R07 Introduction to Linear Regression

1. Simple Linear Regression 2
2. Estimating the Parameters of a Simple Linear Regression 3
2.1 The Basics of Simple Linear Regression 3
2.2 Estimating the Regression Line 4
2.3 Interpreting the Regression Coefficients 5
2.4 Cross-Sectional vs. Time-Series Regressions 5
3. Assumptions of the Simple Linear Regression Model 5
3.1 Assumption 1: Linearity 6
3.2 Assumption 2: Homoskedasticity 7
3.3 Assumption 3: Independence 8
3.4 Assumption 4: Normality 9
4. Analysis of Variance 9
4.1 Breaking down the Sum of Squares Total into Its Components 9
4.2 Measures of Goodness of Fit 10
4.3 ANOVA and Standard Error of Estimate in Simple Linear Regression 11
5. Hypothesis Testing of Linear Regression Coefficients 13
5.1 Hypothesis Tests of the Slope Coefficient 13
5.2 Hypothesis Tests of the Intercept 16
5.3 Hypothesis Tests of Slope When Independent Variable Is an Indicator Variable 17
5.4 Test of Hypotheses: Level of Significance and p-Values 18
6. Prediction Using Simple Linear Regression and Prediction Intervals 19
7. Functional Forms for Simple Linear Regression 20
7.1 The Log-Lin Model 21
7.2 The Lin-Log Model 21
7.3 The Log-Log Model 22
7.4 Selecting the Correct Functional Form 23
Summary 24

Ver 1.0

1. Simple Linear Regression

Financial analysts often need to find whether one variable X can be used to explain another variable Y. Linear regression allows us to examine this relationship.

Suppose an analyst is evaluating the return on assets (ROA) for an industry. He gathers data for six companies in that industry.

Company	ROA (%)
A	6
B	4
C	15
D	20
E	10
F	20

Range = 4% to 20%

Average value = 12.5%

Let Y represents the variable we are trying to explain (in this case ROA), Y_i represents a particular observation, and $\overline{Y}$ represent the mean value. Then the variation of Y can be expressed as:

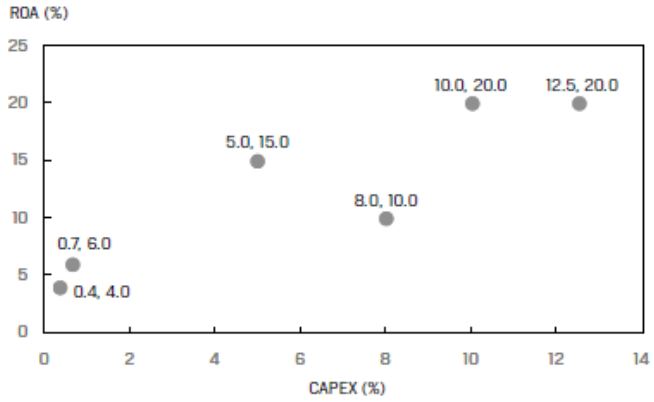
$$\text{Variation of Y} = \sum_{i=1}^{N} (Y_i - \overline{Y})^2$$

The variation of Y is also called **sum of squares total (SST)** or the total sum of squares. Our aim is to understand what explains the variation of Y.

The analyst now wants to check if another variable – CAPEX can be used to explain the variation of ROA. The analyst defines CAPEX as: capital expenditures in the previous period, scaled by the prior period's beginning property, plant, and equipment. He gathers the following data for the six companies:

Company	ROA (%)	CAPEX(%)
A	6	0.7
B	4	0.4
C	15	5.0
D	20	10.0
E	10	8.0
F	20	12.5
Arithmetic mean	12.50	6.10

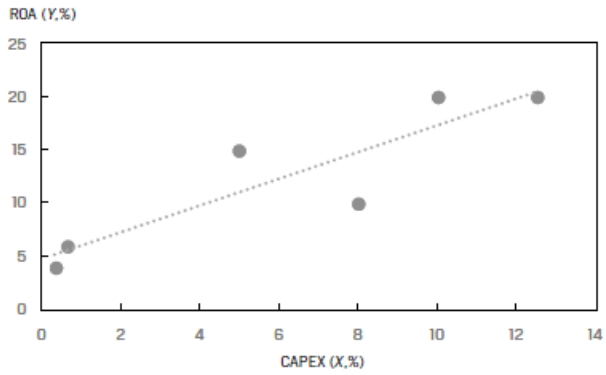
The relation between the two variables can be visualized using a scatter plot.



The variable whose variation we want to explain, also called the **dependent variable** is presented on the vertical axis. It is typically denoted by Y. In our example ROA is the dependent variable.

The explanatory variable, also called the **independent variable** is presented on the horizontal axis. It is typically denoted by X. In our example CAPEX is the independent variable.

A linear regression model computes the best fit line through the scatter plot, which is the line with the smallest distance between itself and each point on the scatter plot. The regression line may pass through some points, but not through all of them.



Regression analysis with only one independent variable is called **simple linear regression (SLR)**. Regression analysis with more than one independent variable is called multiple regression. In this reading we focus on single independent variable, i.e., simple linear regression.

2. Estimating the Parameters of a Simple Linear Regression

2.1 The Basics of Simple Linear Regression

Linear regression assumes a linear relationship between the dependent variable (Y) and independent variable (X). The regression equation is expressed as follows:

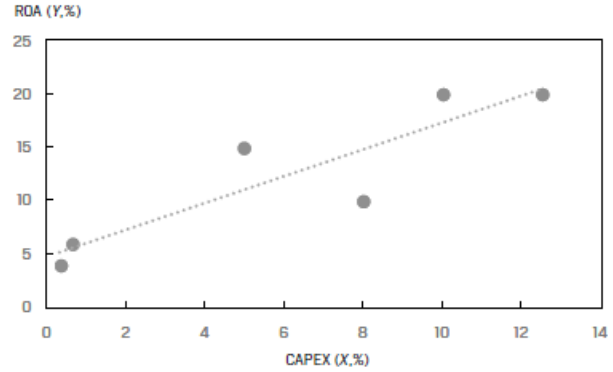
$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

where:
i = 1, …, n
Y = dependent variable
b_0 = intercept
b_1 = slope coefficient
X = independent variable
ε = error term
b_0 and b_1 are called the **regression coefficients.**

The equation shows how much Y changes when X changes by one unit.

2.2 Estimating the Regression Line

Linear regression chooses the estimated values for intercept $\widehat{b_0}$ and slope $\widehat{b_1}$ such that the **sum of the squared errors (SSE)**, i.e., the vertical distances between the observations and the regression line is minimized.



This is represented by the following equation. The error terms are squared so that they don't cancel out each other. The objective of the model is that the sum of the squared error terms should be minimized.

$$\text{SSE} = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{N} \left[Y_i - (\hat{b}_0 + \hat{b}_1 X_i)\right]^2$$

The slope coefficient is calculated as:

$$\hat{b}_1 = \frac{\text{Covairance of Yand X}}{\text{Variance of X}} = \frac{\frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{n-1}}{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{n-1}} = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$$

Note: This formula is similar to correlation which is calculated as:

$$r = \frac{\text{Covairance of Yand X}}{(\text{Std dev of X})(\text{Std dev of Y})}$$

Once we calculate the slope, the intercept can be calculated as:

$$\hat{b}_0 = \overline{Y} - \hat{b}_1\overline{X}$$

Note: On the exam you will most likely be given the values of b_1 and b_0. It is unlikely that you will be asked to calculate these values. Nevertheless, the following table shows how to calculate the slope and intercept.

