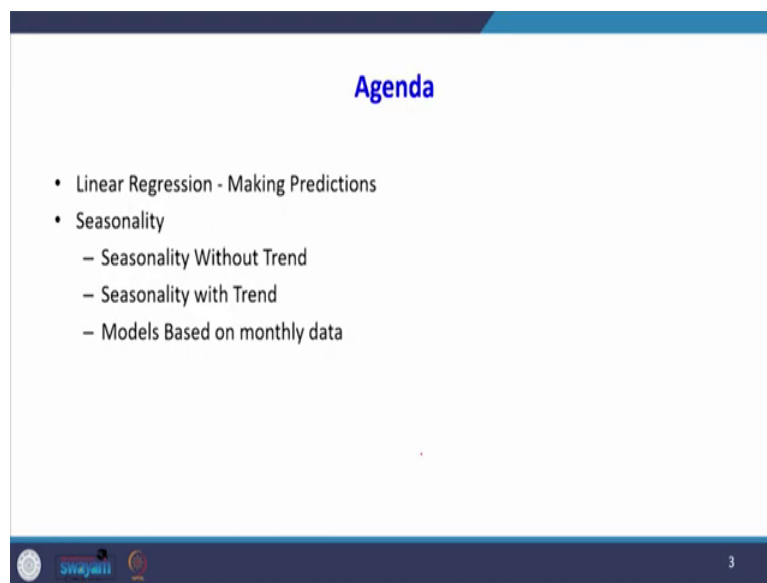


Decision Making with Spreadsheet
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 60
Time Series Analysis and Forecasting – V

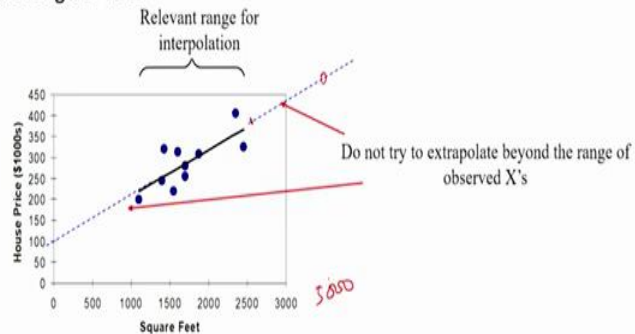
So, dear students, in the previous lecture, I was discussing forecast trends with the help of regression analysis. Also, I have discussed how to build a regression model. In this lecture, I will show you how to use the regression model for prediction. After that, I am going to discuss forecasting seasonality without trends and with the trend.



So, the agenda for this lecture is a linear regression model for making predictions. And how to use regression analysis for predicting seasonality without a trend and seasonality with a trend?

Linear Regression Example Making Predictions

- When using a regression model for prediction, only predict within the relevant range of data



Dear students, in the previous lecture, we constructed a regression model, so, using that regression model, we are going to predict. Suppose you predict the price for a house with 2000 square feet. We know that our regression model is a house price equal to 98.25 plus 0.1098 square feet. If I substitute the square feet value of 2000, I am getting the house price value. 317.85. The predicted price for a house with 2000 square feet is 317.85.

All the values are in terms of thousands, so it will be 317,850 dollars. When using a regression model for prediction, only predict within the relevant range of data. You see that the data is from 1000 to 2500. So, if you are making a prediction just after this point, it is ok. That is a relevant range for interpolation. Suppose if you want to predict here, that the X value is extremely high, maybe 5000 square feet.

So, this model may not be valid. So, the point here is not to try to extrapolate beyond the range of X values observed.

Measures of Variation

Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \bar{Y})^2$	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

where:

- \bar{Y} = Mean value of the dependent variable
- Y_i = Observed values of the dependent variable
- \hat{Y}_i = Predicted value of Y for the given X_i value

6

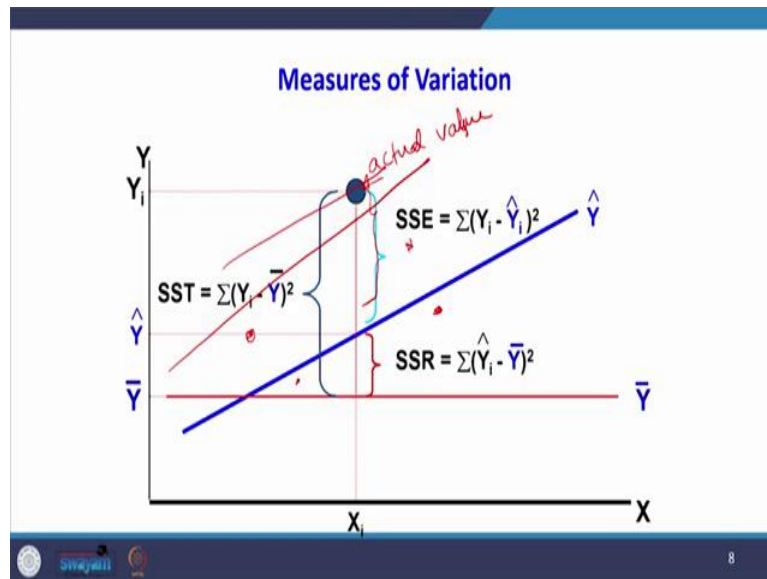
Next, we are going to discuss the measures of variations. This concept is used to find out the goodness of our regression model. So, here is what we are doing: the total variance is made up of two parts. Here, total variance is the total sum of the squares equal to the regression sum of the squares plus the error sum of the squares.

