

1. Introduction .....	2
2. Correlation Analysis .....	2
2.1. Scatter Plots .....	2
2.2. Correlation Analysis .....	3
2.3. Calculating and Interpreting the Correlation Coefficient .....	4
2.4. Limitations of Correlation Analysis.....	6
2.5. Uses of Correlation Analysis .....	6
2.6. Testing the Significance of the Correlation Coefficient .....	8
3. Linear Regression .....	11
3.1. Linear Regression with One Independent Variable .....	11
3.2. Assumptions of the Linear Regression Model.....	13
3.3. The Standard Error of Estimate .....	15
3.4. The Coefficient of Determination.....	17
3.5. Hypothesis Testing.....	19
3.6. Analysis of Variance in a Regression with One Independent Variable.....	25
3.7. Prediction Intervals .....	27
3.8. Limitations of Regression Analysis .....	30
4. Summary .....	31

This document should be read in conjunction with the corresponding reading in the 2017 Level II CFA® Program curriculum. Some of the graphs, charts, tables, examples, and figures are copyright 2016, CFA Institute. Reproduced and republished with permission from CFA Institute. All rights reserved.

Required disclaimer: CFA Institute does not endorse, promote, or warrant the accuracy or quality of the products or services offered by IFT. CFA Institute, CFA®, and Chartered Financial Analyst® are trademarks owned by CFA Institute.

## 1. Introduction

In this reading, we look at two important concepts to examine the relationship between two or more financial variables: correlation analysis and regression analysis. For example, how to determine if there is a relationship between the returns of the U.S. stock market and the Japanese stock market over the past five years, or between unemployment and inflation?

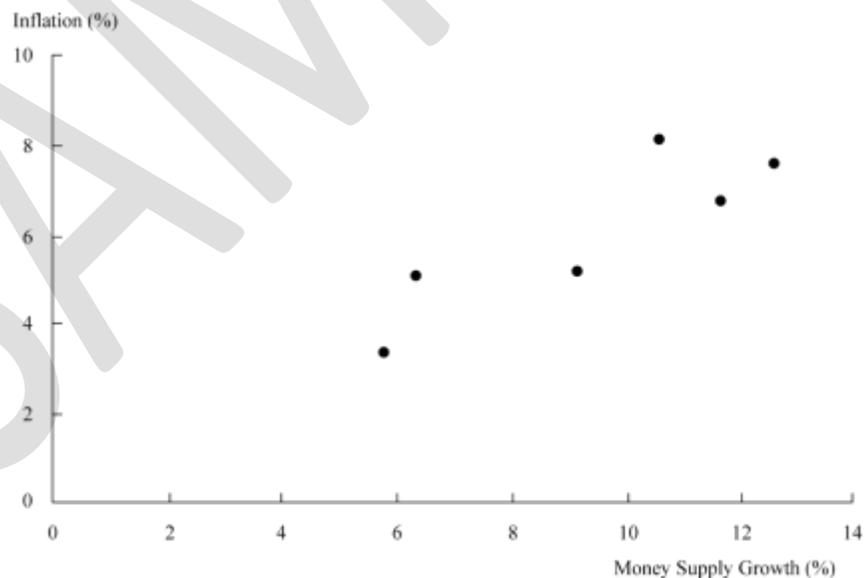
## 2. Correlation Analysis

In this section, we look at two methods to examine how two sets of data are related to each other: scatter plots and correlation analysis.

### 2.1. Scatter Plots

A scatter plot is a graph that shows the relationship between the observations for two data series in two dimensions (x-axis and y-axis). The scatter plot below is reproduced from the curriculum:

**Figure 1.**  
Scatter Plot of  
Annual Money  
Supply Growth Rate  
and Inflation Rate by  
Country, 1970–2001



Interpretation of Figure 1:

- The two data series here are the average growth in money supply (on the x-axis) plotted against the average annual inflation rate (on the y-axis) for six countries.
- Each point on the graph represents (money growth, inflation rate) pair for one country. From the six points, it is evident that there is an increase in inflation as money supply grows.

## 2.2. Correlation Analysis

Correlation analysis is used to measure the strength of the relationship between two variables. It is represented as a number. The correlation coefficient is a measure of how closely related two data series are. In particular, the correlation coefficient measures the direction and extent of **linear** association between two variables.

### Characteristics of the correlation coefficient

A correlation coefficient has no units. The sample correlation coefficient is denoted by  $r$ .

The value of  $r$  is always  $-1 \leq r \leq 1$ .

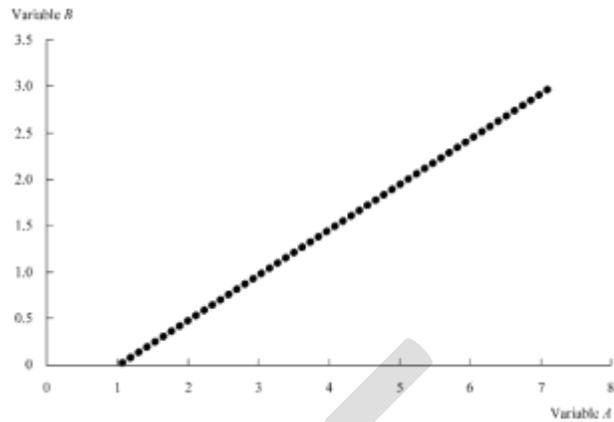
A value of  $r$  greater than 0 indicates a positive linear association between the two variables.

A value of  $r$  less than 0 indicates a negative linear association between the two variables.

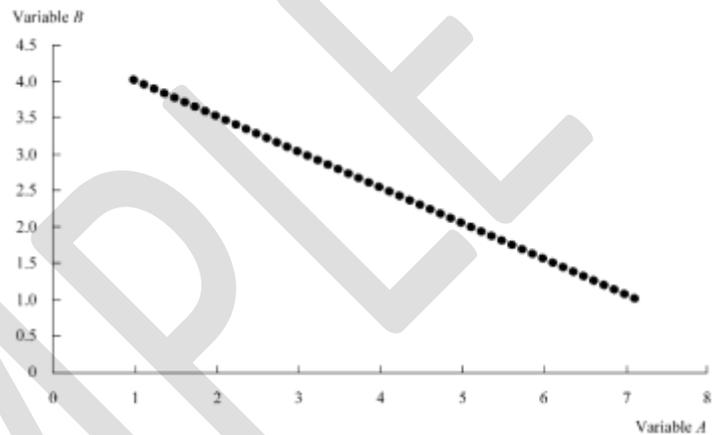
A value of  $r$  equal to 0 indicates no linear relation between the two variables.

The three scatter plots below show a positive linear, negative linear, and no linear relation between two variables A and B. They have correlation coefficients of +1, -1 and 0 respectively.