Company	ROA(Y_i)	CAPEX (X_i)	$(Y_i - \overline{Y})^2$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})(X_i - \overline{X})$
A	6.0	0.7	42.25	29.16	35.10
B	4.0	0.4	72.25	32.49	48.45
C	15.0	5.0	6.25	1.21	-2.75
D	20.0	10.0	56.25	15.21	29.25
E	10.0	8.0	6.25	3.61	-4.75
F	20.0	12.5	56.25	40.96	48.00
Sum	75.0	36.6	239.50	122.64	153.30
Mean	12.5	6.100			

Slope coefficient: $\hat{b}_1 = \frac{153.30}{122.64} = 1.25$

Intercept: $\hat{b}_0 = 12.5 - (1.25 \text{ x } 6.10) = 4.875$

The regression model can thus be expressed as:

$$Y_i = 4.875 + 1.25X_i + \varepsilon_i$$

2.3 Interpreting the Regression Coefficients

The intercept is the value of the dependent variable when the independent variable is zero. In our example, if a company makes no capital expenditures, its expected ROA is 4.875%.

The slope measures the change in the dependent variable for a one-unit change in the independent variable. If the slope is positive, the two variables move in the same direction. If the slope is negative, the two variables move in opposite directions. In our example, if CAPEX increases by one unit, ROA increases by 1.25%.

2.4 Cross-Sectional vs. Time-Series Regressions

Regression analysis can be used for two types of data:

- Cross sectional data: Many observations for different companies, asset classes, investment funds etc. for the same time period.
- Time-series data: Many observations from different time periods for the same company, asset class, investment funds etc.

3. Assumptions of the Simple Linear Regression Model

The simple linear regression model is based on the following four assumptions:

1. Linearity: The relationship between the dependent variable, Y, and the independent

variable, X, is linear.

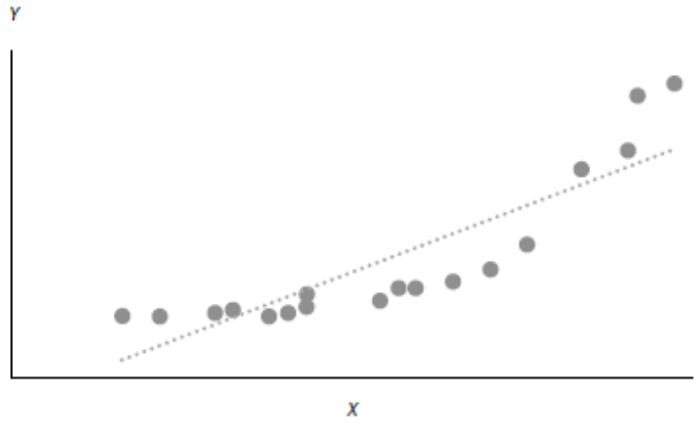
2. Homoskedasticity: The variance of the regression residuals is the same for all observations.
3. Independence: The observations, pairs of Ys and Xs, are independent of one another. This implies the regression residuals are uncorrelated across observations.
4. Normality: The regression residuals are normally distributed.

3.1 Assumption 1: Linearity

Since we are fitting a straight line through a scatter plot, we are implicitly assuming that the true underlying relationship between the two variables is linear.

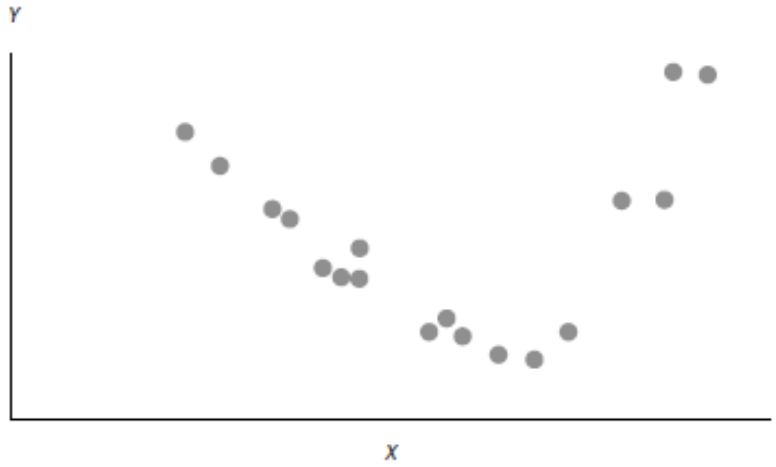
If the relationship between the variables is non-linear, then using a simple linear regression model will produce invalid results.

Exhibit 10 from the curriculum shows two variables that have an exponential relationship. A linear regression line does not fit this relationship well. At lower and higher values of X, the model underestimates Y. Whereas, at middle values of X, the model overestimates Y.



Another point related to this assumption is that the independent variable X should not be random. If X is random, there would be no linear relationship between the two variables.

Also, the residuals of the model should be random. They should not exhibit a pattern when plotted against the independent variable. As shown in Exhibit 11, the residuals from the linear regression model in Exhibit 10 do not appear random.

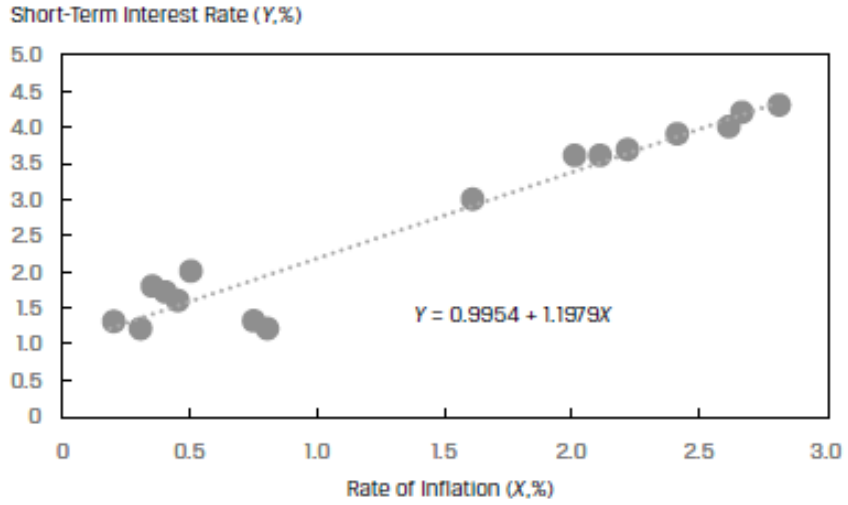


Hence, a linear regression model should not be used for these two variables.