Measures of Variation

- SST = total sum of squares
 - Measures the variation of the Y_i values around their mean \bar{Y}
- SSR = regression sum of squares
 - Explained variation attributable to the relationship between X and Y
- SSE = error sum of squares
 - Variation attributable to factors other than the relationship between X and Y

7

Here, the meaning of SST is the total sum of squares. It measures the variation of Y_i values around their mean. SSR is the regression sum of square explained variations attributable to the relation between X and Y. SSE is the error sum of square variation attributable to factors other than the relationship between X and Y.



So, I discussed how to find the variance. If there is no X value, the easiest way is to predict using its mean value \bar{Y} . So, this point is, for example, the actual data and the actual values of Y. But if we use the mean of the Y value for the prediction, so the total error is you see that the total distance is $(Y_i - \bar{Y})$ that is the error square the error, then we have to sum the error. Why are we summing?

So, this one point is only for illustration purposes. There may be different points there. Some point may be here. Some point may be here, some point here. Some point may be here. So, we are finding the total sum of squares. That is why we use Sigma. So, the total sum of squares is the difference between the actual value minus the mean value, squaring that error, and summing that error, which is our SST.

$$\mathbf{SST = \sum(Y_i - \bar{Y})^2}$$

SSR is a regression sum of the squares, so this distance is $(\hat{Y} - \bar{Y})^2$ the whole square. That is the regression sum of squares, and you see the top here. This portion is the error sum of the squares. What is that value? $(Y_i - \hat{Y})^2$. So, this error sum of squares is an unexplained variance, which is why we are putting sigma. Because there are so many points, we have to add all the errors.

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

So, here, this portion of the regression sum of the square is nothing but the explained variance; suppose our regression model is a good model; for example, it goes like this. Now, what will happen? The value of error, the sum of the squares will be minimum, and the value of the regression sum of the squares will be the maximum larger value, so the model is good. So, if the line goes exactly on this point, we may say that the error, somehow square is 0.

So, here, SSR will be equal to SST. There are no errors at all. That is an ideal regression equation.

Coefficient of Determination, r^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

$$0 \leq r^2 \leq 1$$

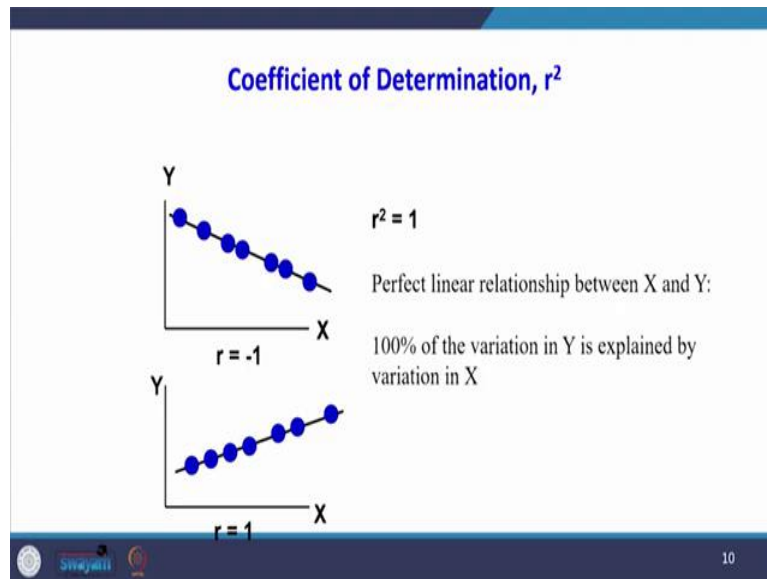
9

Here, the goodness of the model is going to be explained. The term is called the coefficient of determination r^2 . The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable. The coefficient of determination is also called r^2 and is denoted by r^2 . We have seen an explained variance divided by the total variance.

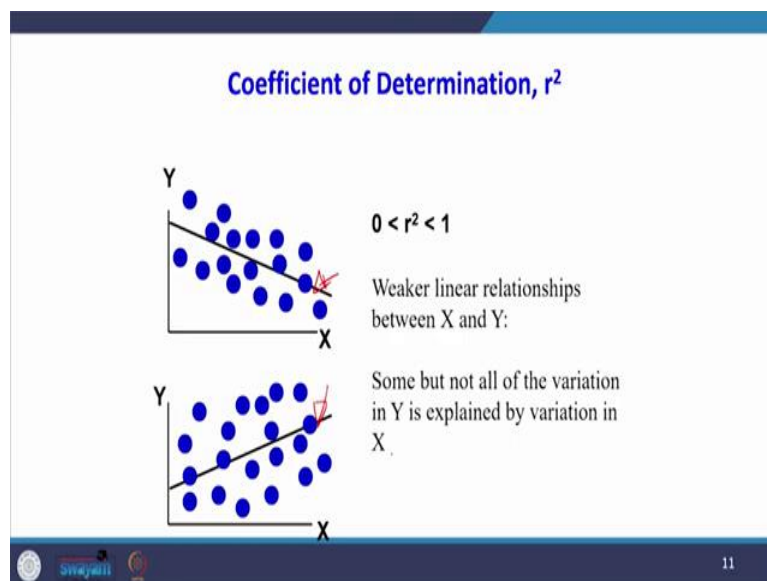
Explained variance is the regression sum of the squares, and total variance is the total sum of the squares. The range of this r^2 is between 0 and 1.

