hypothesis of no difference, concluding that there *is* a statistically significant difference between mean firm betas before and after deregulation.

Keep in mind that we have been describing two distinct hypothesis tests: one about the significance of the difference between the means of two populations and one about the significance of the mean of the differences between pairs of observations. Here are rules for when these tests may be applied:

- The test of the differences in means is used when there are two *independent samples*.
- A test of the significance of the mean of the differences between paired observations is used when the samples are *not independent*.

PROFESSOR'S NOTE
Again, the LOS here says "[c]onstruct hypothesis tests and determine their statistical significance …" We can't believe candidates are expected to memorize these formulas (or that you would be a better analyst if you did). You should instead focus on the fact that both of these tests involve *t*-statistics and depend on the degrees of freedom. Also note that when samples are independent, you can use the difference in means test, and when they are dependent, we must use the paired comparison (mean differences) test. In that case, with a null hypothesis that there is no difference in means, the test statistic is simply the mean of the differences between each pair of observations, divided by the standard error of those differences. This is just a straightforward *t*-test of whether the mean of a sample is zero, which might be considered fair game for the exam.

Value of a Population Variance

The *chi-square test* is used for hypothesis tests concerning the variance of a normally distributed population. Letting σ^2 represent the true population variance and σ_0^2 represent the hypothesized variance, the hypotheses for a two-tailed test of a single population variance are structured as follows:

$$H_0: \sigma^2 = \sigma_0^2 \text{ versus } H_a: \sigma^2 \neq \sigma_0^2$$

The hypotheses for one-tailed tests are structured as follows:

$$H_0: \sigma^2 \leq \sigma_0^2 \text{ versus } H_a: \sigma^2 > \sigma_0^2 \text{ or } H_0: \sigma^2 \geq \sigma_0^2 \text{ versus } H_a: \sigma^2 < \sigma_0^2$$

Hypothesis testing of the population variance requires the use of a chi-square distributed test statistic, denoted χ^2. The chi-square distribution is asymmetrical and approaches the normal distribution in shape as the degrees of freedom increase.

To illustrate the chi-square distribution, consider a two-tailed test with a 5% level of significance and 30 degrees of freedom. As displayed in Figure 8.4, the critical chi-square values are 16.791 and 46.979 for the lower and upper bounds, respectively. These values are obtained from a chi-square table, which is used in the same manner as a *t*-table. A portion of a chi-square table is presented in Figure 8.5.

Note that the chi-square values in Figure 8.5 correspond to the probabilities in the right tail of the distribution. As such, the 16.791 in Figure 8.4 is from the column headed *0.975* because 95% + 2.5% of the probability is to the right of it. The 46.979 is from the column headed *0.025* because only 2.5% probability is to the right of it. Similarly, at a 5% level of significance with 10 degrees of freedom, Figure 8.5 shows that the critical chi-square values for a two-tailed test are 3.247 and 20.483.

Figure 8.4: Decision Rule for a Two-Tailed Chi-Square Test

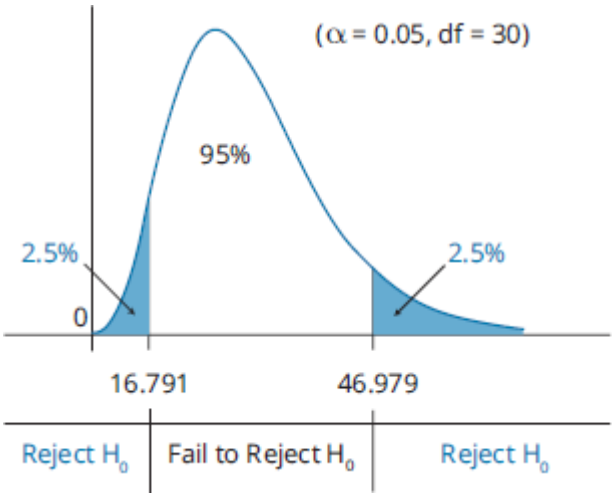


Figure 8.5: Chi-Square Table

	Probability in Right Tail					
Degrees of Freedom	0.975	0.95	0.90	0.1	0.05	0.025
9	2.700	3.325	4.168	14.684	16.919	19.023
10	3.247	3.940	4.865	15.987	18.307	20.483
11	3.816	4.575	5.578	17.275	19.675	21.920
30	16.791	18.493	20.599	40.256	43.773	46.979

The chi-square test statistic, χ^2, with n − 1 degrees of freedom, is computed as follows:

$$\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma_0^2}$$

where:

n = sample size

s^2 = sample variance

σ_0^2 = hypothesized value for the population variance

Similar to other hypothesis tests, the chi-square test compares the test statistic, χ^2_{n-1}, to a critical chi-square value at a given level of significance and n − 1 degrees of freedom. Because the chi-square distribution is bounded below by zero, chi-square values cannot be negative.

EXAMPLE: Chi-square test for a single population variance

Historically, the High-Return Equity Fund has advertised that its monthly returns have a standard deviation equal to 4%. This was based on estimates from the 2005–2013 period. High-Return wants to verify whether this claim still adequately describes the standard deviation of the fund's returns. High-Return collected monthly returns for the 24-month period between 2013 and 2015 and measured a standard deviation of monthly returns of 3.8%. High-Return calculates a test statistic of 20.76. Using a 5% significance level, determine if the more recent standard deviation is different from the advertised standard deviation.

Answer:

The null hypothesis is that the standard deviation is equal to 4% and, therefore, the variance of monthly returns for the population is $(0.04)^2 = 0.0016$. Because High-Return simply wants to test whether the standard deviation has changed, up or down, a two-sided test should be used. The hypothesis test structure takes this form:

$$H_0: \sigma_0^2 = 0.0016 \text{ versus } H_a: \sigma^2 \neq 0.0016$$

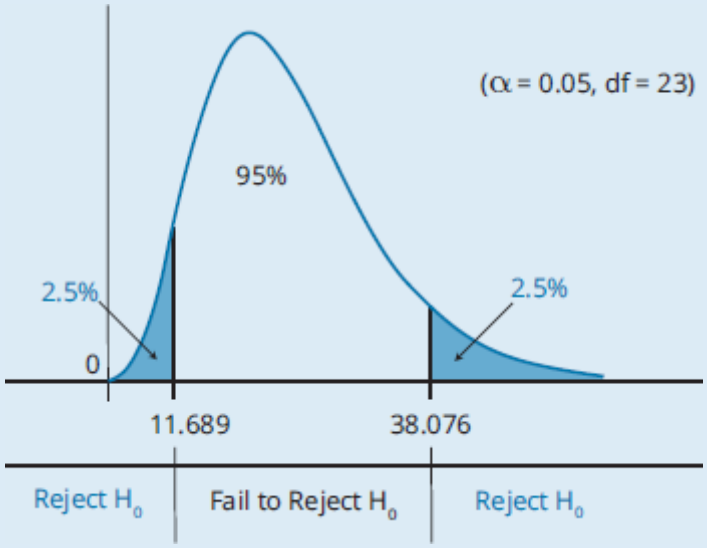
The appropriate test statistic for tests of variance is a chi-square statistic.

With a 24-month sample, there are 23 degrees of freedom. Using the table of chi-square values in Appendix E of this book, for 23 degrees of freedom and probabilities of 0.975 and 0.025, we find two critical values, 11.689 and 38.076. Thus, this is the decision rule:

$$\text{Reject } H_0 \text{ if } \chi^2 < 11.689, \text{ or } \chi^2 > 38.076$$

This decision rule is illustrated in the following figure.

Decision Rule for a Two-Tailed Chi-Square Test of a Single Population Variance



Because the computed test statistic, χ^2, falls between the two critical values, we cannot reject the null hypothesis that the variance is equal to 0.0016. The recently measured standard deviation is close enough to the advertised standard deviation that we cannot say that it is different from 4%, at a 5% level of significance.

Comparing Two Population Variances

The hypotheses concerned with the equality of the variances of two populations are tested with an *F*-distributed test statistic. Hypothesis testing using a test statistic that follows an *F*-distribution is referred to as the *F*-test. The *F*-test is used under the assumption that the populations from which samples are drawn are normally distributed, and that the samples are independent.

If we let σ_1^2 and σ_2^2 represent the variances of normal Population 1 and Population 2, respectively, the hypotheses for the two-tailed *F*-test of differences in the variances can be structured as follows:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ versus } H_a: \sigma_1^2 \neq \sigma_2^2$$

The one-sided test structures can be specified as follows:

$$H_0: \sigma_1^2 \leq \sigma_2^2 \text{ versus } H_a: \sigma_1^2 > \sigma_2^2, \text{ or } H_0: \sigma_1^2 \geq \sigma_2^2 \text{ versus } H_a: \sigma_1^2 < \sigma_2^2$$

The test statistic for the *F*-test is the ratio of the sample variances. The *F*-statistic is computed as follows:

$$F = \frac{s_1^2}{s_2^2}$$

where:

S_1^2 = variance of the sample of n_1 observations drawn from Population 1

S_2^2 = variance of the sample of n_2 observations drawn from Population 2

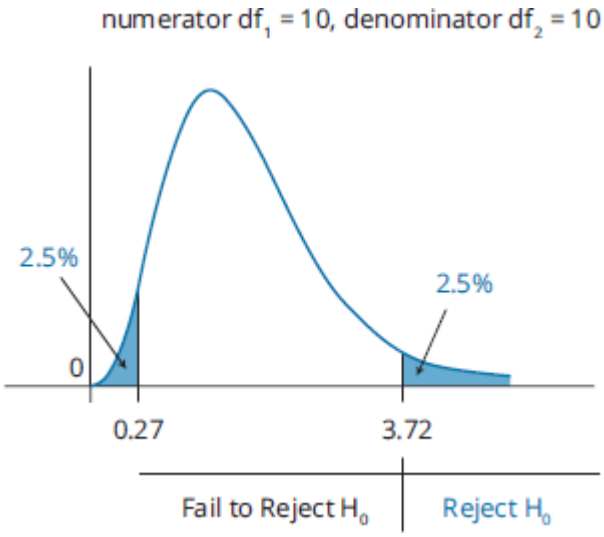
Note that $n_1 - 1$ and $n_2 - 1$ are the degrees of freedom used to identify the appropriate critical value from the *F*-table (provided in this book's Appendix).

PROFESSOR'S NOTE

Always put the larger variance in the numerator (s_1^2). Following this convention means that we only have to consider the critical value for the right-hand tail.

An ***F-distribution*** is presented in Figure 8.6. As indicated, the *F*-distribution is right skewed and is bounded by zero on the left-hand side. The shape of the *F*-distribution is determined by *two separate degrees of freedom*: the numerator degrees of freedom, df_1, and the denominator degrees of freedom, df_2.

Figure 8.6: *F*-Distribution



Note that when the sample variances are equal, the value of the test statistic is 1. The upper critical value is always greater than one (the numerator is significantly greater than the denominator), and the lower critical value is always less than one (the numerator is significantly smaller than the denominator). In fact, the lower critical value is the reciprocal of the upper critical value. For this reason, in practice, we put the larger sample variance in the numerator and consider only the upper critical value.

EXAMPLE: *F*-test for equal variances

Annie Cower is examining the earnings for two different industries. Cower suspects that the variance of earnings in the textile industry is different from the variance of earnings in the paper industry. To confirm this suspicion, Cower has looked at a sample of 31 textile manufacturers and a sample of 41 paper companies. She measured the sample standard deviation of earnings across the textile industry to be $4.30, and that of the paper industry companies to be $3.80. Cower calculates a test statistic of 1.2805. Using a 5% significance level, determine if the earnings of the textile industry have a different standard deviation than those of the paper industry.

Answer:

In this example, we are concerned with whether the variance of earnings for companies in the textile industry is equal to the variance of earnings for companies in the paper industry. As such, the test hypotheses can be appropriately structured as follows:

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ versus } H_a: \sigma_1^2 \neq \sigma_2^2$$

For tests of difference between variances, the appropriate test statistic is:

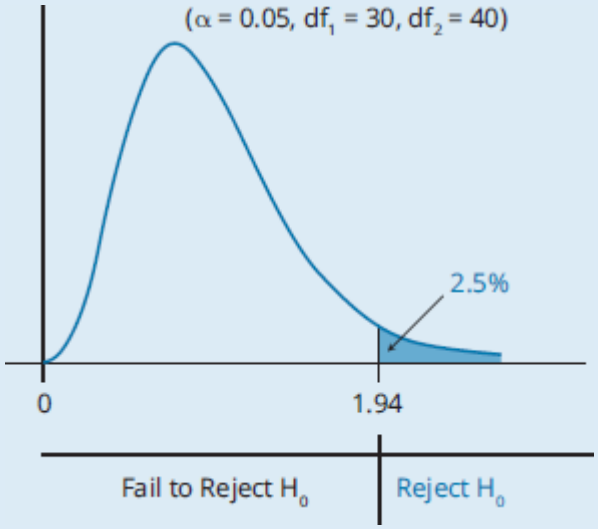
$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 is the larger sample variance.

Using the sample sizes for the two industries, the critical *F*-value for our test is found to be 1.94. This value is obtained from the table of the *F*-distribution for 2.5% in the upper tail, with $df_1 = 30$ and $df_2 = 40$. Thus, if the computed *F*-statistic is greater than the critical value of 1.94, the null hypothesis is rejected. The decision rule, illustrated in the following figure, can be stated as follows:

Reject H_0 if $F > 1.94$

Decision Rule for *F*-Test



Because the calculated *F*-statistic of 1.2805 is less than the critical *F*-statistic of 1.94, Cower cannot reject the null hypothesis. Cower should conclude that the earnings variances of the industries are not significantly different from one another at a 5% level of significance.

LOS 8.c: Compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test.

Parametric tests rely on assumptions regarding the distribution of the population and are specific to population parameters. For example, the *z*-test relies upon a mean and a standard deviation to define the normal distribution. The *z*-test also requires that either the sample is large, relying on the central limit theorem to assure a normal sampling distribution, or that the population is normally distributed.

Nonparametric tests either do not consider a particular population parameter or have few assumptions about the population that is sampled. Nonparametric tests are used when there is concern about quantities other than the parameters of a distribution, or when the assumptions of parametric tests can't be supported. They are also used when the data are not suitable for parametric tests (e.g., ranked observations).

Some situations where a nonparametric test is called for may include the following:

1. The assumptions about the distribution of the random variable that support a parametric test are not met. An example would be a hypothesis test of the mean

value for a variable that comes from a distribution that is not normal and is of small size, so that neither the *t*-test nor the *z*-test is appropriate.

