Learning Module 10: Simple Linear Regression

LOS 10a: describe a simple linear regression model, how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients

Linear regression is a mathematical method used for analyzing how the variation in one variable can explain the variation in another variable.

Let Y be the variable we wish to explain. As such, the observation of this variable is Y_i, and $\bar{Y}$ is the mean of the sample size n. The variation of Y is given by:

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

Our main objective is to explain what causes this variation, usually called the sum of squares total (SST).

By definition of the regression, we need to explain the variation of Y with another variable. Let X be the explanatory variable. As such, the observations of X will be denoted by X_i and $\bar{X}$ sample mean of size n. The variation of X is given by:

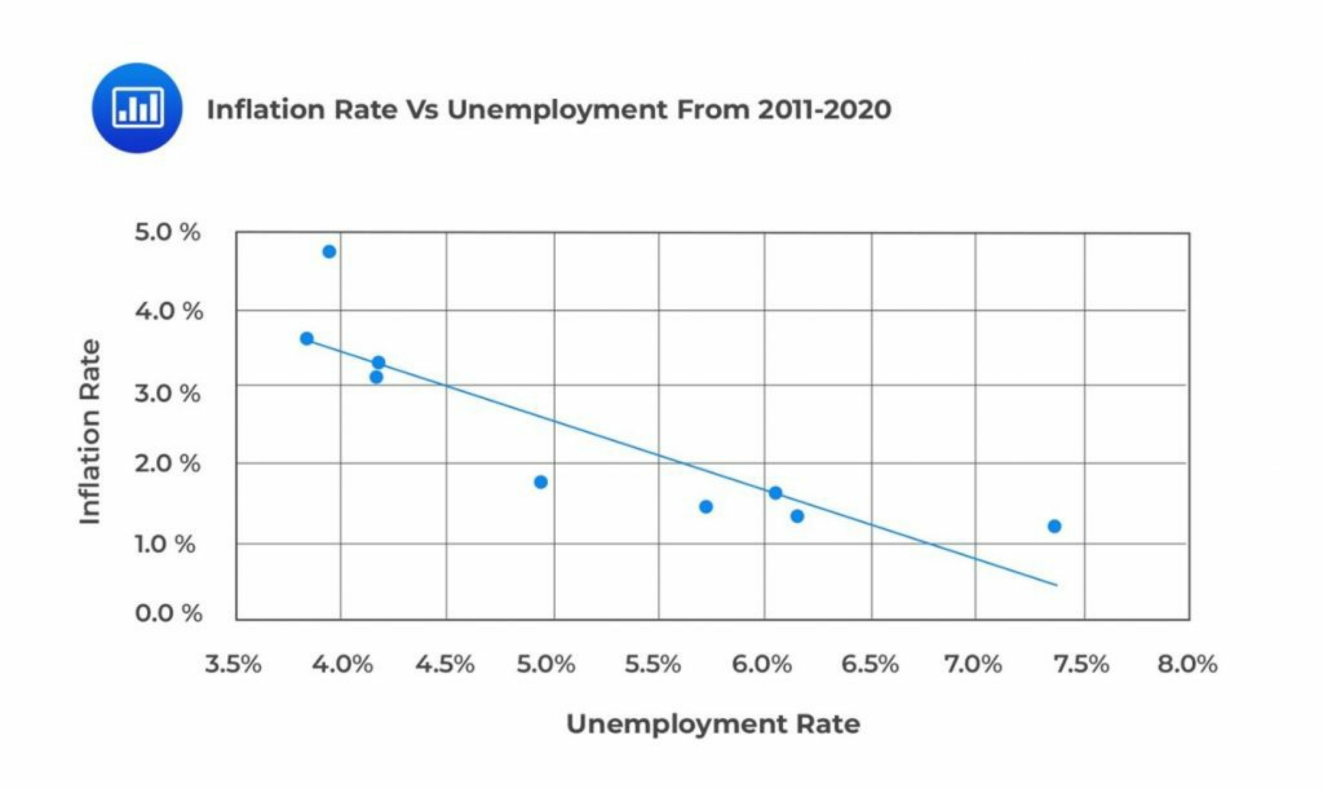
$$\text{Variation of } X = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

To visualize the relationship between variables X and Y, you can use a scatter plot, also known as a scattergram. In this type of plot, the variable you want to explain (Y) is usually plotted on the vertical axis. In contrast, the explanatory variable (X) is placed on the horizontal axis to show the relationship between their variations.

For example, consider the following table. We wish to use linear regression analysis to forecast inflation, given unemployment data from 2011 to 2020.

Year	Unemployment Rate	Inflation Rate
2011	6.1%	1.7%
2012	7.4%	1.2%
2013	6.2%	1.3%
2014	6.2%	1.3%
2015	5.7%	1.4%
2016	5.0%	1.8%
2017	4.2%	3.3%
2018	4.2%	3.1%
2019	4.0%	4.7%
2020	3.9%	3.6%

In this scenario, the Y variable is the inflation rate, and the X axis is the unemployment rate. A scatter plot of the inflation rates against unemployment rates from 2011 to 2020 is shown in the following figure.



Dependent and Independent Variables

A dependent variable, often denoted as YYY, is the variable we want to explain. In contrast, an independent variable, typically denoted as XXX, is used to explain variations in the dependent variable. The independent variable is also referred to as the exogenous, explanatory, or predicting variable.

In our example above, the dependent variable is the inflation rate, and the independent variable is the unemployment rate.

To understand the relationship between dependent and independent variables, we estimate a linear relationship, usually a straight line. When there's one independent variable, we use simple linear regression. If there are multiple independent variables, we use multiple regression.

This reading focuses on linear regression.

Least Squares Criterion

In simple linear regression, we assume linear relationships exist between the dependent and independent variables. The aim is to fit a line to the observations of X (X_is) and Y (Y_is) to minimize the squared deviations from the line. To accomplish this, we use the least squares criterion.

The following is a simple linear regression equation:

$$Y = b_0 + b_1 X_1 + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Where:

Y = Dependent variable.

b_0 = Intercept.

b_1 = Slope coefficient.

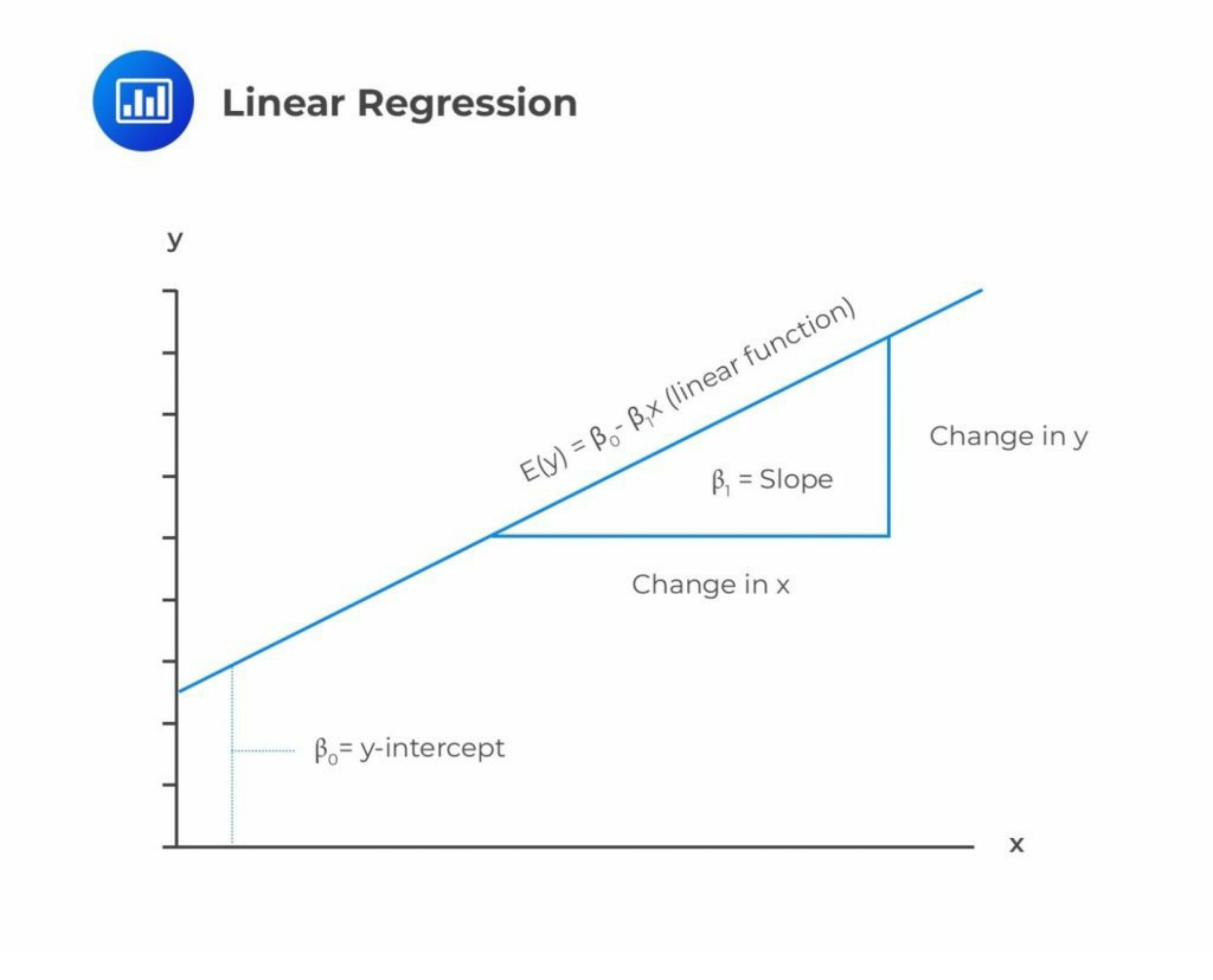
X = Independent variable.

ε = Error term (Noise).