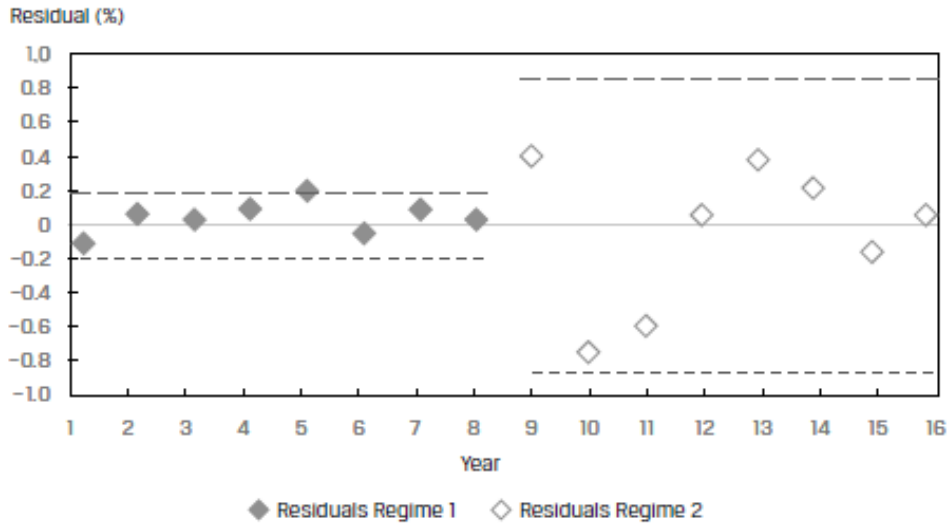
3.2 Assumption 2: Homoskedasticity

Assumption#2 is that the variance of the residuals is constant for all observations. This condition is called homoskedasticity. If the variance of the error term is not constant, then it is called heteroskedasticity.

Exhibit 12 shows a scatter plot of short-term interest rates versus inflation rate for a country. The data represents a total span of 16 years. We will refer to the first eight years of normal rates as Regime 1 and the second eight years of artificially low rates as Regime 2.



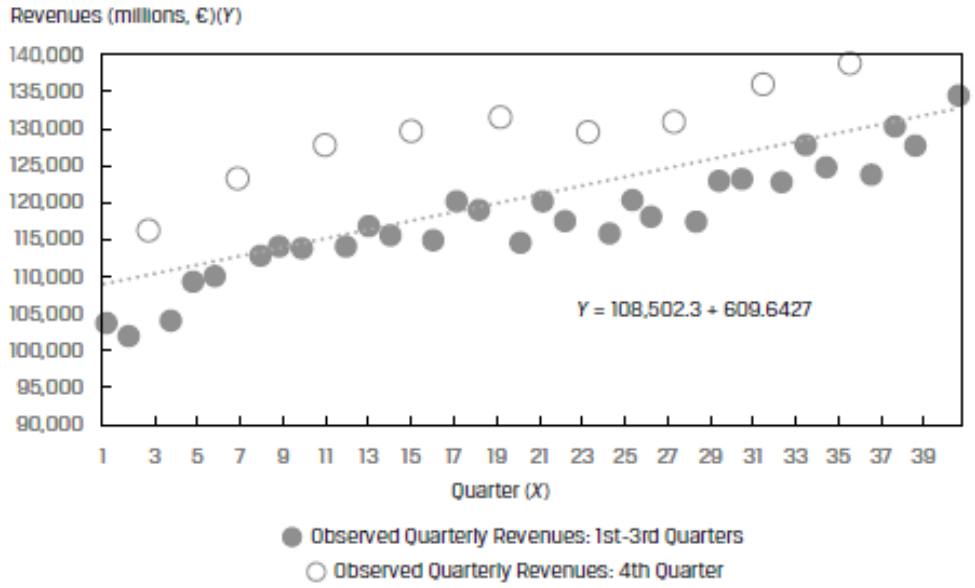
At first glance, the model seems to fit the data well. However, when we plot the residuals of the model against the years, we can see that the residuals for the two regimes appear different. This is shown in Exhibit 13 below. The variation in residuals for Regime 1 is much smaller than the variation in residuals for Regime 2. This is a clear violation of the homoskedasticity assumption and the two regimes should not be clubbed together in a single model.



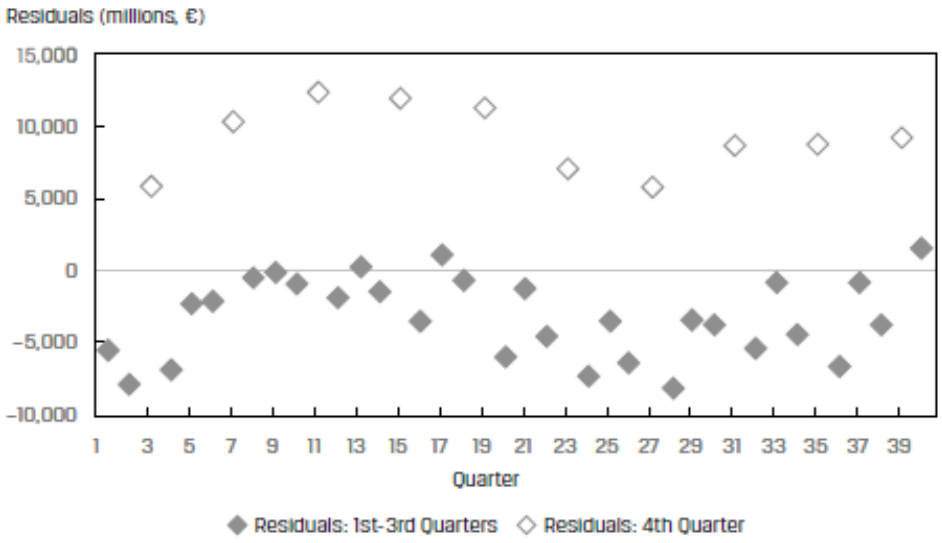
3.3 Assumption 3: Independence

Assumption# 3 states that the residuals should be uncorrelated across observations. If the residuals exhibit a pattern, then this assumption will be violated.

For example, Exhibit 15 from the curriculum plots the quarterly revenues of a company over 40 quarters. The data displays a clear seasonal pattern. Quarter 4 revenues are considerably higher than the first 3 quarters.



A plot of the residuals from this model in Exhibit 16 also helps us see this pattern. The residuals are correlated – they are high in Quarter 4 and then fall back in the other quarters.



Both exhibits show that the assumption of residual independence is violated and the model should not be used for this data.

3.4 Assumption 4: Normality

Assumption#4 states that the residuals should be normally distributed.

Instructor's Note: This assumption does not mean that X and Y should be normally distributed, it only means that the residuals from the model should be normally distributed.

4. Analysis of Variance

4.1 Breaking down the Sum of Squares Total into Its Components

To evaluate how well a linear regression model explains the variation of Y we can break down the total variation in Y (SST) into two components: Sum of square errors (SSE) and regression sum of squares (RSS).

Total sum of squares (**SST**), $\sum_{i=1}^{N} (Y_i - \overline{Y})^2$: This measures the total variation in the dependent variable. SST is equal to the sum of squared distances between the actual values of Y and the average value of Y.