2. A nonparametric test is called for when data are ranks (an ordinal measurement scale) rather than values.
3. The hypothesis does not involve the parameters of the distribution, such as testing whether a variable is normally distributed. We can use a nonparametric test, called a runs test, to determine whether data are random. A runs test provides an estimate of the probability that a series of changes (e.g., +, +, −, −, +, −,....) are random.

MODULE QUIZ 8.2

1. Which of the following assumptions is *least likely* required for the difference in means test based on two samples?
   A. The two samples are independent.
   B. The two populations are normally distributed.
   C. The two populations have known variances.
2. The appropriate test statistic for a test of the equality of variances for two normally distributed random variables, based on two independent random samples, is the:
   A. *t*-test.
   B. *F*-test.
   C. χ^2 test.
3. The appropriate test statistic to test the hypothesis that the variance of a normally distributed population is equal to 13 is the:
   A. *t*-test.
   B. *F*-test.
   C. χ^2 test.

KEY CONCEPTS

LOS 8.a

The hypothesis testing process requires a statement of a null and an alternative hypothesis, the selection of the appropriate test statistic, specification of the significance level, a decision rule, the calculation of a sample statistic, a decision regarding the hypotheses based on the test, and a decision based on the test results.

The null hypothesis is what the researcher wants to reject. The alternative hypothesis is what the researcher wants to support, and it is accepted when the null hypothesis is rejected.

A Type I error is the rejection of the null hypothesis when it is actually true, while a Type II error is the failure to reject the null hypothesis when it is actually false.

The significance level can be interpreted as the probability that a test statistic will reject the null hypothesis by chance when it is actually true (i.e., the probability of a Type I error). A significance level must be specified to select the critical values for the test.

The power of a test is the probability of rejecting the null when it is false. The power of a test = 1 − P(Type II error).

The *p*-value for a hypothesis test is the smallest significance level for which the hypothesis would be rejected.

LOS 8.b

Hypothesis tests of:	Use a:	With degrees of freedom:
One population mean	*t*-statistic	$n - 1$
Two population means	*t*-statistic	$n - 1$
One population variance	Chi-square statistic	$n - 1$
Two population variances	*F*-statistic	$n_1 - 1, n_2 - 1$

LOS 8.c

Parametric tests, like the *t*-test, *F*-test, and chi-square test, make assumptions regarding the distribution of the population from which samples are drawn. Nonparametric tests either do not consider a particular population parameter or have few assumptions about the sampled population. Nonparametric tests are used when the assumptions of parametric tests can't be supported, or when the data are not suitable for parametric tests.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 8.1

1. **C** A Type I error is rejecting the null hypothesis when it is true. The probability of rejecting a false null is [1 − Prob Type II] = [1 − 0.60] = 40%, which is called the power of the test. The other answer choices are not necessarily true, because the null may be false and the probability of rejection unknown. (LOS 8.a)

2. **A** The power of a test is 1 − P(Type II error) = 1 − 0.15 = 0.85. (LOS 8.a)

Module Quiz 8.2

1. **C** The difference in means test does not require the two population variances to be known. (LOS 8.b)

2. **B** The *F*-test is the appropriate test. (LOS 8.b)

3. **C** A test of the population variance is a chi-square test. (LOS 8.b)

READING 9

PARAMETRIC AND NON-PARAMETRIC TESTS OF INDEPENDENCE

MODULE 9.1: TESTS FOR INDEPENDENCE

Video covering this content is available online.

LOS 9.a: Explain parametric and nonparametric tests of the hypothesis that the population correlation coefficient equals zero, and determine whether the hypothesis is rejected at a given level of significance.

Correlation measures the strength of the relationship between two variables. If the correlation between two variables is zero, there is no linear relationship between them. When the sample correlation coefficient for two variables is different from zero, we must address the question of whether the true population correlation coefficient (ρ) is equal to zero. The appropriate test statistic for the hypothesis that the population correlation equals zero, when the two variables are normally distributed, is as follows:

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

where:
r = sample correlation
n = sample size

This test statistic follows a *t*-distribution with $n - 2$ degrees of freedom. Note that the test statistic increases, not only with the sample correlation coefficient, but also with sample size.

EXAMPLE: Test of the hypothesis that the population correlation coefficient equals zero

A researcher computes the sample correlation coefficient for two normally distributed random variables as 0.35, based on a sample size of 42. Determine whether to reject the hypothesis that the population correlation coefficient is equal to zero at a 5% significance level.

Answer:

Our test statistic is $\frac{0.35\sqrt{42-2}}{\sqrt{1-0.35^2}} = 2.363$.

Using the *t*-table with 42 − 2 = 40 degrees of freedom for a two-tailed test and a significance level of 5%, we can find the critical value of 2.021. Because our computed test statistic of 2.363 is greater than 2.021, we reject the hypothesis that the population mean is zero and conclude that it is not equal to zero. That is, the two populations are correlated—in this case, positively.

PROFESSOR'S NOTE

The correlation coefficient we refer to here is the Pearson correlation coefficient, which is a measure of the linear relationship between two variables. There are other correlation coefficients that better measure the strength of any nonlinear relationship between two variables.

The **Spearman rank correlation test**, a nonparametric test, can be used to test whether two sets of ranks are correlated. Ranks are simply ordered values. If there is a tie (equal values), the ranks are shared—so if second and third rank is the same, the ranks are shared, and each gets a rank of (2 + 3) / 2 = 2.5.

The Spearman rank correlation, r_s (when all ranks are integer values), is calculated as follows:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where:

r_S = rank correlation

n = sample size

d_i = difference between two ranks

We can test the significance of the Spearman rank correlation calculated with the formula just listed using the same test statistic we used for estimating the significance of a parametric correlation coefficient:

$$\frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

When the sample size is greater than 30, the test statistic follows a *t*-distribution with n − 2 degrees of freedom.

LOS 9.b: Explain tests of independence based on contingency table data.

A contingency or two-way table shows the number of observations from a sample that have a combination of two characteristics. Figure 9.1 is a contingency table where the characteristics are earnings growth (low, medium, or high) and dividend yield (low, medium, or high). We can use the data in the table to test the hypothesis that the two characteristics, earnings growth and dividend yield, are independent of each other.

Figure 9.1: Contingency Table for Categorical Data

Earnings Growth	Dividend Yield			
	Low	Medium	High	Total
Low	28	53	42	123
Medium	42	32	39	113
High	49	25	14	88
Total	119	110	95	324

We index our three categories of earnings growth from low to high with i = 1, 2, or 3, and our three categories of dividend yield from low to high with j = 1, 2, or 3. From the table, we see in Cell 1,1 that 28 firms have both low earnings growth and low dividend yield. We see in Cell 3,2 that 25 firms have high earnings growth and medium dividend yields.

For our test, we are going to compare the actual table values to what the values would be if the two characteristics were independent. The test statistic is a chi-square test statistic calculated as follows:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

where:

O_{ij} = the number of observations in Cell i,j: Row i and Column j (i.e., observed frequency)

E_{ij} = the expected number of observations for Cell i,j

r = the number of row categories

c = the number of column categories

The degrees of freedom are (r − 1) × (c − 1), which is 4 in our example for dividend yield and earnings growth.

E_{ij}, the expected number of observations in Cell i,j, is:

$$\frac{\text{total for Row i} \times \text{total for Column j}}{\text{total for all columns and rows}}$$

The expected number of observations for Cell 2,2 is:

$$\frac{110 \times 113}{324} = 38.4$$

In calculating our test statistic, the term for Cell 2,2 is:

$$\frac{(32-38.4)^2}{38.4} = 1.0667$$

Figure 9.2 shows the expected frequencies for each pair of categories in our earnings growth and dividend yield contingency table.

Figure 9.2: Contingency Table for Expected Frequencies

Earnings Growth	Dividend Yield		
	Low	Medium	High
Low	45.2	41.8	36.1
Medium	41.5	38.4	33.1
High	32.3	29.9	25.8

For our test statistic, we sum, for all nine cells, the squared difference between the expected frequency and observed frequency, divided by the expected frequency. The resulting sum is 27.47. Figure 9.3 shows the results for each cell in calculating the test statistic.

Figure 9.3: Squared Differences from Contingency Table

Earnings Growth	Dividend Yield		
	Low	Medium	High
Low	6.5451	3.0010	0.9643
Medium	0.0060	1.0667	1.0517
High	8.6344	0.8030	5.3969
	Sum = 27.4691		

Our degrees of freedom are (3 − 1) × (3 − 1) = 4. The critical value for a significance level of 5% (from the chi-square table in the Appendix) with 4 degrees of freedom is 9.488. Based on our sample data, we can reject the hypothesis that the earnings growth and dividend yield categories are independent.

MODULE QUIZ 9.1

1. The test statistic for a Spearman rank correlation test for a sample size greater than 30 follows a:
   A. *t*-distribution.
   B. normal distribution.
   C. chi-square distribution.
2. A contingency table can be used to test:
   A. a null hypothesis that rank correlations are equal to zero.
   B. whether multiple characteristics of a population are independent.
   C. the number of *p*-values from multiple tests that are less than adjusted critical values.
3. For a parametric test of whether a correlation coefficient is equal to zero, it is *least likely* that:
   A. degrees of freedom are n − 1.
   B. the test statistic follows a *t*-distribution.
   C. the test statistic increases with a greater sample size.

KEY CONCEPTS

LOS 9.a

To test a hypothesis that a population correlation coefficient equals zero, the appropriate test statistic is a *t*-statistic with n − 2 degrees of freedom, calculated as $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, where *r* is the sample correlation coefficient.

A nonparametric test of correlation can be performed when we have only ranks (e.g., deciles of investment performance). The Spearman rank correlation test examines whether the ranks for multiple periods are correlated. The rank correlation is $r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$, where d_i^2 is the sum of the squared differences in pairs of ranks and *n* is the number of sample periods. The test statistic follows a *t*-distribution for samples sizes greater than 30.

LOS 9.b

A contingency table can be used to test the hypothesis that two characteristics (categories) of a sample of items are independent. A contingency table shows the number of the sample items (e.g., firms that have both of two characteristics). The test statistic follows a chi-square distribution and is calculated as follows:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

where:

O_{ij} = the number of observations in Cell i,j: Row *i* and Column *j* (i.e., observed frequency)

E_{ij} = the expected number of observations for Cell i,j of the contingency table with independence

r = the number of row categories

c = the number of column categories

The degrees of freedom are (r − 1) × (c − 1). If the test statistic is greater than the critical chi-square value for a given level of significance, we reject the hypothesis that the two characteristics are independent.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 9.1

1. **A** The test statistic for a Spearman rank correlation test follows a *t*-distribution. (LOS 9.a)

2. **B** A contingency table is used to determine whether two characteristics of a group are independent. (LOS 9.b)

3. **A** Degrees of freedom are n − 2 for a test of the hypothesis that correlation is equal to zero. The test statistic increases with sample size (degrees of freedom increase) and follows a *t*-distribution. (LOS 9.a)

READING 10

SIMPLE LINEAR REGRESSION

MODULE 10.1: LINEAR REGRESSION BASICS

Video covering this content is available online.

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

The purpose of **simple linear regression** is to explain the variation in a dependent variable in terms of the variation in a single independent variable. Here, the term *variation* is interpreted as the degree to which a variable differs from its mean value. Don't confuse *variation* with *variance*—they are related, but they are not the same.

$$\text{variation in Y} = \sum_{i}^{n}(Y_i - \overline{Y})^2$$

- The **dependent variable** is the variable whose variation is explained by the independent variable. We are interested in answering the question, "What explains fluctuations in the dependent variable?" The dependent variable is also referred to as the terms *explained variable, endogenous variable,* or *predicted variable*.
- The **independent variable** is the variable used to explain the variation of the dependent variable. The independent variable is also referred to as the terms *explanatory variable, exogenous variable,* or *predicting variable*.

EXAMPLE: Dependent vs. independent variables

Suppose you want to predict stock returns with GDP growth. Which variable is the independent variable?

Answer:

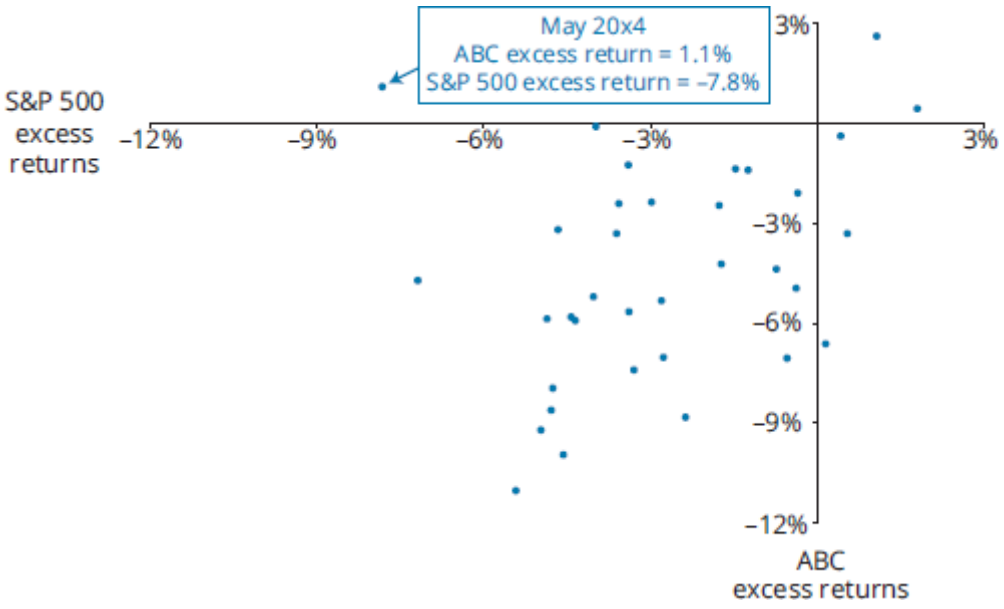
Because GDP is going to be used as a predictor of stock returns, stock returns are being *explained* by GDP. Hence, stock returns are the dependent (explained) variable, and GDP is the independent (explanatory) variable.

Suppose we want to use excess returns on the S&P 500 (the independent variable) to explain the variation in excess returns on ABC common stock (the dependent variable).

For this model, we define excess return as the difference between the actual return and the return on 1-month Treasury bills.

We would start by creating a scatter plot with ABC excess returns on the vertical axis and S&P 500 excess returns on the horizontal axis. Monthly excess returns for both variables from June 20X2 to May 20X5 are plotted in Figure 10.1. For example, look at the point labeled May 20X4. In that month, the excess return on the S&P 500 was −7.8%, and the excess return on ABC was 1.1%.