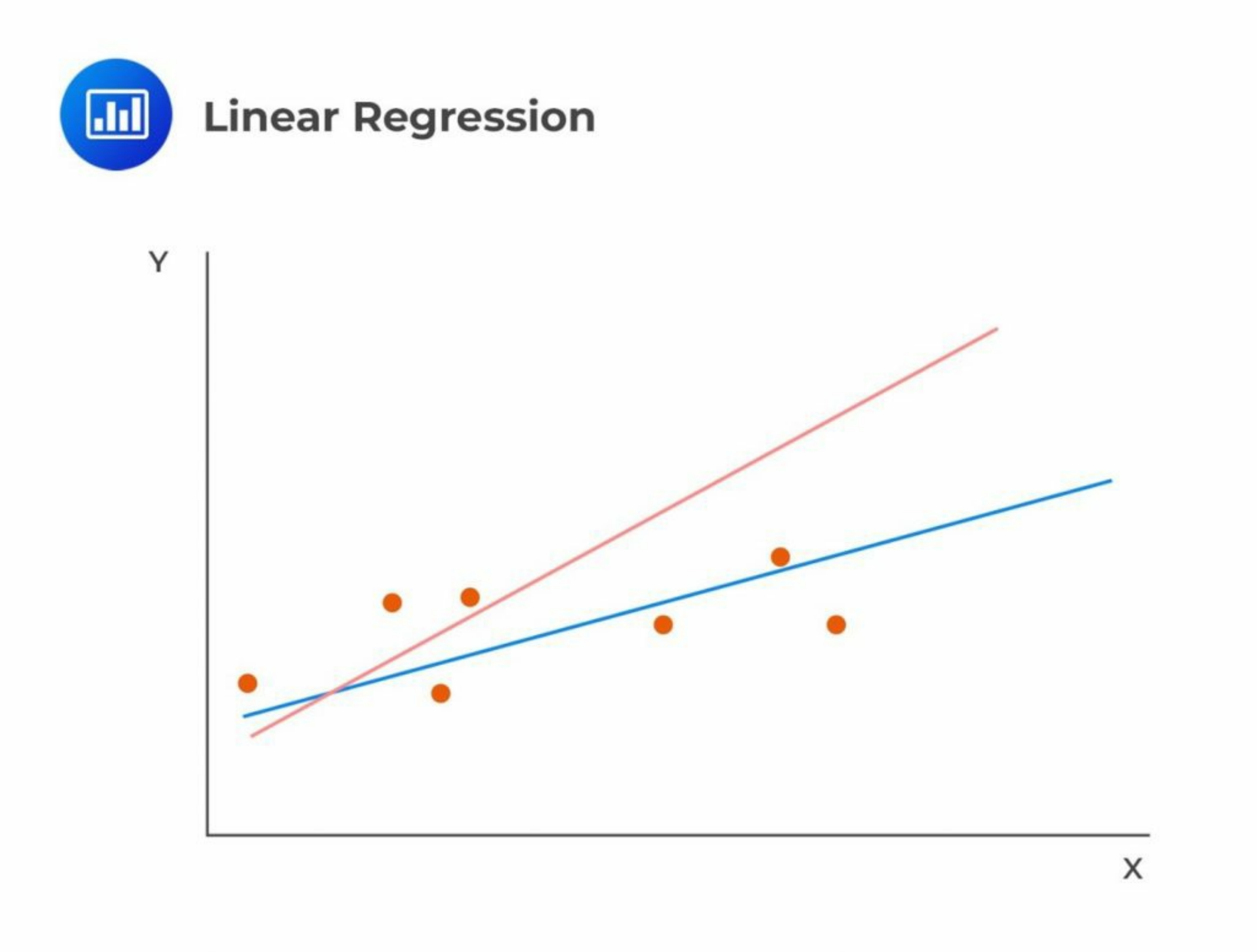
b_0 and b_1 are known as **regression coefficients**. The equation above implies that the dependent is equivalent to the intercept (b_0) plus the product of the slope coefficient (b_1) and the independent variable plus the error term.

The error term is equal to the difference between the observed value of Y and the one expected from the underlying population relation between X and Y

Below is an illustration of a simple linear regression model.



As stated earlier, linear regression calculates a line that best fits the observations. In the following image, the line that best fits the regression is clearly the blue one:



Note that we cannot directly observe the population parameters b_0 and b_1. As such, we observe their estimates, $\hat{b}_0$ and $\hat{b}_1$. They are the estimated parameters of the population using a sample. In simple linear regression, $\hat{b}_0$ and $\hat{b}_1$ are such that the sum of squared vertical distances is minimized.

Specifically, we concentrate on the sum of the squared differences between observations Y_i and the respective estimated value $\hat{Y}_i$ on the regression line, also called the sum of squares error (SSE).

Note that,

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i + e_i^2$$

As such,

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Note that the residual for the ith observation ($e_i = Y_i - \hat{Y}_i$) is different from the error term (ε_i). The error term is based on the underlying population, while the residual term results from regression analysis on a sample.

Conventionally, the sum of the residuals is zero. As such, the aim is to fit the regression line in a simple linear regression that minimizes the sum of squared residual terms.

Estimation and Interpretation of Regression Coefficients

The Slope Coefficient $\hat{\beta}_1$

For a simple linear regression, the slope coefficient is estimated as the ratio of the $Cov(X, Y)$ and $Var(X)$:

$$\hat{\beta}_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{n-1}}{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The slope coefficient is defined as the change in the dependent variable caused by a one-unit change in the value of the independent variable.

The Intercept $\hat{\beta}_0$

The intercept is estimated using the mean of X and Y as follows:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Where:

$\hat{Y}$ = Mean of Y.

$\hat{X}$ = Mean of X.

The intercept is the estimated value of the dependent variable when the independent variable is zero. The fitted regression line passes through the point equivalent to the means of the dependent and the independent variables in a linear regression model.

Example: Estimating Regression Line

Let us consider the following table. We wish to estimate a regression line to forecast inflation given unemployment data from 2011 to 2020.

Year	Unemployment Rate% (X_is)	Inflation Rate% (Y_is)
2011	6.1	1.7
2012	7.4	1.2
2013	6.2	1.3
2014	6.2	1.3
2015	5.7	1.4
2016	5.0	1.8
2017	4.2	3.3
2018	4.2	3.1
2019	4.0	4.7
2020	3.9	3.6

We can create the following table:

Year	Unemployment Rate% (X_is)	Inflation Rate% (Y_is)	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})(X_i - \bar{X})$
2011	6.1	1.7	0.410	0.656	−0.518
2012	7.4	1.2	1.300	4.452	−2.405
2013	6.2	1.3	1.082	0.828	−0.946
2014	6.2	1.3	1.082	0.828	−0.946
2015	5.7	1.4	0.884	0.168	−0.385
2016	5.0	1.8	0.292	0.084	0.157
2017	4.2	3.3	0.922	1.188	−1.046
2018	4.2	3.1	0.578	1.188	−0.828
2019	4.0	4.7	5.570	1.664	−3.044
2020	3.9	3.6	1.588	1.932	−1.751
Sum	52.90	23.4	13.704	12.989	−11.716
Arithmetic Mean	5.29	2.34			

From the table above, we estimate the regression coefficients:

$$\hat{b}_1 = \frac{Cov(X, Y)}{Var(X)} = \frac{\sum_{i=1}^{n} (Y_i - Y)(X_i - X)}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{-11.716}{12.989} = -0.9020$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} = 2.34 - (-0.9020) \times 5.29 = 7.112$$

As such, the regression model is given by:

$$\hat{Y} = 7.112 - 0.9020 X_i + \varepsilon_i$$

From the above regression model, we can note the following:

- The inflation rate is 7.112% if the unemployment rate is 0% (theoretically speaking).
- If the unemployment rate increases (decreases) by one unit–say, from 2% to 3%–the inflation rate decreases(increases) by 0.9020%.

In general,

- If the slope is positive, a unit increase(decrease) in the independent variable results in an increase(decrease) in the dependent variable.
- If the slope is negative, a one-unit increase(decrease) in the independent variable results in a decrease(increase) in the dependent variable.

Furthermore, with the estimated regression model, we can predict the values of the dependent variable based on the value of the independent variable. For instance, if the unemployment rate is 4.5%, then the predicted value of the dependent variable is:

$$\hat{Y} = 7.112 - 0.9020 \times 4.5 = 3.05\%$$

In practice, analysts use statistical functions in software like Excel, statistical tools like R, or programming languages such as Python to perform regression analysis.

Cross-sectional and Time Series Regressions

Regression analysis is commonly used with cross-sectional and time series data. In cross-sectional analysis, you compare X and Y observations from different entities, like various companies in the same time period. For instance, you might analyze the link between a company's R&D spending and its stock returns across multiple firms in a single year.

Time-series regression analysis involves using data from various time periods for the same entity, like a company or an asset class. For instance, an analyst might examine how a company's quarterly dividend payouts relate to its stock price over multiple years.

Question

The independent variable in a regression model is *most likely* the:

A. Predicted variable.

B. Predicting variable.

C. Endogenous variable.

Solution

The correct answer is B.

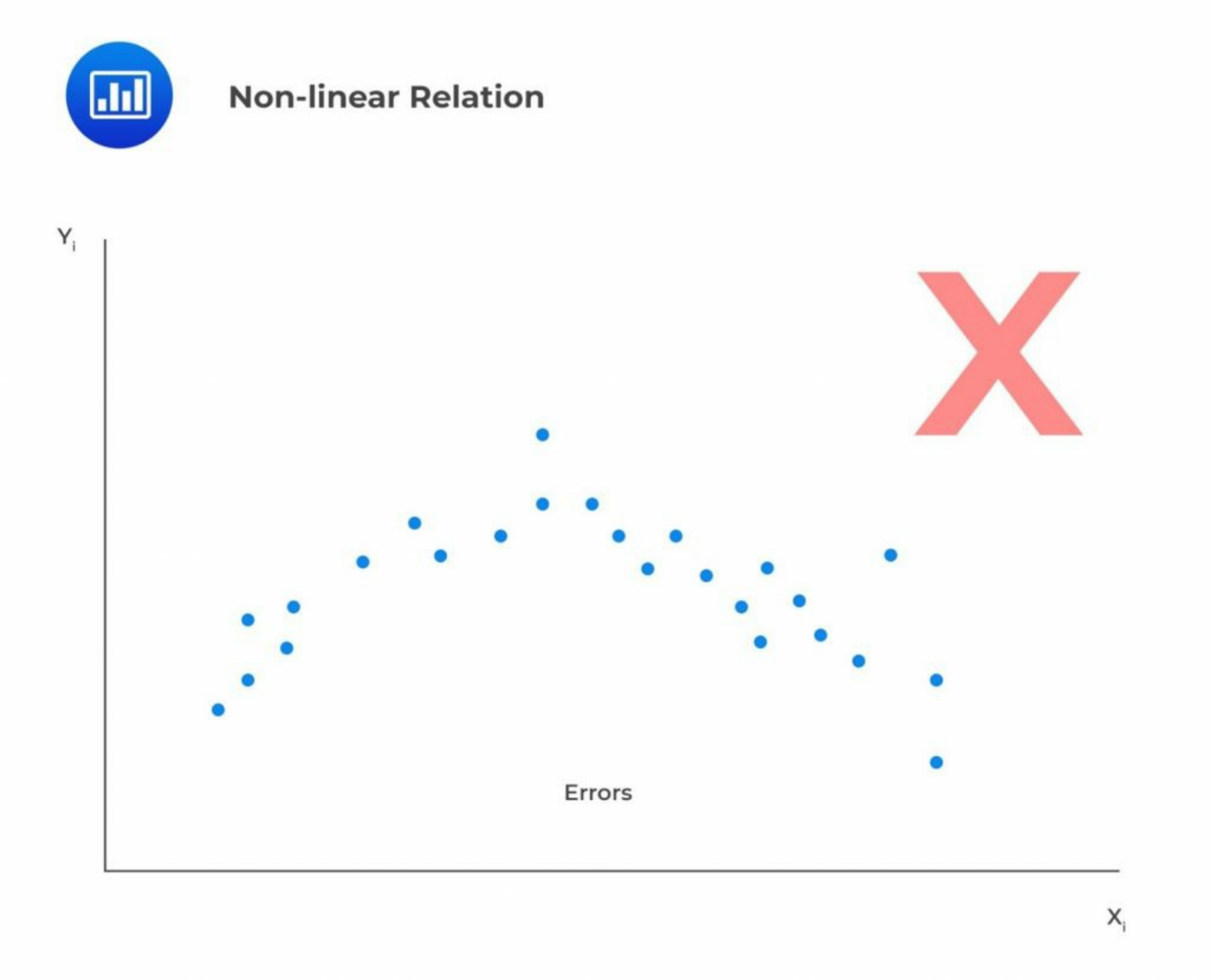
An independent variable explains the variation of the dependent variable. It is also called the explanatory variable, exogenous variable, or the **predicting variable**.