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

$$0 \leq r^2 \leq 1$$



So, the coefficient of determination r^2 is 1 when there is a perfect linear relationship between X and Y. What is the meaning of that? 100% of the variation in Y is explained by the variation in the X. You may see that on the X axis, r is there, and $r = -1$. So, it has a perfect negative relationship, but the value of $r^2 = 1$. The bottom figure says $r = 1$. It has a perfect positive relationship. Because the line passes through all the points, but the r^2 value = 1.



Now, if the value of r^2 is between 0 and 1, it will not be perfect, but it will be the actual points, the blue points scattered around this line. Similarly, the bottom is also scattered around the thick line, this line. So, there is a weaker linear relationship between X and Y, some by the bottom, but not all the variations in Y are explained by the variation in X. So, this is somewhat good, but it is not that good because there are more variations here.

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

$n-k-1$
 $k = \text{no of}$
 independent
 variables

Where

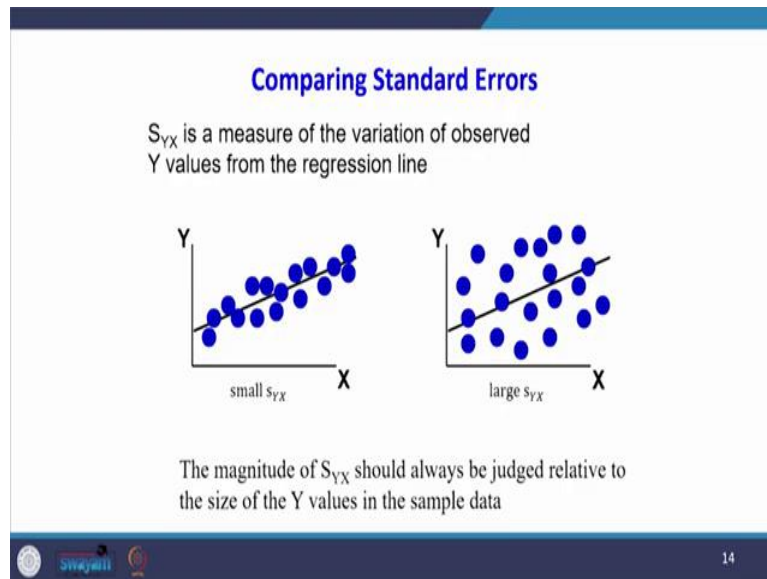
SSE = error sum of squares
n = sample size

You see when $r^2 = 0$, there is no linear relationship between X and Y; the value of Y is not related to X. That is, none of the variation in Y is explained by variation in X. Apart from the coefficient of determination, another way to test the goodness of the model is the accuracy of our model, which is the standard error of the estimate. SSE estimates the standard deviation of the variation of observations around the regression line.

Error sum of the square upon $(n - 2)$, why it is $(n - 2)$? Actually, it is $(n - k - 1)$, so k represents a number of the independent variables. In our example, the number of independent variables is the square feet of the land. So, we know that already explained how to find out SSE is $(Y_i - \hat{Y})^2 / (n - 2)$. Here, n is the sample size.

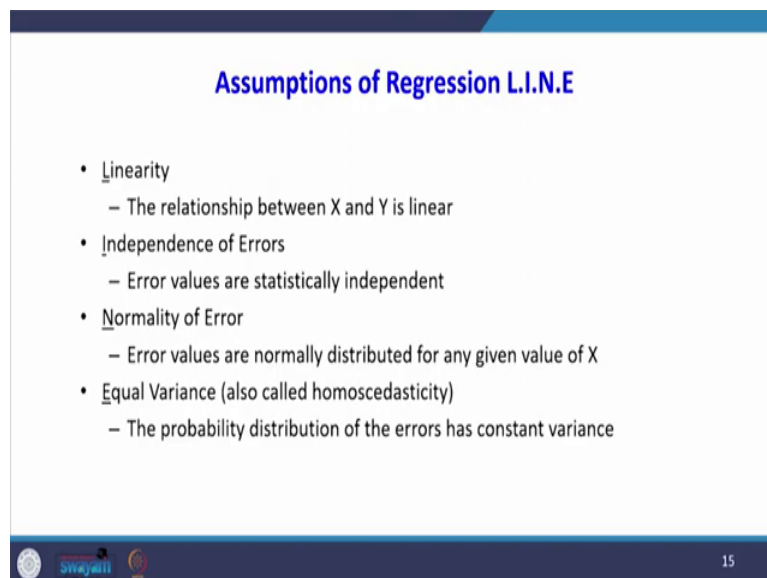
This can be represented in the following ways:

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$



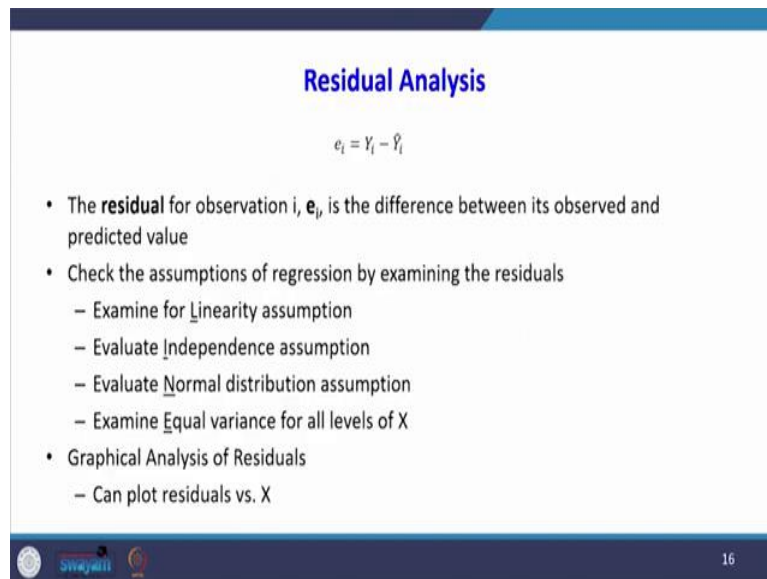
Comparing standard errors, S_{YX} is the measure of the variation of observed Y values from the regression line. See the figure, which is on the left-hand side. It has a small standard error, but you see the figure on the right hand, the side which has a large standard error. The magnitude of S_{YX} should always be judged relative to the size of Y values in the sample data.

If you compare 2 models, if the standard error is minimum, we can say that the model is a good model.



And other things, what are the assumptions for the regression line? We can say in the short form itself L. I N.E line. The first one is linearity. The relationship between X and Y is linear. The second one is independent of errors. Error values are statistically independent. The third one is the normality of the error. Error-values are normally distributed for any given value of X. The last assumption is equal variance, also called homoscedasticity.

The probability of distribution of the error has constant variance. That is the meaning of your equal variance.



The slide is titled "Residual Analysis" in blue text. Below the title is the equation $e_i = Y_i - \hat{Y}_i$. A bulleted list follows, detailing the definition of a residual and the assumptions of regression that can be checked by examining residuals. The list includes: the residual for observation i , e_i , is the difference between its observed and predicted value; check the assumptions of regression by examining the residuals (with sub-points for Linearity, Independence, Normal distribution, and Equal variance); and Graphical Analysis of Residuals (with a sub-point for plotting residuals vs. X). The slide footer contains logos for "UWagyeji" and "16".

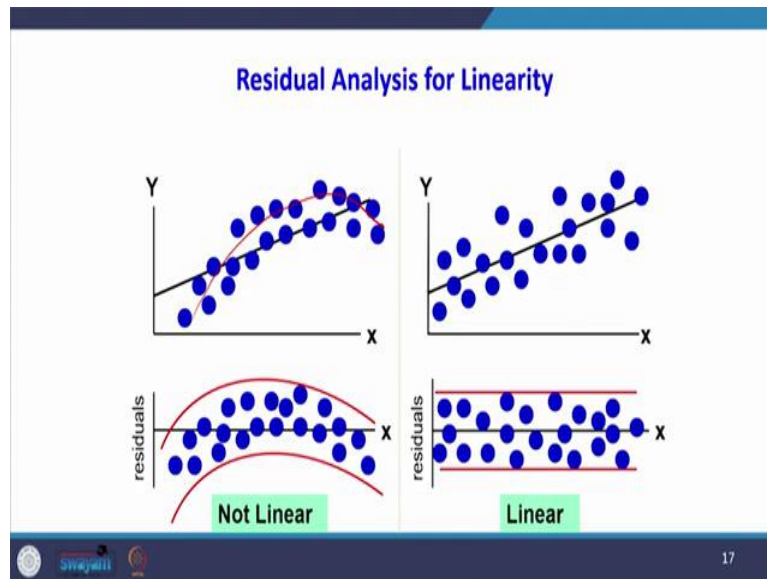
Residual Analysis

$$e_i = Y_i - \hat{Y}_i$$

- The **residual** for observation i , e_i , is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
 - Examine for Linearity assumption
 - Evaluate Independence assumption
 - Evaluate Normal distribution assumption
 - Examine Equal variance for all levels of X
- Graphical Analysis of Residuals
 - Can plot residuals vs. X