Figure 10.1: Scatter Plot of ABC Excess Returns vs. S&P 500 Index Excess Returns



The two variables in Figure 10.1 appear to be positively correlated: excess ABC returns tended to be positive (negative) in the same month that S&P 500 excess returns were positive (negative). This is not the case for all the observations, however (for example, May 20X4). In fact, the correlation between these variables is approximately 0.40.

Simple Linear Regression Model

The following linear regression model is used to describe the relationship between two variables, *X* and *Y*:

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, \ldots i = 1, \ldots, n$$

where:

Y_i = *i*th observation of the dependent variable, *Y*
X_i = *i*th observation of the independent variable, *X*
b_0 = regression intercept term
b_1 = regression slope coefficient
ε_i = **residual** for the *i*th observation (also referred to as the disturbance term or error term)

Based on this regression model, the regression process estimates an equation for a line through a scatter plot of the data that "best" explains the observed values for *Y* in terms of the observed values for *X*.

The linear equation, often called the line of best fit or **regression line**, takes the following form:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i, i = 1,2,3...,n$$

where:

$\hat{Y}_i$ = estimated value of Y_i given X_i

$\hat{b}_0$ = estimated intercept term

$\hat{b}_1$ = estimated slope coefficient

PROFESSOR'S NOTE

The hat "^" above a variable or parameter indicates a predicted value.

The regression line is just one of the many possible lines that can be drawn through the scatter plot of *X* and *Y*. The criteria used to estimate this line is the essence of linear regression. The regression line is the line that minimizes the sum of the squared differences (vertical distances) between the *Y*-values predicted by the regression equation ($\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$) and the actual *Y*-values, Y_i. The sum of the squared vertical distances between the estimated and actual *Y*-values is referred to as the **sum of squared errors (SSE)**.

Thus, the regression line is the line that minimizes the SSE. This explains why simple linear regression is frequently referred to as **ordinary least squares (OLS)** regression, and the values determined by the estimated regression equation, $\hat{Y}_i$, are called least squares estimates.

The estimated **slope coefficient** ($\hat{b}_1$) for the regression line describes the change in *Y* for a one-unit change in *X*. It can be positive, negative, or zero, depending on the relationship between the regression variables. The slope term is calculated as follows:

$$\hat{b}_1 = \frac{Cov_{XY}}{\sigma_X^2}$$

The intercept term ($\hat{b}_0$) is the line's intersection with the *Y*-axis at X = 0. It can be positive, negative, or zero. A property of the least squares method is that the intercept term may be expressed as follows:

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

where:

$\overline{Y}$ = mean of Y

$\overline{X}$ = mean of X

The intercept equation highlights the fact that the regression line passes through a point with coordinates equal to the mean of the independent and dependent variables (i.e., the point X, Y).

EXAMPLE: Computing the slope coefficient and intercept term

Compute the slope coefficient and intercept term using the following information:

Cov(S&P 500, ABC) = 0.000336 Mean return, S&P 500 = −2.70%

Var(S&P 500) = 0.000522 Mean return, ABC = −4.05%

Answer:

The slope coefficient is calculated as $\hat{b}_1 = 0.000336 / 0.000522 = 0.64$.

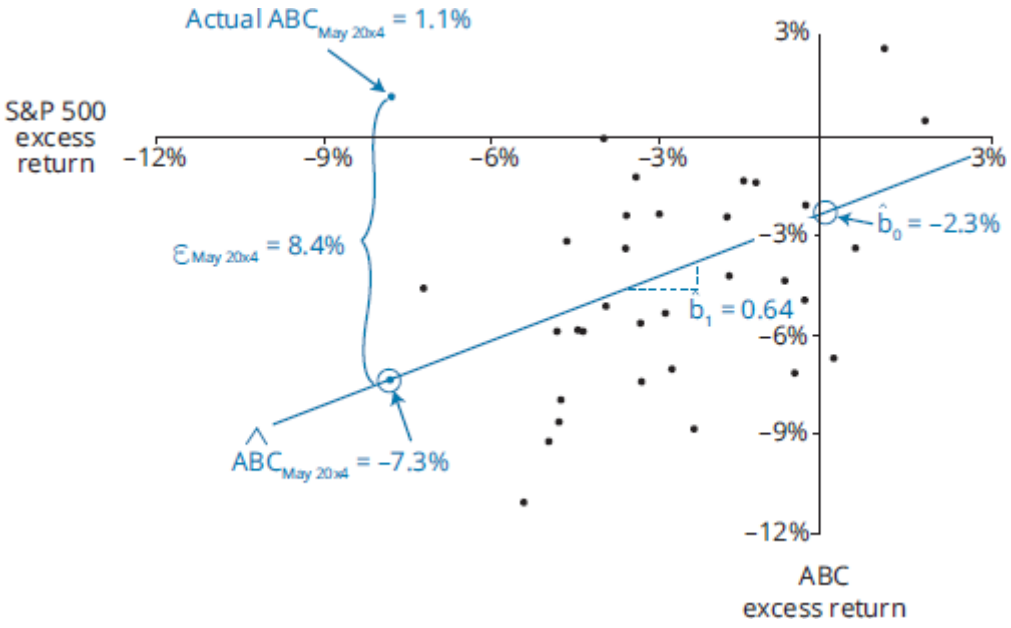
The intercept term is calculated as follows:

$$\hat{b}_0 = \overline{ABC} - \hat{b}_1 \overline{S\&P\ 500} = -4.05\% - 0.64(-2.70\%) = -2.3\%$$

The estimated regression line that minimizes the SSE in our ABC stock return example is shown in Figure 10.2.

This regression line has an intercept of −2.3% and a slope of 0.64. The model predicts that if the S&P 500 excess return is −7.8% (May 20X4 value), then the ABC excess return would be −2.3% + (0.64)(−7.8%) = −7.3%. The residual (error) for the May 20X4 ABC prediction is 8.4%—the difference between the actual ABC excess return of 1.1% and the predicted return of −7.3%.

Figure 10.2: Estimated Regression Equation for ABC vs. S&P 500 Excess Returns



Interpreting a Regression Coefficient

The estimated intercept represents the value of the dependent variable at the point of intersection of the regression line and the axis of the dependent variable (usually, the vertical axis). In other words, the intercept is an estimate of the dependent variable when the independent variable is zero.

We also mentioned earlier that the estimated slope coefficient is interpreted as the expected change in the dependent variable for a one-unit change in the independent variable. For example, an estimated slope coefficient of 2 would indicate that the dependent variable is expected to change by two units for every one-unit change in the independent variable.

EXAMPLE: Interpreting regression coefficients

In the previous example, the estimated slope coefficient was 0.64 and the estimated intercept term was −2.3%. Interpret each coefficient estimate.

Answer:

The slope coefficient of 0.64 can be interpreted to mean that when excess S&P 500 returns increase (decrease) by 1%, ABC excess returns is expected to increase (decrease) by 0.64%.

The intercept term of −2.3% can be interpreted to mean that when the excess return on the S&P 500 is zero, the expected return on ABC stock is −2.3%.

PROFESSOR'S NOTE

The slope coefficient in a regression of the excess returns of an individual security (the *y*-variable) on the return on the market (the *x*-variable) is called the stock's beta, which is an estimate of systematic risk of ABC stock. Notice that ABC is less risky than the average stock, because its returns tend to increase or decrease by less than the overall change in the market returns. A stock with a beta (regression slope coefficient) of 1 has an average level of systematic risk, and a stock with a beta greater than 1 has more-than-average systematic risk. We will apply this concept in the Portfolio Management topic area.

Keep in mind, however, that any conclusions regarding the importance of an independent variable in explaining a dependent variable are based on the statistical significance of the slope coefficient. The magnitude of the slope coefficient tells us nothing about the strength of the linear relationship between the dependent and independent variables. A hypothesis test must be conducted, or a confidence interval must be formed, to assess the explanatory power of the independent variable. Later in this reading we will perform these hypothesis tests.

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

Linear regression is based on numerous assumptions. Most of the major assumptions pertain to the regression model's residual term (ε). Linear regression assumes the following:

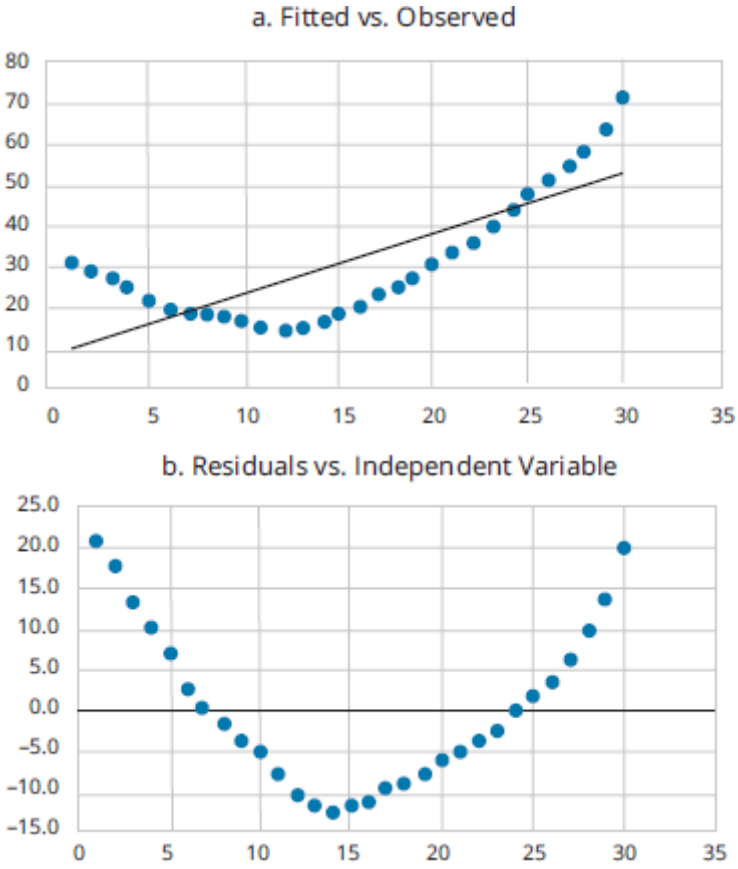
1. A linear relationship exists between the dependent and the independent variables.
2. The variance of the residual term is constant for all observations (homoskedasticity).
3. The residual term is independently distributed; that is, the residual for one observation is not correlated with that of another observation (or, the paired *x* and *y* observations are independent of each other).

4. The residual term is normally distributed.

Linear Relationship

A linear regression model is not appropriate when the underlying relationship between *X* and *Y* is nonlinear. In Panel A of Figure 10.3, we illustrate a regression line fitted to a nonlinear relationship. Note that the prediction errors (vertical distances from the dots to the line) are positive for low values of *X*, then increasingly negative for higher values of *X*, and then turning positive for still-greater values of *X*. One way of checking for linearity is to examine the model residuals (prediction errors) in relation to the independent regression variable. In Panel B, we show the pattern of residuals over the range of the independent variable: positive, negative, then positive.

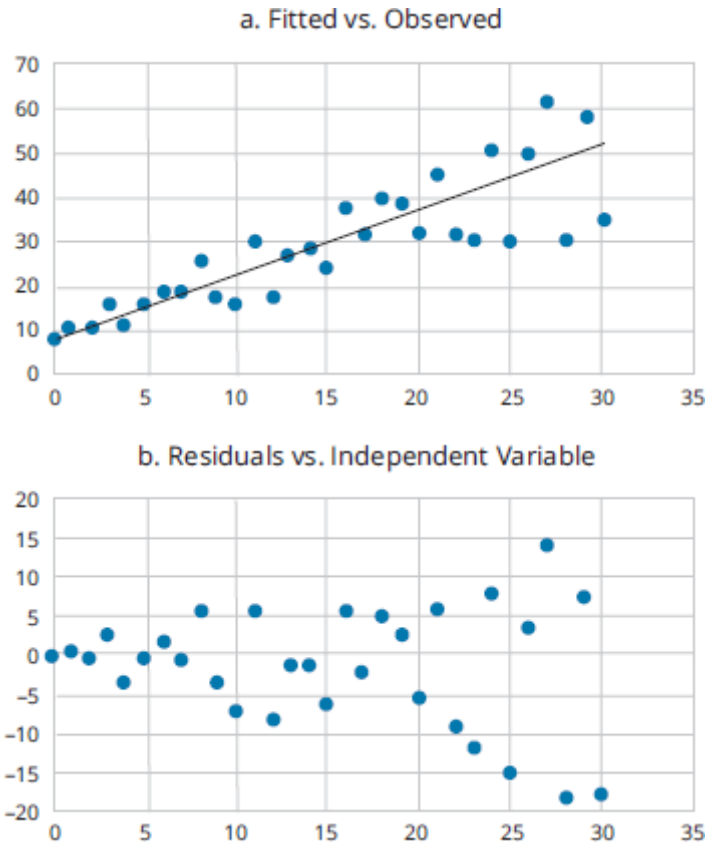
Figure 10.3: Nonlinear Relationship



Homoskedasticity

Homoskedasticity refers to the case where prediction errors all have the same variance. **Heteroskedasticity** refers to the situation when the assumption of homoskedasticity is violated. Figure 10.4, Panel A shows a scatter plot of observations around a fitted regression line where the residuals (prediction errors) increase in magnitude with larger values of the independent variable *X*. Panel B shows the residuals plotted versus the value of the independent variable, and it also illustrates that the variance of the error terms is not likely constant for all observations.