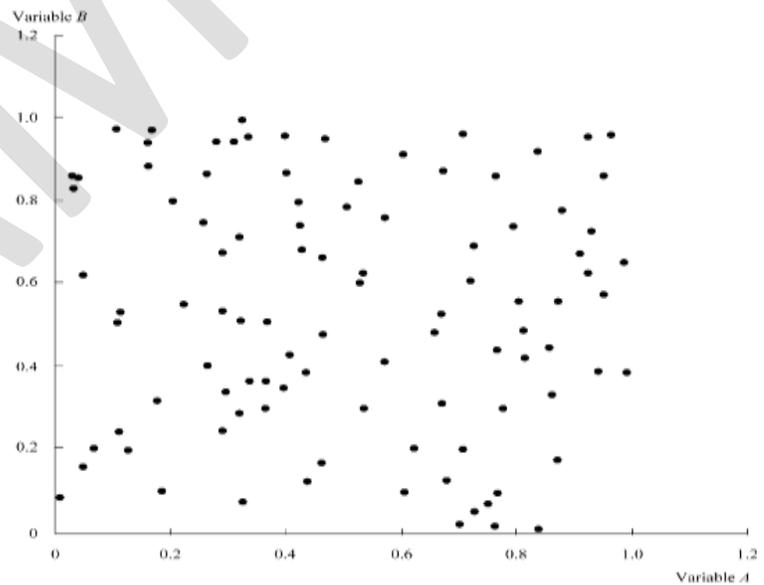
**Figure 2.**  
Variables with a  
Correlation of 1.



**Figure 3.**  
Variables with a  
Correlation of -1.



**Figure 4:**  
Variables with a  
Correlation of 0.



### 2.3. Calculating and Interpreting the Correlation Coefficient

In order to calculate the correlation coefficient between two variables, X and Y, we need the

following:

1. Covariance between X and Y, denoted by Cov (X,Y)
2. Standard deviation of X, denoted by  $s_x$
3. Standard deviation of Y, denoted by  $s_y$

This is the formula for computing the **sample covariance of X and Y**:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

The table below illustrates how to apply the covariance formula. Our data is the money supply growth rate ( $X_i$ ) and the inflation rate ( $Y_i$ ) for six different countries.  $\bar{X}$  represents the average money supply growth rate and  $\bar{Y}$  represents the average inflation rate.

Country	$X_i$	$Y_i$	Cross-Product	Squared Deviations	Squared Deviations
Australia	0.1166	0.0676	0.000169	0.000534	0.000053
Canada	0.0915	0.0519	0.000017	0.000004	0.000071
New Zealand	0.106	0.0815	0.000265	0.000156	0.000449
Switzerland	0.0575	0.0339	0.000095	0.001296	0.000697
United Kingdom	0.1258	0.0758	0.000501	0.001043	0.00024
United States	0.0634	0.0509	0.000283	0.000906	0.000088
Sum	0.5608	0.3616	0.002185	0.003939	0.001598
Average	0.0935	0.0603			
Covariance			0.000437		
Variance				0.000788	0.00032
Standard deviation				0.028071	0.017889

Notes:

1. Divide the cross-product sum by  $n - 1$  (with  $n = 6$ ) to obtain the covariance of X and Y.
2. Divide the squared deviations sums by  $n - 1$  (with  $n = 6$ ) to obtain the variances of X and Y.

Source: International Monetary Fund.

Given the covariance between X and Y and the two standard deviations, the sample correlation can be easily calculated.

**The following equation** shows the formula for computing the **sample correlation of X and Y**:

$$r = \frac{\text{Cov}(X, Y)}{s_x * s_y}$$

$$r = \frac{\text{Cov}(X, Y)}{\sigma_x * \sigma_y} = \frac{0.000437}{0.028071 * 0.017889} = 0.870236$$

LO.a: Calculate and interpret a sample covariance and a sample correlation coefficient; and interpret a scatter plot.

## 2.4. Limitations of Correlation Analysis

The correlation analysis has certain limitations:

- Two variables can have a strong non-linear relation and still have a very low correlation. Recall that correlation is a measure of the linear relationship between two variables.
- The correlation can be unreliable when outliers are present.
- The correlation may be spurious. **Spurious correlation** refers to the following situations:
  - The correlation between two variables that reflects chance relationships in a particular data set.
  - The correlation induced by a calculation that mixes each of two variables with a third variable.
  - The correlation between two variables arising not from a direct relation between them, but from their relation to a third variable. Ex: shoe size and vocabulary of school children. The third variable is age here. Older shoe sizes simply imply that they belong to older children who have a better vocabulary.

LO.b: Describe the limitations to correlation analysis.

## 2.5. Uses of Correlation Analysis

The uses of correlation analysis are highlighted through six examples in the curriculum. Instead of reproducing the examples, the specific scenarios where they are used are listed below:

- **Evaluating economic forecasts:** Inflation is often predicted using the change in the consumer price index (CPI). By plotting actual vs predicted inflation, analysts can

determine the accuracy of their inflation forecasts.

- **Style analysis correlation:** Correlation analysis is used in determining the appropriate benchmark to evaluate a portfolio manager's performance. For example, assume the portfolio managed consists of 200 small value stocks. The Russell 2000 Value Index and the Russell 2000 Growth Index are commonly used as benchmarks to measure the small-cap value and small-cap growth equity segments, respectively. If there is a high correlation between the returns to the two indexes, then it may be difficult to distinguish between small-cap growth and small-cap value as different styles.
- **Exchange rate correlations:** Correlation analysis is also used to understand the correlations among many asset returns. This helps in asset allocation, hedging strategy and diversification of the portfolio to reduce risk. Historical correlations are used to set expectations of future correlation. For example, suppose an investor who has an exposure to foreign currencies. He needs to ascertain whether to increase his exposure to the Canadian dollar or to Japanese Yen. By analyzing the historical correlations between USD returns to holding the Canadian dollar and USD returns to holding the Japanese yen, he will be able to come to a conclusion. If they are not correlated, then holding both the assets helps in reducing risk.
- **Correlations among stock return series:** Analyzing the correlations among the stock market indexes such as large-cap, small-cap and mid-cap helps in asset allocation and diversifying risk. For instance, if there is a high correlation between the returns to the large-cap index and the small-cap index, then their combined allocation may be reduced to diversify risk.
- **Correlations of debt and equity returns:** Similarly, the correlation among different asset classes, such as equity and debt, is used in portfolio diversification and asset allocation. For example, high-yield corporate bonds may have a high correlation to equity returns, whereas long-term government bonds may have a low correlation to equity returns.
- **Correlations among net income, cash flow from operations, and free cash flow to the firm:** Correlation analysis shows if an analyst's decision to value a firm based only on NI and ignore CFO and FCFE is correct. FCFE is the cash flow available to debt holders and

shareholders after all operating expenses have been paid and investments in working and fixed capital have been made. If there is a low correlation between NI and FCFF, then the analyst's decision to use NI instead of FCFF/CFO to value a company is questionable.