A and C are incorrect. A dependent variable is a variable predicted by the independent variable. It is also known as the **predicted variable**, explained variable, or **endogenous variable**.

LOS 10b: explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated

Assume that we have samples of size n for dependent variable Y and independent variable X. We wish to estimate the simple regression of Y and X. The classic normal linear regression model assumptions are as follows:

I. **Linearity**: A linear relationship implies that the change in Y due to a one-unit change in X is constant, regardless of the value X takes. If the relationship between the two is not linear, the regression model will not accurately capture the trend, resulting in inaccurate predictions. The model will be biased and underestimate or overestimate Y at various points. For example, the model $Y = b_0 + b_1 e^{b_1 x}$ is nonlinear in b_1. For this reason, we should not attempt to fit a linear model between X and Y. It also follows that the independent variable, X, must be non-stochastic (must not be random). A random independent variable rules out a linear relationship between the dependent and independent variables.In addition, linearity means the residuals should not exhibit an observable pattern when plotted against the independent variable. Instead, they should be completely random. In the example below, we're looking at a scenario where the residuals appear to show a pattern when plotted against the independent variable, X. This effectively indicates a nonlinear relation.

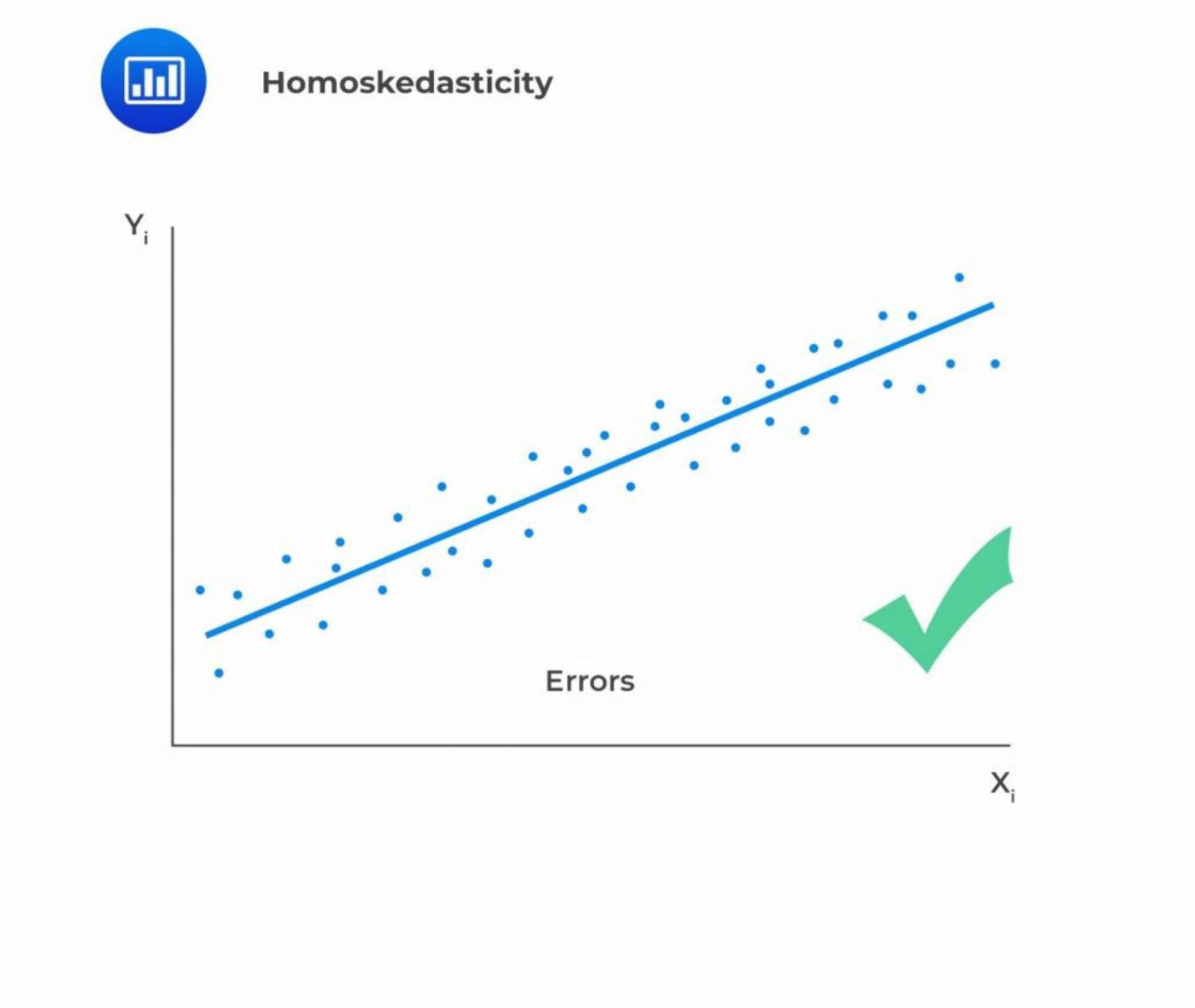


II. **Normality Assumption**: This assumption implies that the error terms (residuals) must follow a normal distribution. It's important to note that this doesn't mean the dependent and independent variables must be normally distributed. However, it's crucial to check the distribution of the dependent and independent variables to identify any outliers. A histogram of the residuals can be used to detect if the error term is normally distributed. A symmetric bell-shaped histogram indicates that the normality assumption is likely to be true.

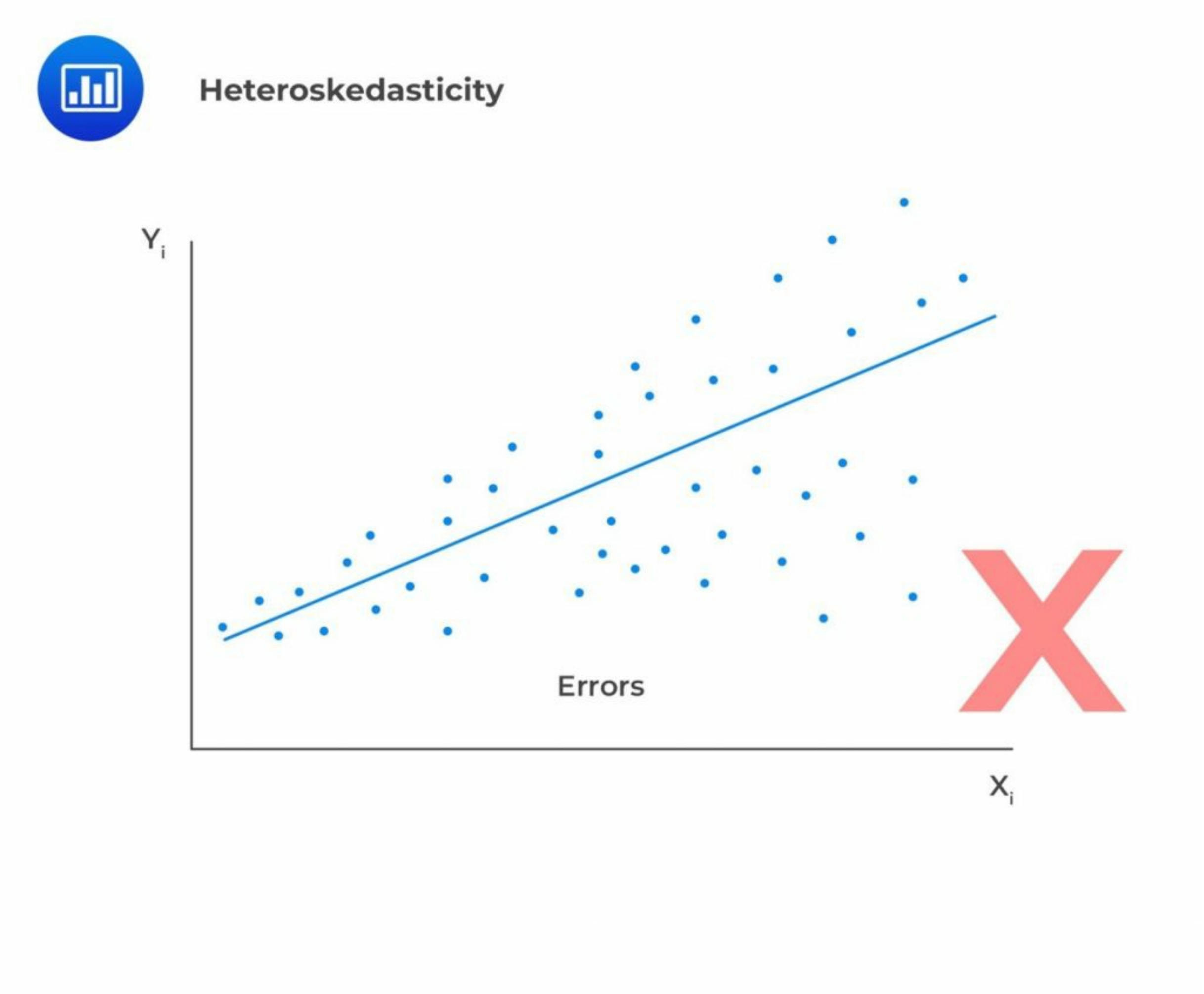
III. **Homoskedasticity**: Homoskedasticity implies that the variance of the error terms is **constant** across all observations. Mathematically, this is expressed as:

$$E\left(\epsilon_i^2\right) = \sigma_\epsilon^2, \quad i = 1, 2, \ldots, n$$

If the variance of residuals varies across observations, then we refer to this as heteroskedasticity (not homoscedasticity). We plot the least square residuals against the independent variable to test for heteroscedasticity. If there is an evident pattern in the plot, that is a manifestation of heteroskedasticity.



In case residuals and the predicted values increase simultaneously, then such a situation is known as **heteroscedasticity** (or heteroskedasticity).



IV. **Independence Assumption**: The independence assumption implies that the observations X_i and Y_i are independent of each other. Failure to satisfy this assumption implies the variables are not independent, and thus, residuals will be correlated. To ascertain this assumption, we visually and statistically analyze the residuals to check whether residual shows exhibit a pattern.