Figure 10.4: Heteroskedasticity

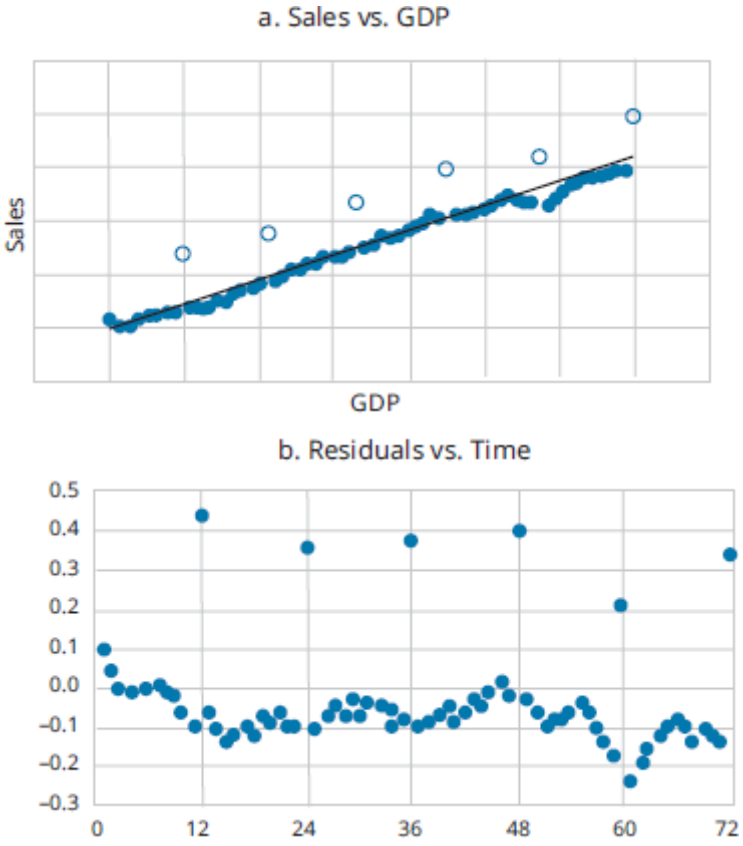


Another type of heteroskedasticity results if the variance of the error term changes over time (rather than with the magnitude of the independent variable). We could observe this by plotting the residuals from a linear regression model versus the dates of each observation and finding that the magnitude of the errors exhibits a pattern of changing over time. To illustrate this, we could plot the residuals versus a time index (as the *x*-variable). Residuals would exhibit a pattern of increasing over time.

Independence

Suppose we collect a company's monthly sales and plot them against monthly GDP as in Figure 10.5, Panel A, and observe that some prediction errors (the unfilled dots) are noticeably larger than others. To investigate this, we plot the residuals versus time, as in Panel B. The residuals plot illustrates that there are large negative prediction errors every 12 months (in December). This suggests that there is seasonality in sales such that December sales (the unfilled dots in Figure 10.5) are noticeably farther from their predicted values than sales for the other months. If the relationship between *X* and *Y* is not independent, the residuals are not independent, and our estimates of the model parameters' variances will not be correct.

Figure 10.5: Independence



Normality

When the residuals (prediction errors) are normally distributed, we can conduct hypothesis testing for evaluating the goodness of fit of the model (discussed later). With a large sample size, based on the central limit theorem, our parameter estimates may be valid, even when the residuals are not normally distributed.

Outliers are observations (one or a few) that are far from our regression line (have large prediction errors or *X* values that are far from the others). Outliers will influence our parameter estimates so that the OLS model will not fit the other observations well.

MODULE QUIZ 10.1

1. What is the *most appropriate* interpretation of a slope coefficient estimate equal to 10.0?
   A. The predicted value of the dependent variable when the independent variable is 0 is 10.0.
   B. For every 1-unit change in the independent variable, the model predicts that the dependent variable will change by 10 units.
   C. For every 1-unit change in the independent variable, the model predicts that the dependent variable will change by 0.1 units.
2. Which of the following is *least likely* a necessary assumption of simple linear regression analysis?
   A. The residuals are normally distributed.
   B. There is a constant variance of the error term.

C. The dependent variable is uncorrelated with the residuals.

MODULE 10.2: ANALYSIS OF VARIANCE (ANOVA) AND GOODNESS OF FIT

Video covering this content is available online.

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

LOS 10.d: Describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the standard error of estimate in a simple linear regression.

Analysis of variance (ANOVA) is a statistical procedure for analyzing the total variability of the dependent variable. Let's define some terms before we move on to ANOVA tables:

- The **total sum of squares (SST)** measures the total variation in the dependent variable. SST is equal to the sum of the squared differences between the actual *Y*-values and the mean of *Y*:

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

- The **sum of squares regression (SSR)** measures the variation in the dependent variable that is explained by the independent variable. SSR is the sum of the squared distances between the predicted *Y*-values and the mean of *Y*:

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

- The **mean square regression (MSR)** is the SSR divided by the number of independent variables. A simple linear regression has only one independent variable, so in this case, MSR = SSR.

PROFESSOR'S NOTE
Multiple regression (i.e., with more than one independent variable) is addressed in the Level II CFA curriculum.

- The **sum of squared errors (SSE)** measures the unexplained variation in the dependent variable. It's also known as the sum of squared residuals or the residual sum of squares. SSE is the sum of the squared vertical distances between the actual *Y*-values and the predicted *Y*-values on the regression line:

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

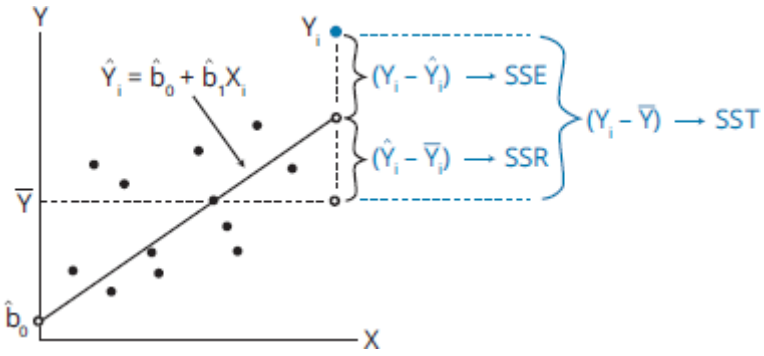
- The **mean squared error (MSE)** is the SSE divided by the degrees of freedom, which is $n - 1$ minus the number of independent variables. A simple linear regression has only one independent variable, so in this case, degrees of freedom are $n - 2$.

You probably will not be surprised to learn the following:

$$\text{total variation} = \text{explained variation} + \text{unexplained variation}$$
or:
$$\text{SST} = \text{SSR} + \text{SSE}$$

Figure 10.6 illustrates how the total variation in the dependent variable (SST) is composed of SSR and SSE.

Figure 10.6: Components of Total Variation



The output of the ANOVA procedure is an ANOVA table, which is a summary of the variation in the dependent variable. ANOVA tables are included in the regression output of many statistical software packages. You can think of the ANOVA table as the source of the data for the computation of many of the regression concepts discussed in this reading. A generic ANOVA table for a simple linear regression (one independent variable) is presented in Figure 10.7.

Figure 10.7: ANOVA Table for a Simple Linear Regression

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	SSR	$\text{MSR} = \frac{\text{SSR}}{k} = \frac{\text{SSR}}{1} = \text{SSR}$
Error (unexplained)	$n - 2$	SSE	$\text{MSE} = \frac{\text{SSE}}{n-2}$
Total	$n - 1$	SST	

Standard Error of Estimate (SEE)

The SEE for a regression is the standard deviation of its residuals. The lower the SEE, the better the model fit:

$$\text{SEE} = \sqrt{\text{MSE}}$$

Coefficient of Determination (R^2)

The **coefficient of determination** (R^2) is defined as the percentage of the total variation in the dependent variable explained by the independent variable. For example, an R^2 of 0.63 indicates that the variation of the independent variable explains 63% of the variation in the dependent variable:

$$R^2 = \text{SSR} / \text{SST}$$

PROFESSOR'S NOTE
For simple linear regression (i.e., with one independent variable), the coefficient of determination, R^2, may be computed by simply squaring the correlation coefficient, *r*. In other words, $R^2 = r^2$ for a regression with one independent variable.

EXAMPLE: Using the ANOVA table

Given the following ANOVA table based on 36 observations, calculate the R^2 and the standard error of estimate (SEE).

Completed ANOVA table for ABC regression

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	0.0076	0.0076
Error (unexplained)	34	0.0406	0.0012
Total	35	0.0482	

Answer:

$$R^2 = \frac{\text{explained variation (SSR)}}{\text{total variation (SST)}} = \frac{0.0076}{0.0482} = 0.158 \text{ or } 15.8\%$$

$$\text{SEE} = \sqrt{\text{MSE}} = \sqrt{0.0012} = 0.035$$

The *F*-Statistic

An *F*-test assesses how well a set of independent variables, as a group, explains the variation in the dependent variable.

The *F*-statistic is calculated as follows:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/k}{\text{SSE}/n - k - 1}$$

where:

MSR = mean regression sum of squares
MSE = mean squared error
Important: This is always a one-tailed test!

For simple linear regression, there is only one independent variable, so the *F*-test is equivalent to a *t*-test of the statistical significance of the slope coefficient:

$$H_0: b_1 = 0 \text{ versus } H_a: b_1 \neq 0$$

To determine whether b_1 is statistically significant using the *F*-test, the calculated *F*-statistic is compared with the critical *F*-value, F_c, at the appropriate level of significance. The degrees of freedom for the numerator and denominator with one independent variable are as follows:

$$df_{numerator} = k = 1$$
$$df_{denominator} = n - k - 1 = n - 2$$

where:

n = number of observations

The decision rule for the *F*-test is to reject H_0 if $F > F_c$.

Rejecting the null hypothesis that the value of the slope coefficient equals zero at a stated level of significance indicates that the independent variable and the dependent variable have a significant linear relationship.

EXAMPLE: Calculating and interpreting the *F*-statistic

Use the ANOVA table from the previous example to calculate and interpret the *F*-statistic. Test the null hypothesis at the 5% significance level that the slope coefficient is equal to 0.

Answer:

$$F = \frac{MSR}{MSE} = \frac{0.0076}{0.0012} = 6.33$$
$$df_{numerator} = k = 1$$
$$df_{denominator} = n - k - 1 = 36 - 1 - 1 = 34$$

The null and alternative hypotheses are h_0: $b_1 = 0$ versus h_a: $b_1 \neq 0$. The critical *F*-value for 1 and 34 degrees of freedom at a 5% significance level is approximately 4.1. (Remember, it's a one-tailed test, so we use the 5% *F*-table.) Therefore, we can reject the null hypothesis and conclude that the slope coefficient is significantly different than zero.

Hypothesis Test of a Regression Coefficient

A *t*-test may also be used to test the hypothesis that the true slope coefficient, b_1, is equal to a hypothesized value. Letting $\hat{b}_1$ be the point estimate for b_1, the appropriate test statistic with n – 2 degrees of freedom is:

$$t_{b_1} = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

The decision rule for tests of significance for regression coefficients is:

$$\text{Reject } H_0 \text{ if } t > +t_{critical} \text{ or } t < -t_{critical}$$

Rejection of the null supports the alternative hypothesis that the slope coefficient is *different* from the hypothesized value of b_1. To test whether an independent variable explains the variation in the dependent variable (i.e., it is statistically significant), the null hypothesis is that the true slope is zero ($b_1 = 0$). The appropriate test structure for the null and alternative hypotheses is:

$$H_0: b_1 = 0 \text{ versus } H_a: b_1 \neq 0$$

EXAMPLE: Hypothesis test for significance of regression coefficients

The estimated slope coefficient from the ABC example is 0.64 with a standard error equal to 0.26. Assuming that the sample has 36 observations, determine if the estimated slope coefficient is significantly different than zero at a 5% level of significance.

Answer:

The calculated test statistic is:

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}} = \frac{0.64 - 0}{0.26} = 2.46$$

The critical two-tailed *t*-values are ±2.03 (from the *t*-table with df = 36 − 2 = 34). Because $t > t_{critical}$ (i.e., 2.46 > 2.03), we reject the null hypothesis and conclude that the slope is different from zero.

Note that the *t*-test for a simple linear regression is equivalent to a *t*-test for the correlation coefficient between *x* and *y*:

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

MODULE QUIZ 10.2

1. Consider the following statement: "In a simple linear regression, the appropriate degrees of freedom for the critical *t*-value used to calculate a confidence interval around both a parameter estimate and a predicted *Y*-value is the same as the number of observations minus two." This statement is:
   A. justified.
   B. not justified, because the appropriate degrees of freedom used to calculate a confidence interval around a parameter estimate is the number of observations.
   C. not justified, because the appropriate degrees of freedom used to calculate a confidence interval around a predicted *Y*-value is the number of observations.

2. What is the appropriate alternative hypothesis to test the statistical significance of the intercept term in the following regression?

   $Y = a_1 + a_2(X) + \varepsilon$

   A. $H_A: a_1 \neq 0$.
   B. $H_A: a_1 > 0$.
   C. $H_A: a_2 \neq 0$.

3. The variation in the dependent variable explained by the independent variable is measured by the:
   A. mean squared error.
   B. sum of squared errors.

C. regression sum of squares.

MODULE 10.3: PREDICTED VALUES AND FUNCTIONAL FORMS OF REGRESSION

Video covering this content is available online.

LOS 10.e: 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.

Predicted values are values of the dependent variable based on the estimated regression coefficients and a prediction about the value of the independent variable. They are the values that are *predicted* by the regression equation, given an estimate of the independent variable.

For a simple regression, this is the predicted (or forecast) value of *Y*:

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

where:

$\hat{Y}$ = predicted value of the dependent variable
X_p = forecasted value of the independent variable

EXAMPLE: Predicting the dependent variable

Given the ABC regression equation as follows:

$$\widehat{ABC} = -2.3\% + (0.64)(\widehat{S\&P\ 500})$$

Calculate the predicted value of ABC excess returns if forecast S&P 500 excess returns are 10%.

Answer:

The predicted value for ABC excess returns is determined as follows:

$$\widehat{ABC} = -2.3\% + (0.64)(10\%) = 4.1\%$$

Confidence Intervals for Predicted Values

This is the equation for the confidence interval for a predicted value of *Y*:

$$\hat{Y} \pm (t_c \times s_f) \Rightarrow [\hat{Y} - (t_c \times s_f) < Y < \hat{Y} + (t_c \times s_f)]$$

where:

t_c = two-tailed critical *t*-value at the desired level of significance with df = n − 2
s_f = standard error of the forecast

The challenge with computing a confidence interval for a predicted value is calculating s_f. On the Level I exam, it's highly unlikely that you will have to calculate the standard

error of the forecast (it will probably be provided if you need to compute a confidence interval for the dependent variable). However, if you do need to calculate s_f, it can be done with the following formula for the variance of the forecast:

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

where:

SEE^2 = variance of the residuals = the square of the standard error of estimate

s_x^2 = variance of the independent variable

X = value of the independent variable for which the forecast was made

EXAMPLE: Confidence interval for a predicted value

Calculate a 95% prediction interval on the predicted value of ABC excess returns from the previous example. Assume the standard error of the forecast is 3.67, and the forecast value of S&P 500 excess returns is 10%.