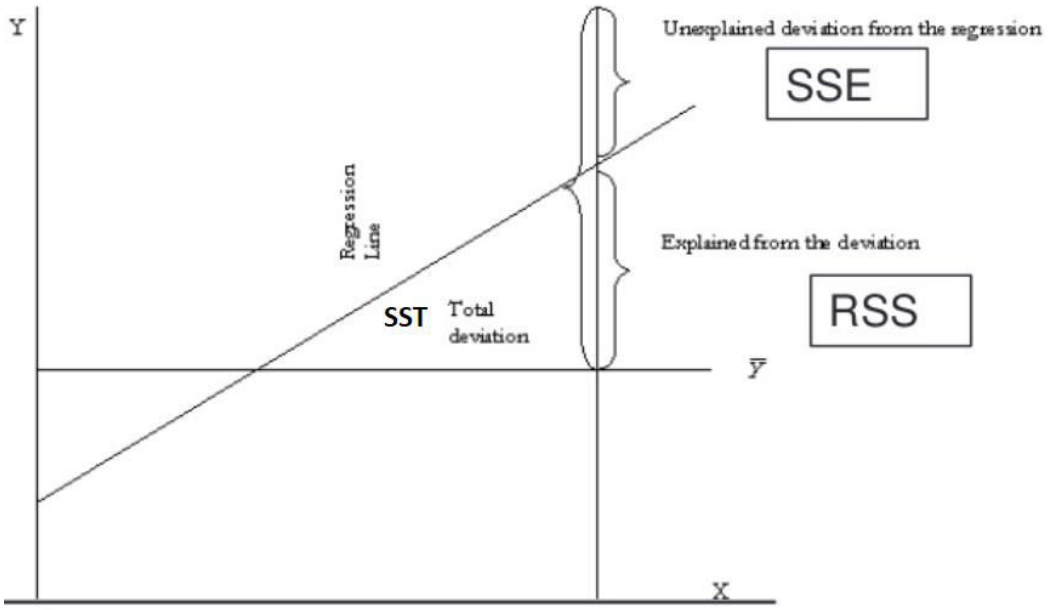
The **regression sum of squares** (**RSS**), $\sum_{i=1}^{N} (\hat{Y}_i - \overline{Y})^2$. This is the amount of total variation in Y that is explained in the regression equation. RSS is the sum of squared distances between the predicted values of Y and the average value of Y.

The **sum of squared errors** or residuals (**SSE**), $\sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2$.This is also known as residual sum of squares. It measures the unexplained variation in the dependent variable. SSE is the sum of the vertical distances between the actual values of Y and the predicted values of Y on the regression line.

SST = RSS + SSE

i.e., Total variation = explained variation + unexplained variation

These concepts are illustrated in the figure below:



4.2 Measures of Goodness of Fit

Goodness of fit indicates how well the regression model fits the data. Several measures are used to evaluate the goodness of fit.

Coefficient of determination: The coefficient of determination, denoted by R^2, measures the fraction of the total variation in the dependent variable that is explained by the independent variable.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Regression sum of squares (RSS)}}{\text{Sum of squares total (SST)}}$$

Characteristics of coefficient of determination, R^2:

The higher the R^2, the more useful the model. R^2 has a value between 0 and 1. For example, if R^2 is 0.03, then this explains only 3% of the variation in the independent variable and the explanatory power of the model is low. However, if R^2 is 0.85, the model explains over 85% of the variation and the explanatory power of the model is high.

It tells us how much better the prediction is by using the regression equation rather than just $\bar{y}$ (average value) to predict y.

With only one independent variable, R^2 is the square of the correlation between X and Y.

The correlation, r, is also called the "multiple-R".

F-test: For a meaningful regression model, the slope coefficients should be non-zero. This is

determined through the F-test which is based on the F-statistic. The F-statistic tests whether all the slope coefficients in a linear regression are equal to 0. In a regression with one independent variable, this is a test of the null hypothesis H_0: $b_1 = 0$ against the alternative hypothesis H_a: $b_1 \neq 0$.

The F-statistic also measures how well the regression equation explains the variation in the dependent variable. The four values required to construct the F-statistic for null hypothesis testing are:

- The total number of observations (n)
- The total number of independent variables (k)
- The regression sum of squares (RSS)
- The sum of squared errors or residuals (SSE)

The F-statistic is calculated as:

$$F = \frac{\frac{RSS}{k}}{\frac{SSE}{n-(k+1)}} = \frac{\text{Mean square regression}}{\text{Mean square error}} = \frac{MSR}{MSE}$$

Interpretation of F-test statistic:

- The higher the F-statistic, the better.
- A high F-statistic implies that the regression model does a good job of explaining the variation in the dependent variable.
- A low F-statistic implies that the regression model does not do a good job of explaining the variation in the dependent variable.
- An F-statistic of 0 indicates that the independent variable does not explain variation in the dependent variable.

4.3 ANOVA and Standard Error of Estimate in Simple Linear Regression

Analysis of variance or ANOVA is a statistical procedure of dividing the total variability of a variable into components that can be attributed to different sources. We use ANOVA to determine the usefulness of the independent variable or variables in explaining variation in the dependent variable.

ANOVA table

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-statistic
Regression (explained variation)	k	RSS	$MSR = \frac{RSS}{k}$	$F = \frac{MSR}{MSE}$
Error (unexplained variation)	n - 2	SSE	$MSE = \frac{SSE}{n-k-1}$	

Total variation	n – 1	SST		

n represents the number of observations and k represents the number of independent variables. With one independent variable, k = 1. Hence, MSR = RSS and MSE = SSE / (n - 2).

Information from the ANOVA table can be used to compute the standard error of estimate (SEE). The **standard error of estimate (SEE)** measures how well a given linear regression model captures the relationship between the dependent and independent variables. It is the standard deviation of the prediction errors. A low SEE implies an accurate forecast.

The formula for the standard error of estimate is given below:

$$\text{Standard error of estimate (SEE)} = \sqrt{\text{MSE}}$$

A low SEE implies that the error (or residual) terms are small and hence the linear regression model does a good job of capturing the relationship between dependent and independent variables.

The following example demonstrates how to interpret an ANOVA table.

Example: Using ANOVA Table Results to Evaluate a Simple Linear Regression
(*This is based on Example 5 from the curriculum.*)

You are provided the following ANOVA table:

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression	576.1485	1	576.1485
Error	1,873.5615	98	19.1180
Total	2,449.7100		

1. What is the coefficient of determination for this regression model?
2. What is the standard error of the estimate for this regression model?
3. At a 5% level of significance, do we reject the null hypothesis of the slope coefficient equal to zero if the critical F-value is 3.938?
4. Based on your answers to the preceding questions, evaluate this simple linear regression model.

Solution to 1: Coefficient of determination (R^2) = RSS / SST = 576.148/2,449.71 = 23.52%.

Solution to 2: Standard error of estimate (SEE) $= \sqrt{\text{MSE}} = \sqrt{19.1180} = 4.3724$

Solution to 3:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{576.1485}{19.1180} = 30.1364$$

Since the calculated F-stat is higher than the critical value of 3.938, we can conclude that the slope coefficient is statistically different from 0.

Solution to 4: The coefficient of determination indicates that the model explains 23.52% of

the variation in Y. Also, the F-stat confirms that the model's slope coefficient is statistically different from 0. Overall, the model seems to fit the data reasonably well.

5. Hypothesis Testing of Linear Regression Coefficients

5.1 Hypothesis Tests of the Slope Coefficient

In order to test whether an estimated slope coefficient is statistically significant, we use hypothesis testing.

Continuing with our previous example of a simple linear regression with ROA as the dependent variable and CAPEX as the independent variable. Suppose we want to test whether the slope coefficient of CAPEX is different from zero.

The steps are:

Step 1: State the hypothesis

H_0: $b_1 = 0$; H_a: $b_1 \neq 0$

Step 2: Identify the appropriate test statistic

To test the significance of individual slope coefficients we use the t-statistic. It is calculated by subtracting the hypothesized population slope (B_1) from the estimated slope coefficient ($\widehat{b_1}$) and then dividing this difference by the standard error of the slope coefficient, $s_{\widehat{b_1}}$:

$$t = \frac{\widehat{b_1} - B_1}{s_{\widehat{b_1}}} = \frac{\text{Estimated value} - \text{Hypothesized value}}{\text{standard error}}$$

with $n - 2 = 6 - 2 = 4$ degrees of freedom

Step 3: Specify the level of significance

Typically, a 5% level of significance is selected.