Question

A regression model with one independent variable requires several assumptions for valid conclusions. Which of the following statements *most likely* violates those assumptions?

A. The independent variable is random.
B. The error term is distributed normally.
C. There exists a linear relationship between the dependent variable and the independent variable.

Solution

The correct answer is A.

Linear regression assumes that the independent variable, X, is NOT random. This ensures that the model produces the correct estimates of the regression coefficients.

B is incorrect. The assumption that the error term is distributed normally allows us to easily test a particular hypothesis about a linear regression model.

C is incorrect. Essentially, the assumption that the dependent and independent variables have a linear relationship is the key to a valid linear regression. If the parameters of the dependent and independent variables are not linear, then the estimation of that relation can yield invalid results.

LOS 10c: calculate and interpret measures of fit and formulate and evaluate tests of fit and of regression coefficients in a simple linear regression

The sum of Squares Total (SST) and Its Components

The sum of Squares Total (total variation) is a measure of the total variation of the dependent variable. It is the sum of the squared differences of the **actual** y-value and **mean** of y-observations.

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

The Sum of Squares Total contains two parts:

i. The Sum of Square Regression (SSR).

ii. The sum of Squares Error (SSE).

i. **The sum of Squares Regression (SSR)**: The sum of squares regression measures the **explained** variation in the dependent variable. It is given by the sum of the squared differences of the **predicted** y-value $\hat{Y}_i$, and **mean** of y-observations, $\bar{Y}$:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

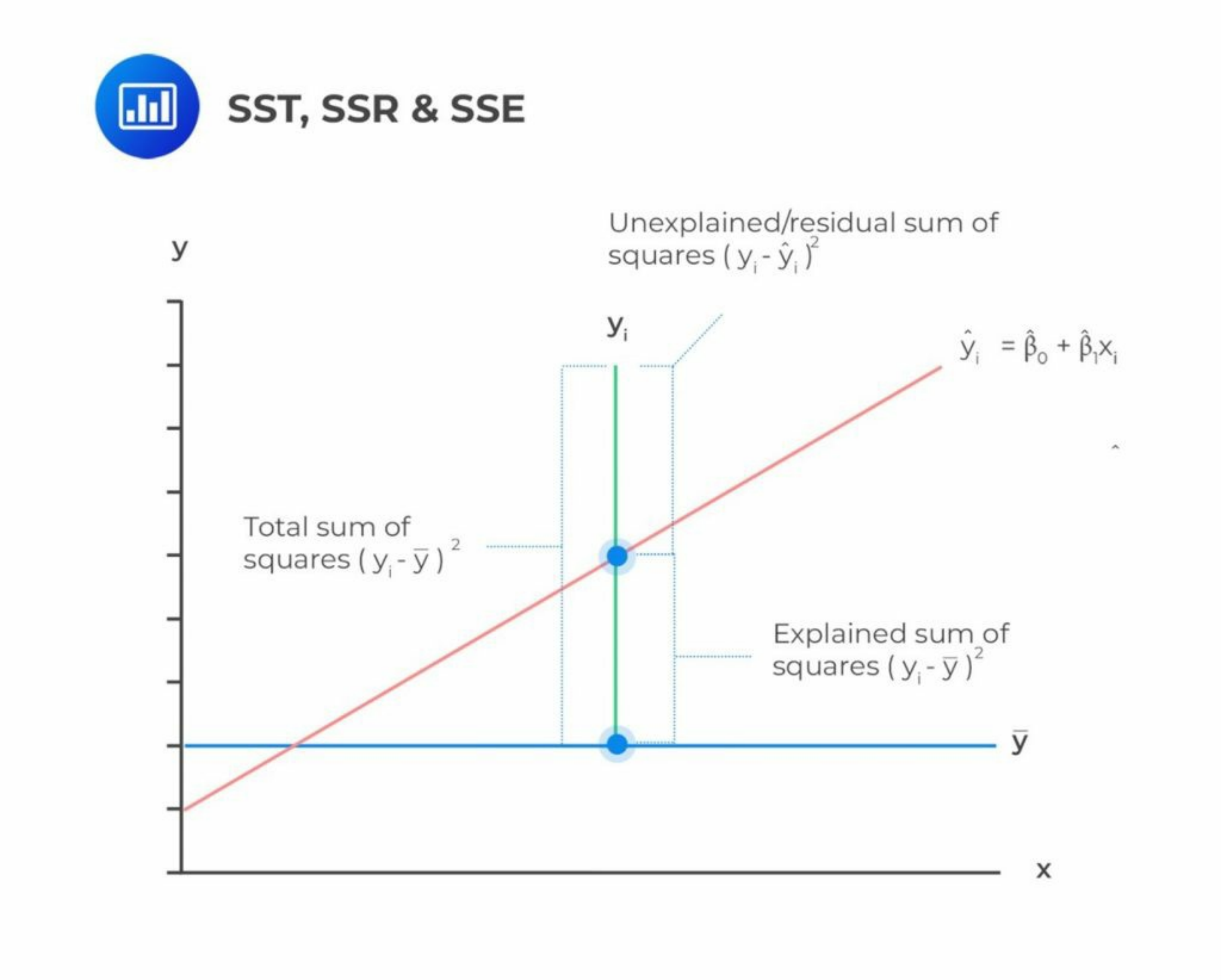
ii. **The Sum of Squared Errors (SSE)**: The sum of squared errors is also called the residual sum of squares. It is defined as the variation of the dependent variable **unexplained** by the independent variable. SSE is given by the sum of the squared differences of the **actual** y-value (Y_i) and the **predicted** y-values, $\hat{Y}_i$.

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

Therefore, the sum of squares total is given by:

$$\begin{aligned}\text{Sum of Squares Total} &= \text{Explained Variation} + \text{Unexplained Variation}\\ &= \text{SSR} + \text{SSE}\end{aligned}$$

The components of the total variation are shown in the following figure.


SST, SSR & SSE
y
Unexplained/residual sum of squares $(y_i - \hat{y}_i)^2$
y_i
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Total sum of squares $(y_i - \bar{y})^2$
Explained sum of squares $(y_i - \bar{y})^2$
$\bar{y}$
x

For example, consider the following table. We wish to use linear regression analysis to forecast inflation, given unemployment data from 2011 to 2020.

Year	Unemployment Rate (%)	Inflation Rate (%)
2011	6.1	1.7
2012	7.4	1.2
2013	6.2	1.3
2014	6.2	1.3
2015	5.7	1.4
2016	5.0	1.8
2017	4.2	3.3
2018	4.2	3.1
2019	4.0	4.7
2020	3.9	3.6

Remember that we had estimated the regression line to be $\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$. As such, we can create the following table:

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
2013	6.2	1.3	1.520	1.082	0.048	0.673
2014	6.2	1.3	1.520	1.082	0.048	0.673
2015	5.7	1.4	1.971	0.884	0.326	0.136
2016	5.0	1.8	2.602	0.292	0.643	0.069
2017	4.2	3.3	3.324	0.922	0.001	0.967
2018	4.2	3.1	3.324	0.578	0.050	0.967
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

From the table above, we can calculate the following:

$$\text{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 13.704$$

$$\text{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = 10.568$$

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 3.136$$

Measures of Goodness of Fit

We use the following measures to analyze the goodness of fit of simple linear regression:

I. Coefficient of determination.

II. F-statistic for the test of fit.

III. Standard error of the regression.

Coefficient of Determination

The coefficient of determination (R^2) measures the proportion of the total variability of the dependent variable explained by the independent variable. It is calculated using the formula below:

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{Sum of Squares Regression (SSR)}}{\text{Sum of Squares Total (SST)}}$$
$$= \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

Intuitively, we can think of the above formula as:

$$R^2 = \frac{\text{Total Variation} - \text{Unexplained Variation}}{\text{Total Variation}}$$
$$= \frac{\text{Sum of Squares Total (SST)} - \text{Sum of Squared Errors (SSE)}}{\text{Sum of Squares Total}}$$

Simplifying the above formula gives:

$$R^2 = 1 - \frac{\text{Sum of Squared Errors (SSE)}}{\text{Sum of Squares Total (SST)}}$$

In the above example, the coefficient of determination is:

Explained Variation

$$
\begin{aligned}
R^2 &= \frac{\text{Explained Variation}}{\text{Total Variation}} \\
&= \frac{\text{Sum of Squares Regression (SSR)}}{\text{Sum of Squares Total (SST)}} \\
&= \frac{10.568}{13.794} = 76.61\%
\end{aligned}
$$

Features of Coefficient of Determination (R^2)

R^2 lies between 0% and 100%. A high R^2 explains variability better than a low R^2. If R^2=1%, only 1% of the total variability can be explained. On the other hand, if R^2=90%, over 90% of the total variability can be explained. In a nutshell, the higher the R^2, the higher the model's explanatory power.

For simple linear regression (R^2) is calculated by squaring the correlation coefficient between the dependent and the independent variables:

$$
r^2 = R^2 = \left(\frac{Cov(X, Y)}{\sigma_X \sigma_Y}\right)^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
$$

Where:

($Cov(X, Y)$) = Covariance between two variables, X and Y.

(σ_X) = Standard deviation of X.

(σ_Y) = Standard deviation of Y.

Example: Calculating Coefficient of Determination (R^2)

An analyst determines that ($\sum_{i=1}^{6} (Y_i - \bar{Y})^2 = 13.704$) and ($\sum_{i=1}^{6} (Y_i - \hat{Y}_i)^2 = 3.136$) from the regression analysis of inflation rates on unemployment rates. The coefficient of determination ((R^2)) is *closest to*:

Solution

$$
\text{Sum of Squares Total (SST)} - \text{Sum of Squared Errors (SSE)}
$$

$$R^2 = \frac{\text{Sum of Squares Total (SST)} - \text{Sum of Squared Errors (SSE)}}{\text{Sum of Squares Total (SST)}}$$

$$= \frac{\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y})^2\right)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{13.704 - 3.136}{13.704}$$

$$= 0.7712 = 77.12\%$$

F-statistic in Simple Regression Model

Note that the coefficient of determination discussed above is just a descriptive value. To check the statistical significance of a regression model, we use the F-test. The F-test requires us to calculate the F-statistic.

In simple linear regression, the F-test confirms whether the slope (denoted by (b_1)) in a regression model is equal to zero. In a typical simple linear regression hypothesis, the null hypothesis is formulated as: $(H_0 : b_1 = 0)$ against the alternative hypothesis $(H_1 : b_1 \neq 0)$. The null hypothesis is rejected if the confidence interval at the desired significance level excludes zero.

The Sum of Squares Regression (SSR) and Sum of Squares Error (SSE) are employed to calculate the F-statistic. In the calculation, the Sum of Squares Regression (SSR) and Sum of Squares Error (SSE) are adjusted for the degrees of freedom.