## 2.6. Testing the Significance of the Correlation Coefficient

The objective of a significance test is to assess whether there is really a correlation between random variables, or if it is a coincidence. If it can be ascertained that the relationship is not a result of chance, then one variable can be used to predict another variable using the correlation coefficient.

A t-test is used to determine whether the correlation between two variables is significant. The population correlation coefficient is denoted by  $\rho$  (rho). As long as the two variables are distributed normally, we can use hypothesis testing to determine whether the null hypothesis should be rejected using the sample correlation,  $r$ . The formula for the t-test is:

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

The test statistic has a t-distribution with  $n - 2$  degrees of freedom.

$n$  denotes the number of observations.

### How to use the t-test to determine significance:

1. Write the null hypothesis  $H_0$  i.e. ( $\rho = 0$ ), and the alternative hypothesis  $H_a$  i.e. ( $\rho \neq 0$ ). Since the alternative hypothesis is to test the correlation is *not equal to zero*, it is a two-tailed test.
2. Specify the level of significance. Determine the degrees of freedom.
3. Determine the critical value,  $t_c$  for the given significance level and degrees of freedom.
4. Calculate the test statistic,  $t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$
5. Make a decision to reject the null hypothesis  $H_0$ , or fail to reject  $H_0$ . If absolute value of  $t > t_c$ , then reject  $H_0$ . If absolute value of  $t \leq t_c$ , then fail to reject  $H_0$ .
6. Interpret the decision:
  - a. If you reject  $H_0$ , then there is a significant linear correlation.

- b. If you fail to reject  $H_0$ , then it can be concluded that there is statistically no significant linear correlation.

Important points from Examples 7 through 10 in the curriculum are summarized below:

**Example 7: Testing the correlation between money supply growth rate and inflation**

Data given: The sample correlation between long-term supply growth and long-term inflation in six countries during the 1970 - 2001 period is 0.8702. The sample has six observations.

Test the null hypothesis that the true correlation in the population is zero ( $\rho = 0$ ).

**Solution:**

Compute the test statistic:  $t = \frac{0.8702 * \sqrt{6 - 2}}{\sqrt{1 - 0.8702^2}} = 3.532$

Critical value:  $t_c = 2.776$  at 0.05 significance level with  $n - 2 = 6 - 2 = 4$  degrees of freedom

Decision rule: If the test statistic is greater than 2.776 or less than - 2.776, then we can reject the null hypothesis.

Conclusion: Since the test statistic 3.532 is greater than 2.776, we can conclude that there is a strong relationship between long-term money supply growth and long-term inflation in six countries.

**Example 8: Testing the Krona - Yen Return Correlation**

Data given: Sample correlation between the USD monthly returns to Swedish kronor and the Japanese yen for the period from January 1990 to December 1999 is 0.2860.

Test the null hypothesis that the true correlation in the population is zero ( $\rho = 0$ ).

**Solution:**

Number of observations = 120 months (Jan. 1990 – Dec. 1999)

Test statistic  $t = \frac{0.2860 * \sqrt{120 - 2}}{\sqrt{1 - 0.2860^2}} = 3.242$

Critical value  $t_c = 1.98$  (significance level = 0.05, degrees of freedom = 118)

Decision rule: If the test statistic is greater than 1.98 or less than - 1.98, then we can reject the null hypothesis.

Conclusion: Since the test statistic 3.242 is greater than 1.98, we can reject the null hypothesis and conclude that there is a correlation between the USD monthly return to Swedish kronor and the Japanese yen. The correlation coefficient is smaller at 0.2860 but is still significant because the sample is larger. As  $n$  increases, the critical value decreases and  $t$  increases.

### Example 10: Testing the correlation between net income and free cash flow to the firm

Data given: The sample correlation between NI and FCF for six women's clothing stores was 0.4045 in 2001. The sample has six observations. Test the null hypothesis that the true correlation in the population is 0 against the alternative hypothesis that the correlation in the population is different from 0.

#### Solution:

$$\text{The test statistic } t = \frac{0.4045 * \sqrt{6 - 2}}{\sqrt{1 - 0.4045^2}} = 0.8846$$

Critical test statistic = 2.776 (at 0.05 significance level with 4 degrees of freedom)

Decision rule: If the test statistic is greater than 2.776 or less than -2.776, then reject the null hypothesis.

Conclusion: Since  $t = 0.8846$  is less than 2.776, we cannot reject the null hypothesis. For this sample of women's clothing stores, there is no statistically significant correlation between NI and FCF.

Some inferences based on the above examples:

- As  $n$  increases, the absolute value of the critical value  $t_c$  decreases.
- As  $n$  increases, the absolute value of the numerator increases resulting in large  $t$ -value.
- Large absolute values of the correlation coefficient indicate strong linear relationships.

- Relatively small correlation coefficients can be statistically significant (**Example 8**).

LO.c: Formulate a test of the hypothesis that the population correlation coefficient equals zero, and determine whether the hypothesis is rejected at a level of significance.

## 3. Linear Regression

### 3.1. Linear Regression with One Independent Variable

Linear regression allows us to use one variable to make predictions about another, test hypotheses about the relation between two variables, and quantify the strength of the relationship between the two variables. Linear regression assumes a **linear relationship** between the **dependent** and **independent** variables.

In simple terms, regression analysis uses the historical relationship between the independent variable and the dependent variable to predict the values of the dependent variable. The regression equation is expressed as follows:

$$Y_i = b_0 + b_1X_i + \varepsilon_i$$

where

$$i = 1, \dots, n$$

Y = dependent variable

$b_0$  = intercept

$b_1$  = slope coefficient

X = independent variable

$\varepsilon$  = error term

$b_0$  and  $b_1$  are called the **regression coefficients**.

Dependent variable is the variable being predicted. It is denoted by Y in the equation. The

variable used to explain changes in the dependent variable is the independent variable. It is denoted by X. The equation shows how much Y changes when X changes by one unit. Regression analysis uses two types of data: **cross-sectional** and **time-series**.

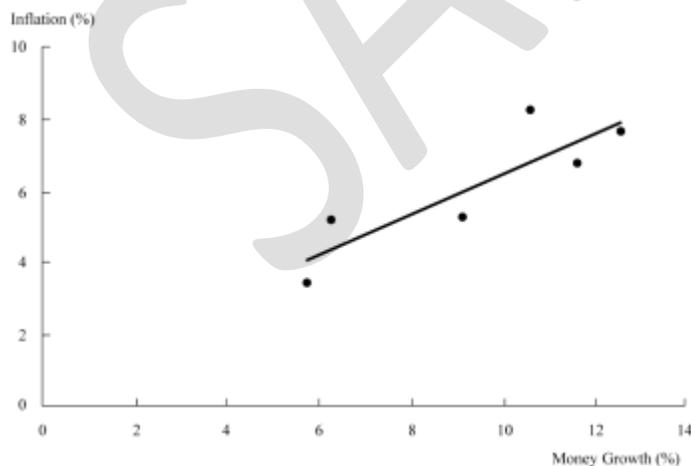
A **linear regression** model or **linear least squares** method, computes the best-fit line through the scatter plot, or the line with the smallest distance between itself and each point on the scatter plot. The regression line may pass through some points, but not through all of them. The vertical distances between the observations and the regression line are called error terms, denoted by  $\varepsilon_i$ .

