But actually, in general this variance expression that we have written in terms of V of Epsilon given X that is basically a variance covariance metrics, because later on in the course, we will relax these assumptions and then we will see that these off diagonal elements can be replaced by certain statistical measures. And these are all covariance measures and wait for some time to see that complicated picture.

Now, what is the implication of this normality assumption, if we introduce it in the classical linear regression model. Well, if you remember when we discussed the case of normal distribution, we spoke about the linearity property of normal distributions. So, let me remind you again, if you have forgotten.

So, any linear function of a normally distributed random variable is itself normally distributed. So, if I take simple variable X, which follows normal distribution and if I now take transformed variable Y equals two alpha plus beta times X where Y is another continuous variable. So here, Y will also follow the normal distribution that is what basically linear T property of normal distribution says.

So, here in the classical linear regression model context, as U is normally distributed by assumption and Y is a linear function of U by assumption the regression vector Y has an invidious normal distribution and this normal distribution or multivariate normal distribution will have a conditional mean vector, and that is given very expected value of Y given X equal to X times beta and conditional co-variants, variance, covariance matrix given by V Y given, X the data matrix and that is equal to sigma square times the identity metrics of Nth order.

So, then you know what to do? If you remember next step would be to apply the least squares principle and get the estimates of this slope coefficient and the intercept coefficients. And here, as you have K number of explanatory variables it is indeed very difficult to even apply Crammer’s rule.

Crammer’s rule will save you up to say three explanatory variables, but not more than that. After that it is going to be extremely clumsy. So, what is the remedy? And you have no other remedy, but to apply matrix algebra. So, if you now apply matrix algebra the process of taking derivatives to minimize the error sum of squares is extremely messy and it is too theoretical so I am going to skip all these steps, but I am going to show you the final outcome, because if you remember then that will become very handy. So, here I am going to show you the final expression that comes out of the sum of square error residuals minimization exercise, and that coefficient vector B is given by these matrix formula X transpose, X inverse X transpose Y.

So, here this capital T is denoting transports, but not that in many textbooks you can see the use of our dash to denote transpose. I am not going to show you any example here to help you with matrix transpose and inverse and all, because I have set the prerequisite for the course, and I am assuming that you have some knowledge of matrix algebra at the high school level. So, there you should have done matrix inversion and taking transports of a matrix and all. So, I equate you to consult some of the previous study materials on matrix algebra.

So, once the co-efficient vector B is estimated, are we done? The answer is no, we are not done. Why? Because when you are writing your classical linear regression model, then there is another population parameter that you have written. Many times, see that students forget about this the population parameter when I give them a data set and ask them to compute the values of the coefficients and the unknown population parameters, they think that the unknown population parameters are only the coefficients that are there in the regression equation, but how can you forget that Epsilon, which actually follows some distribution.

I mean, here in this lecture, I have introduced the concept of normal distribution that is an assumption, but you do not have to assume normality. It can follow any arbitrary distribution with zero mean, and constant variance, but that unknown constant variance is also a population parameter. So, you have to get a proxy or an estimate for that as well. So, when I am seeing that, you have to get the estimates of your classical linear regression model or classical normal linear regression model do not forget to get the value of the sigma square hat or the estimate for sigma squared the unknown population error variance.

So here, the estimator of the constant error variance sigma square can be obtained by dividing the estimated SSE the sum of square errors and that is given by transpose E by 8 degrees of freedom capital N minus K.