The Sum of Squares Regression(SSR) is divided by the number of independent variables (k) to get the Mean Square Regression (MSR). That is:

$$MSR = \frac{SSR}{k} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{k}$$

Since we only have $(k = 1)$, in a simple linear regression model, the above formula changes to:

$$MSR = \frac{SSR}{1} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Therefore, in the Simple Linear Regression Model, MSR = SSR.

Also, the Sum of Squares Error (SSE) is divided by degrees of freedom given by $(n - k - 1)$ (this translates to $(n - 2)$ for simple linear regression) to arrive at Mean Square Error (MSE). That is,

$$\text{MSE} = \frac{\text{Sum of Squares Error (SSE)}}{n - k - 1} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2}{n - k - 1}$$

For a simple linear regression model,

$$\text{MSE} = \frac{\text{Sum of Squares Error(SSE)}}{n - 2} = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2}{n - 2}$$

Finally, to calculate the F-statistic for the linear regression, we find the ratio of MSR to MSE. That is,

$$F - \text{statistic} = \frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\text{SSR}}{k}}{\frac{\text{SSE}}{n-k-1}} = \frac{\frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{k}}{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2}{n-k-1}}$$

For simple linear regression, this translates to:

$$F - \text{statistic} = \frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\text{SSR}}{k}}{\frac{\text{SSE}}{n-k-1}} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2}{\frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2}{n-2}}$$

The F-statistic in simple linear regression is F-distributed with (1) and $(n - 2)$ degrees of freedom. That is,

$$\frac{\text{MSR}}{\text{MSE}} \sim F_{1, n-2}$$

Note that the F-test regression analysis is a one-side test, with the rejection region on the right side. This is due to the fact that the objective is to test whether the variation in Y explained (the numerator) is larger than the variation in Y unexplained (the denominator).

Interpretation of F-test Statistic

A large F-statistic value proves that the regression model effectively explains the variation in the dependent variable and vice versa. On the contrary, an F-statistic of 0 indicates that the independent variable does not explain the variation in the dependent variable.

We reject the null hypothesis if the calculated value of the F-statistic is greater than the critical F-value.

It is worth mentioning that F-statistics are not commonly used in regressions with one independent variable. This is because the F-statistic is equal to the square of the t-statistic for the slope coefficient, which implies the same thing as the t-test.

Standard Error of Estimate

Standard Error of Estimate, S_e or SEE, is alternatively referred to as the root mean square error or standard error of the regression. It measures the distance between the observed dependent variables and the dependent variables the regression model predicts. It is calculated as follows:

$$\text{Standard Error of Estimate}\,(S_e) = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}}$$

The standard error of estimate, coefficient of determination, and F-statistic are the measures that can be used to gauge the goodness of fit of a regression model. In other words, these measures tell the extent to which a regression model syncs with data.

The smaller the Standard Error of Estimate is, the better the fit of the regression line. However, the Standard Error of Estimate does not tell us ***how well*** the independent variable explains the variation in the dependent variable.

Hypothesis Tests of Regression Coefficients

Hypothesis Test on the Slope Coefficient

Note that the F-statistic discussed above is used to test whether the slope coefficient is significantly different from 0. However, we may also wish to test whether the population slope differs from a specific value or is positive. To accomplish this, we use the t-distributed test.

The process of performing the t-distributed test is as follows:

1. **State the hypothesis**: For instance, typical hypothesis statements include:
   - $H_0 : b_1 = 0$ versus $H_a : b_1 \neq 0$
   - $H_0 : b_1 \leq 0$ versus $H_a : b_1 > 0$
2. **Identify the appropriate test statistic**: The test statistic for the t-distributed test on slope coefficient is given by:

   $$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}$$

   Where:B_1 = Hypothesized slope coefficient.$\hat{b}_1$ = Point estimate for $b_1$$s_{\hat{b}_1}$ = Standard error of the slope coefficient.The test statistic is t-distributed with $n - k - 1$ degrees of freedom. Since we are dealing with simple linear regression, we will deal with $n - 2$ degrees of freedom. The standard error of the slope coefficient ($s_{\hat{b}_1}$) is calculated as the ratio of the standard error of estimate (s_e) and the square root of the variation of the independent variable:

   $$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

   Where:

   $$s_e = \sqrt{MSE}$$

3. **Specify the level of significance**: Note the level of significance level, usually denoted by alpha, α. A typical significance level might be $\alpha = 5\%$
4. **State the decision rule**: Using the significance level, find the critical values. You can use the t-table or spreadsheets such as Excel, statistical software such as R, or programming languages such as Python. In an exam situation, such critical values will be

provided. Compare the t-statistic value to the critical t-value (t_c). Reject the null hypothesis if the absolute t-statistic value is greater than the upper critical t-value or less than the lower critical value, i.e., $t > +t_{\text{critical}}$ or $t < -t_{\text{critical}}$

5. **Calculate the test statistic**: Using the formula above, calculate the test statistic. Intuitively, you might need to calculate the standard error of the slope coefficient ($s_{\hat{b}_1}$) first.
6. **Make a decision**: Make a decision whether to reject or fail to reject the null hypothesis.

Example: Hypothesis Test Concerning Slope Coefficient

Recall the example where we regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

The estimated regression model is

$$\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$$

Assume that we need to test whether the slope coefficient of the unemployment rates is positive at a 5% significance level.

The hypotheses are as follows:

- $H_0 : b_1 < 0$ versus $H_a : b_1 \geq 0$

Next, we need to calculate the test statistic given by:

- $t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}$

Where:

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Recall that,

$$s_e = \sqrt{MSE} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}} = \sqrt{\frac{3.136}{8}} = 0.6261$$

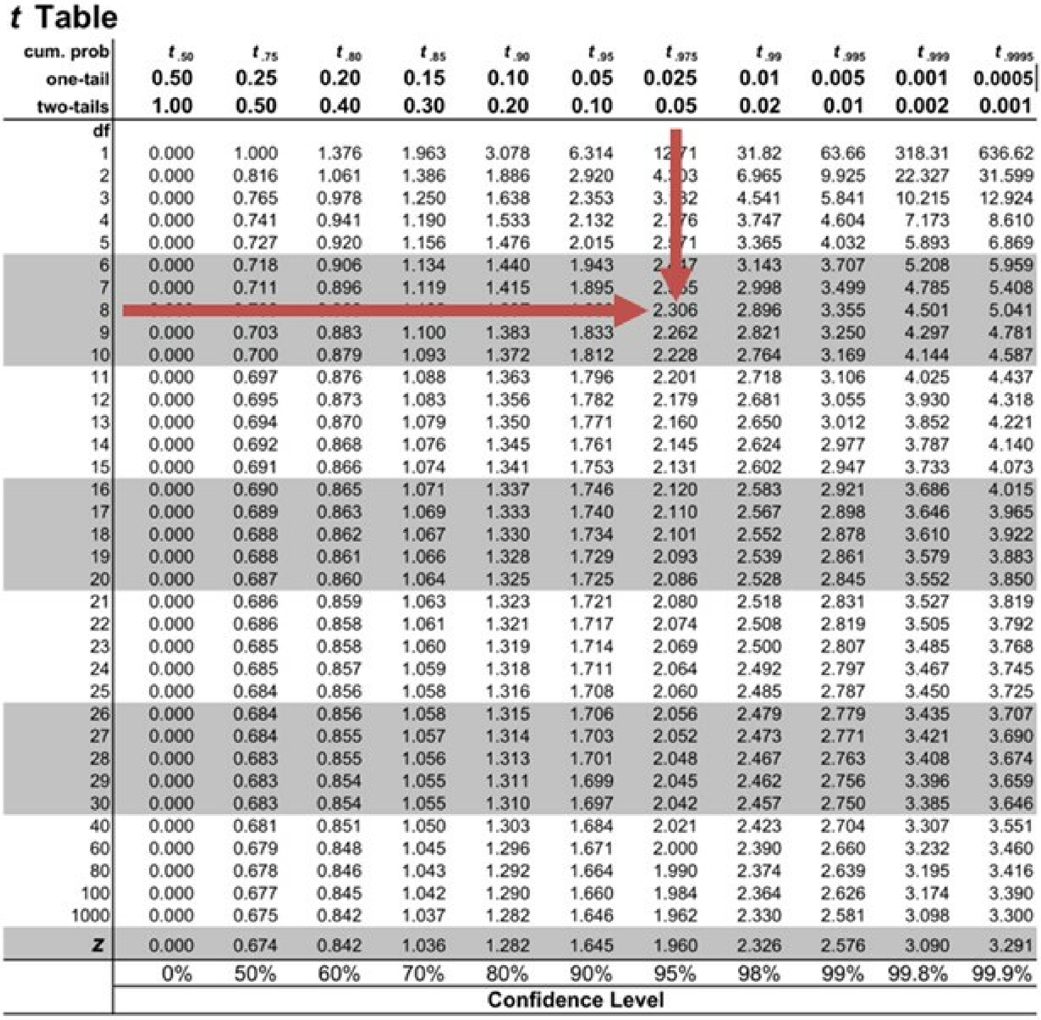
So that,

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{0.6261}{\sqrt{12.989}} = 0.1737$$

Therefore,

$$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}} = \frac{-0.9020 - 0}{0.1737} = -5.193$$

Next, we need to find critical t-values. Note that this is a one-sided test. As such, we need to find t_{8}, 0.05. We will use the t-table:



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

From the table, $t_8, 0.05 = 1.860$. We fail to reject the null hypothesis since the calculated test statistic is less than the critical t-value (?5.193 < 1.860). There is sufficient evidence to indicate that the slope coefficient is not positive.

Relationship between the Hypothesis Test of Correlation and Slope Coefficient

In simple linear regression, a distinct characteristic exists: the t-test statistic checks if the slope coefficient equals zero. This t-test statistic is the same as the test-statistic used to determine if the pairwise correlation is zero.

This feature is true for two-sided tests ($H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ and $H_0 : b_1 = 0$ versus $H_a : \rho \neq 0$) and one-sided test ($H_0 : \rho \leq 0$ versus $Ha : \rho > 0$ and

$H_0 : b_1 \leq 0$ versus $H_a : \rho > 0$ or $H_0 : \rho > 0$ versus $H_a : \rho \leq 0$ and $H_0 : b_1 > 0$ versus $H_a : \rho \leq 0$).

Note that the test -statistic to test whether the correlation is equal to zero is given by:

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

The above test statistic is t-distributed with $(n-2)$ degrees of freedom.

Consider our previous example, where we regressed inflation rates against unemployment rates from 2011 to 2020. Assume we want to test whether the pairwise correlation between the unemployment and inflation rates equals zero.