Answer:

This is the predicted value for ABC excess returns:

$$\widehat{ABC} = -2.3\% + (0.64)(10\%) = 4.1\%$$

The 5% two-tailed critical *t*-value with 34 degrees of freedom is 2.03. This is the prediction interval at the 95% confidence level:

$$\widehat{ABC} \pm (t_c \times s_f) \Rightarrow [4.1\% \pm (2.03 \times 3.67\%)] = 4.1\% \pm 7.5\%$$

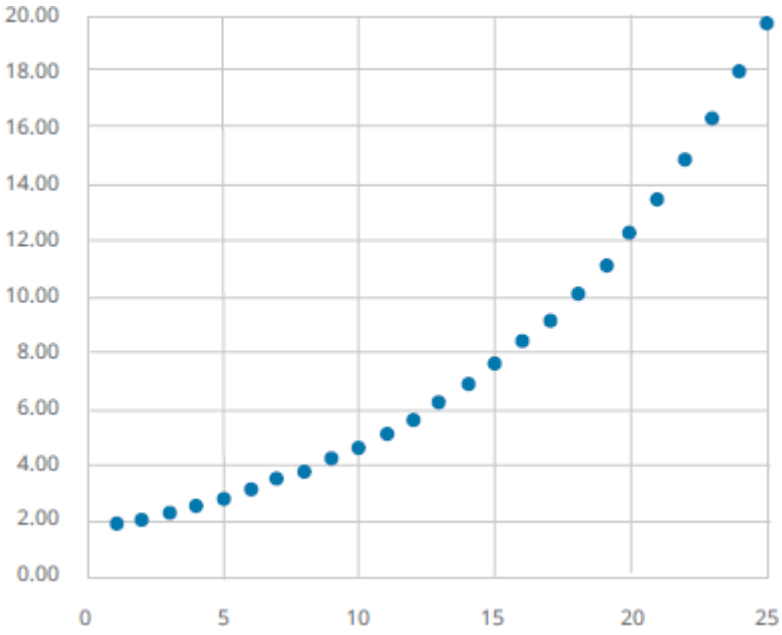
Or, −3.4% to 11.6%.

We can interpret this range to mean that, given a forecast value for S&P 500 excess returns of 10%, we can be 95% confident that the ABC excess returns will be between −3.4% and 11.6%.

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

One of the assumptions of linear regression is that the relationship between *X* and *Y* is linear. What if that assumption is violated? Consider *Y* = EPS for a company and *X* = time index. Suppose that EPS is growing at approximately 10% annually. Figure 10.8 shows the plot of actual EPS versus time.

Figure 10.8: Nonlinear Relationship



In such a situation, transforming one or both of the variables can produce a linear relationship. The appropriate transformation depends on the relationship between the two variables. One often-used transformation is to take the natural log of one or both of the variables. Here are some examples:

- **Log-lin model.** This is if the dependent variable is logarithmic, while the independent variable is linear.
- **Lin-log model.** This is if the dependent variable is linear, while the independent variable is logarithmic.
- **Log-log model.** Both the dependent variable and the independent variable are logarithmic.

Selecting the correct functional form involves determining the nature of the variables and evaluating the goodness-of-fit measures (e.g., R^2, SEE, *F*-stat).

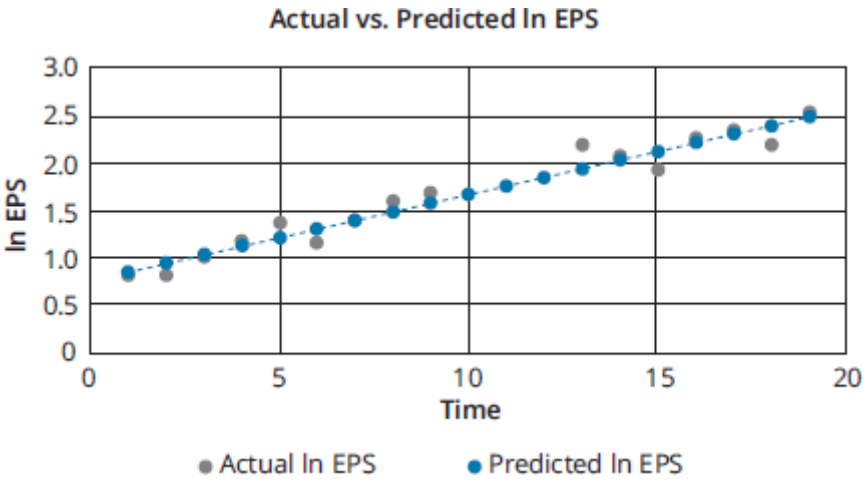
Log-Lin Model

Taking the natural logarithm of the dependent variable, our model now becomes this:

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

In this model, the slope coefficient is interpreted as the *relative* change in dependent variable for an absolute change in the independent variable. Figure 10.9 shows the results after taking the natural log of EPS, and fitting that data using a log-lin model.

Figure 10.9: Log-Lin Model, EPS Data



Lin-Log Model

Taking the natural logarithm of the independent variable, our model now becomes this:

$$Y_i = b_0 + b_1 \ln(X)_i + \varepsilon_i$$

In this model, the slope coefficient is interpreted as the *absolute* change in dependent variable for a *relative* change in the independent variable.

Log-Log Model

Taking the natural logarithm of both variables, our model now becomes this:

$$\ln Y_i = b_0 + b_1 \ln(X)_i + \varepsilon_i$$

In this model, the slope coefficient is interpreted as the relative change in dependent variable for a relative change in the independent variable.

MODULE QUIZ 10.3

1. For a regression model of Y = 5 + 3.5X, the analysis (based on a large data sample) provides the standard error of the forecast as 2.5 and the standard error of the slope coefficient as 0.8. A 90% confidence interval for the estimate of *Y* when the value of the independent variable is 10 is *closest* to:

   A. 35.1 to 44.9.
   B. 35.6 to 44.4.
   C. 35.9 to 44.1.

2. The appropriate regression model for a linear relationship between the relative change in an independent variable and the absolute change in the dependent variable is a:

   A. log-lin model.
   B. lin-log model.
   C. lin-lin model.

KEY CONCEPTS

LOS 10.a

Linear regression provides an estimate of the linear relationship between an independent variable (the explanatory variable) and a dependent variable (the predicted variable).

The general form of a simple linear regression model is as follows:

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

The least squares model minimizes the sum of squared errors:

- $\hat{b}_0$ = fitted intercept = $\overline{Y} - \hat{b}_1 \overline{X}$
- $\hat{b}_1$ = fitted slope coefficient = Cov (X,Y) / variance of X

The estimated intercept, $\hat{b}_0$, represents the value of the dependent variable at the point of intersection of the regression line and the axis of the dependent variable (usually, the vertical axis). The estimated slope coefficient, $\hat{b}_1$, is interpreted as the change in the dependent variable for a one-unit change in the independent variable.

LOS 10.b

Assumptions made regarding simple linear regression include the following:

1. A linear relationship exists between the dependent and the independent variable.
2. The variance of the residual term is constant (homoskedasticity).
3. The residual term is independently distributed (residuals are uncorrelated).
4. The residual term is normally distributed.

LOS 10.c

The total sum of squares (SST) measures the total variation in the dependent variable and equals the sum of the squared differences between its actual values and its mean.

The sum of squares regression (SSR) measures the variation in the dependent variable that is explained by the independent variable, and equals the sum of the squared distances between the predicted values and the mean of the dependent variable.

The mean square regression (MSR) is the SSR divided by the number of independent variables. For a simple linear regression, MSR = SSR.

The sum of squared errors (SSE) measures the unexplained variation in the dependent variable and is the sum of the squared vertical distances between the actual and the predicted values of the dependent variable.

The mean squared error (MSE) is the SSE divided by the degrees of freedom (n − 2 for a simple linear regression).

The coefficient of determination, R^2, is the proportion of the total variation of the dependent variable explained by the regression:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

In simple linear regression, because there is only one independent variable (k = 1), the *F*-test tests the same null hypothesis as testing the statistical significance of b_1 using

the *t*-test: H_0: $b_1 = 0$ versus H_a: $b_1 \neq 0$. With only one independent variable, *F* is calculated as follows:

$$\text{F-stat} = \frac{\text{MSR}}{\text{MSE}} \text{ with 1 and n} - 2 \text{ degrees of freedom}$$

LOS 10.d

ANOVA Table for Simple Linear Regression (k = 1)

Source of Variation	Degrees of Freedom (df)	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	SSR	$\text{MSR} = \frac{\text{SSR}}{\text{k}} = \frac{\text{SSR}}{1} = \text{SSR}$
Error (unexplained)	n − 2	SSE	$\text{MSE} = \frac{\text{SSE}}{\text{n} - 2}$
Total	n − 1	SST	

The standard error of the estimate in a simple linear regression is calculated as follows:

$$\text{SEE} = \sqrt{\frac{\text{SSE}}{\text{n} - 2}}$$

LOS 10.e

A predicted value of the dependent variable, $\hat{Y}$, is determined by inserting the predicted value of the independent variable, X_p, in the regression equation and calculating $\hat{Y}_p = \hat{b}_0 + \hat{b}_1 \hat{X}_p$.

The confidence interval for a predicted *Y*-value is $\left[\hat{Y} - (t_c \times s_f) < Y < \hat{Y} + (t_c \times s_f)\right]$ where s_f is the standard error of the forecast.

LOS 10.f

Dependent Variable	Independent Variable	Model	Slope Interpretation
Logarithmic	Linear	Log-lin	*Relative* change in dependent variable for an *absolute* change in the independent variable
Linear	Logarithmic	Lin-log	*Absolute* change in dependent variable for a *relative* change in the independent variable
Logarithmic	Logarithmic	Log-log	*Relative* change in dependent variable for a *relative* change in the independent variable

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 10.1

1. **B** The slope coefficient is best interpreted as the predicted change in the dependent variable for a 1-unit change in the independent variable; if the slope coefficient estimate is 10.0 and the independent variable changes by 1 unit, the dependent variable is expected to change by 10 units. The intercept term is best interpreted

as the value of the dependent variable when the independent variable is equal to zero. (LOS 10.a)

2. **C** The model does not assume that the dependent variable is uncorrelated with the residuals. It does assume that the independent variable is uncorrelated with the residuals. (LOS 10.b)

Module Quiz 10.2

1. **A** In simple linear regression, the appropriate degrees of freedom for both confidence intervals is the number of observations in the sample (n) minus two. (LOS 10.c)

2. **A** In this regression, a_1 is the intercept term. To test the statistical significance means to test the null hypothesis that a_1 is equal to zero, versus the alternative that a_1 is not equal to zero. (LOS 10.c)

3. **C** The regression sum of squares measures the amount of variation in the dependent variable explained by the independent variable (i.e., the explained variation). The sum of squared errors measures the variation in the dependent variable not explained by the independent variable. The mean squared error is equal to the sum of squared errors divided by its degrees of freedom. (LOS 10.d)

Module Quiz 10.3

1. **C** The estimate of Y, given $X = 10$, is $Y = 5 + 3.5(10) = 40$. The critical value for a 90% confidence interval with a large sample size (z-statistic) is approximately 1.65. Given the standard error of the forecast of 2.5, the confidence interval for the estimated value of Y is 40 +/− 1.65(2.5) = 35.875 to 44.125. (LOS 10.e)

2. **B** The appropriate model would be a lin-log model, in which the values of the dependent variable (Y) are regressed on the natural logarithms of the independent variable (X): $Y = b_0 + b_1 \ln(X)$. (LOS 10.f)

READING 11

INTRODUCTION TO BIG DATA TECHNIQUES

MODULE 11.1: INTRODUCTION TO FINTECH

Video covering this content is available online.

LOS 11.a: Describe aspects of "fintech" that are directly relevant for the gathering and analyzing of financial data.

The term **fintech** refers to developments in technology that can be applied to the financial services industry. Companies that are in the business of developing technologies for the finance industry are often referred to as fintech companies.

Some of the primary areas where fintech is developing include the following:

- Increasing functionality to handle large sets of data that may come from many sources and exist in various forms
- Tools and techniques such as artificial intelligence for analyzing very large datasets

LOS 11.b: Describe Big Data, artificial intelligence, and machine learning.

Big Data is a widely used expression that refers to all the potentially useful information that is generated in the economy. This includes not only data from traditional sources, such as financial markets, company financial reports, and government economic statistics, but also **alternative data** from nontraditional sources. Some of these nontraditional sources are as follows:

- Individuals generate usable data such as social media posts, online reviews, email, and website visits.
- Businesses generate potentially useful information such as bank records and retail scanner data. These kinds of data are referred to as **corporate exhaust**.
- Sensors, such as radio frequency identification chips, are embedded in numerous devices such as smartphones and smart buildings. The broad network of such devices is referred to as the **Internet of Things**.

Characteristics of Big Data include its volume, velocity, and variety.

The *volume* of data continues to grow by orders of magnitude. The units in which data can be measured have increased from megabytes and gigabytes to terabytes (1,000 gigabytes) and even petabytes (1,000 terabytes).

Velocity refers to how quickly data are communicated. Real-time data such as stock market price feeds are said to have low **latency**. Data that are only communicated periodically or with a lag are said to have high latency.

The *variety* of data refers to the varying degrees of structure in which data may exist. These range from structured forms (e.g., spreadsheets and databases), to semistructured forms (e.g., photos and web page code), to unstructured forms (e.g., video).

The field of **data science** concerns how we extract information from Big Data. Data science describes methods for processing and visualizing data. Processing methods include the following:

- *Capture*. This is collecting data and transforming it into usable forms.
- *Curation*. This is assuring data quality by adjusting for bad or missing data.
- *Storage*. This is archiving and accessing data.
- *Search*. This is examining stored data to find needed information.
- *Transfer*. This is moving data from their source or a storage medium to where they are needed.

Visualization techniques include the familiar charts and graphs that display structured data. To visualize less structured data requires other methods. Some examples of these are word clouds that illustrate the frequency with which words appear in a sample of text, or mind maps that display logical relations among concepts.

Taking advantage of Big Data presents numerous challenges. Analysts must ensure that the data they use are of high quality, accounting for the possibilities of outliers, bad or missing data, or sampling biases. The volume of data collected must be sufficient and appropriate for its intended use.

The need to process and organize data before using it can be especially problematic with qualitative and unstructured data. This is a process to which **artificial intelligence**, or computer systems that can be programmed to simulate human cognition, may be applied usefully. **Neural networks** are an example of artificial intelligence in that they are programmed to process information in a way similar to the human brain.

An important development in the field of artificial intelligence is **machine learning**. In machine learning, a computer algorithm is given inputs of source data, with no assumptions about their probability distributions, and may be given outputs of target data. The algorithm is designed to learn, without human assistance, how to model the output data based on the input data, or to learn how to detect and recognize patterns in the input data.