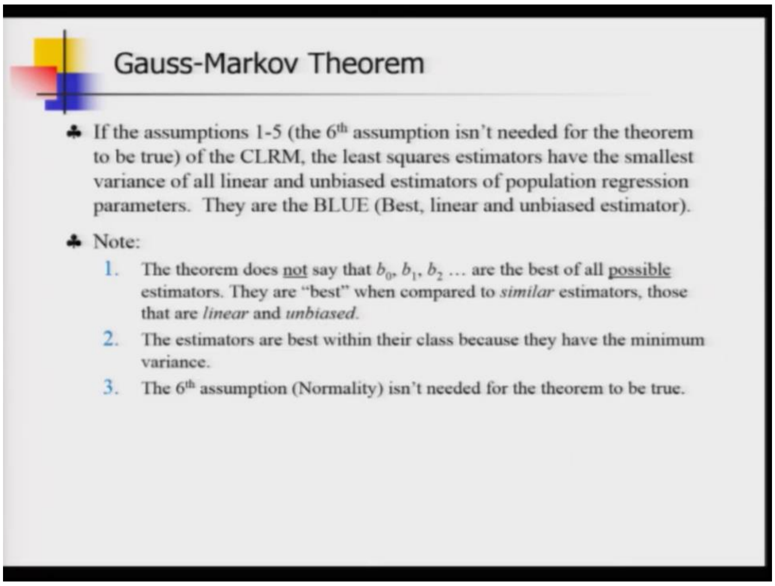
Now, what do I mean by this expression E transpose E. So, here E is basically the sample residual. So, epsilon is basically an error which is a population concept. And when you get a data set, you apply OLS least squares principles or maximum likelihood whatever estimation principle to the data set and you derive the values of these slope coefficients and the intercept coefficients.

And then you generate the residuals from the regulation, which is basically E equals two Y minus Y hat. So, Y hat is basically the fitted value that comes from the regression equation that you have just estimated. So, that E now you can make use of to get the estimator for the unknown population error variance. And you have to basically take that residual square it and then sum that squared residual for all N number of observations and that is basically, this E transpose E is giving you in the matrix form.

And this is basically just a scalar number. And you have to divide this by its degrees of freedom, capital N minus K. Why we have to divide by the degrees of freedom and that is N minibus K. Note that you start with N number of observations in the data set, and then you apply your OLS technique to it.

And how many unknown parameters you are estimating from the model before you actually, tried to find an estimator for these unknown population in variance. You have to get K number of explanatory variables, correspondingly, slope coefficients and then there is also an intercept coefficient. So, small k plus 1 equal two K. So, you have to actually calculate or derive the guess estimate or proxy values for these many unknown population parameters before you actually take up the task of finding an estimator for the unknown population error variance. Hence you need to divide these some of squares error residuals by capital N minus K, which is basically the number of observations which are independently left to the sample.

(Refer Slide Time: 29:48)



Gauss-Markov Theorem

- ♣ If the assumptions 1-5 (the 6th assumption isn't needed for the theorem to be true) of the CLRM, the least squares estimators have the smallest variance of all linear and unbiased estimators of population regression parameters. They are the BLUE (Best, linear and unbiased estimator).
- ♣ Note:
  1. The theorem does not say that b_0, b_1, b_2 … are the best of all possible estimators. They are "best" when compared to *similar* estimators, those that are *linear* and *unbiased*.
  2. The estimators are best within their class because they have the minimum variance.
  3. The 6th assumption (Normality) isn't needed for the theorem to be true.

Now, we are going to talk about a very important theorem, and that is called Gauss-Markov Theorem. Now, before I state it formally from a slide, let me tell you why we required this theoretical result.

Well, we have got our OLS estimators for the intercept coefficient, the slope coefficient, and the population error variance, but we have to know judge how good or bad they are. Now, what do I mean by that? So, if I say that some estimator formula is good then it should obey certain properties.

So, the Gauss-Markov Theorem actually talks about certain properties of the OLS estimators. So, here we say that if the first five assumptions of the classical linear regression model are holding, then the OLS estimators have the smallest variance of all linear and unbiased estimators of population regression parameters. And if these Gauss-Markov Theorem holds, then we say that the OLS estimators are blue. And by blue, I mean, based linear unbiased estimator.

So, now I will have three notes that are related to these Gauss-Markov Theorem. Note that this hearing does not say that this beta naught, beta 1, beta 2 and other estimators of unknown population parameters are based of all possible estimators that you can theoretically propose. They are based when compared to similar estimators.