In the example, the correlation between the unemployment rates and inflation rates is -0.8782. As such, the test- statistic to test whether the correlation is equal to zero is

$$t = \frac{-0.8782\sqrt{10-2}}{\sqrt{1-(-0.8782)^2}} \approx -5.19$$

Note this is equal to the test statistic t-test statistic used to perform the hypothesis test whether the slope coefficient is zero:

$$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}} = \frac{-0.9020 - 0}{0.1737} = -5.193$$

Hypothesis Test of the Intercept Coefficient

Similar to the slope coefficient, we may also want to test whether the population intercept is equal to a certain value. The process is similar to that of the slope coefficient. However, the test statistic for t-distributed test on slope coefficient is given by:

$$t = \frac{\hat{b}_0 - B_0}{s_{\hat{b}_0}}$$

Where:

B_1 = Hypothesized intercept coefficient.

$\hat{b}_1$ = Point estimate for b_1.

$s_{\hat{b}_0}$ = Standard error of the intercept.

The formula for the standard error of the intercept $s_{\hat{b}_0}$ is given by:

$$s_{\hat{b}_0} = \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Recall the example where regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

The estimated regression model is

$$\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$$

Assume that we need to test whether the intercept is greater than 1 at a 5% significance level.

The hypotheses are as follows:

$$H_0 : b_0 \leq 1 \text{ versus } H_a : b_0 > 1$$

Next, we need to calculate the test statistic given by:

$$t = \frac{\hat{b}_0 - B_0}{s_{\hat{b}_0}}$$

Where:

$$s_{\hat{b}_0} = \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{1}{10} + \frac{5.29^2}{12.989}} = 1.501$$

Therefore,

$$t = \frac{7.112 - 1}{1.501} = 4.0719$$

Note that this is a one-sided test. From the table, $t_8, 0.05 = 1.860$. Since the calculated test statistic is less than the critical t-value (4.0179 > 1.860), we reject the null hypothesis. There is sufficient evidence to indicate that the intercept is greater than 1.

Hypothesis Tests Concerning Slope Coefficient When Independent Variable is an Indicator Variable

Dummy variables, also known as indicator variables or binary variables, are used in regression analysis to represent categorical data with two or more categories. They are particularly useful for including qualitative information in a model that requires numerical input variables.

Example: Regression Analysis With Indicator Variables

Assume we aim to investigate if a stock's inclusion in an Environmental, Social, and Governance (ESG) focused fund affects its monthly stock returns. In this case, we'll analyze the monthly returns of a stock over a 48-month period.

We can use a simple linear regression model to explore this. In the model, we regress monthly returns, denoted as R, on an indicator variable, ESG. This indicator takes the value of 0 if the stock isn't part of an ESG-focused fund and 1 if it is.

$$R = b_0 + b_1 ESG + \varepsilon_i$$

Note that we estimate the simple linear regression in a way similar to if the independent variable was continuous.

The intercept β_0 is the predicted value when the indicator variable is 0. On the other hand, the slope when the indicator variable is 1 is the difference in the means if we grouped the observations by the indicator variable.

Assume that the following table is the results of the above regression analysis:

	Estimated Coefficients	Standard Error of Coefficients	Calculated Test Statistic
Intercept	0.5468	0.0456	9.5623
ESG	1.1052	0.1356	9.9532

Additionally, we have the following information regarding the means and variances of the variables.

	Monthly returns of ESG Focused Stocks	Monthly Returns of Non-ESG Stocks	Difference in Means
Mean	1.6520	0.5468	1.1052
Variance	1.1052	0.1356	
Observations	10	38	

From the above tables, we can see that:

- The intercept (0.5468) is equal to the mean of the returns for the non-ESG stocks.
- The slope coefficient (1.1052) is the difference in means of returns between ESG-focused stocks and non-ESG stocks.

Now, assume that we want to test whether the slope coefficient is equal to 0 at a 5% significance level. Therefore, the hypothesis is $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. Note that the degrees of freedom in $48 - 2 = 46$. As such, the critical t-values (usually given in the table above) is $t_{46, 0.025} = \pm 2.013$.

From the first table above, the calculated test statistic for the slope is greater than the critical t-value ($9.9532 > 2.013$). As a result, we reject the null hypothesis that the slope coefficient is equal to zero.

p-Values and Level of Significance

The p-value is the smallest level of significance level at which the null hypothesis is rejected. Therefore, the smaller the p-value, the smaller the probability of rejecting the true null hypothesis (type I error) and, hence, the greater the validity of the regression model.

Software packages commonly offer p-values for regression coefficients. These p-values help test a null hypothesis that the true parameter equals 0 versus the alternative that it's not equal to zero.

We reject the null hypothesis if the p-value corresponding to the calculated test statistic is less than the significance level.

Example: Hypothesis Testing of Slope Coefficients

An analyst generates the following output from the regression analysis of inflation on unemployment:

	Regression Statistics	
	R Square	0.7684
	Standard Error	0.0063
	Observations	10

	Coefficients	Standard Error	t-Stat
Intercept	0.0710	0.0094	7.5160
Forecast (Slope)	−0.9041	0.1755	−5.1516

At the 5% significant level, test the null hypothesis that the slope coefficient is significantly different from one, that is,

$$H_0 : b_1 = 1 \text{ vs. } H_a : b_1 \neq 1$$

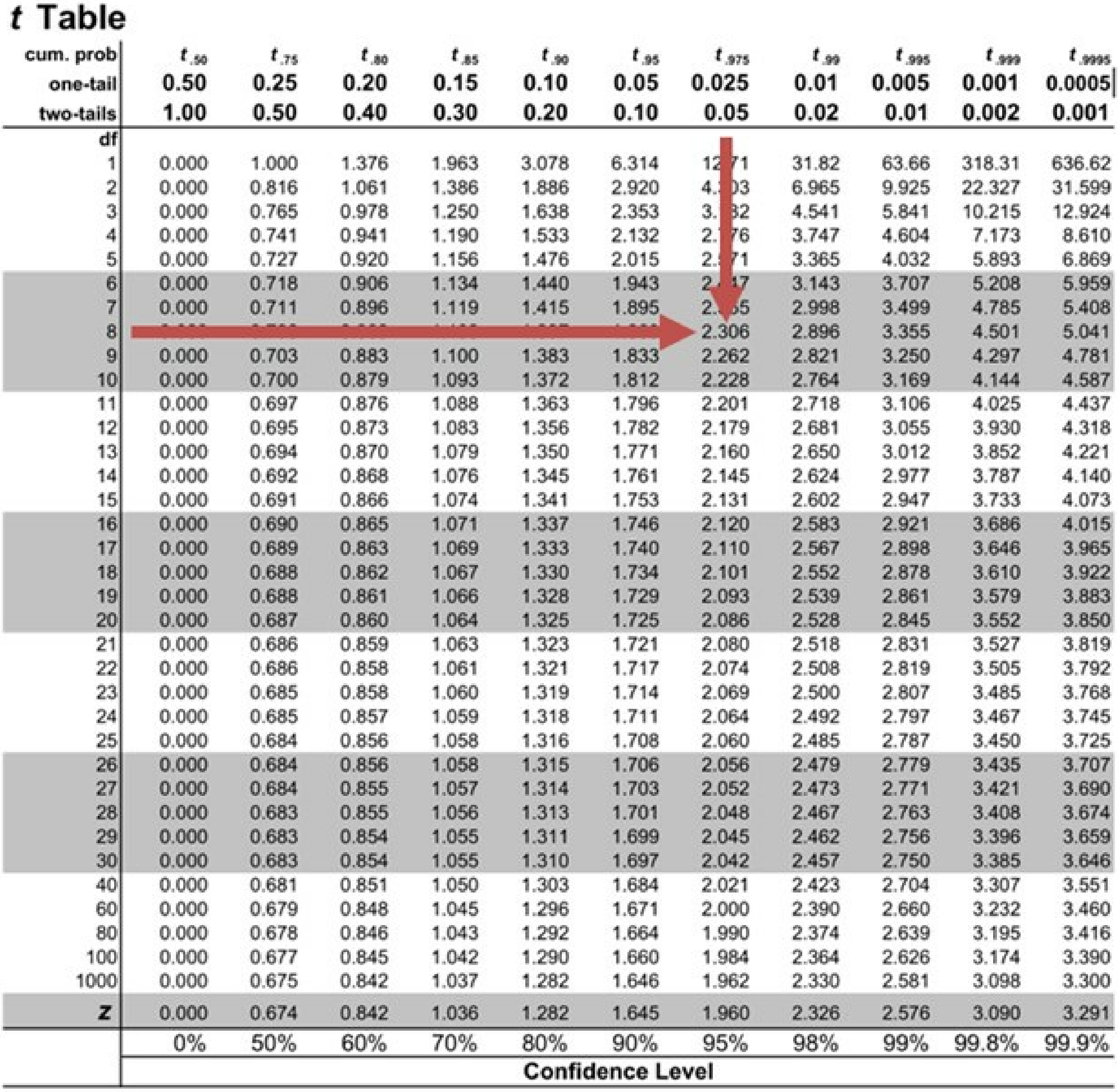
Solution

The calculated t-statistic, $t = \frac{\hat{b}_1 - b_1}{\hat{S}_{b_1}}$ is equal to:

$$t = \frac{-0.9041 - 1}{0.1755} = -10.85$$

The critical two-tail t-values from the table with $n - 2 = 8$ degrees of freedom are:

$$t_c = \pm 2.306$$



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Notice that $|t| > t_c$ i.e., $(10.85 > 2.306)$

Therefore, we reject the null hypothesis and conclude that the estimated slope coefficient is statistically different from one.

Note that we used the confidence interval approach and arrived at the same conclusion.

Question 1

Samantha Lee, an investment analyst, is studying monthly stock returns. She focuses on companies listed in a Renewable Energy Index across various economic conditions. In her analysis, she performed a simple regression. This regression explains how stock returns vary concerning the indicator variable RENEW. RENEW equals 1 when there's a positive policy change towards renewable energy during that month, and 0 if not. The total variation in the dependent variable amounted to 220.34. Of this, 94.75 is the part explained by the model. Samantha's dataset includes 36 monthly observations.

Calculate the coefficient of determination, F-statistic, and standard deviation of monthly stock returns of companies listed in a Renewable Energy Index.

A. R^2=43.00%;F=26.07;Standard deviation=2.51.

B. R^2=53.00%;F=26.41;Standard deviation=2.55.

C. R^2=33.00%;F=36.07;Standard deviation=3.55.

Solution

The correct answer is A.

Coefficient of determination:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{94.75}{220.34} \approx 43\%$$

F-statistic:

$$F = \frac{\frac{\text{Explained variation}}{k}}{\frac{\text{Unexplained variation}}{n-2}} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-2}} = \frac{\frac{94.75}{1}}{\frac{220.34-94.75}{34}} = 26.07$$

Standard deviation:

Note that,

$$\text{Total Variation} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 220.34$$

And the standard deviation is given by:

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}$$

As such,

$$\text{Standard deviation} = \sqrt{\frac{\text{Total variation}}{n-1}} = \sqrt{\frac{220.34}{n-1}} = 2.509$$

Question 2

Neeth Shinu, CFA, is forecasting the price elasticity of supply for a specific product. Shinu uses the quantity of the product supplied for the past 5months as the dependent variable and the price per unit of the product as the independent variable. The regression results are shown below.