Machine learning typically requires vast amounts of data. A typical process begins with a *training* dataset in which the algorithm looks for relationships. A *validation* dataset is

then used to refine these relationship models, which can then be applied to a *test* dataset to analyze their predictive ability.

In **supervised learning**, the input and output data are labeled, the machine learns to model the outputs from the inputs, and then the machine is given new data on which to use the model. In **unsupervised learning**, the input data are not labeled, and the machine learns to describe the structure of the data. **Deep learning** is a technique that uses layers of neural networks to identify patterns, beginning with simple patterns and advancing to more complex ones. Deep learning may employ supervised or unsupervised learning. Some of the applications of deep learning include image and speech recognition.

Machine learning can produce models that overfit or underfit the data. **Overfitting** occurs when the machine learns the input and output data too exactly, treats noise as true parameters, and identifies spurious patterns and relationships. In effect, the machine creates a model that is too complex. **Underfitting** occurs when the machine fails to identify actual patterns and relationships, treating true parameters as noise. This means that the model is not complex enough to describe the data. A further challenge with machine learning is that its results can be a "black box," producing outcomes based on relationships that are not readily explainable.

LOS 11.c: Describe applications of Big Data and Data Science to investment management.

Applications of fintech that are relevant to investment management include text analytics, natural language processing, risk governance, and algorithmic trading.

Text analytics refers to the analysis of unstructured data in text or voice forms. An example of text analytics is analyzing the frequency of words and phrases. In the finance industry, text analytics have the potential to partially automate specific tasks such as evaluating company regulatory filings.

Natural language processing refers to the use of computers and artificial intelligence to interpret human language. Speech recognition and language translation are among the uses of natural language processing. Possible applications in finance could be to check for regulatory compliance in an examination of employee communications, or to evaluate large volumes of research reports to detect more subtle changes in sentiment than can be discerned from analysts' recommendations alone.

Risk governance requires an understanding of a firm's exposure to a wide variety of risks. Financial regulators require firms to perform risk assessments and stress testing. The simulations, scenario analysis, and other techniques used for risk analysis require large amounts of quantitative data along with a great deal of qualitative information. Machine learning and other techniques related to Big Data can be useful in modeling and testing risk, particularly if firms use real-time data to monitor risk exposures.

Algorithmic trading refers to computerized securities trading based on a predetermined set of rules. For example, algorithms may be designed to enter the optimal execution instructions for any given trade based on real-time price and volume

data. Algorithmic trading can also be useful for executing large orders by determining the best way to divide the orders across exchanges. Another application of algorithmic trading is **high-frequency trading** that identifies and takes advantage of intraday securities mispricings.

MODULE QUIZ 11.1

1. Fintech is *most accurately* described as the:
   A. application of technology to the financial services industry.
   B. replacement of government-issued money with electronic currencies.
   C. clearing and settling of securities trades through distributed ledger technology.
2. Which of the following technological developments is *most likely* to be useful for analyzing Big Data?
   A. Machine learning.
   B. High-latency capture.
   C. The Internet of Things.

KEY CONCEPTS

LOS 11.a

Fintech refers to developments in technology that can be applied to the financial services industry. Companies that develop technologies for the finance industry are referred to as fintech companies.

LOS 11.b

Big Data refers to the potentially useful information that is generated in the economy, including data from traditional and nontraditional sources. Characteristics of Big Data include its volume, velocity, and variety.

Artificial intelligence refers to computer systems that can be programmed to simulate human cognition. Neural networks are an example of artificial intelligence.

Machine learning is programming that gives a computer system the ability to improve its performance of a task over time and is often used to detect patterns in large sets of data.

LOS 11.c

Applications of fintech to investment management include text analytics, natural language processing, risk governance, and algorithmic trading.

Text analytics refers to analyzing unstructured data in text or voice forms. Natural language processing is the use of computers and artificial intelligence to interpret human language. Algorithmic trading refers to computerized securities trading based on predetermined rules.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 11.1

1. **A** Fintech is the application of technology to the financial services industry and to companies that are involved in developing and applying technology for financial services. Cryptocurrencies and distributed ledger technology are examples of fintech-related developments. (LOS 11.a)

2. **A** Machine learning is a computer programming technique useful for identifying and modeling patterns in large volumes of data. The Internet of Things is the network of devices that is one of the sources of Big Data. Capture is one aspect of processing data. Latency is the lag between when data is generated and when it is needed. (LOS 11.b)

TOPIC QUIZ: QUANTITATIVE METHODS

You have now finished the Quantitative Methods topic section. Please log into your Schweser online dashboard and take the Topic Quiz on this section. The Topic Quiz provides immediate feedback on how effective your study has been for this material. Questions are more exam-like than typical Module Quiz or QBank questions; a score of less than 70% indicates that your study likely needs improvement. These tests are best taken timed; allow 1.5 minutes per question.

READING 12

FIRMS AND MARKET STRUCTURES

MODULE 12.1: BREAKEVEN, SHUTDOWN, AND SCALE

Video covering this content is available online.

LOS 12.a: Determine and interpret breakeven and shutdown points of production, as well as how economies and diseconomies of scale affect costs under perfect and imperfect competition.

In economics, we define the **short run** for a firm as the time period over which some factors of production are fixed. Typically, we assume that capital is fixed in the short run so that a firm cannot change its scale of operations (plant and equipment) over the short run. All factors of production (costs) are variable in the **long run**. The firm can let its leases expire and sell its equipment, thereby avoiding costs that are fixed in the short run.

Shutdown and Breakeven Under Perfect Competition

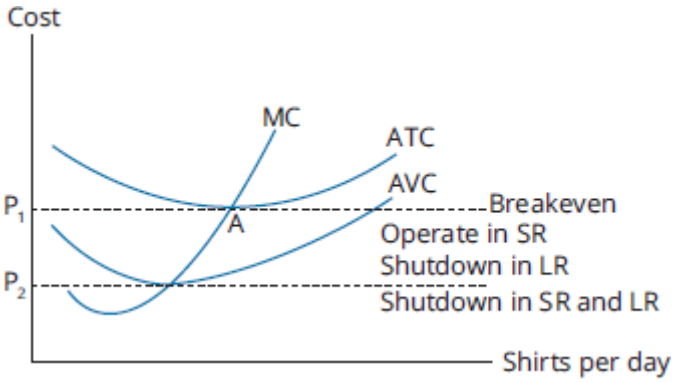
As a simple example of shutdown and breakeven analysis, consider a retail store with a one-year lease (fixed cost) and one employee (quasi-fixed cost), so that variable costs are simply the store's cost of merchandise. If the total sales (total revenue) just cover both fixed and variable costs, price equals both average revenue and average total cost —so we are at the breakeven output quantity, and economic profit equals zero.

During the period of the lease (the short run), as long as items are being sold for more than their variable cost, the store should continue to operate to minimize losses. If items are being sold for less than their average variable cost, losses would be reduced by shutting down the business in the short run.

In the long run, a firm should shut down if the price is less than average total cost, regardless of the relation between price and average variable cost.

For a firm under perfect competition (a price-taker), we can use a graph of cost functions to examine the profitability of the firm at different output prices. In Figure 12.1, at price P_1, price and average revenue equal average total cost. At the output level of Point A, the firm is making an economic profit of zero. At a price above P_1, economic profit is positive, and at prices less than P_1, economic profit is negative (the firm has economic losses).

Figure 12.1: Shutdown and Breakeven



Because some costs are fixed in the short run, it will be better for the firm to continue production in the short run as long as average revenue is greater than average variable costs. At prices between P_1 and P_2 in Figure 12.1, the firm has losses, but the losses are smaller than would occur if all production were stopped. As long as total revenue is greater than total variable cost, at least some of the firm's fixed costs are covered by continuing to produce and sell its product. If the firm were to shut down, losses would be equal to the fixed costs that still must be paid. As long as price is greater than average variable costs, the firm will minimize its losses in the short run by continuing in business.

If average revenue is less than average variable cost, the firm's losses are greater than its fixed costs, and it will minimize its losses by shutting down production in the short run. In this case (a price less than P_2 in Figure 12.1), the loss from continuing to operate is greater than the loss (total fixed costs) if the firm is shut down.

In the long run, all costs are variable, so a firm can avoid its (short-run) fixed costs by shutting down. For this reason, if price is expected to remain below minimum average total cost (Point A in Figure 12.1) in the long run, the firm will shut down rather than continue to generate losses.

To sum up, if average revenue is less than average variable cost in the short run, the firm should shut down. This is its **short-run shutdown point**. If average revenue is greater than average variable cost in the short run, the firm should continue to operate, even if it has losses. In the long run, the firm should shut down if average revenue is less than average total cost. This is the **long-run shutdown point**. If average revenue is just equal to average total cost, total revenue is just equal to total (economic) cost, and this is the firm's **breakeven point**.

- If AR ≥ ATC, the firm should stay in the market in both the short and long run.
- If AR ≥ AVC, but AR < ATC, the firm should stay in the market in the short run but will exit the market in the long run.
- If AR < AVC, the firm should shut down in the short run and exit the market in the long run.

Shutdown and Breakeven Under Imperfect Competition

For price-searcher firms (those that face downward-sloping demand curves), we could compare average revenue to ATC and AVC, just as we did for price-taker firms, to identify shutdown and breakeven points. However, marginal revenue is no longer equal to price.

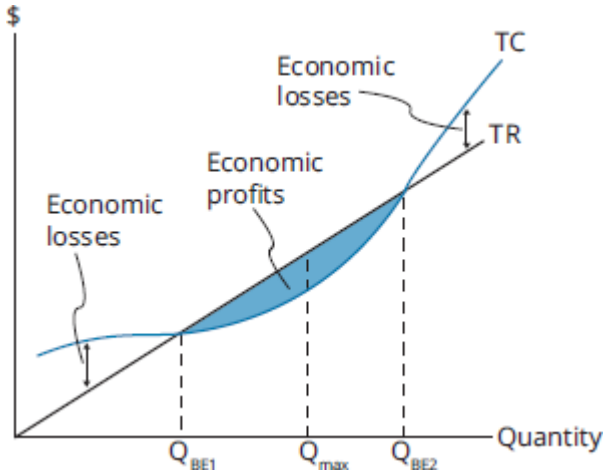
We can, however, still identify the conditions under which a firm is breaking even, should shut down in the short run, and should shut down in the long run in terms of total costs and total revenue. These conditions are as follows:

- TR = TC: break even
- TC > TR > TVC: firm should continue to operate in the short run but shut down in the long run
- TR < TVC: firm should shut down in the short run and the long run

Because price does not equal marginal revenue for a firm in imperfect competition, analysis based on total costs and revenues is better suited for examining breakeven and shutdown points.

The previously described relations hold for both price-taker and price-searcher firms. We illustrate these relations in Figure 12.2 for a price-taker firm (TR increases at a constant rate with quantity). Total cost equals total revenue at the breakeven quantities Q_{BE1} and Q_{BE2}. The quantity for which economic profit is maximized is shown as Q_{max}.

Figure 12.2: Breakeven Point Using the Total Revenue/Total Cost Approach



If the entire TC curve exceeds TR (i.e., no breakeven point), the firm will want to minimize the economic loss in the short run by operating at the quantity corresponding to the smallest (negative) value of TR − TC.

> **EXAMPLE: Short-run shutdown decision**
>
> For the last fiscal year, Legion Gaming reported total revenue of $700,000, total variable costs of $800,000, and total fixed costs of $400,000. Should the firm

continue to operate in the short run?

Answer:

The firm should shut down. Total revenue of $700,000 is less than total costs of $1,200,000, and it is also less than total variable costs of $800,000. By shutting down, the firm will lose an amount equal to fixed costs of $400,000. This is less than the loss of operating, which is TR − TC = $500,000.

EXAMPLE: Long-run shutdown decision

Suppose, instead, that Legion Gaming reported total revenue of $850,000. Should the firm continue to operate in the short run? Should it continue to operate in the long run?

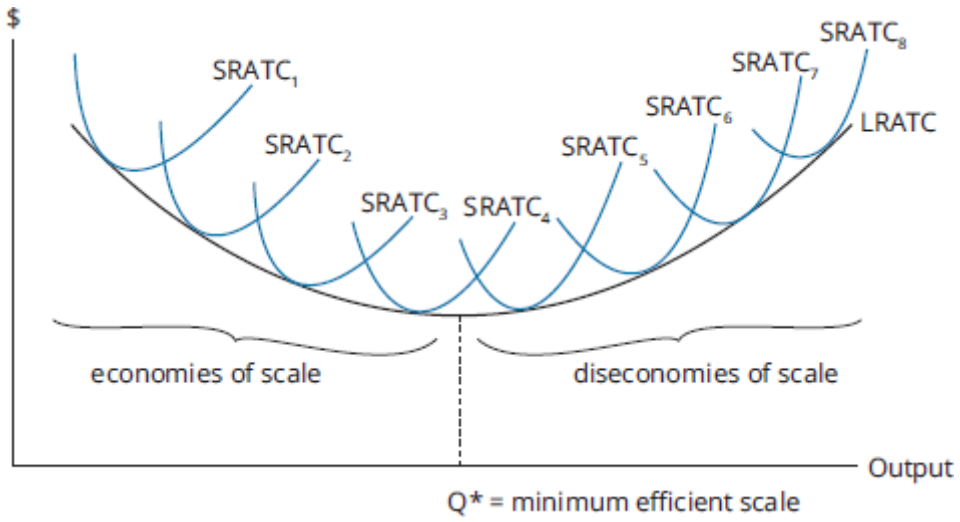
Answer:

In the short run, TR > TVC, and the firm should continue operating. The firm should consider exiting the market in the long run, as TR is not sufficient to cover all of the fixed costs and variable costs.

Economies and Diseconomies of Scale

While plant size is fixed in the short run, in the long run, firms can choose their most profitable scale of operations. Because the long-run average total cost (LRATC) curve is drawn for many different plant sizes or scales of operation, each point along the curve represents the minimum ATC for a given plant size or scale of operations. In Figure 12.3, we show a firm's LRATC curve along with short-run average total cost (SRATC) curves for many different plant sizes, with $SRATC_{n+1}$ representing a larger scale of operations than $SRATC_n$.