Linear regression chooses the estimated values for intercept  $\widehat{b}_0$  and slope  $\widehat{b}_1$  such that the sum of the squared 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.

$$\sum_{i=1}^N (Y_i - \widehat{b}_0 - \widehat{b}_1 X_i)^2$$

Where  $\widehat{b}_0$  and  $\widehat{b}_1$  are estimated parameters.

*Note:* The predictions in a regression model are based on the population parameter values of  $\widehat{b}_0$  and  $\widehat{b}_1$  and not on  $b_0$  or  $b_1$ . **Figure 5** below shows the regression line.



Interpretation of the graph:

- The graph plots money growth rate, the independent variable, on the x-axis against inflation rate, the dependent variable, on the y-axis.
- The equation estimates the value of the long-term inflation rate as  $b_0 + b_1$  (long-term rate of money growth) +  $\varepsilon$ .
- The error term or regression residual is the distance between the line and each of the six points. This is also equal to the difference between the actual value of the dependent variable and the predicted value of the dependent variable.

LO.d: Distinguish between the dependent and independent variables in a linear regression.

### 3.2. Assumptions of the Linear Regression Model

The classic linear regression model ( $Y_i = b_0 + b_1X_i + \varepsilon_i$  where  $i = 1, \dots, n$ ) is based on the following six assumptions:

1. The relationship between Y and X is linear in the parameters  $b_0$  and  $b_1$ . This means that  $b_0$ ,  $b_1$ , and X can only be raised to the power of 1.  $b_0$  and  $b_1$  cannot be multiplied or divided by another regression parameter. This assumption is critical for a valid linear regression.
2. X is not random.
3. The expected value of the error term is zero. This ensures the model produces correct estimates for  $b_0$  and  $b_1$ .
4. The variance of the error term is constant for all observations. This is called homoskedasticity. If the variance of the error term is not constant, then it is called heteroskedasticity.
5. The error term,  $\varepsilon$ , is uncorrelated across observations.
6. The error term,  $\varepsilon$ , is normally distributed.

#### Example 11: A summary of how economic forecasts are evaluated

This example discusses the importance of making accurate and unbiased forecasts. It is based on the following premise: if forecasts are accurate, every prediction of change in an economic variable will be equal to the actual change. For an unbiased forecast, the expected value of the error term is zero and  $E(\text{actual change} - \text{predicted change}) = 0$

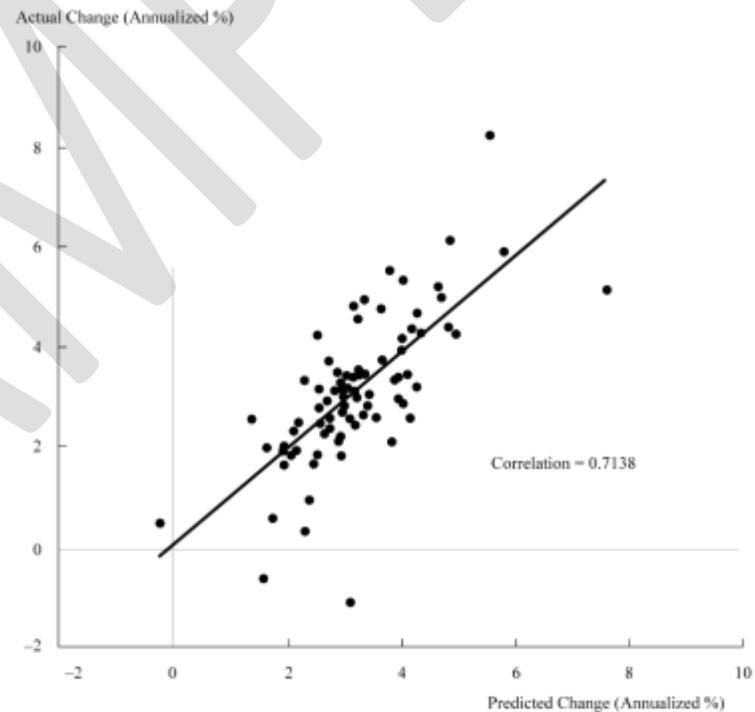
**Figure 6** below shows a scatter plot of the mean forecast of current-quarter percentage change in the CPI from the previous quarter and the actual percentage change in the CPI. It shows the fitted regression line for the following equation:

$$\text{Actual percentage change} = b_0 + b_1 (\text{predicted percentage change}) + \varepsilon$$

To prove that the forecasts are unbiased, the following must be true:

- Intercept  $b_0 = 0$ ; slope  $b_1 = 1$ . If the values are different from 0 and 1, then the error term will have an expected value different from zero.
- $E[\text{actual change} - \text{predicted change}] = 0$
- The expected value of the error term given by  $E[\text{actual change} - b_0 - b_1 (\text{predicted change})] = 0$

**Figure 6.**



What we get:

The fitted regression line is drawn based on the equation:

Actual change =  $-0.0140 + 0.9637$  (predicted change)

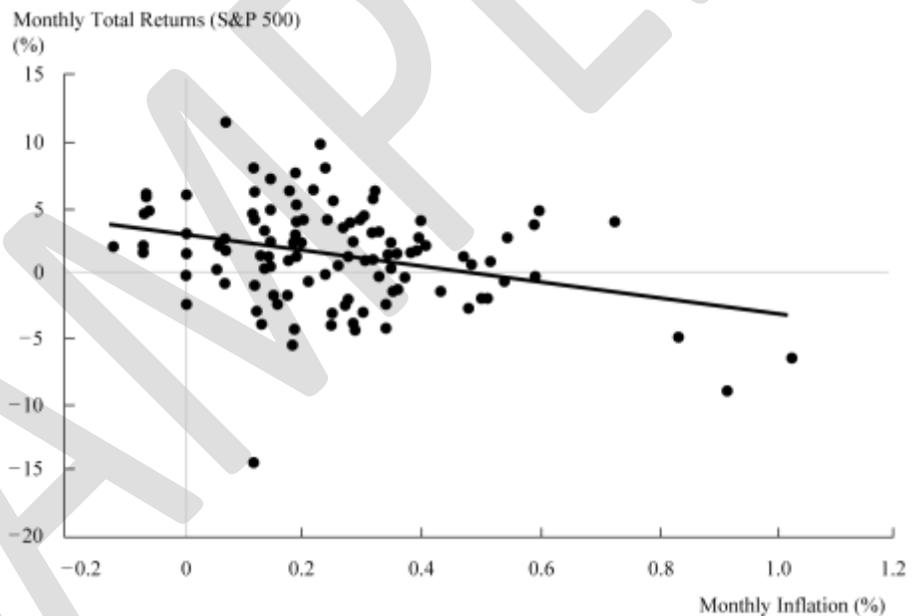
Since  $b_0$  and  $b_1$  are close to 0 and 1 respectively, we can conclude that the forecast is unbiased.