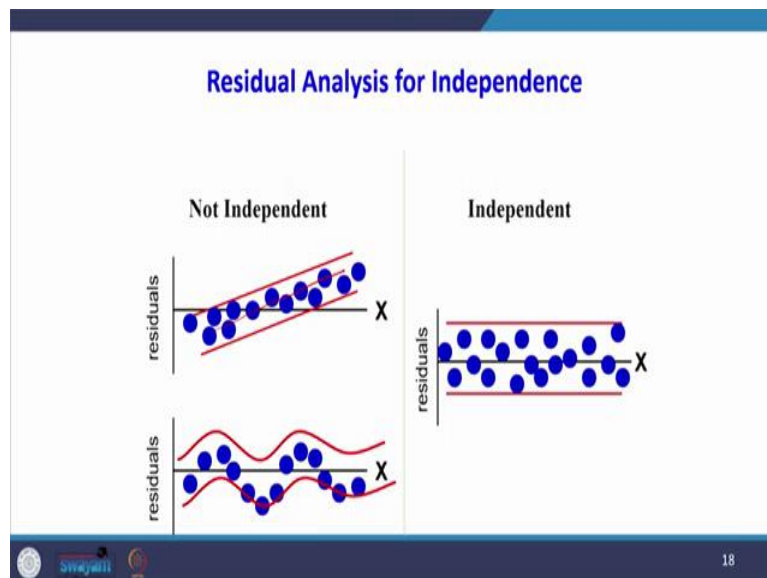
So, finding the r square and the standard error is important. Standard apart from the standard, error, and coefficient of determination, another way to judge the goodness of the model is residual analysis. That is, the error analysis error is the actual minus predicted value. The residual for observation i , e_i is the difference between its observed and predicted value. Check the assumptions of regression by examining the residuals.

Examine for linearity assumption, evaluate independence assumption, evaluate normal distribution assumption, and examine equal variances for all levels of X. So, the graphical analysis of residuals will help you to check this assumption. We can plot residuals that are errors versus the value of independent variables.



Now, look at the residual analysis in the X axis; we have X in the Y axis, and we have residuals. Now, when you plot the error, you see that it is following a curved shape. See that it follows the curved shape, we can say that our assumption of linearity is violated. Generally, this will happen if the actual data follow a nonlinear pattern. Look at the right-hand side if it is linear; it is equally scattered around our predicted regression line.

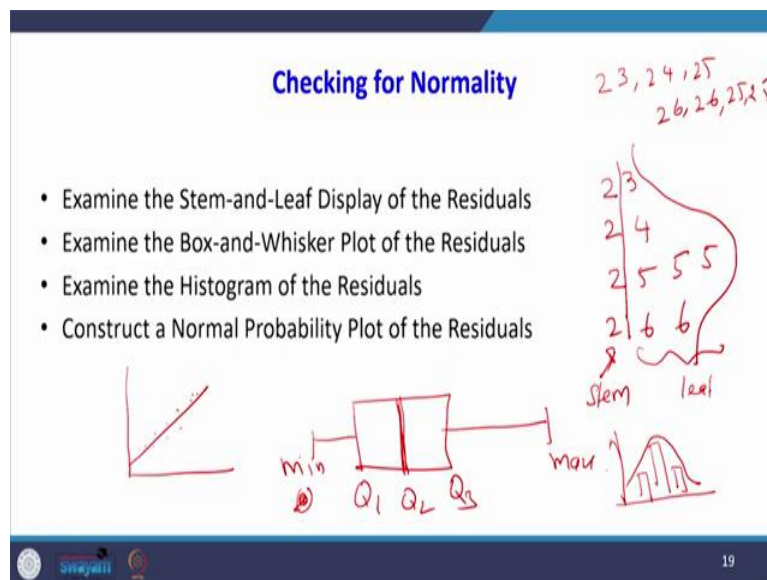
So, the right-hand side is the plot for error for a linear function, and the left-hand side figures for a non-linear function.



Now, the next assumption is that the error should be independent. Look at the left-hand side here. There is a trend on the error, so that means there is a dependency on the error, but you see the bottom one if the residuals follow certain patterns. If there is a pattern in the error, we

can say that errors are not independent. There is a dependency that that is a violation of assumption.

When you plot the residuals, you see that there is an almost equal number of points above the X and below the X . So, we can say that the errors are randomly distributed. So, this assumption can also be verified with the help of this plot.



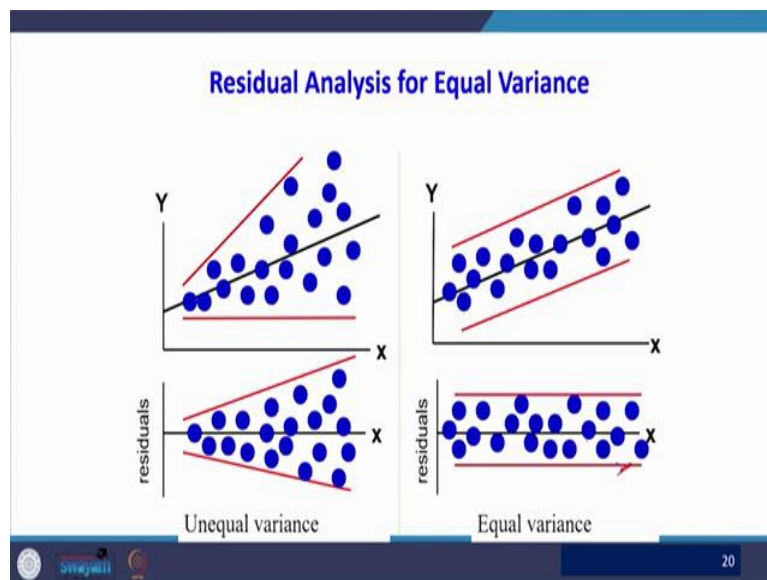
Then, we can check the normality of the error with the help of Stem-and-Leaf Display of the residuals. What is Stem-and-Leaf? So, what will happen? Suppose you have an error of 23, 24, 25, 26, again 26, again 25, again 25, again 25, again 25, again 25, for example. The stem is the first digit 2, so this is 2, 3, 2, 4, 2, 5, following 2, 6, again 2, 6, again, 2, 5, again 2, 5, again. So, this pattern, this portion is called the first digit's stem.