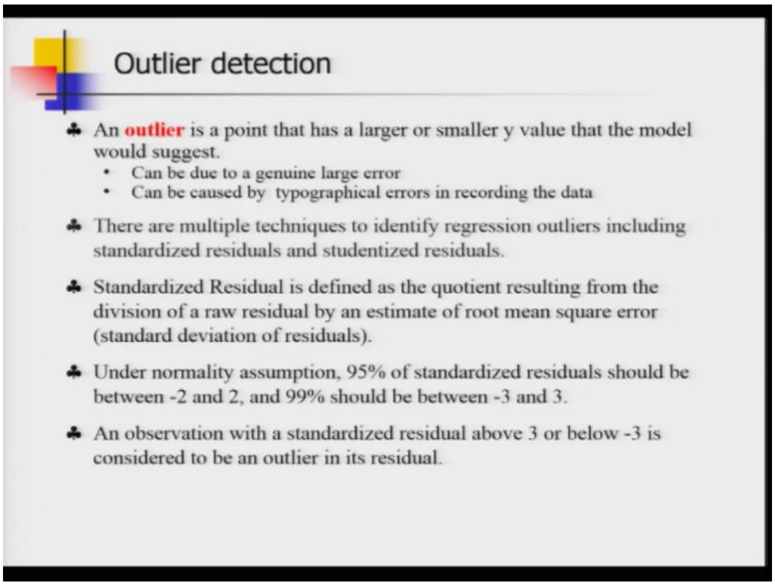
So, what do we mean by similar? So, the estimator has to be first unbiased and then it has to be linear. Now, we all understand what do we mean by unbiasedness and the linear estimator, so I am skipping that discussion. Now, out of this special class of estimators, which are unbiased and linear in nature then what will make them so special that we will call them the best? So, it can be shown that these OLS estimators have minimum variance out of all linear and unbiased estimators in this context and hence OLS estimators are called blue.

So, note that in this Gauss-Markov Theorem also we have not made use of the normality assumption, which is the sixth assumption that we made in this lecture. Once we conduct in OLS regression technique, and we get the values of the unknown population parameters then we can get the fitted values. And of course, one can try to draw a scatter plot and they are in one axis, the person can plot the actual observed values of Y and on the other axis, one can plot the fitted values of Y, Y hat.

And if the scatter plot is showing that points are close of each another then you can say that the fit is good, but there could be some cases where you see that there is huge deviance between the fitted Y hat and the actually observed value Y. So basically, we are talking about high residuals in these cases. So, we should not stop after getting linear regression equation which is given by the OLS. We should also look at the fit of our regression model. And the first way to look at the regression and diagnostic checking is by looking at the outlier.

So, if there is an outlet or not. There could be difference between the fitted value and the actual value, but the gap should not be very high. If the gap is very high, then there could be actually some issue of outliers. And in the next slide, we are going to discuss outlier detection and all, and this is very handy for applied economic and statistical research.

(Refer Slide Time: 34:22)



Outlier detection

- An **outlier** is a point that has a larger or smaller y value that the model would suggest.
  - Can be due to a genuine large error
  - Can be caused by typographical errors in recording the data
- There are multiple techniques to identify regression outliers including standardized residuals and studentized residuals.
- Standardized Residual is defined as the quotient resulting from the division of a raw residual by an estimate of root mean square error (standard deviation of residuals).
- Under normality assumption, 95% of standardized residuals should be between -2 and 2, and 99% should be between -3 and 3.
- An observation with a standardized residual above 3 or below -3 is considered to be an outlier in its residual.

So, what is an outlier? Formally, we can define an outlet as a point that has a larger or smaller Y value that the model would suggest. Now, of course, this is in a relative sense and this can be due to two different reasons mostly. And the number one reason could be that there is a genuine large error and otherwise there could be some typographical errors in data entry process and all.