LO.e: Describe the assumptions underlying linear regression, and interpret regression coefficients.

### 3.3. The Standard Error of Estimate

**Figure 7** below shows a fitted regression line to the scatter plot on monthly returns to the S&P 500 Index and the monthly inflation rate in the United States during the 1990s. The monthly inflation rate is the independent variable. The stock returns is the dependent variable as stock returns vary with the inflation rate.

**Figure 7:**



In contrast to the fitted regression line in Figure 6, many actual observations in Figure 7 are farther from the regression line. This may imply the predictions were not strong, resulting in an inaccurate forecast. The **standard error of estimate (SEE)** measures how well a given linear regression model captures the relationship between the dependent and independent variables. In simple terms, it is the standard deviation of the prediction errors.

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

$$SEE = \left( \sum_{i=1}^N \frac{(Y_i - \widehat{b}_0 - \widehat{b}_1 X_i)^2}{n-2} \right)^{\frac{1}{2}} = \left( \frac{\sum_{i=1}^N (\varepsilon_i)^2}{n-2} \right)^{\frac{1}{2}}$$

where

Numerator: Regression residual,  $\varepsilon_i$  = dependent variable's actual value for each observation – predicted value for each observation

Denominator: degrees of freedom =  $n - 2$  =  $n$  observations – two estimated parameters  $\widehat{b}_0$  and  $\widehat{b}_1$ . The reason 2 is subtracted is because SEE describes characteristics of two variables.

**A low SEE implies an accurate forecast.**

### Example 12:

This example illustrates how to compute the standard error of estimate based on money supply and inflation data, and regression equation:  $Y_i = 0.0084 + 0.5545X_i$

Table 7. Computing the Standard Error of Estimate

Country	Money Supply Growth Rate $X_i$	Inflation rate $Y_i$	Predicted Inflation Rate	Regression Residual	Squared Residual
Australia	0.1166	0.0676	0.0731	- 0.0055	0.00003
Canada	0.0915	0.0519	0.0591	- 0.0072	0.000052
New Zealand	0.106	0.0815	0.0672	0.0143	0.000204
Switzerland	0.0575	0.0339	0.0403	- 0.0064	0.000041
United Kingdom	0.1258	0.0758	0.0782	- 0.0024	0.000006
United States	0.0634	0.0509	0.0436	0.0073	0.000053
Sum					0.000386

Source: International Monetary Fund.

Data given:

Column 2, 3: Values are given for money supply growth rate (the independent variable) and an

inflation rate (the dependent variable).

Calculated values: Let us see how to calculate the values for each column:

Column 4: Predicted value calculated for each observation from the regression equation

Column 5: Residual = actual value – predicted value =  $Y_i - (\widehat{b}_0 + \widehat{b}_1 X_i)$

Column 6: Squared residual and finally, the sum of squared residuals, is calculated as 0.000386

Using Equation 5, SEE is calculated as  $\left(\frac{0.000386}{6-2}\right)^{\frac{1}{2}} = 0.009823$ . SEE is about 0.98 percent.

*Note: From an exam point of view, the testability of a question like this is low that asks you to calculate SEE given x and y data.*

### 3.4. The Coefficient of Determination

The SEE gives some indication of how certain we can be about a particular prediction of Y using the regression equation; it still does not tell us how well the independent variable explains variation in the dependent variable. The coefficient of determination does exactly this: it measures the fraction of the total variation in the dependent variable that is explained by the independent variable. It is denoted by  $R^2$ .

There are two methods to calculate the coefficient of determination:

**Method 1:** Square the correlation coefficient between the dependent and the independent variable. The drawback of this method is that it cannot be used when there is more than one independent variable.

**Method 2:** The percent of variation that can be explained by the regression equation.

Total variation = unexplained variation + explained variation

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

where

Total variation =  $\sum_{i=1}^N (y_i - \bar{y})^2$  = actual value of  $y_i$  and the average value of  $y_i$ . If there was no regression equation, then the predicted value of any observation would be mean of  $y_i$ ,  $\bar{y}_1$ .

Unexplained variation from the regression =  $\sum_{i=1}^N (y_i - \hat{y})^2$

### 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.

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

With only one independent variable,  $R^2$  is the square of the correlation.

The correlation,  $r$ , is also called the “multiple -  $R$ ”.

### Example 13: Inflation rate and growth in money supply

Unexplained variation in the regression = sum of squared residuals = 0.000386

The table below from the curriculum shows the computation for total variation in the dependent variable.

Table 8. Computing Total Variation

Country	Money Supply Growth Rate $X_i$	Inflation Rate $Y_i$	Deviation from Mean	Squared Deviation
Australia	0.1166	0.0676	0.0073	0.000053
Canada	0.0915	0.0519	-0.0084	0.000071
New Zealand	0.106	0.0815	0.0212	0.000449
Switzerland	0.0575	0.0339	-0.0264	0.000697
United Kingdom	0.1258	0.0758	0.0155	0.00024
United States	0.0634	0.0509	-0.0094	0.000088
	Average:	0.0603	Sum:	0.001598

Source: International Monetary Fund.

Calculate the coefficient of determination.

**Solution:**

Since there is only one independent variable, both the methods can be used and they should yield the same answer.

Method 1: in section 2.3 we had calculated the correlation coefficient for the same data to be 0.870236. Squaring it, we get coefficient of determination as 0.7573.

Method 2: using Equation 5,

$$R^2 = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = \frac{0.001598 - 0.000386}{0.001598} = 0.7584.$$

This method is to be used if there is more than one independent variable.

LO.f: Calculate and interpret the standard error of the estimate, the coefficient of determination, and a confidence interval for a regression coefficient.

### 3.5. Hypothesis Testing

A parameter tells us something about a population. A **confidence interval** is an interval of values that we believe includes the true parameter value,  $b_1$ , with a given degree of confidence. Simply put, it is used to estimate a population parameter.

Hypothesis testing is used to assess whether the evidence supports the claim about a population.

The three things required to perform a hypothesis test using the confidence interval approach are listed below:

- The estimated parameter value,  $\widehat{b}_0$  or  $\widehat{b}_1$
- The hypothesized value of the parameter,  $b_0$  or  $b_1$
- A confidence interval around the estimated parameter

There are two ways to perform hypothesis testing using either the confidence interval or testing

for a significance level, that we will see in the following example.