Step 4: State the decision rule

Since the alternate hypothesis contains a '≠' sign, this is a two tailed test. For a 5% level of significance, 4 degrees of freedom and a two tailed test the critical t-values are ±2.776.

We reject the null hypothesis if the calculated t-statistic is less than −2.776 or greater than +2.776.

Step 5: Calculate the test statistic

Instructor's Note: On the exam, the value of the standard error will most likely be given to you. You are unlikely to be asked to calculate this value.

Let's say that the standard error given is 0.31

$$t = \frac{\widehat{b_1} - B_1}{s_{\widehat{b_1}}} = \frac{1.25 - 0}{0.31} = 4$$

Instructor's Note: In an extreme case if the value of the standard error is not given to you. It can be calculated as shown below:

The standard error of the slope coefficient is calculated as the ratio of the model's standard error of the estimate (SEE) to the square root of the variation of the independent variable:

$$s_{\widehat{b_1}} = \frac{SEE}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

The slope coefficient is 1.25.

The mean square error is 11.96875.

The variation of CAPEX is 122.640

SEE = $\sqrt{11.96875}$ = 3.459588

$$s_{\widehat{b_1}} = \frac{SEE}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$= \frac{3.459588}{\sqrt{122.640}} = 0.312398$$

Step 6: Make a decision

Reject the null hypothesis of a zero slope. There is sufficient evidence to indicate that the slope is different from zero.

What if we want to test if the slope coefficient is statistically different from 1.0?

Instead of using a hypothesized value of zero for the slope coefficient, what if we want to test if the slope coefficient is statistically different from 1.0?

The hypotheses become H_0: $b_1 = 1$ and H_a: $b_1 \neq 1$.

The calculated t-statistic is:

$$t = \frac{1.25 - 1}{0.31} = 0.8$$

This calculated test statistic falls within the critical values, ±2.776, so we fail to reject the null hypothesis: There is not sufficient evidence to indicate that the slope is different from 1.0.

What if we want to test if the slope coefficient is positive?

The hypotheses become H_0: $b_1 \leq 0$ and H_a: $b_1 > 0$.

The critical values change because this is a one tailed test. For a 5% level of significance, 4 degrees of freedom and a one tailed test the critical t-value is +2.132.

All other steps stay the same.

The calculated test stat is greater than 2.132. Therefore, we can reject the null hypothesis. There is sufficient evidence to indicate that the slope is greater than zero.

Testing the correlation:

We can also use hypothesis testing to test the significance of the correlation between the two variables. The process is the same except that the hypothesis are written and the test statistic is calculated differently.

This is demonstrated using the ROA and CAPEX example. The regression software provided us an estimated correlation of 0.8945. Suppose we want to test if this correlation is statistically different from zero.

The steps are:

Step 1: State the hypothesis

Here the null and alternate hypothesis are: H_0: ρ = 0 versus H_a: ρ ≠ 0.

Step 2: Identify the appropriate test statistic

The test-statistic is calculated as:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

with 6 – 2 = 4 degrees of freedom

Step 3: Specify the level of significance

α = 5%.

Step 4: State the decision rule

Critical t-values = ±2.776.

Reject the null if the calculated t-statistic is less than −2.776 or greater than +2.776.

Step 5: Calculate the test statistic

$$t = \frac{0.8945\sqrt{6-2}}{\sqrt{1-0.8001}} = 4.00131$$

Step 6: Make a decision

Reject the null hypothesis of no correlation. There is sufficient evidence to indicate that the correlation between ROA and CAPEX is different from zero.

In a simple linear regression, the t-statistic used to test the slope coefficient and the t-statistic used to test the correlation will have the same value. And we will get the same results using either hypothesis tests. This is demonstrated in the table below:

	Test of the Slope	**Test of the Correlation**

Null and alternate hypothesis	H_0: $b_1 = 0$ versus H_a: $b_1 \neq 0$	H_0: $\rho = 0$ versus H_a: $\rho \neq 0$.
Critical values based on the level of significance and degrees of freedom	α = 5% and dof =4: CV = ±2.776	α = 5% and dof =4: CV = ±2.776
Test statistic	$t = \frac{\widehat{b_1} - B_1}{s_{\widehat{b_1}}} = \frac{1.25 - 0}{0.312398} = 4.00131$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$ $= \frac{0.8945\sqrt{6-2}}{\sqrt{1-0.8001}} = 4.00131$
Decision	Since calculated t-stat is greater than the critical value, reject the null hypothesis of zero slope.	Since calculated t-stat is greater than the critical value, reject the null hypothesis of zero correlation.

Another feature of simple linear regression is that the F-stat is simply the square of the t-stat for the slope/correlation. For our example, the F-stat is $4.00131^2 = 16.0104$

5.2 Hypothesis Tests of the Intercept

For the ROA regression example, the intercept is 4.875%. Say you want to test if the intercept is statistically greater than 3%. This will be a one-tailed hypothesis test and the steps are:

Step 1: State the hypothesis

H_0: $b_0 \leq 3\%$ versus H_a: $b_0 > 3\%$

Step 2: Identify the appropriate test statistic

To test whether the population intercept is a specific value we can use the following t-stat:

$$t_{\text{intercept}} = \frac{\widehat{b_0} - B_0}{s_{\widehat{b_0}}}$$

with 6 – 2 = 4 degrees of freedom

The standard error of the intercept is calculated as:

$$s_{\widehat{b_0}} = \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Instructor's Note: On the exam, the value of the standard error will most likely be given to you. You are unlikely to be asked to calculate this value.

Step 3: Specify the level of significance

$\alpha = 5\%$.

Step 4: State the decision rule

Critical t-value = 2.132.

Reject the null if the calculated t-statistic is greater than 2.132.

Step 5: Calculate the test statistic

$$s_{\widehat{b_0}} = \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{1}{6} + \frac{6.1^2}{122.64}} = 0.68562$$

$$t_{intercept} = \frac{\widehat{b_0} - B_0}{s_{\widehat{b_0}}} = \frac{4.875 - 3.0}{0.68562} = 2.73475$$

Step 6: Make a decision

Since the calculated t-stat is greater than the critical value, we can reject the null hypothesis and conclude that the intercept is greater than 3%.

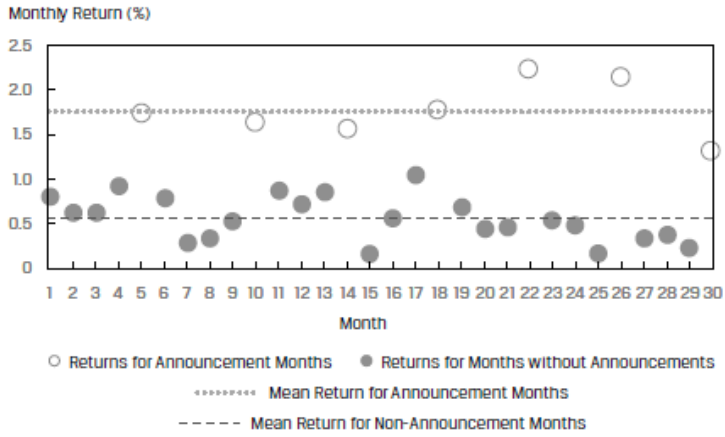
5.3 Hypothesis Tests of Slope When Independent Variable Is an Indicator Variable

An indicator variable or a dummy variable can only take values of 0 or 1. An independent variable is set up as an indicator variable in specific cases. Say we want to evaluate if a company's quarterly earnings announcement influences its monthly stock returns. Here the monthly returns RET would be regressed on the indicator variable, EARN, that takes on a value of 0 if there is no earnings announcement that month and 1 if there is an earnings announcement.