Regression Statistics				
R Square	0.9941			
Standard Error	3.6515			
Observations	5			
	Coefficients	Standard Error	t Stat	P-value
Intercept	−159	10.520	(15.114)	0.001
Slope	0.26	0.012	22.517	0.000

Which of the following most likely reports the correct value of the t-statistic for the slope and most accurately evaluates its statistical significance with 95% confidence?

A. $t = 21.67$; the slope is significantly different from zero.

B. $t = 3.18$; the slope is significantly different from zero.

C. $t = 22.57$; the slope is not significantly different from zero.

Solution

The correct answer is A.

The t-statistic is calculated using the formula:

$$t = \frac{\hat{b}_1 - b_1}{\hat{S}_{b_1}}$$

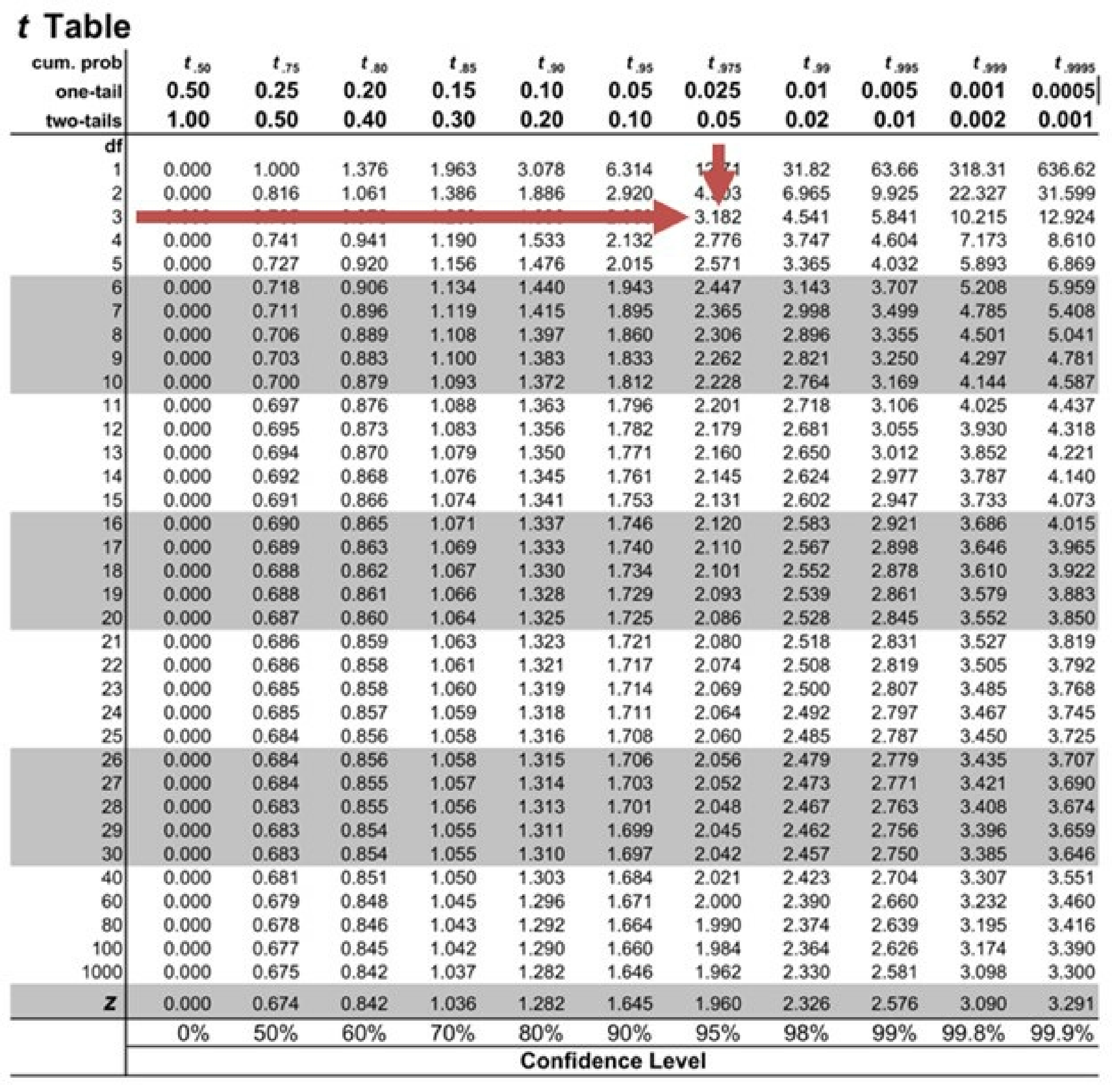
Where:

- b_1 = True slope coefficient.
- $\hat{b}_1$ = Point estimator for B_1.
- $\hat{S}_{b_1}$ = Standard error of the regression coefficient.

$$t = \frac{0.26 - 0}{0.012} = 21.67$$

The critical two-tail t-values from the t-table with $n - 2 = 3$ degrees of freedom are:

$$t_c = \pm 3.18$$



***t* Table**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Notice that $|t| > t_c$ (i.e., 21.67 > 3.18).

Therefore, the null hypothesis can be rejected. Further, we can conclude that the estimated slope coefficient is statistically different from zero.

LOS 10d: 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

The sum of squares of a regression model is usually represented in the Analysis of Variance (ANOVA) table. The ANOVA table contains the sum of squares (SST, SSE, and SSR), the degrees of freedom, the mean squares (MSR and MSE), and F-statistics.

The typical format of ANOVA is as shown below:

Source	Sum of Squares	Degrees of Freedom	Mean square	F-statistic
Regression (Explained)	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Residual (explained)	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$	
Total	$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n - 1$		

Standard Error of EstimateStandard Error of Estimate, S_e or SEE, is referred to as the root mean square error or standard error of the regression. It measures the distance between the observed and dependent variables predicted by the regression model. The Standard Error of Estimate is easily calculated from the ANOVA table using the following formula:

$$\text{Standard Error of Estimate } (S_e) = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}}$$

The standard error of estimate, coefficient of determination, and F-statistic are the measures that can be used to gauge the goodness of fit of a regression model. In other words, these measures are used to tell the extent to which a regression model syncs with data.

The smaller the Standard Error of Estimate is, the better the fit of the regression line. However, the Standard Error of Estimate does not tell us how well the independent variable explains the variation in the dependent variable.

Example: Calculating and Interpreting F-Statistic

The completed ANOVA table for the regression model of the inflation rate against the unemployment rate over 10 years is given below:

Source	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F-Statistic
Regression	10.568	1	10.568	?
Error	3.136	8	0.392	
Total	13.704	9		

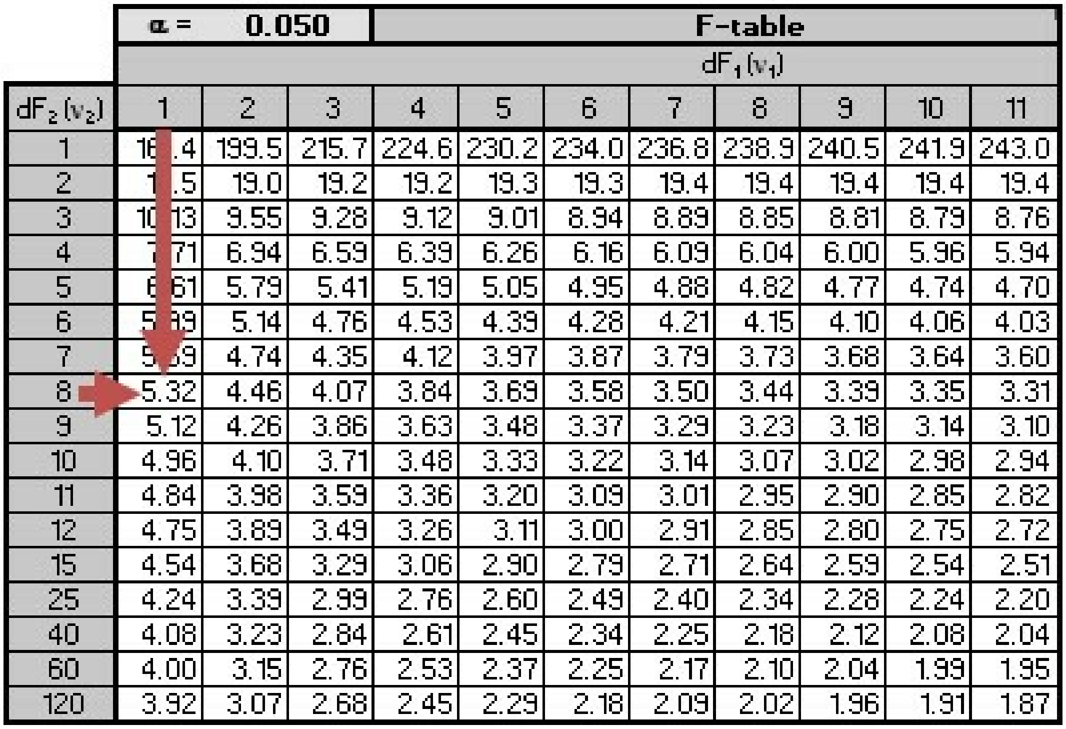
a. Use the above ANOVA table to calculate the F-statistic.

b. Test the hypothesis that the slope coefficient is equal to a 5% significance level.

Solution

a. $= \frac{\text{Mean Regression Sum of Squares (MSR)}}{\text{Mean Squared Error(MSE)}} = \frac{10.568}{0.392} = 26.960$

b. We are testing the null hypothesis $H_0 : b_1 = 0$ against the alternative hypothesis $H_1 : b_1 \neq 0$. The critical F-value for $k = 1$ and $n - 2 = 8$ degrees of freedom at a 5% significance level is roughly 5.32. Note that this is a one-tail test, and therefore, we use the 5% F-table.



α = 0.050	F-table										
	dF₁ (v₁)										
dF₂ (v₂)	1	2	3	4	5	6	7	8	9	10	11
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87

Remember that the null hypothesis is rejected if the calculated value of the F-statistic is greater than the critical value of F. Since $26.960 > 5.32$, we reject the null hypothesis and conclude that the slope coefficient is significantly different from zero.Notice that we also rejected the null hypothesis in the previous examples. We did so because the 95% confidence interval did not include zero.

An F-test duplicates the t-test in regard to the slope coefficient significance for a linear regression model with one independent variable. In this case, $t^2 = 2.306^2 \approx 5.32$. Since the F-statistic is the square of the t-statistic for the slope coefficient, its inferences are the same as the t-test. However, this is not the case for multiple regressions.