Suppose we regress a stock's returns on a stock market index's returns and find that the slope coefficient is 1.5 with a standard error estimate of 0.200. Assume we need 60 monthly observations in our regression analysis. The hypothesized value of the parameter is 1.0, which is the market average slope coefficient. Define the null and alternative hypothesis. Compute the test statistic. At a 0.05 level of significance, should we reject  $H_0$ ?

**Solution:**

**Method 1: The confidence interval approach**

Step 1: Select the significance level for the test. For instance, if we are testing at a significance level of 0.05, then we will construct a 95 percent confidence interval.

Step 2: Define the null hypothesis. For this example, the probability that the confidence interval includes  $b_1$  is 95 percent, will be the null hypothesis.  $H_0: b_1 = 1.0$ ;  $H_a: b_1 \neq 1.0$

Step 3: Determine the critical t value  $t_c$  for a given level of significance and degrees of freedom. In this regression with one independent variable, there are two estimated parameters, the intercept and the coefficient on the independent variable.

Number of degrees of freedom = number of observations – number of estimated parameters = 60 – 2 = 58.

$t_c = 2.00$  at (significance level of 0.05 and 58 degrees of freedom)

Step 4: Construct the confidence interval which will span from  $\widehat{b}_1 - t_c s_{\widehat{b}_1}$  to  $\widehat{b}_1 + t_c s_{\widehat{b}_1}$  where  $t_c$  is the critical t value.

Confidence interval =  $1.5 - 2(0.2)$  to  $1.5 + 2(0.2) = 1.10$  to  $1.90$

Step 5: Conclusion: Make a decision to reject or fail to reject the null hypothesis.

Since the hypothesized value of 1.0 for the slope coefficient falls outside the confidence interval, we can reject the null hypothesis. This means we are 95 percent confident that the interval for the slope coefficient does not contain 1.0.

**Method 2: Using the t-test of significance**

Step 1: Construct the null hypothesis.  $H_0: b_1 = 1.0$

Step 2: Compute the t-statistic using the formula below:

$$t = \frac{\widehat{b}_1 - b_1}{s_{\widehat{b}_1}}$$

The t-statistic is for a t-distribution with  $(n - 2)$  degrees of freedom because two parameters are estimated in the regression.

$$t = \frac{1.5 - 1.0}{0.2} = 2.5$$

Step 3: Compare the absolute value of the t-statistic to  $t_c$ , and make a decision to reject or fail to reject the null hypothesis. If the absolute value of  $t$  is greater than  $t_c$ , then reject the null hypothesis. Since  $2.5 > 2$ , we can reject the null hypothesis that  $b_1 = 1.0$ .

Notice that both the approaches give the same result.

**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  $p\text{-value} \leq \alpha$ , then reject the null hypothesis  $H_0$ .  $\alpha$  is the significance level.

**Example 14: Estimating beta for General Motors stock**

Given the table below that shows the results of the regression analysis, test the null hypotheses

that  $\beta$  for GM stock equals 1 against the alternative hypotheses that  $\beta$  does not equal 1.

The regression equation is  $Y = 0.0036 + 1.1958X + \varepsilon$

Regression Statistics			
Multiple R	0.5549		
R-squared	0.3079		
Standard error of estimate	0.0985		
Observations	60		
	Coefficients	Standard Error	t-Statistic
Alpha	0.0036	0.0127	0.284
Beta	1.1958	0.2354	5.0795

Source: Ibbotson Associates and Bloomberg L.P.

### Solution:

We will test the null hypotheses using the confidence interval and t-test approaches.

**Method 1:** See if the hypothesized value falls within the confidence interval. We follow the same steps as the previous example.

- The critical value of the test statistic at the 0.05 significance level with 58 degrees of freedom  $t_c \approx 2$ .
- Construct the 95 percent confidence interval for the data for any hypothesized value of  $\beta$ :  
 $\hat{\beta} \pm t_c * s_{\hat{\beta}} = 1.1958 \pm 2 (0.2354) = 0.7250$  to  $1.6666$
- Conclusion: Since the hypothesized value of  $\beta = 1$  falls in this confidence interval, we cannot reject the hypothesis at the 0.05 significance level. This also means that we cannot reject the hypotheses that GM stock has the systematic risk as the market.

**Method 2:** Using the t-statistic to test the significance level.

- Compute the t-statistic for GM using equation 7:  $t = \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} = \frac{1.1958 - 1.0}{0.2354} = 0.8318$
- Since t-statistic is less than the critical value of 2, we fail to reject the null hypothesis.

Other observations from the data given:

- A  $R^2$  of 0.3079 indicates only 31 percent of the total variation in the excess return to GM stock can be explained by excess return to the market portfolio. The remaining 69 percent comes from non-systematic risk.
- Smaller standard errors of an estimated parameter, result in tighter confidence intervals.

### Example 15: Explaining the company value based on returns to invested capital

This example shows a regression hypothesis test with a one-sided alternative. The regression equation  $\frac{EV_i}{IC_i} = b_0 + b_1(\text{ROIC}_i - \text{WACC}_i) + \varepsilon_i$  where the subscript  $i$  is an index to identify the company, tests the relationship between  $\frac{EV}{IC}$  and  $\text{ROIC} - \text{WACC}$ . Our null hypothesis is  $H_0: b_1 \leq 0$ ; the significance level is 0.05. Use hypothesis testing to test the relationship between  $\frac{EV}{IC}$  and  $(\text{ROIC} - \text{WACC})$ . The results of the regression are given:

Regression Statistics			
Multiple R	0.9469		
R-squared	0.8966		
Standard error of estimate	0.7422		
Observations	9		
	Coefficients	Standard Error	t-Statistic
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson (2003).

### Solution:

The t-statistic for a coefficient reported by software programs assumes the hypothesized value to be 0. If you are testing for a different null hypothesis, as in the previous example, then the t-statistic must be computed. In this example, however, we are testing for  $b_1 = 0$ . Since the t-statistic of 7.7928 is greater than  $t_c$ , we can reject the null hypothesis and conclude there is a statistically significant relationship between  $\text{EV}/\text{IC}$  and  $(\text{ROIC} - \text{WACC})$ .  $R^2$  of 0.8966 implies  $\text{ROIC} - \text{WACC}$  explains about 90 percent of the variation in  $\text{EV}/\text{IC}$ .