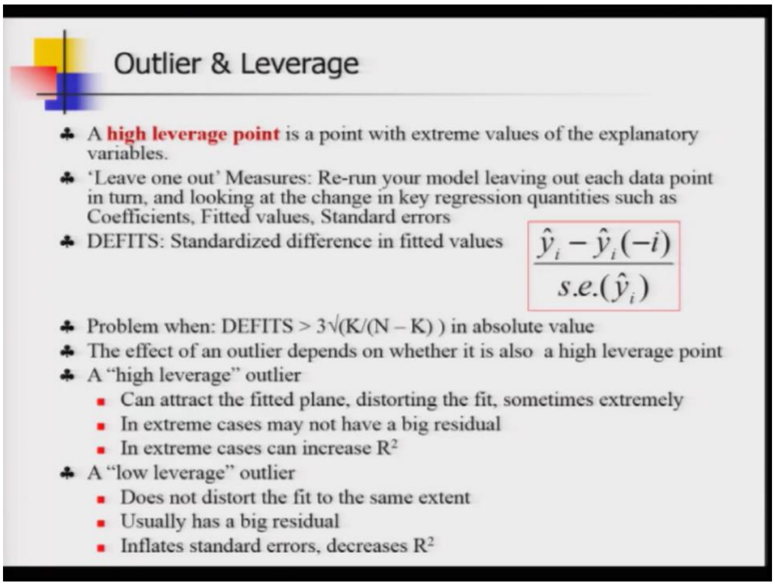
But we do not know absolutely, what is the reason. What has generated that outlier? Now, there are multiple techniques, statistical techniques that are available in a statisticians’ toolkit to detect outliers. We do not have time to go through all of them, we are only going to concentrate on one particular measure.

So, here in this course, we are going to focus on the standardized residual as the measure for outlier detection. And this is defined as the quotient resulting from a division of the raw value of the residual by an estimate of the root mean square error that is basically the standard deviation of residuals. So, once you generated the values of residuals that is basically a variable, so you can also compute the standard deviation of that variable. So that is basically the root mean square error.

As we have introduced the assumption of normality in this lecture. So, we can now feel free to assume normality. And by the assumption of normality, we can say that 95percent of these standardized residuals should be between minus 2 and plus 2 and 99percent should be between minus 3 and plus 3.

So, if we find an observation with a standardized residual value above three or below minus 3, then it can be considered as an outlier. Now, outlet is also related with the concept which is called leverage. What is leverage? We are not going to get into details of this concept, leverage because they are not actually that much used in the field of econometrics, but as they are heavily used in the field of applied statistics, I am going to briefly introduce this concept here. And I will just give you some brief idea how leverage and outlier can actually interact with each other.

(Refer Slide Time: 37:12)



Outlier & Leverage

- A **high leverage point** is a point with extreme values of the explanatory variables.
- 'Leave one out' Measures: Re-run your model leaving out each data point in turn, and looking at the change in key regression quantities such as Coefficients, Fitted values, Standard errors
- DEFITS: Standardized difference in fitted values $\frac{\hat{y}_i - \hat{y}_i(-i)}{s.e.(\hat{y}_i)}$
- Problem when: DEFITS > $3\sqrt{(K/(N - K)}$) in absolute value
- The effect of an outlier depends on whether it is also a high leverage point
- A "high leverage" outlier
  - Can attract the fitted plane, distorting the fit, sometimes extremely
  - In extreme cases may not have a big residual
  - In extreme cases can increase R^2
- A "low leverage" outlier
  - Does not distort the fit to the same extent
  - Usually has a big residual
  - Inflates standard errors, decreases R^2

So formally speaking, a high leverage point is a point with extreme values of the explanatory variables. And how to detect leverage from the regression exercise. There are, again different measures, but we are only going to talk about one particular measure, but note that all these different measures that are available there to find out leverage they follow one principle and that is called leave one out principle.

So, what do I mean by leave one out principle? So here, I am going to focus on one particular measure, and I am going to explain it this leave one out principle. So basically, what you have to do, you have to redone your regression model several times. Suppose you have capital N number of observations, so you are interested to find the leverage of the small nth observation.

So, you actually have to now leave it from the sample, so you have to exclude it from the sample and then basically you have to run the regression. And if you are now interested on N plus 1 th data point, then you have to bring nth observation back in the sample and you have to exclude n plus 1 th sample out from the sample, and then you have to run the regression again.

So, this is the way you have to run multiple times of regression analysis, but let us look at the measure. We are only going to concentrate on the simplest possible measure.

So, here the measure is called DEFITS, and the full form is standardized difference in fitted values. So, that is defined as the difference of actually fitted value, which is Y hat i, that comes from my original regression and the fitted value for that particular observation. So here, the DEFITS is defined as the difference between Y hat i and Y had i given minus i, so I have to exclude that observation i from the sample, and then I have to run the regression. I have to get the coefficients.