Question

Consider the following analysis of variance (ANOVA) table:

Source	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression	1	$1,701,563$	$1,701,563$
Error (Unexplained)	3	$106,800$	$13,350$
Total	4	$1,808,363$	

The value of R^2 and the F-statistic for the test of fit of the regression model are *closest to*:

A. 6% and 16.

B. 94% and 127.

C. 99% and 127.

Solution

The correct answer is B.

$$R^2 = \frac{\text{Sum of Squares Regression (SSR)}}{\text{Sum of Squares Total (SST)}} = \frac{1,701,563}{1,808,363} = 0.94 = 94\%$$

$$\begin{aligned} F &= \frac{\text{Mean Regression Sum of Squares (MSR)}}{\text{Mean Squared Error (MSE)}} \\ &= \frac{1,701,563}{13,350} = 127.46 \approx 127 \end{aligned}$$

LOS 10e: 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

We calculate the predicted value of the dependent variable, Y, by inserting the estimated value of the independent variable, X, into the regression equation. The predicted value of the dependent variable, Y, is determined using the following formula:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X$$

Where:

$\hat{Y}$ = Predicted value of the dependent variable.

X = Estimated value of the independent variable.

Example: Calculating the Predicted Value of a Dependent Variable

Refer to the example of regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

The estimated regression model is illustrated below.

$$\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$$

Calculate the predicted inflation rate value if the forecasted value of the unemployment rate is

4.5%.

Solution

The predicted value of the inflation rate is determined as follows:

$$\hat{Y} = 7.112 - 0.9020 \times 4.5 = 3.053\%$$

Confidence Interval for Predicted Values

The calculation of the confidence interval for the predicted value of a dependent variable is the same as that of the confidence interval for regression coefficients. The confidence interval for a predicted value of the dependent variable is given by:

$$\text{Prediction Interval} = \hat{Y} \pm t_c s_f$$

Where:

t_c= Two-tailed critical t-value at the given significance level with $n-2$ df.

$\hat{Y}$ = Predicted value of a dependent variable.

s_f^2= The estimated variance of the prediction error.

$$s_f^2 = s_e^2\left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)\,s_x^2}\right] = s_e^2\left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]$$

Where:

s_e^2 = The squared standard error of the estimate.

n = Number of observations.

s_X^2= Variance of the independent variable.

X_f = Value of the independent variable.

We can, therefore, calculate the standard error of forecast as shown below:

$$s_f = s_e\sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

From the formula above, we can observe that:

- A better fit of the regression analysis leads to a smaller standard error of the estimate (s_e), subsequently resulting in a lower standard error of the forecast.
- When the sample size (n) in the regression calculation increases, it directly corresponds to a reduction in the standard error of the forecast.
- If the forecasted independent variable (X_f) approaches the mean of the independent variable ($\bar{X}$) utilized in the regression analysis, it decreases the standard error of the forecast.

Example: Calculating the Confidence Interval of the Predicted Value

Refer to the example of regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

Consider the results of the regression analysis of inflation rates on unemployment rates:

Regression Statistics	
R Square	0.7711
Standard Error	0.6261
Observations	10

ANOVA				
	df	Sum of Squares	Mean Square	F
Regression	1	10.568	10.568	26.9565
Residual	8	3.136	0.392	
Total	9	13.704		

	Coefficients	Standard Error	t Stat	p-value
Intercept	7.112	0.940	7.565	0.000
Unemployment rate (%)	−0.902	0.174	−5.192	0.001

Calculate the 95% confidence interval of the predicted value of the inflation rate, given that the forecasted unemployment rate is 4.5%.

Solution

$$\text{Prediction Interval} = \hat{Y} \pm t_c s_f$$

The estimated variance of the prediction error is:

$$\begin{aligned} s_f^2 &= s_e^2\left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)\, s_X^2}\right] \\ &= s_e^2\left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right] \\ &= 0.6261^2\,[1 + \frac{1}{10} + \frac{(4.5 - 5.29)^2}{12.989}] = 0.450 \end{aligned}$$

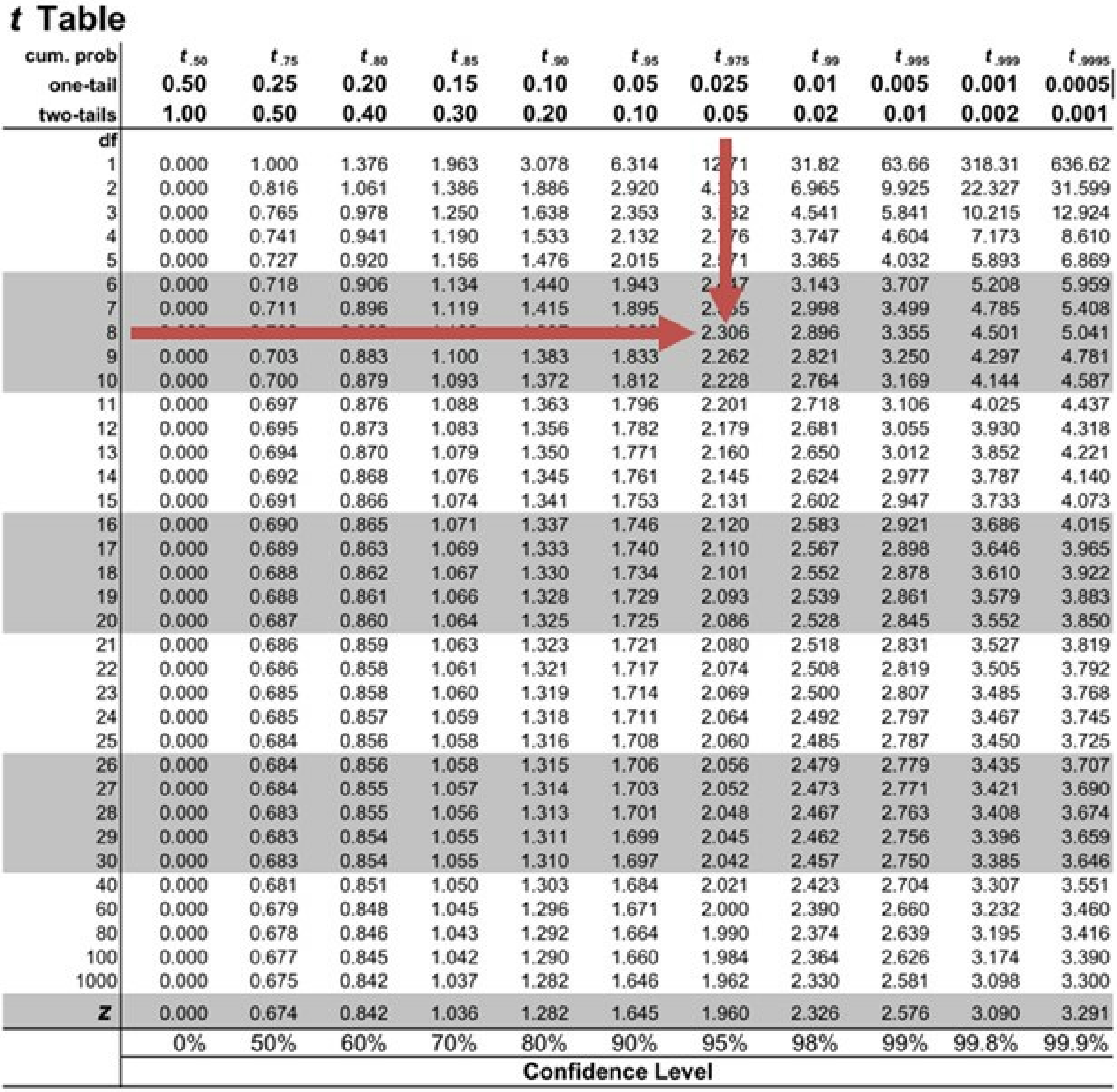
As such, the standard error of forecast is:

$$s_f = \sqrt{0.450} = 0.6708$$

The predicted value of the inflation rate given an unemployment rate of 4.5% is:

$$\hat{Y} = 7.112 - 0.9020 \times 4.5 = 3.05\%$$

The two-tailed critical t-value with 8 (n − 2) degrees of freedom at the 5% significance level is 2.306.



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

The prediction interval at the 95% confidence level is:

$$\text{Prediction Interval (PI)} = \hat{Y} \pm t_c s_f$$

$$\text{PI} = 3.05 \pm 2.306 \times 0.6708 = 1.50\% \text{ to } 4.60\%$$

Interpretation

Given an unemployment rate of 4.5%, we are **95% confident** that **the inflation rate will lie between 1.50%** and **4.60%**.

Question 1

The regression equation of the quantity of goods against the price is given by:

$$Y = -159 + 0.26X$$

Where:

Y = Quantity supplied.

X = Price per unit of the product.

The predicted value of the quantity supplied when the price equals 1,200 is *closest to*:

A. 153.
B. 155.
C. 471.

The correct answer is A.

$$Y = -159 + 0.26 \times 1,200 = 153$$

LOS 10f: Describe different functional forms of simple linear regressions

To address non-linear relationships, we employ various functional forms to potentially convert the data for linear regression. Here are three commonly used log transformation functional forms:

1. **Log-lin model:**In this log transformation, the dependent variable is logarithmic, while the independent variable is linear. It is represented as shown below.

   $$\ln Y = b_0 + b_1 X_i.$$

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

   When utilizing a log-lin model, caution must be exercised when making forecasts. For example, in the predicted regression equation like $Y = -3 + 5X$, if X is equal to 1, the $\ln Y = -3$, then,

   $$Y = e^{-3} = 0.0498$$

   Moreover, the lin-lin model cannot be compared with the log-lin model without the transformation. As such, we need to transform R^2 and F-statistic.

2. **Lin-log model:**In this case, the dependent variable is linear, while the independent variable is logarithmic. It is represented as follows:
   $Y_i = b_0 + b_1 \ln X_i$.

   The slope coefficient in the lin-log model is responsible for the absolute change in the dependent variable for a relative change in the independent variable.

3. **Log-log model:**In this log transformation, both the dependent and independent variables are logarithmic. It is represented as $\ln Y_i = b_0 + b_1 \ln X_i$. The slope coefficient in the log-log model is the relative change in the dependent variable for a relative change in the independent variable. In other words, if X increases by 1%, Y will change by b_1.

Selecting the Correct Functional Form

To settle on the correct functional form, consider the following goodness of fit measures:

I. Coefficient of determination (R^2). A high value is better.

II. F-statistic. The high value of the F-statistic is better.

III. Standard error of the estimate (S_e). A low value of S_e is better.

Aside from the factors cited above, the patterns in residuals can also be analyzed when evaluating a model. Residuals are random and uncorrelated in a good model.

Question 1

Which of the following statements about the log-lin model is *most likely* correct:

A. The dependent variable is linear, while the independent variable is logarithmic.

B. Both the dependent and independent variables are logarithmic

C. The dependent variable is logarithmic, while the independent variable is linear.

The correct answer is c.

In the log-lin model, the dependent variable (Y) is logarithmic, as represented by

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

While the independent variable (X) is linear.

A is incorrect. It describes the lin-log model, where the dependent variable is linear and the independent variable is logarithmic.

B is incorrect. It describes the log-log model, where both the dependent and independent variables are logarithmic.