The second digit is a leaf. So, this pattern follows a normal distribution. This is one way to check the normality. Next, we can check the normality using the Box-and-Whisker Plot of the residuals. When you plot the Box-and-Whisker Plot, it will be like this. This is your Q1, this is your Q2, this is Q3. This is the minimum; this is the maximum. So, what will happen? If this is the middle line, it is the middle of the box; then we can say that data that an error follows a normal distribution.

If this line on the left-hand side says that it follows a skewed distribution, that means the right is secured. If the middle line is on the right side, we can say that the data has left skewed data. Another way to test the normality is a Histogram of the Residuals. So, when you plot

the histogram, for example, suppose I am plotting like this. If it is when you plot that it is following the bell-shaped curve, then we can say that the error follows a normal distribution.

Then we can construct a normal Probability Plot of the residuals PP plot. So, the PP plot will be like this. If it is exactly all the values are falling exactly on this diagonal, then we can say that the error term follows a normal distribution.

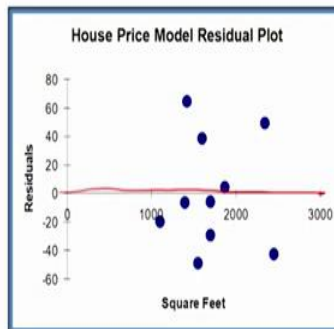


Then, how do we test the equal variance assumptions? The left-hand side figure says that there is unequal variance. You see that the variances keep on increasing when the value of X increases. So, this is a violation of our equal variance assumptions. But look at the right-hand side of the variance. There may be an increase in trend. But it is around our predictor line. So that is at the bottom. also, it is level data.

But the variance is around the predicted line, we can say the equal variance assumptions are accepted.

Linear Regression Example Excel Residual Output

RESIDUAL OUTPUT		
	Predicted House Price	Residuals
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Does not appear to violate any regression assumptions

Another way to test whether the errors are independent or not is to plot the residuals. So, in our example, when you plot the residuals, you see that there is an equal number of points above the mean above this point above this standardized residual value, and below the standardized residual values. So, what is happening? It does not appear to violate any regression assumptions because the error terms are randomly distributed.

Inferences About the Slope

$$\hat{y} = b_0 + b_1x$$

House Price in \$1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098 (\text{sq.ft.})$$

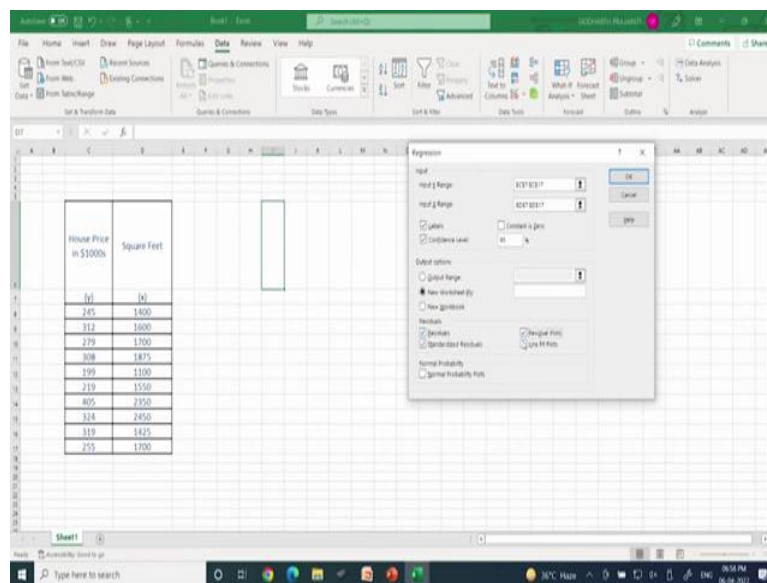
The slope of this model is 0.1098

Is there a relationship between the square footage of the house and its sales price?

Now, we are going to discuss inferences about the slope. What is the meaning of inferences about the slope? The regression equation has a house price equal to 98.25 plus 0.1098 (square feet). So, this is nothing but, in the form, $\hat{Y} = b_0 + b_1x$. Now, here the slope is b_1 . The slope value for this model is 0.1098. Now, we have to see this regression equation is for our sample data.