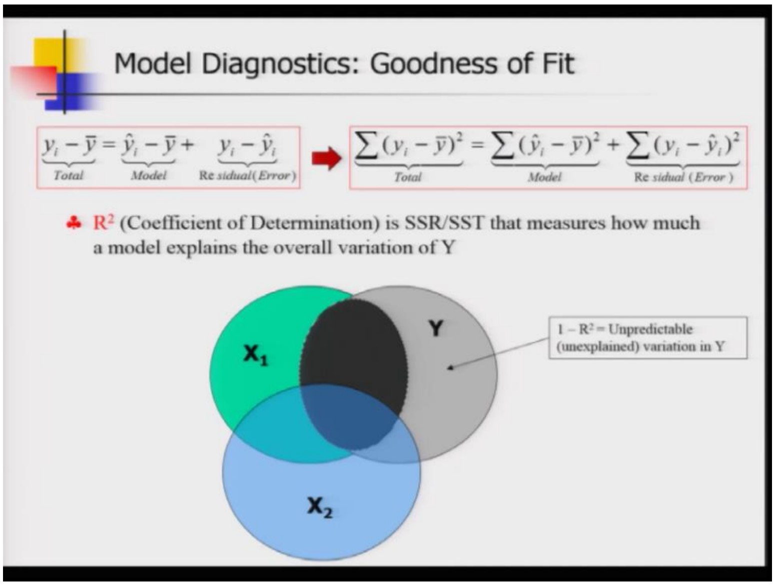
And then, if I apply those coefficients from the second revision, I will get, again a value of Y hat i or the fitted value for the ith observation. And I am talking about that difference. And that difference should be divided by the standard error of Y hat i which comes from the original regression.

So, how to detect the problem? So, here if for any observation I see that DEFITS take value greater than three times square root of k divided by N minus k in absolute terms then there is a problem. Now, in the last part of this slide, I am going to talk about the linkage between leverage and the outlier.

So, let us first look at a high leverage outlier. What is a high leverage outlier? So, it can actually affect the fitted plane and distorts the feet sometimes extremely and in extreme cases it may not have a big residual value, but as the leverage is very high for that particular observation from the explanatory variables side there could be a problem. And in extreme cases, it can increase the R square value, but that is really the case.

Now, what is the low leverage outlier. So, in that that case, we see that, that observation does not distort the fit to some extent. Usually, it has a big residual and I know it inflates the standard errors so it decreases the R square value. Now, note that here I actually have given you a very simple introduction to leverage. If we really want to explore the interaction between leverage and the outlier, then it will take up a full class and that is not the objective of the course also. So, I end here at this point. Let me move on to the other concepts, which are much more relevant for econometrics.

(Refer Slide Time: 41:43)



Model Diagnostics: Goodness of Fit

$$\underbrace{y_i - \bar{y}}_{Total} = \underbrace{\hat{y}_i - \bar{y}}_{Model} + \underbrace{y_i - \hat{y}_i}_{\mathrm{Re}\,sidual(Error)} \Rightarrow \underbrace{\sum (y_i - \bar{y})^2}_{Total} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{Model} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\mathrm{Re}\,sidual\,(Error)}$$

♣ R^2 (Coefficient of Determination) is SSR/SST that measures how much a model explains the overall variation of Y

1 – R² = Unpredictable (unexplained) variation in Y

So, we again, visit the concept of goodness of fit as the model diagnostic issue. We have already seen this goodness of fit previously, but let me repeat this again, because now we have a multiple linear regression model case. So, I will start with that decomposition of the variance thing that we have done in and over, but I will first start with the simple identity, which has this Yi equal to Y hat i plus the regression residual. So, to have an idea about the goodness of fit tissue you concentrate on the first-grade box that is given on the Northwest corner of the slide.

So, here you see that I have actually expressed the total variation in the explanatory variable that is split in two components. And the first component is the variation that is captured by the model and that is Y hat i minus Y bar and then there is, of course, residual or error, which is Y i minus Y hat i.