The simple linear regression model can be expressed as:

$$RET_i = b_0 + b_1 EARN_i + \epsilon_i$$

Say we run the regression analysis over a 30-month period. The observations and regression results are shown in Exhibit 28.



Clearly the returns for announcement months are different from the returns for months without announcement.

The results of the regression are given in Exhibit 29.

	Estimated Coefficients	**Standard Error of Coefficients**	**Calculated Test Statistic**
Intercept	0.5629	0.0560	10.0596
EARN	1.2098	0.1158	10.4435

We can draw the following inferences this table:

- The t-stats for both the intercept and the slope are high hence we can conclude that both the intercept and the slope are statistically significant.
- The intercept is the predicted value of Y when X = 0. Therefore, the intercept (0.5629) is the mean of the returns for non-earnings-announcement months.
- The slope coefficient (1.2098) is the difference in means of returns between earnings-announcement and non-announcement months.

5.4 Test of Hypotheses: Level of Significance and p-Values

p-value: At times financial analysts report the *p*-value or probability value for a particular hypothesis. The p-value is the smallest level of significance at which the null hypothesis can be rejected. It allows the reader to interpret the results rather than be told that a certain hypothesis has been rejected or accepted. In most regression software packages, the p-values printed for regression coefficients apply to a test of the null hypothesis that the true parameter is equal to 0 against the alternative that the parameter is not equal to 0, given the estimated coefficient and the standard error for that coefficient.

Here are a few important points connecting t-statistic and p-value:

- Higher the t-statistic, smaller the p-value.
- The smaller the p-value, the stronger the evidence to reject the null hypothesis.
- Given a p-value, if p-value ≤ α, then reject the null hypothesis H_0. α is the significance

level. For example, if we are given a p value of 0.03 or 3%, we can reject the null hypothesis at the 5% significance level, but not at the 1% significance level.

6. Prediction Using Simple Linear Regression and Prediction Intervals

We use regression equations to make predictions about a dependent variable. Let us consider the regression equation: $Y = b_0 + b_1X$. The predicted value of $\widehat{Y} = \widehat{b_0} + \widehat{b_1}X$.

The two sources of uncertainty to make a prediction are:

1. The error term
2. Uncertainty in predicting the estimated parameters b_0 and b_1

The estimated variance of the prediction error is given by:

$$s_f^2 = s^2 * \left[1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{(n-1)s_x^2}\right]$$

Note: You need not memorize this formula, but understand the factors that affect s_f^2, like higher the n, lower the variance, and the better it is.

The estimated variance depends on:

- the squared standard error of estimate, s^2
- the number of observations, n
- the value of the independent variable, X
- the estimated mean $\overline{X}$
- variance, s^2, of the independent variable

Once the variance of the prediction error is known, it is easy to determine the confidence interval around the prediction. The steps are:

1. Make the prediction.
2. Compute the variance of the prediction error.
3. Determine t_c at the chosen significance level α.
4. Compute the (1-α) prediction interval using the formula below:
   $\widehat{Y} \pm t_c * s_f$

In our ROA regression model, if a company's CAPEX is 6%, its forecasted ROA is:

$$\widehat{Y} = \widehat{b_0} + \widehat{b_1}X = 4.875 + 1.25 \times 6 = 12.375\%$$

Assuming a 5% significance level (α), two sided, with n − 2 degrees of freedom (so, df = 4), the critical values for the prediction interval are ±2.776.

The standard error of the forecast is:

$$s_f = 3.49588 * \sqrt{\left[1 + \frac{1}{6} + \frac{(6 - 6.1)^2}{122.640}\right]} = 3.736912$$

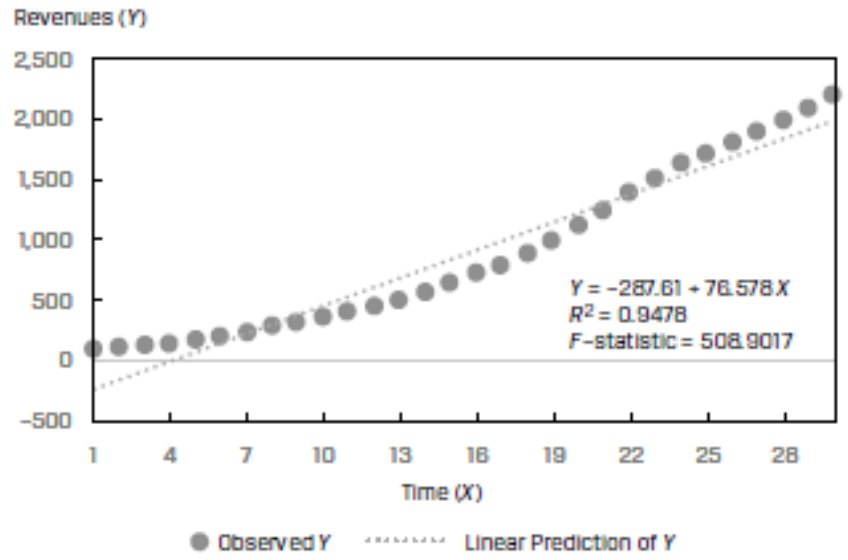
The 95% prediction interval is:

12.375 ± 2.776 (3.736912)

$2.0013 < \hat{Y} < 22.7487$

7. Functional Forms for Simple Linear Regression

Economic and financial data often exhibit non-linear relationships. For example, consider a plot of revenues of a company as the dependent variable (Y) and time as the independent variable (X). Such data will often show exponential growth. Using a simple linear model will not fit this data well as illustrated in Exhibit 33 below.



To make the simple linear regression model fit well, we will have to modify either the dependent or the independent variable. The modification process is called 'transformation' and the different types of transformations are:

- Using the log of the dependent variable
- Using the log of the independent variable
- Using the square of the independent variable
- Differencing the independent variable

In the subsequent sections, we will discuss three commonly used functional forms based on log transformations:

- Log-lin model: The dependent variable is logarithmic but the independent variable is linear.
- Lin-log model: The dependent variable is linear but the independent variable is logarithmic.
- Log-log model: Both the dependent and independent variables are in logarithmic form.