Figure 12.3: Economies and Diseconomies of Scale



We draw the LRATC curve as U-shaped. Average total costs first decrease with larger scale, but eventually begin to increase with larger scale. The lowest point on the LRATC corresponds to the scale or plant size at which the average total cost of production is at a minimum. This scale is sometimes called the **minimum efficient scale**. Under perfect competition, firms must operate at minimum efficient scale in long-run equilibrium, and LRATC will equal the market price. Recall that under perfect competition, firms earn zero economic profit in long-run equilibrium. Firms that have chosen a different scale of operations with higher average total costs will have economic losses and must either leave the industry or change to the minimum efficient scale.

The downward-sloping segment of the LRATC curve presented in Figure 12.3 indicates that **economies of scale** (or *increasing returns to scale*) are present. Economies of scale result from factors such as labor specialization, mass production, and investment in more efficient equipment and technology. In addition, the firm may be able to negotiate lower input prices with suppliers as it increases in size and purchases more resources. A firm operating with economies of scale can increase its competitiveness by expanding production and reducing costs.

The upward-sloping segment of the LRATC curve indicates that **diseconomies of scale** are present. Diseconomies of scale may result as the increasing bureaucracy of larger firms leads to inefficiency, problems with motivating a larger workforce, and greater barriers to innovation and entrepreneurial activity. A firm operating under diseconomies of scale will want to decrease output and move back toward the minimum efficient scale. The U.S. auto industry is an example of an industry that has exhibited diseconomies of scale.

There may be a relatively flat portion at the bottom of the LRATC curve that exhibits *constant returns to scale*, or relatively constant costs across a range of plant sizes.

MODULE QUIZ 12.1

1. In a purely competitive market, economic losses indicate that:
   A. price is below average total costs.
   B. collusion is occurring in the marketplace.
   C. firms need to expand output to reduce costs.

2. A firm is likely to operate in the short run as long as price is at least as great as:
   A. marginal cost.
   B. average total cost.
   C. average variable cost.

3. A firm's average revenue is greater than its average variable cost and less than its average total cost. If this situation is expected to persist, the firm should:
   A. shut down in the short run and in the long run.
   B. shut down in the short run, but operate in the long run.
   C. operate in the short run, but shut down in the long run.

4. If a firm increases its plant size by 10% and its minimum average total cost increases by 10%, the firm is experiencing:
   A. constant returns to scale.
   B. diseconomies of scale.

C. economies of scale.

MODULE 12.2: CHARACTERISTICS OF MARKET STRUCTURES

Video covering this content is available online.

LOS 12.b: Describe characteristics of perfect competition, monopolistic competition, oligopoly, and pure monopoly.

Recall from the prerequisite readings that perfect competition results in firm demand that is horizontal (perfectly elastic) at the market price. The firm demand curves for the three other market structures we discuss are all downward sloping. When a firm's demand curve slopes downward, marginal revenue (MR) is less than price. For both horizontal and downward-sloping demand curves, a firm will maximize profits by producing the quantity for which MR is just equal to marginal cost.

While it may not be true in every case, it may be useful to think of firms under pure monopoly as having the steepest demand curves. Firms under monopolistic competition typically have relatively elastic downward-sloping demand curves, while firms in an oligopoly market will face downward-sloping demand curves somewhere between these two extremes.

We can analyze where a market falls along the spectrum from perfect competition to pure monopoly by examining five factors:

1. Number of firms and their relative sizes
2. Degree to which firms differentiate their products
3. Bargaining power of firms with respect to pricing
4. Barriers to entry into or exit from the industry
5. Degree to which firms compete on factors other than price

Perfect competition refers to a market in which many firms produce identical products, barriers to entry into the market are very low, and firms compete for sales only on the basis of price. Firms face perfectly elastic (horizontal) demand curves at the price determined in the market because no firm has a large enough portion of the overall market to affect the market price of the good. The market for wheat in a region is a good approximation of such a market. Overall market supply and demand determine the price of wheat, and each producer can sell all that they choose to at that price.

Monopolistic competition differs from perfect competition in that products are not identical. Each firm differentiates its product(s) from those of other firms through some combination of differences in product quality, product features, and marketing. The demand curve faced by each firm is downward sloping (i.e., neither perfectly elastic nor perfectly inelastic). Prices that producers charge are not identical because of perceived differences among their products, and typically, barriers to entry are low. The market for toothpaste is a good example of monopolistic competition. Firms differentiate their products through features and marketing with claims of more attractiveness, whiter teeth, fresher breath, and even actually cleaning your teeth and preventing decay. If the

price of your personal favorite increases, you are not likely to immediately switch to another brand as we assume under perfect competition. Some customers may switch brands in response to a 10% increase in price, and some may not. This is why firm demand is downward sloping rather than perfectly elastic.

The most important characteristic of an **oligopoly** market is that only a few firms are in the industry. In such a market, each firm must consider the actions and responses of other firms in setting price and business strategy. We say that such firms are *interdependent*. While the firms' products are typically good substitutes for each other, they may be either quite similar or differentiated through features, branding, marketing, and quality. Barriers to entry are typically high, often because of economies of scale in production or marketing, which accounts for the existence of a small number of firms with relatively large market shares. Demand can be more or less elastic than for firms in monopolistic competition. The automobile market is dominated by a small number of large firms and can be characterized as an oligopoly. The product and pricing decisions of Toyota certainly affect those of Ford, and vice versa. Automobile makers compete based on price, but they also compete through marketing, product features, and quality, which are often signaled strongly through the brand name. The oil industry also has a few firms with large market shares, but their products are, in most cases, good substitutes for each other.

A **monopoly** market is characterized by a single seller of a product with no good substitutes. This fact alone means that the firm faces a downward-sloping demand curve (the market demand curve) and has the power to choose the price at which it sells its product. High barriers to entry protect a monopoly producer from competition. One source of monopoly power is the protection offered by copyrights and patents. Another possible source of monopoly power is control over a resource specifically needed to produce the product. Most frequently, monopoly power is supported by specific laws or government regulation (e.g., a local electric utility).

Figure 12.4 shows the key features of each market structure.

Figure 12.4: Characteristics of Market Structures

	Perfect Competition	Monopolistic Competition	Oligopoly	Monopoly
Number of sellers	Many firms	Many firms	Few firms	Single firm
Barriers to entry	Very low	Low	High	Very high
Nature of substitute products	Very good substitutes	Good substitutes, but differentiated	Good substitutes or differentiated	No good substitutes
Nature of competition	Price only	Price, marketing, features	Price, marketing, features	Advertising
Pricing power	None	Some	Some to significant	Significant

LOS 12.c: Explain supply and demand relationships under monopolistic competition, including the optimal price and output for firms as well as pricing strategy.

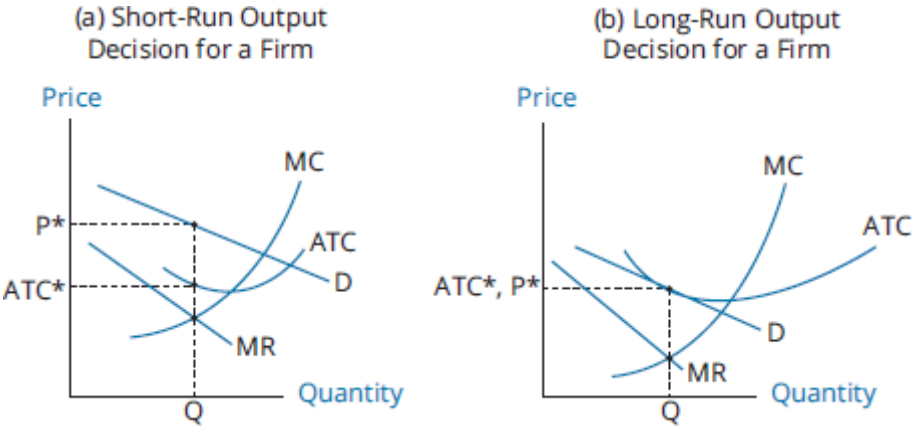
Monopolistic competition has the following market characteristics:

- *A large number of independent sellers*. (1) Each firm has a relatively small market share, so no individual firm has any significant power over price. (2) Firms only need to pay attention to average market price, not the prices of individual competitors. (3) There are too many firms in the industry for collusion (price-fixing) to be possible.
- *Differentiated products*. Each producer has a product that is, in some way, different from those of its competitors (in the minds of consumers). The competing products are considered close substitutes for one another.
- *Firms compete less on price and more on marketing, perceived quality, and differences in features*. Firms must make price and output decisions because they face downward-sloping demand curves.
- *Low barriers to entry*. The cost of entering the market and exiting the market are relatively low.

Think about the market for toothpaste. Brands of toothpaste are quite similar, and it is reasonable to assume that toothpaste is not too difficult or costly to produce. But brands are differentiated based on specific features, on influential advertising and marketing, and on the reputations of the producers.

The price/output decision for monopolistic competition is illustrated in Figure 12.5. Panel A of Figure 12.5 illustrates the short-run price/output characteristics of monopolistic competition for a single firm. As indicated, firms in monopolistic competition maximize economic profits by producing where marginal revenue (MR) equals marginal cost (MC), and by charging the price for that quantity from the demand curve, *D*. Here, the firm earns positive economic profits because price, *P**, exceeds average total cost, *ATC**. Due to low barriers to entry, competitors can enter the market in pursuit of these economic profits.

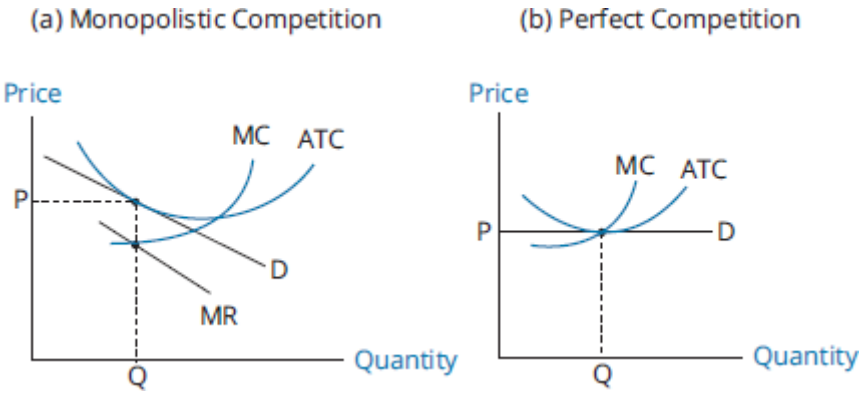
Figure 12.5: Short-Run and Long-Run Output Under Monopolistic Competition



Panel B of Figure 12.5 illustrates long-run equilibrium for a *representative* firm after new firms have entered the market. As indicated, the entry of new firms shifts the demand curve faced by each individual firm down to the point where price equals average total cost (P* = ATC*), such that economic profit is zero. At this point, there is no longer an incentive for new firms to enter the market, and long-run market equilibrium is established. A firm in monopolistic competition continues to produce at the quantity where MR = MC, but it no longer earns positive economic profits. We can get to a similar long-run equilibrium even without entry of new firms if each firm increases its marketing spending (a component of ATC) to either increase or defend its market share until ATC has increased to that shown in Panel B. If all firms compete in this way, each firm will produce Q* and sell at P* but earn no economic profit because marketing costs have increased ATC to ATC*, which is equal to price. Advertising expenses are often relatively high for firms in monopolistic competition.

Figure 12.6 illustrates the differences between long-run equilibrium in markets with monopolistic competition and markets with perfect competition. Note that with monopolistic competition, price is greater than MC (i.e., producers can realize an economic profit), average total cost is not at a minimum for the quantity produced (suggesting excess capacity, or an inefficient scale of production), and the price is slightly higher than under perfect competition. The point to consider here, however, is that perfect competition is characterized by no product differentiation. The question of the efficiency of monopolistic competition becomes, "Is there an economically efficient amount of product differentiation?"

Figure 12.6: Firm Output Under Monopolistic and Perfect Competition



In a world with only one brand of toothpaste, clearly, average production costs would be lower. That fact alone probably does not mean that a world with only one brand or type of toothpaste would be a better world. While product differentiation has costs, it also has benefits to consumers. Consider the market for a pharmaceutical that reduces blood pressure to prolong life. There may be several competing drugs that are more or less effective for, or well or poorly tolerated by, different groups of patients. In this case, we may find that firm demand curves are relatively steep compared to those for brands of toothpaste, indicating that the competing drugs are not considered good substitutes for many patients.

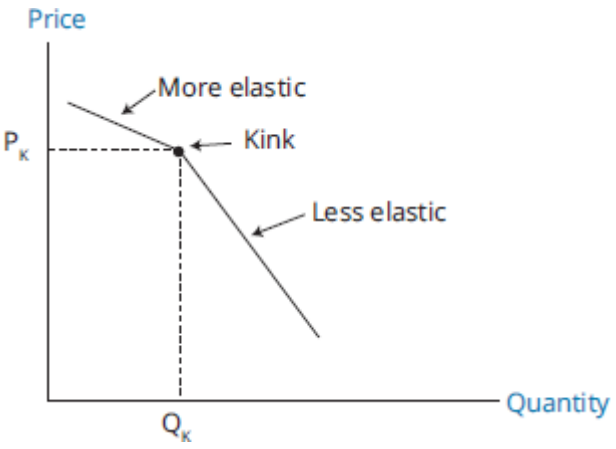
LOS 12.d: Explain supply and demand relationships under oligopoly, including the optimal price and output for firms as well as pricing strategy.