Now, that equation can be used to write another equation or expression. Now, this time you have to make use of the sum of squares’ concept. So, you just squared these differences, and then you sum over all observations, and then you get the first component here and that is basically sum of square of the difference between the actual observation value for ith observation and the grand mean Y bar that is there in the sample. So, that is called the total variation. So, that is basically no one component of the sample variance formula that we have dealt with previously.

Now, that is now divided into two further components. And first component is basically the model components, so that is basically called the sum of squares regression or sour squares model. And the second component is basically called the sum of squares residual or sum of square error.

Now, R square, the coefficient of determination or multiple coefficient of determination as we are in the multiple linear regression lecture that we have defined previously and that is basically defined as the SSR divided by SST or 1 minus SSE divided by SST and that measures how much a model explains the overall variation of Y.

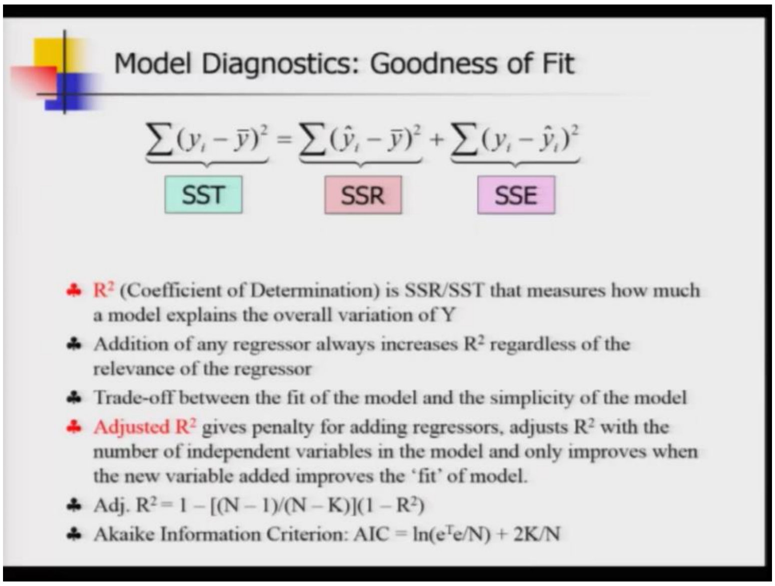
So, let us quickly have a recap of these multiple coefficient of determination concept. So here, I am showing you a Venn diagram. So, here I have three circles. So, this is a simplified case, of course, so I have one Y the dependent variable and I have two Xs, explanatory variables X1 and X2.

And note that both X1 and X2 explains some bit of variation of Y and a actually these two circles X1 and X2 they are variations inter have some common part and all these three circles intersect with each other. So, here you see that there are intersections of these three circles, and there are common areas in these window diagram.

So, this common area, which is given by the X which is given by the black color, that basically is basically the common contribution of X1 and X2 variables variation in explaining the variation of Y. But there is also another component, which is basically an intersection of the X2 circle and Y circle that is in light blue color that also contributes to the R square computation.

So basically, what is left? That left area is basically colored in a gray, and let us have a look at it. So here the part of the Y circle, so this circle gives you the total variation of Y. So, this area, which is colored in gray that part is actually not explained by X1 and X2 the explanatory variable. So, this is the unpredicted or unexplained variation in Y. And that gives me the 1 minus R square area.

(Refer Slide Time: 46:42)



Model Diagnostics: Goodness of Fit

$$\underbrace{\sum (y_i - \bar{y})^2}_{SST} = \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{SSE}$$

♣ R^2 (Coefficient of Determination) is SSR/SST that measures how much a model explains the overall variation of Y

♣ Addition of any regressor always increases R^2 regardless of the relevance of the regressor

♣ Trade-off between the fit of the model and the simplicity of the model

♣ Adjusted R^2 gives penalty for adding regressors, adjusts R^2 with the number of independent variables in the model and only improves when the new variable added improves the 'fit' of model.