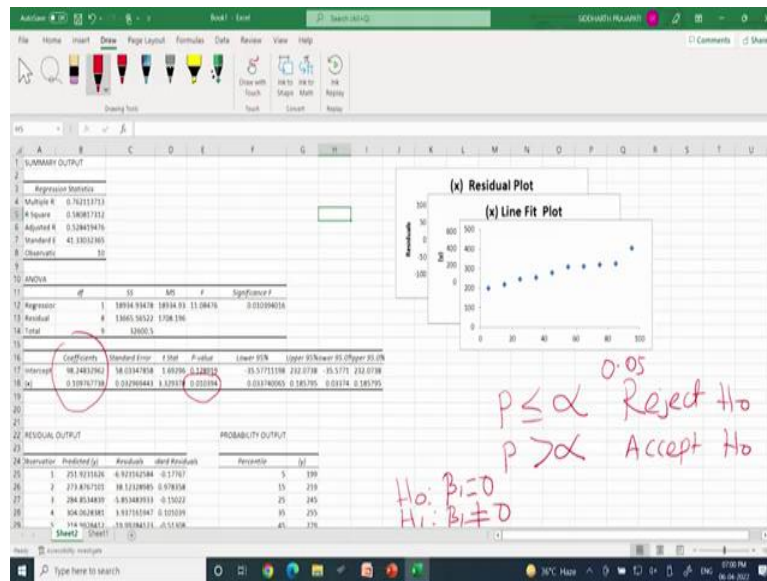
If you expand, you extrapolate for the population whether this relationship between square feet and the house price is significant or not, which we can test with the help of inferences about the slope. So, this slope, value 0.1098, is only for the slope that we got from the sample data. Now, we are going to test whether there is a relationship between the square footage of the house and its sales price.

That is how I am going to do? With the help of Excel, I am going to show whether the model is significant or not.



So, I am going to open excel sheet. I have taken house price and square fit value, so go to data, data analysis, and regression. Select Y value, including Y. Then you select X value. There is a label. Check the label and check the confidence level. Then you can see that here, there is a residual, the standardized result is there, a residual plot is there, a line fit plot is there, and a normal probability plot is there.

When you check these boxes, you will get residuals also. So, we are going to test whether the model is significant or not. So, Press, ok.



Now, you see C17. Now, I am going to interpret this output. So, what is there? Look at these values here: I have intercept and X value, and I have the standard error, and we are going to test whether this model is significant or not by looking at this P value. If the P value is less than alpha so, we have to be less than or equal to alpha, reject H_0 . If the P value is greater than alpha, accept H_0 .

So, in our model, the H_0 , the null hypothesis assumption is that beta 1 is equal to 0. The alternate hypothesis is beta 1 not equal to 0. So, if I accept my null hypothesis, there is no relation between X and Y at the population level. There may be a relation between X and Y for the sample data. When I look at the P value, the value of the P value is supposed to be alpha, which is a significant level of 0.05.

So that means I have a 95% confidence level. So, the significant level is 0.05. So, here, the P value is less than 0.05; it is only point 0.1. So, what will you do? I will reject my null hypothesis. So, when I reject the null hypothesis, I am inferring there is a relation between X and Y. We will go back to ppt.

Confidence Interval Estimate for the Slope

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{n-2} S_{b_1} \quad \text{d.f.} = n - 2$$

(X) ± 2σ_m
+

Excel Printout for House Prices:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-55.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

At the 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

23

Confidence intervals estimate for the slope: We have realized that the slope is significant.

Suppose I want to find out the confidence interval for the slope. So, it is

$$b_1 \pm t_{n-2} S_{b_1}$$

For example, in hypothesis testing, if you are predicting a confidence interval for the mean.

So, what is the formula?

Similar to that here, b_1 is our coefficient. Here, the sigma by root n is the standard error similar to that here. We can see the standard error here, S_{b_1} 0.03297.

$p \leq \alpha$ 0.05 Reject H_0
 $p > \alpha$ Accept H_0
 $H_0: B_1 = 0$
 $H_1: B_1 \neq 0$

So, I will go back to Excel. I will show you what is the standard error for the independent variable. See that it is 0.0329, that is, the standard error, will go back; so, we know we can from the Excel output directly we can take the b_1 value, then I can get the S_{b_1} value and the

t_{n-2} . So, the t_{n-2} where n is the number of data sets that value. I can get it from the table. But here you need not worry about the fact that we have a lower limit.

After calculating the minus value, we got a lower value, 0.03; the upper value is 0.18. This t_{n-2} we can get it from the statistical table or with the help of Excel. You can do that, but here, the Excel itself provides the confidence interval estimate for the slope. So, the actual value is 0.10, but it can be down to 0.03, and it can be increased up to 0.18. So, at the 95 % confidence level, the confidence interval for the slope is 0 points. This value is taken from here, 0.0337 and 0.1858.

Confidence Interval Estimate for the Slope

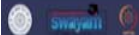
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Since the units of the house price variable is \$1000s, you are 95% confident that the mean change in sales price is between \$33.74 and \$185.80 per square foot of house size

This 95% confidence interval does not include 0.

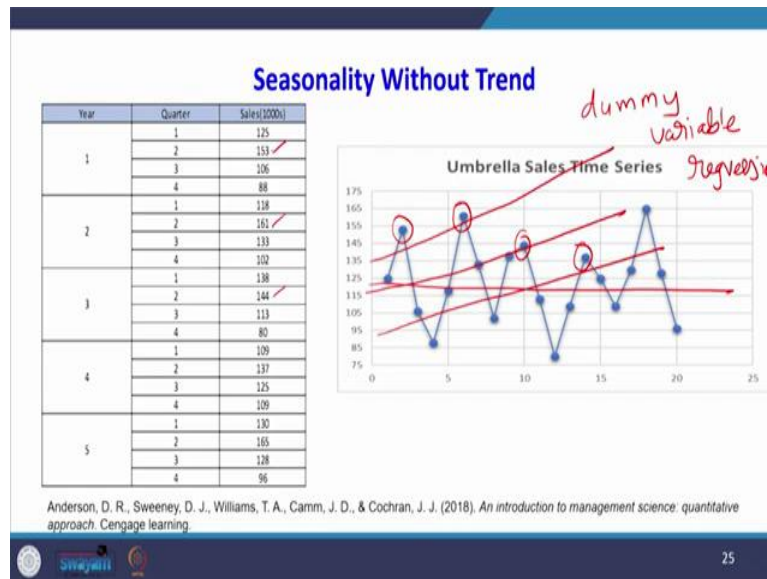
Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