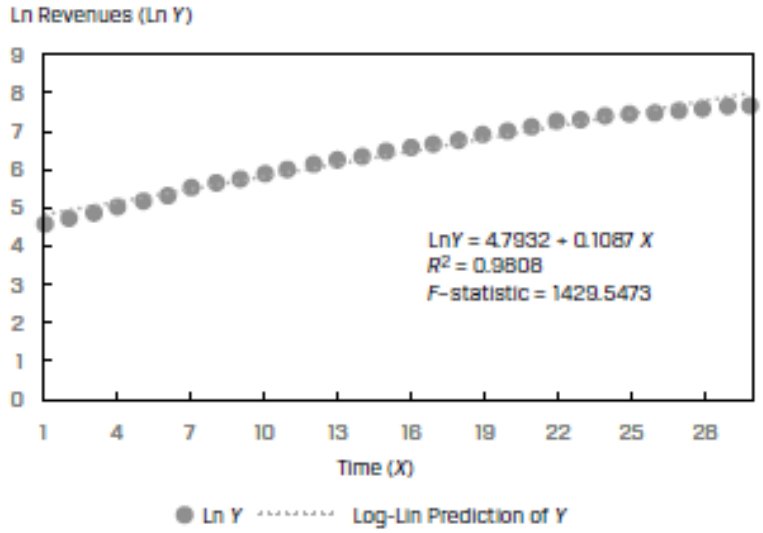
7.1 The Log-Lin Model

In the log-lin model, the dependent variable is logarithmic but the independent variable is linear. The regression equation is expressed as:

$$\ln Y_i = b_0 + b_1 X_i$$

The slope coefficient in this model is the relative change in the dependent variable for an absolute change in the independent variable.

We can transform the Y variable (revenues) in Exhibit 33 into its natural log and then fit the regression line. This is demonstrated in Exhibit 34 below:



We can see that the log-lin model is a better fitting model than the simple linear model for data with exponential growth.

Example: Making forecasts with a log-lin model

Suppose the regression model is: ln Y = −7 + 2X. If X is 2.5% what is the forecasted value of Y?

Solution:

Ln Y = -7 + 2*2.5 = -2

$Y = e^{-2} = 0.135335$

7.2 The Lin-Log Model

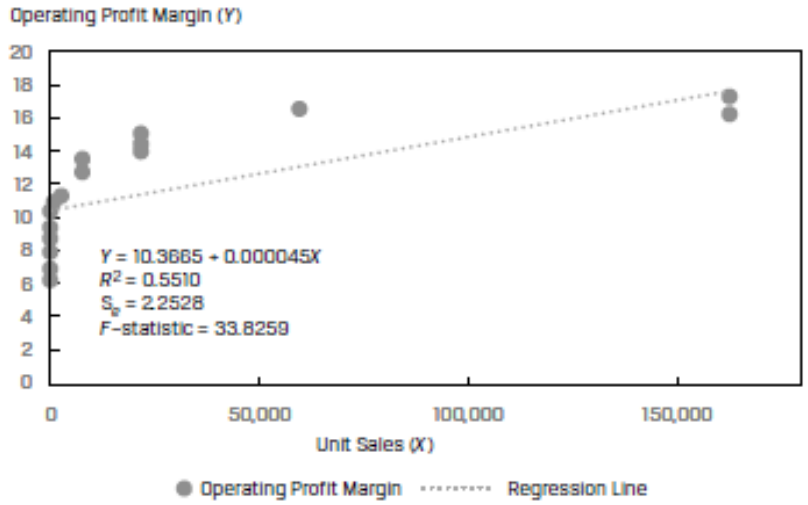
In the lin-log model, the dependent variable is linear but the independent variable is logarithmic. The regression equation is expressed as:

$$Y_i = b_0 + b_1 \ln X_i$$

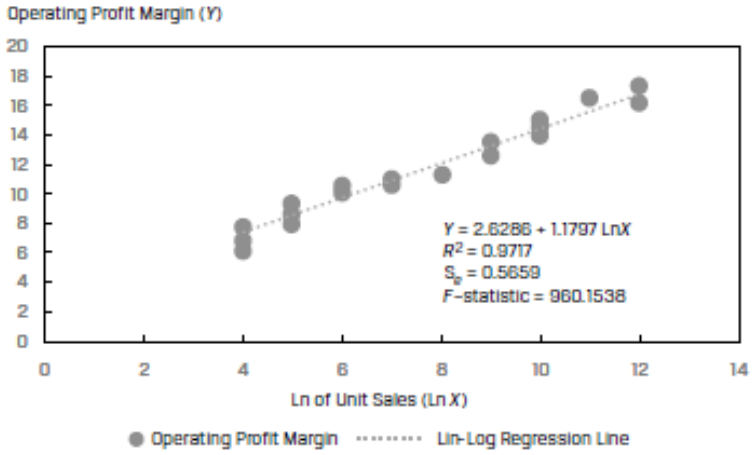
The slope coefficient in this model is the absolute change in the dependent variable for a

relative change in the independent variable.

Consider a pot of operating profit margin as the dependent variable (Y) and unit sales as the independent variable (X). The scatter plot and regression line for a sample of 30 companies is shown in Exhibit 35.



Instead of using the unit sales directly, if we transform the variable and use the natural log of unit sales as the independent variable, we get a much better fit. This is shown in Exhibit 36.



The R^2 of the model jumps to 97.17% from 55.10%. For this data, the lin-log model has a significantly higher explanatory power as compared to simple linear model.

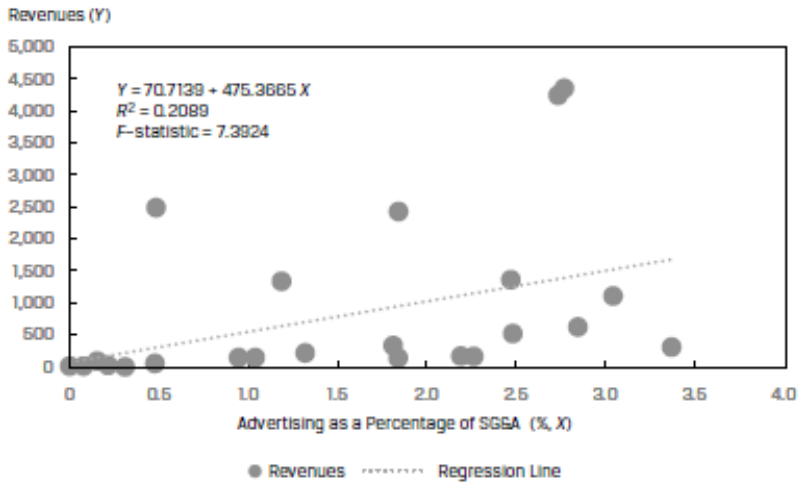
7.3 The Log-Log Model

In the log-log model, both the dependent and independent variables are in logarithmic form. It is also called the 'double-log' model. The regression equation is expressed as:

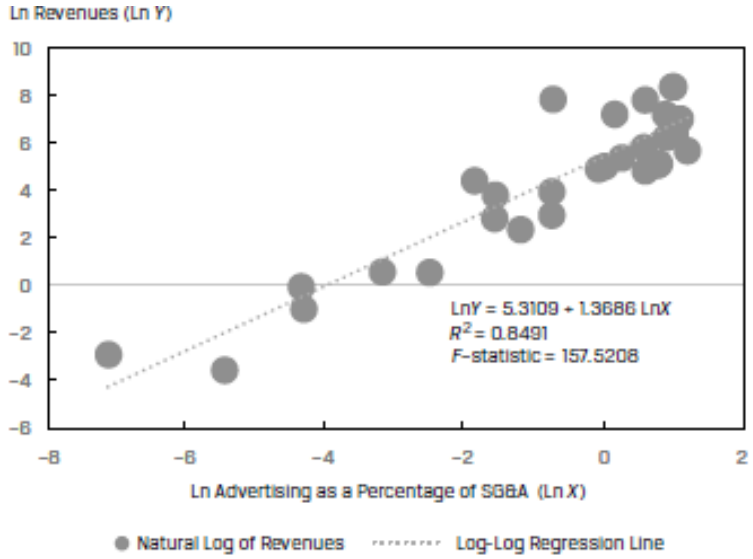
$$\ln Y_i = b_0 + b_1 \ln X_i$$

The slope coefficient in this model is the relative change in the dependent variable for a relative change in the independent variable. The model is often used to calculate elasticities.