♣ Adj. $R^2 = 1 - [(N-1)/(N-K)](1-R^2)$

♣ Akaike Information Criterion: $AIC = \ln(e^Te/N) + 2K/N$

Let us have some brief discussion of R square. So, you see, as we are dealing with k number of explanatory variables addition of any regressor always increases R square value, regardless of its relevance to the explanatory variables, relevance to the variation of the dependent variable.

Now, does it mean that as addition of explanatory variables will not be creased my R square value. In fact in all probability, it is going to increase the R square value. And we know that higher R square value is good for my model fit. I can say my model is performing better. Shall I keep on adding more and more explanatory variables in the linear regression equation? No, not really.

Actually, there is a trade-off between the parsimony or the simplicity of a regression model and the know this goodness of fit. I mean, you can have more explanatory variables, and Jack up the R square or the multiple coefficient W termination value, but that will come at the expense of the simplicity or the parsimony of the model.

Note that, when you are conducting a regression analysis just think from a policy perspective or a practical perspective. Say a farm manager wants to do something with the degression equation that you have estimated or a policy maker wants to frame a policy or decided about a very serious policy issue by looking at the regression equation that you have reported to the authority.

Now, if you actually have a 27 variable regression model or a very complicated functional forms involving in a square, cube log, square root and whatnot, then that messy model that complicated model will be of not much help to the person who is actually going to use the model or who was willing to use your model. So, from the practical standpoint there has to be a limit up to which you can add variables in your linear regression model. So, where to stop that is the question.

Now, there is a measure provided by statisticians. Actually, there are multiple majors, not one, but related to R square there is only one measure that is proposed by statisticians which will help you to identify where you can stop.

So, this adjacent R square measure takes care of or addresses these trade-off between the fit of the model and the simplicity of the model. So, this concept, adjusted R squared, gives the penalty for adding extra regressors. So, it adjusts the original R square concept with the number of independent variables in the model, and it only improves when the new variable added improvement to the fit of the model.

So, basically what happens suppose you have one particular variable that is coming to your mind. And the question is that whether you want to add in the regression equation or not. And if you add in the regression equation probably it will give you a higher R squared value, maybe marginally high.

But then the question is that will you keep that variable in the regression equation. You have to judge by looking at the adjusted R square, because adjusted R square actually now charges you a penalty for throwing an additional explanatory variable in the regression equation and make it further complicated. So, if the adjusted R squared value goes up by incorporating this additional variable then you should keep that variable in the model.

So, now let us look at the formula of the adjusted R square. It is a little bit complicated compared to the original R square formula, so it is given by 1 minus capital N minus 1, divided by capital N minus k, and this is whole multiplied by 1 minus the original R square.

So, many software’s routinely report this are adjusted R square. You do not have to even ask the software to calculate it separately. So, the best way to go for a model selection is not by looking at the value of the R square, you should look at the value of the adjusted R square. So, you keep on adding variables in the model, as you see that adjustment R-squared is increasing, but you stop when you do not see any improvement in there adjusted R square measure after throwing an additional variable in the regression equation.

So, now we are going to look at the final measure of the goodness of fit, and that is called the Akaike Information Criteria. So, in the Akaike Information Criteria, what happens, we still make use of the summer square error residuals as we make use of it in the R square formula.

So, that is why you see that E transpose E is given in the formula, and that is basically giving me the measure of misfit in my model the measurement of deviation in the model. And of course, I expect that as I introduce more number of explanatory variables in my regression model, this E transpose E number will come down, so that is good for you, but actually you have to now change the penalty for adding this explanatory variable, because you are sacrificing the simplicity or the parsimony of the model. So, you have to add the penalty component.

And if you now look at the second component of this formula, which is 2K divided by N, so that is basically the penalty component in the formula. So, as you throw more and more explanatory variables in the linear regression equation E transpose E falls, but then although the first component is overall falling the second component actually is rising because 2 and N are fixed number, so if you throw or add more and more degreasers in the regression equations, k will increase, so the second component of the formula increase.

So, you see here one component is falling and then the other component is rising. So there has to be a balance to strike. And generally, what happens you go for a model selection using AIC when you go for the minimum AIC criteria. So, this is the end of today's lecture. In the next lecture, we are going to introduce the case of hypothesis testing in the context of multiple linear regression model. See you then. Thank you.