### Example 16: Testing whether inflation forecasts are unbiased

In this example, we test the null hypothesis for two parameters: a slope of 0 and a slope coefficient of 1. In Example 11, we saw that for an unbiased forecast, the expected value of the error term is zero and  $E(\text{actual change} - \text{predicted change}) = 0$ . For the average forecast error to be 0, the value of  $b_0$  (the intercept) should be 0 and the value of  $b_1$  (slope) should be 1 in this regression equation:

$$\text{Actual percentage change} = b_0 + b_1 (\text{predicted change}) + \varepsilon_i$$

The following data is given:

<b>Regression Statistics</b>			
Multiple R	0.7138		
R-squared	0.5095		
Standard error of estimate	1.0322		
Observations	80		
	<b>Coefficients</b>	<b>Standard Error</b>	<b>t-Statistic</b>
Intercept	-0.0140	0.3657	-0.0384
Forecast (slope)	0.9637	0.1071	9.0008

Sources: Federal Reserve Banks of Philadelphia and St. Louis.

Using hypothesis testing, determine if the forecasts are unbiased.

### Solution:

1. Define the null hypotheses. First null hypothesis  $H_0: b_0 = 0$ ; alternative hypothesis  $H_a: b_0 \neq 0$ ; second null hypothesis  $H_0: b_1 = 1$ ;  $b_1 \neq 1$ .
2. Select the significance level. We choose a significance level of 0.05.
3. Determine the critical t value. The t-statistic at 0.05 significance level and 78 degrees of freedom is 1.99.
4. For the first null hypothesis, see if the hypothesized value of  $b_0 = 0$  falls within a 95 percent confidence interval. The confidence interval is from  $\widehat{b}_0 - t_c s_{\widehat{b}_0}$  to  $\widehat{b}_0 + t_c s_{\widehat{b}_0}$ . The estimated value of  $b_0$  is -0.0140. 95 percent confidence interval = -0.0140 - 1.99 (0.3657) to -0.0140 + 1.99 (0.3657) = -0.7417 to 0.7137. Since  $b_0 = 0$  falls within this confidence interval, we cannot reject the null hypothesis that  $b_0 = 0$ .

5. For the second null hypothesis, see if the hypothesized value of  $b_1 = 1$  falls within a 95 percent confidence interval. The confidence interval is from  $\widehat{b}_1 - t_c s_{\widehat{b}_1}$  to  $\widehat{b}_1 + t_c s_{\widehat{b}_1}$ . The estimated value of  $b_1$  is 0.9637. 95 percent confidence interval =  $0.9637 - 1.99 (0.1071)$  to  $+ 1.99 (0.3657) = 0.7506$  to  $1.1768$ . Since  $b_1 = 1$  falls within this confidence interval, we cannot reject the null hypothesis that  $b_1 = 1$ .
6. Conclusion: We cannot reject the null hypotheses that the forecasts of CPI change were unbiased.

LO.g: Formulate a null and alternative hypothesis about a population value of a regression coefficient, and determine the appropriate test statistic and whether the null hypothesis is rejected at a level of significance.

### 3.6. Analysis of Variance in a Regression with One Independent Variable

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. In simple terms, ANOVA explains the variation in the dependent variable for different levels of the independent variables.

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 parameters to be estimated (self-assessment: how many parameters are estimated in a one-independent variable regression?)
- The sum of squared errors or residuals (SSE):  $\sum_{i=1}^N (Y_i - \widehat{Y}_i)^2$ . This is also known as residual sum of squares.

- The regression sum of squares (RSS),  $\sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2$ . This is the amount of total variation in Y that is explained in the regression equation.  $TSS = SSE + RSS$

The F-statistic is the ratio of the average regression sum of squares to the average sum of the squared errors. Average regression sum of squares (RSS) is the amount of variation in Y explained by the regression equation.

$$F = \frac{\frac{RSS}{1}}{\frac{SSE}{n-2}} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}}$$

where

$$\text{Mean regression sum of squares} = \frac{\text{regression sum of squares}}{\text{number of slope parameters estimated}}$$

$$\text{Mean squared error} = \frac{\text{sum of squared errors}}{\text{number of observations } n - \text{total number of parameters estimated}}$$

### Interpretation and application of F-test statistic:

- The higher the F-statistic, the better it is as the regression model does a good job of explaining the variation in the dependent variable.
- The F-statistic is 0 if the independent variable explains no variation in the dependent variable.
- The F-statistic is used to evaluate whether a fund has generated positive alpha. The null hypothesis for this test will be:  $H_0: \alpha = 0$  versus  $H_a: \alpha \neq 0$ .
- F-statistic is usually not used in a regression with one independent variable because the F-statistic is the square of the t-statistic for the slope coefficient.

### Example 17: Performance evaluation: The Dreyfus Appreciation Fund

This example illustrates how the F-test reveals nothing more than what the t-test already does. The ANOVA table and results of regression to evaluate the performance of the Dreyfus appreciation fund is given below:

Regression Statistics			
Multiple R	0.928		

R-squared	0.8611		
Standard error of estimate	0.0174		
Observations	60		
<b>ANOVA</b>	<b>Degrees of Freedom (df)</b>	<b>Mean Sum of Squares (MSS)</b>	<b>F</b>
Regression	1	0.1093	359.64
Residual	58	0.0003	
Total	59		

	<b>Coefficients</b>	<b>Standard Error</b>	<b>t-Statistic</b>
Alpha	0.0009	0.0023	0.4036
Beta	0.7902	0.0417	18.9655

*Source:* Center for Research in Security Prices, University of Chicago.

*Note:* From an exam perspective, be familiar with how to read the values from a table like this and understand what values matter in coming to a conclusion.

Test the null hypothesis that  $\alpha = 0$ , i.e. the fund did not have a significant excess return beyond the return associated with the market risk of the fund.

### **Solution:**

Recall that the software tests the null hypothesis for slope coefficient = 0, so we can use the given t-statistic. Since the t-statistic of 18.9655 is high, the probability that the coefficient is 0 is very small.

We get the same result using the F-statistic as well.

$F = \frac{\frac{0.1093}{1}}{\frac{0.0176}{60-2}} = 360.19$ . The p-value for F-statistic is less than 0.0001. This is much lower than the significance level.

LO.j: describe the use of analysis of variance (ANOVA) in regression analysis; interpret ANOVA results, calculate and interpret the F-statistic;

### **3.7. Prediction Intervals**

We use regression intervals to make predictions about a dependent variable. Earlier we discussed the construction of a confidence interval, which is a range of values that is likely to contain the value of an unknown population parameter. With a 95 percent confidence interval, are we 95 percent confident that, say  $\beta$ , will have a value of 1?

Prediction intervals focus on the accuracy. It represents an interval of values associated with a parameter that we believe includes the true parameter value,  $b_1$ , with a specified probability.

Let us consider the regression equation:  $Y = b_0 + b_1X$ . The predicted value of  $\hat{Y} = \hat{b}_0 + \hat{b}_1X$ .