H₀: $\beta_1 = 0$


24

Since the unit of the house price variable is in terms of thousands, you are 95% confident. That means a change in the sales price is between if you multiply thousand 33.74 dollars and 185.80 dollars per square foot of the house size. One thing you have to remember when you look at the confidence interval is That the 95 percent confidence interval does not include 0; either both the values are positive, or it may be both the values are negative also.

But you see that here, both are positive, so we are not capturing the 0 value. But our null hypothesis is $\beta_1 = 0$. So, you see that here, the lower limit is 0.03, and the upper limit is 0.180. We are not able to capture 0. So, there is no chance that β_1 will be equal to 0, but it is always β_1 will not be equal to 0. So, we are rejecting your null hypothesis. So, what you are concluding is that there is a significant relationship between house prices and the square feet at the 5 % level of significance.

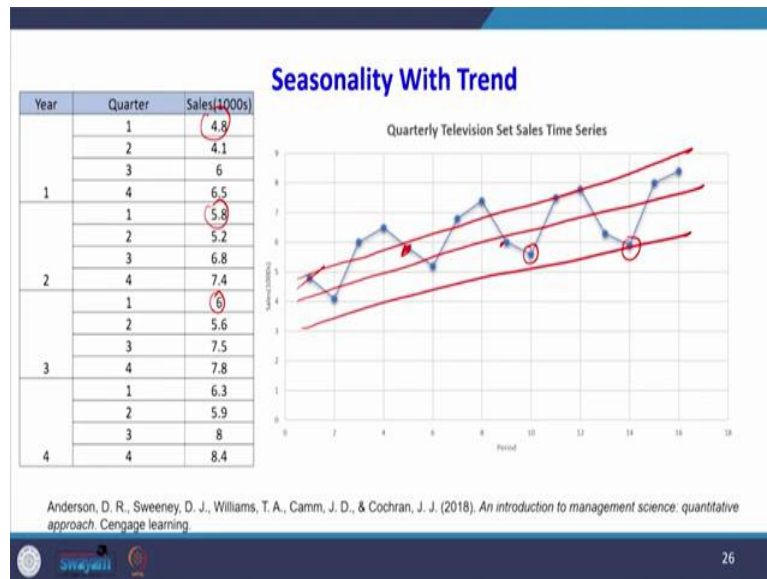


Dear students, so far, I have discussed forecasting trend data with the help of regression analysis. Now, I am going to discuss seasonality without a trend. Look at the table on the left-hand side. There is a quarter 1, the sale is 125, quarter 2, quarter 3, and quarter 4 for year 1, similarly, for years, 2 year, 3 year, 4, year 5. When I plot the data, it looks like trend data, but when you look closely at it, you see a cyclic variation.

For example, look at quarter 2 in year 1, so, again, in this data, the values in the second quarter are always higher; for example, here it is 153, here it is 161, for the third year also it is 144. So, in the second quarter, there is a cyclic pattern, but there is no trend. So, if instead of using a regression analysis for the level data for each quarter, you can have a separate regression line.

For example, this is for quarter 1, sorry, quarter 2, this is for quarter 2, this is for quarter 3, so this kind of regression model can be built with the help of a regression technique called dummy variable regression. So, when you use dummy variable regression, you can forecast values for each quarter; instead of considering them as a single data set, you can consider you can forecast for each quarter.

So, what will happen when you forecast for each quarter? The overall mean squared error will be decreased. So that will provide a more accurate model.



The next one is seasonality with trends. Now look at the year 1 data for quarter 1, so it is 4.8; in year 2, it is 5.8; in year 3, it is 6. So, what is happening? There is a cyclic pattern, and you see, it keeps increasing the quarter 1 data every year. So, here, there is not only a cyclic pattern, but there is also a trend there. So, instead of doing a regression model for the trend data, what can we do?

For example, here, this is 1, 1, 2, 3, 4. This is 1 2 3 4, 1, 1 2 3 4, 1, 1 2 3 4, 1 2 3 4, 1 2 3 4, again 1 2 3 4. So, instead of predicting only one trend, you can have a trend regression line for each quarter. So, what will happen so that the model will be more accurate instead of forecasting only trends? So, in the seasonality, with the trend model, we are not only predicting the trend, but we are also predicting the trend in the seasonality.

So, here, you also have to use the dummy variable regression to increase the accuracy of the model. Dear students, In this lecture, I explained how to use the regression model for prediction and how to test the assumptions of the regression model using residual analysis. Then, I explained how to test the significance of the regression model with the help of a spreadsheet. I have also discussed forecasting seasonality without trend and with the trend. Thank you.