Consider a plot of company revenues as the dependent variable (Y) and the advertising spend as a percentage of SG&A, ADVERT, as the independent variable (X). The scatter plot and regression line for a sample company is shown in Exhibit 37 below:



The R^2 of this model is only 20.89%. If we use the natural logs of both the revenues and ADVERT variables, we get a much better fit. This is shown in Exhibit 38.



The R^2 of the model increases by more than 4 times to 84.91% and the F-stat jumps from 7.39 to 157.52. For this data, the log-log model results in a much better fit as compared to the simple linear model.

7.4 Selecting the Correct Functional Form

To select the correct functional form, we can examine the goodness of fit measures:

- Coefficient of determination (R^2)
- F-statistic
- Standard error of estimate (SEE)

A model with a high R^2, high F-stat and low SEE is better.

In addition to these fitness measures, we can also look at the plots of residuals. A good model should show random residuals i.e. the residuals should not be correlated.

Example: Comparing Functional Forms
(*This is based on Example 8 of the curriculum.*)

An analyst is evaluating the relationship between the annual growth in consumer spending (CONS) in a country and the annual growth in the country's GDP (GGDP). He estimates the following two models:

	Model 1	**Model 2**
	$GGDP_i = b_0 + b_1 CONS_i + \varepsilon_i$	$GGDP_i = b_0 + b_1 \ln(CONS_i) + \varepsilon_i$
Intercept	1.040	1.006
Slope	0.669	1.994
R^2	0.788	0.867
SEE	0.404	0.320
F-stat	141.558	247.040

1. Identify the functional form used in these models.
2. Explain which model has better goodness-of-fit with the sample data.

Solution to 1:

Model 1 is a simple linear regression with no transformations. Model 2 is a lin-log model.

Solution to 2:

Model 2 has a higher R^2, higher F-stat, and a lower SEE as compared to Model 1. Therefore Model 2 has a better goodness-of-fit with the sample data.

Summary

LO.a: Describe a simple linear regression model and the roles of the dependent and independent variables in the model.

A linear regression model computes the best fit line through the scatter plot, which is the line with the smallest distance between itself and each point on the scatter plot.

The variable whose variation we want to explain is called the dependent variable.

The explanatory variable is called the independent variable.

LO.b: Describe the least squares criterion, how it is used to estimate regression coefficients, and their interpretation.

Linear regression chooses the estimated values for intercept $\widehat{b_0}$ and slope $\widehat{b_1}$ such that the sum of the squared errors (SSE), i.e., the vertical distances between the observations and the regression line is minimized.

The intercept is the value of the dependent variable when the independent variable is zero.

The slope measures the change in the dependent variable for a one-unit change in the independent variable. If the slope is positive, the two variables move in the same direction. If the slope is negative, the two variables move in opposite directions.

LO.c: Explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated.

The simple linear regression model is based on the following four assumptions:

1. Linearity: The relationship between the dependent variable, Y, and the independent variable, X, is linear.
2. Homoskedasticity: The variance of the regression residuals is the same for all observations.
3. Independence: The observations, pairs of Ys and Xs, are independent of one another. This implies the regression residuals are uncorrelated across observations.
4. Normality: The regression residuals are normally distributed.

LO.d: Calculate and interpret the coefficient of determination and the F-statistic in a simple linear regression.

The coefficient of determination, denoted by R^2, measures the fraction of the total variation in the dependent variable that is explained by the independent variable.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Regression sum of squares (RSS)}}{\text{Sum of squares total (SST)}}$$

The F-statistic tests whether all the slope coefficients in a linear regression are equal to 0. In a regression with one independent variable, this is a test of the null hypothesis H_0: $b_1 = 0$

against the alternative hypothesis H_a: $b_1 \neq 0$.

The F-statistic is calculated as:

$$F = \frac{\frac{RSS}{k}}{\frac{SSE}{n-(k+1)}} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}} = \frac{MSR}{MSE}$$

LO.e: Describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the standard error of estimate in a simple linear regression.

Analysis of variance or ANOVA is a statistical procedure of dividing the total variability of a variable into components that can be attributed to different sources.

Source of variation	Degrees of freedom	Sum of squares	Mean sum of squares	F-statistic
Regression (explained variation)	k	RSS	$MSR = \frac{RSS}{k}$	$F = \frac{MSR}{MSE}$
Error (unexplained variation)	n - 2	SSE	$MSE = \frac{SSE}{n-k-1}$	
Total variation	n – 1	SST		

The standard error of estimate (SEE) measures how well a given linear regression model captures the relationship between the dependent and independent variables. It is the standard deviation of the prediction errors. A low SEE implies an accurate forecast.

$$\text{Standard error of estimate (SEE)} = \sqrt{MSE}$$

LO.f: Formulate a null and an alternative hypothesis about a population value of a regression coefficient, and determine whether the null hypothesis is rejected at a given level of significance.

In order to test whether an estimated slope coefficient is statistically significant, we use hypothesis testing.

The steps are:

Step 1: Define the null and alternate hypothesis. Typically used definitions are: H_0: $b_1 = 0$; H_a: $b_1 \neq 0$

Step 2: Specify the level of significance. Typically, a 5% level of significance is selected. Since the alternate hypothesis contains a '≠' sign, this is a two tailed test. For a 5% level of significance, and a two tailed test the critical t-values are ±2.776.

Step 3: Compute the test statistic.

The t-statistic is calculated by subtracting the hypothesized population slope (B_1) from the estimated slope coefficient ($\widehat{b_1}$) and then dividing this difference by the standard error of the slope coefficient, $s_{\widehat{b_1}}$:

$$t = \frac{\widehat{b_1} - B_1}{s_{\widehat{b_1}}} = \frac{\text{Estimated value} - \text{Hypothesized value}}{\text{standard error}}$$

The standard error of the slope coefficient is calculated as the ratio of the model's standard error of the estimate (SEE) to the square root of the variation of the independent variable:

$$s_{\widehat{b_1}} = \frac{SEE}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}}$$

Step 4: Make a decision. Compare the absolute value of the t-statistic with the critical value and make a decision to accept or reject the null hypothesis. For a 5% significance level, our decision rule is: reject the null hypothesis if the calculated t-statistic is less than −2.776 or greater than +2.776.

LO.g: Calculate and interpret the predicted value for the dependent variable, and a prediction interval for it, given an estimated linear regression model and a value for the independent variable.

The estimated variance of the prediction error is given by:

$$s_f^2 = s^2 * \left[1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{(n-1)s_x^2}\right]$$

Once the variance of the prediction error is known, it is easy to determine the confidence interval around the prediction. The steps are:

1. Make the prediction.
2. Compute the variance of the prediction error.
3. Determine t_c at the chosen significance level α.
4. Compute the (1-α) prediction interval using the formula below:
   $\widehat{Y} \pm t_c * s_f$

LO.h: Describe different functional forms of simple linear regressions.

The three commonly used functional forms based on log transformations:

- Log-lin model: The dependent variable is logarithmic but the independent variable is linear.
- Lin-log model: The dependent variable is linear but the independent variable is logarithmic.
- Log-log model: Both the dependent and independent variables are in logarithmic form.