The two sources of uncertainty in regression analysis using the estimated parameters 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:

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

*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  $\bar{X}$
- 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  $\alpha$ .
4. Compute the  $(1-\alpha)$  prediction interval using the formula below:

$$\hat{Y} \pm t_c * s_f$$

This part is not covered in the curriculum. But, it may help in understanding how a prediction interval is different from a confidence interval, as both are essentially a range of values:

- It is wider than the confidence interval.
- The interval accounts for movements in  $y$ , away from its mean value, in the future.
- Any random future value of  $y$ .

### Example 18: Predicting the ratio of enterprise value to invested capital

Given the results of a regression analysis below, predict the ratio of enterprise value to invested capital for the company at a 95 percent confidence interval. The return spread between ROIC and WACC is given as 10 percentage points.

Regression Statistics			
Multiple R	0.9469		
R-squared	0.8966		
Standard error of estimate	0.7422		
Observations	9		
	Coefficients	Standard Error	t-Statistic
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson (2003).

### Solution:

1. Expected EV/IC =  $b_0 + b_1 (\text{ROIC-WACC}) + \varepsilon$

Plugging in values from the table, we have:

$$\text{Expected EV/IC} = 1.3478 + 30.0169 (0.1) = 4.3495$$

This means that if the return spread is 10 percent, then the EV/IC ratio will be 4.3495.

2. Compute the variance of the prediction error using equation 7.

$$s_f^2 = s^2 * \left[ 1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_x^2} \right] = 0.7422^2 * \left[ 1 + \frac{1}{9} + \frac{(0.1 - 0.0647)^2}{(9-1)0.004641} \right]$$

$$= 0.630556$$

$$S_f = 0.7941$$

3.  $t_c$  at the 95 percent confidence level and 7 degrees of freedom is 2.365.  
 4. 95 percent prediction interval using equation 8 is:

$$4.3495 - 2.365 (0.7941) \text{ to } 4.3495 + 2.365 (0.7941) = 2.4715 \text{ to } 6.2275.$$

LO.i: Calculate the predicted value for the dependent variable, given an estimated regression model and a value for the independent variable.

LO.j: Calculate and interpret a confidence interval for the predicted value of the dependent variable.

### 3.8. Limitations of Regression Analysis

Following are the limitations of regression analysis:

- Regression relations can change over time as do correlations. This is called parameter instability. This characteristic is observed in both cross-series and time-series regression relationships.
- Regression analysis is often difficult to apply because of specification issues and assumption violations.
  - Specification issues: Identifying the independent variables and the dependent variable, and formulating the regression equation may be challenging.
  - Assumptions violations: Often there is an uncertainty in determining if an assumption has been violated.
- Public knowledge of regression relationships may limit their usefulness in the future. For example, is low P/E stocks in the sugar industry have had historically high returns during Oct-Jan every year, then this knowledge may cause other analysts to buy the stock which

will push their prices up.

LO.k: Describe the limitations of the regression analysis.

## 4. Summary

Below is a summary of the important points discussed in this reading:

- Correlation analysis is used to measure the strength of the relationship between two variables. If there are two population data sets A and B, a sample is drawn from each one of them and we attempt to find a correlation between the samples.
- The sample correlation is denoted by  $r$ , and the population correlation by  $\rho$ . The correlation coefficient has no units. It can take a value between -1 and 1.
- The sample covariance is 
$$\text{Cov}(X,Y) = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$
- The sample correlation coefficient of two variables X and Y is 
$$\frac{\text{Cov}(X,Y)}{S_X * S_Y}$$
- Significance tests through hypothesis testing allow us to assess whether apparent relationships between random variables are the result of chance. The null hypothesis is  $H_0 = 0$  and alternate hypothesis is  $H_a \neq 0$ .
- A t-test is used to determine whether the correlation between two variables is significant. As long as the two variables are distributed normally, we can test to determine whether the null hypothesis should be rejected using the sample correlation,  $r$ . The formula for the t-test is:
 
$$t = \frac{r\sqrt{n-2}}{1-r^2}$$
- How to use the t-test: calculate the t-statistic, and compare it with  $t_c$ . If absolute value of  $t > t_c$ , then reject  $H_0$ . If you reject  $H_0$ , then there is a significant linear correlation.
- Linear regression allows us to use one variable to make predictions about another, test hypotheses about the relation between two variables, and quantify the strength of the relationship between the two variables. Linear regression assumes a linear relationship

between the dependent and the independent variables.

- A simple linear regression using one independent variable can be expressed as:

$$Y_i = b_0 + b_1X_i + \varepsilon_i \text{ where } i = 1, 2, \dots, n$$

- The regression model chooses the estimated values for  $\widehat{b}_0$  and  $\widehat{b}_1$  such that this term  $\sum_{i=1}^N (Y_i - \widehat{b}_0 - \widehat{b}_1X_i)^2$  is minimized.
- The variable used to explain the dependent variable is the independent variable, X. The variable being explained is the dependent variable, Y.

- Standard error of estimate:  $SEE = \left( \sum_{i=1}^N \frac{(Y_i - \widehat{b}_0 - \widehat{b}_1X_i)^2}{n-2} \right)^{\frac{1}{2}} =$

$$\left( \frac{\sum_{i=1}^N (\widehat{\varepsilon}_i)^2}{n-2} \right)^{\frac{1}{2}}$$

- Coefficient of determination

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

- The higher the  $R^2$ , the more useful the model.  $R^2$  has a value between 0 and 1.
- Correlation coefficient, r, is also called multiple-R.
- A **confidence interval** is an interval of values that we believe includes the true parameter value,  $b_1$ , with a given degree of confidence.
- The confidence interval is given by  $\widehat{b}_1 - t_c s_{\widehat{b}_1}$  to  $\widehat{b}_1 + t_c s_{\widehat{b}_1}$  where  $t_c$  is the critical t value for a given level of significance and degrees of freedom and  $\widehat{b}_1$  is the estimated slope coefficient. If the hypothesized value is outside the confidence interval, then reject the null hypothesis.

- Alternatively, compare t-statistic =  $t = \frac{\widehat{b}_1 - b_1}{s_{\widehat{b}_1}}$  with  $t_c$ ; if  $t > t_c$ , then reject the null

hypothesis.

- The p-value is the smallest level of significance at which the null hypothesis can be rejected. Higher the t-statistic, smaller the p-value and the stronger the evidence to reject the null hypothesis.

- Analysis of variance is a statistical procedure for dividing the 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.
- The F-statistic measures how well the regression equation explains the variation in the dependent variable.

$$F = \frac{\frac{RSS}{1}}{\frac{SSE}{n-2}} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}}$$

where

$$\text{Mean regression sum of squares} = \frac{\text{regression sum of squares}}{\text{number of slope parameters estimated}}$$

$$\text{Mean squared error} = \frac{\text{sum of squared errors}}{\text{number of observations } n - \text{total number of parameters estimated}}$$

- The prediction interval is given by:  $\hat{Y} \pm t_c * S_f$
- The limitation of regression analysis is that relationships can change over time.