Compared to monopolistic competition, an oligopoly market has higher barriers to entry and fewer firms. The other key difference is that the firms are interdependent; a price change by one firm can be expected to be met by a price change by its competitors in response. This means that the actions of another firm will directly affect a given firm's demand curve for the product. Given this complicating fact, models of oligopoly pricing and profits must make numerous important assumptions. In the following, we describe four of these models and their implications for price and quantity:

1. Kinked demand curve model
2. Cournot duopoly model
3. Nash equilibrium model
4. Stackelberg dominant firm model

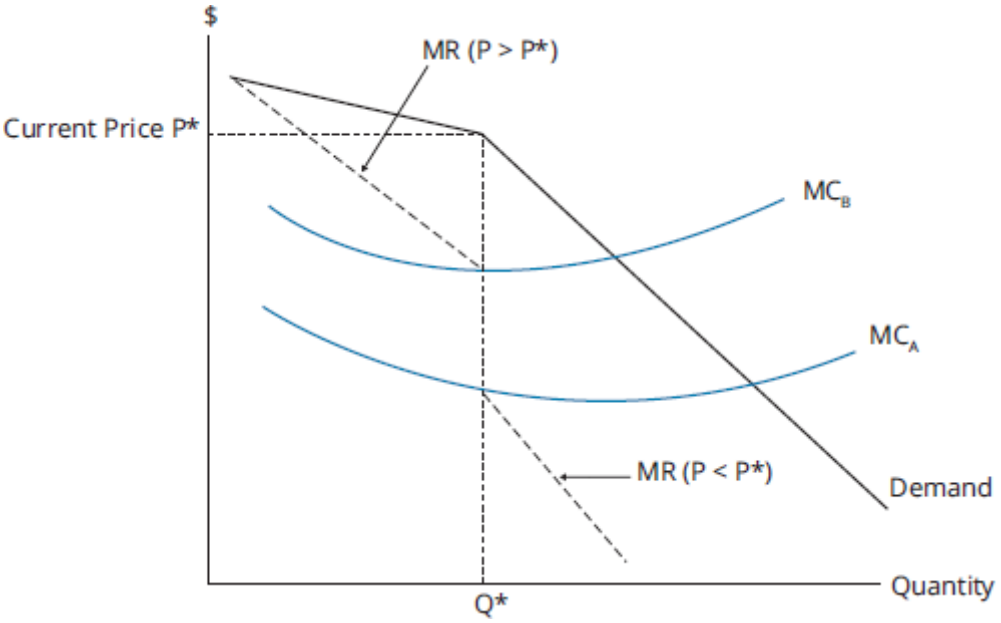
One traditional model of oligopoly, the **kinked demand curve model**, is based on the assumption that competitors are unlikely to match a price increase by a competitor, but very likely to match a price decrease by a competitor. This results in a kink in the demand curves faced by each producer, at the current market price. Each firm believes that it faces a demand curve that is more elastic (flatter) above the current price (the kink in the demand curve) than it is below the given price. The kinked demand curve model is illustrated in Figure 12.7. The kink price is at price P_K, where a firm produces Q_K. A firm believes that if it raises its price above P_K, its competitors will remain at P_K, and it will lose market share because it has the highest price. Above P_K, the demand curve is considered to be relatively elastic (i.e., a small price increase will result in a large decrease in demand). On the other hand, if a firm decreases its price below P_K, other firms will match the price cut, and all firms will experience a relatively small increase in sales relative to any price reduction. Therefore, Q_K is the profit-maximizing level of output.

Figure 12.7: Kinked Demand Curve Model



With a kink in the demand curve, we also get a gap in the associated MR curve, as shown in Figure 12.8. For any firm with a MC curve passing through this gap, the price where the kink is located is the firm's profit-maximizing price.

Figure 12.8: Marginal Revenue With Kinked Demand Curve



We say that the decisions of firms in an oligopoly are interdependent; that is, the pricing decision of one firm depends on the pricing decisions of other firms. Some models of market price equilibrium have a set of rules for the actions of oligopolists. These rules assume they choose prices based on the choices of the other firms. By specifying the decision rules that each firm follows, we can design a model that allows us to determine the equilibrium prices and quantities for firms operating in an oligopoly market.

An early model of oligopoly pricing decisions is the **Cournot model**. In Cournot's duopoly model, two firms with identical MC curves each choose their preferred selling price based on the price the other firm chose in the previous period. Firms assume that the competitor's price will not change. The long-run equilibrium for an oligopoly with two firms (duopoly), in the Cournot model, is for both firms to sell the same quantity, dividing the market equally at the equilibrium price. The equilibrium price is less than the price that a single monopolist would charge, but greater than the equilibrium price that would result under perfect competition. With a greater number of producers, the long-run market equilibrium price moves toward the competitive price.

Another model, the **Stackelberg model**, uses a different set of rules and produces a different result. While the Cournot model assumes the competitors choose price simultaneously each period, the Stackelberg model assumes pricing decisions are made sequentially. One firm, the "leader," chooses its price first, and the other firm chooses a price based on the leader's price. In long-run equilibrium, under these rules, the leader charges a higher price and receives a greater proportion of the firms' total profits.

These models are early versions of *rules-based models*, which fall under the heading of what are now generally termed *strategic games*. Strategic games comprise decision

models in which the best choice for a firm depends on the expected actions (reactions) of other firms.

A more general model of strategic games was developed by Nobel Prize winner John Nash, who developed the concept of a **Nash equilibrium**. A Nash equilibrium is reached when the choices of all firms are such that there is no other choice that makes any firm better off (increases profits or decreases losses). The Cournot model results in a Nash equilibrium. In equilibrium, neither competitor can increase profits by changing the price they charge.

The concept of a Nash equilibrium can be illustrated with the situation presented in Figure 12.9, which shows the choices and resulting profits for two firms. Each firm can charge either a high price or a low price. If both firms charge a high price, Firm A earns 1,000 and Firm B earns 600. While Firm A would not charge the low price (it would earn less regardless of Firm B's decision), Firm B can increase profits to 700 by charging a low price. With Firm A charging a high price and Firm B charging a low price, neither firm can increase profits by unilaterally changing its price strategy. Thus, we can identify the Nash equilibrium in this scenario as Firm B charging a low price and Firm A charging a high price.

Figure 12.9: Nash Equilibrium

	Firm B high price	Firm B low price
Firm A high price	A earns 1,000 B earns 600	A earns 600 B earns 700
Firm A low price	A earns 160 B earns 0	A earns 100 B earns 140

Thus far, we have assumed that firms act competitively to maximize their individual profits. We can illustrate the potential of cooperative behavior among producers to increase the total profits of all firms in the market. **Collusion** refers to competitors making a joint agreement to charge a given price—or, alternatively, to agree to specific levels of output. In Figure 12.9, the greatest joint profits (1,600) are earned when both firms charge a high price. If Firm A offers to pay Firm B 200 for charging a high price, Firm A's profits increase from 600 to 1,000. After paying 200 to Firm B, Firm A still gains 200. Firm B's profits (including the payment of 200) increase from 700 to 800. Collusion, in this case, increases the profits of both firms, compared to the Nash equilibrium. If firms can enter into and enforce an agreement regarding pricing and output, often they can all benefit. Such agreements among producers are illegal in many countries because they reduce competition.

An example of a collusive agreement is the OPEC **cartel**. Cartel-member countries agree to restrict their oil production to increase the world price of oil. Members sometimes choose to "cheat" on the cartel agreement by producing more than the amount of oil they have agreed to produce. If members of a cartel do not adhere to the agreement (taking advantage of the higher market price but failing to restrict output to the agreed-upon amount), the agreement can quickly break down.

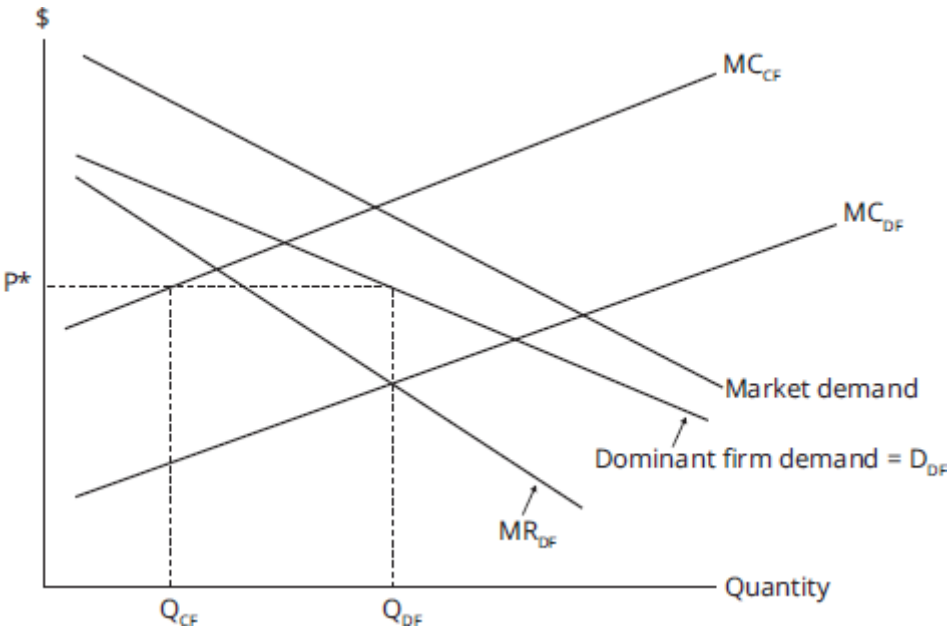
In general, collusive agreements to increase price in an oligopoly market will be more successful (have less cheating) under the following conditions:

- There are fewer firms.
- Products are more similar (less differentiated).
- Cost structures are more similar.
- Purchases are relatively small and frequent.
- Retaliation by other firms for cheating is more certain and more severe.
- There is less actual or potential competition from firms outside the cartel.

A final model of oligopoly behavior to consider is the **dominant firm model**. In this model, a single firm has a significantly large market share because of its greater scale and lower cost structure—the dominant firm (DF). In such a model, the market price is essentially determined by the DF, and the other competitive firms (CFs) take this market price as given.

The DF believes that the quantity supplied by the other firms decreases at lower prices, so that the DF's demand curve is related to the market demand curve, as shown in Figure 12.10. Based on this demand curve (D_{DF}) and its associated marginal revenue (MR_{DF}) curve, the firm will maximize profits at a price of P^*. The CFs maximize profits by producing the quantity for which their marginal cost (MC_{CF}) equals P^*, quantity Q_{CF}.

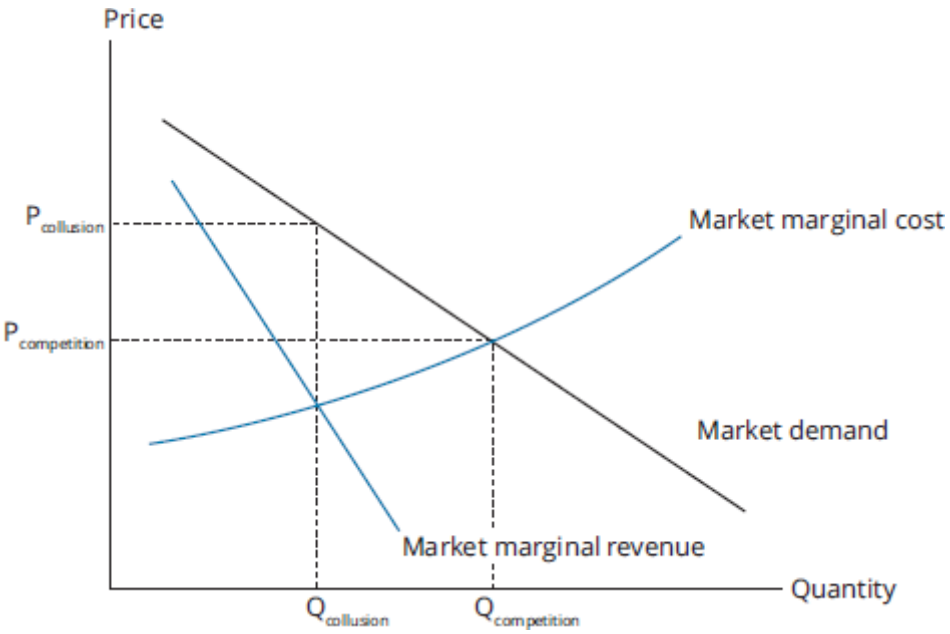
Figure 12.10: Dominant Firm Oligopoly



A price decrease by one of the CFs, which increases Q_{CF} in the short run, will lead to a decrease in price by the DF, and CFs will decrease output or exit the industry in the long run. The long-run result of such a price decrease by competitors below P^* would then be to decrease the overall market share of competitor firms and increase the market share of the DF.

Clearly, oligopoly markets exhibit many possible outcomes that depend on the characteristics of the firms and of the market itself. The important point is that the firms' decisions are interdependent so that the expected reactions of other firms are an important consideration. Overall, the resulting price will be somewhere between the price based on perfect collusion that would maximize total profits to all firms in the market (which is actually the monopoly price), and the price that would result from perfect competition and generate zero economic profits in the long run. These two limiting outcomes are illustrated in Figure 12.11 as $P_{\text{collusion}}$ with $Q_{\text{collusion}}$ for perfect collusion, and $P_{\text{competition}}$ and $Q_{\text{competition}}$ for perfect competition.

Figure 12.11: Collusion vs. Perfect Competition



MODULE QUIZ 12.2

1. The demand for products from monopolistic competitors is relatively elastic due to:
   A. high barriers to entry.
   B. the availability of many close substitutes.
   C. the availability of many complementary goods.

2. Compared to a perfectly competitive industry, in an industry characterized by monopolistic competition:
   A. both price and quantity are likely to be lower.
   B. price is likely to be higher, and quantity is likely to be lower.
   C. quantity is likely to be higher, and price is likely to be lower.

3. A firm will *most likely* maximize profits at the quantity of output for which:
   A. price equals marginal cost.
   B. price equals marginal revenue.
   C. marginal cost equals marginal revenue.

4. An oligopolistic industry has:
   A. few barriers to entry.
   B. few economies of scale.
   C. a great deal of interdependence among firms.

5. Consider a firm in an oligopoly market that believes the demand curve for its product is more elastic above a certain price than below this price. This belief fits *most*

appropriately to which of the following models?

A. Cournot model.
B. Dominant firm model.
C. Kinked demand model.

6. Consider an agreement between France and Germany that will restrict wine production so that maximum economic profit can be realized. The possible outcomes of the agreement are presented in the following table.

	Germany complies	Germany defaults
France complies	France gets €8 billion Germany gets €8 billion	France gets €2 billion Germany gets €10 billion
France defaults	France gets €10 billion Germany gets €2 billion	France gets €4 billion Germany gets €4 billion

Based on the concept of a Nash equilibrium, the *most likely* strategy followed by the two countries with respect to whether they comply with or default on the agreement will be:

A. both countries will default.
B. both countries will comply.
C. one country will default and the other will comply.

MODULE 12.3: IDENTIFYING MARKET STRUCTURES

Video covering this content is available online.

LOS 12.e: Identify the type of market structure within which a firm operates and describe the use and limitations of concentration measures.

We can use the characteristics we outlined earlier to identify the type of market structure within which a firm is operating. Our earlier table is repeated here in Figure 12.12. For an analyst attempting to determine the degree of pricing power that firms in the industry have, the focus is on the number of firms in the industry, its barriers to entry, the nature of substitute products, and the nature of industry competition. Significant interdependence among firm pricing and output decisions is a characteristic of all oligopoly markets, although some interdependence may be present under monopolistic competition—even with many more firms than for an oligopoly structure.

The following table illustrates the differences in characteristics among the various market structures.

Figure 12.12: Characteristics of Market Structures

	Perfect Competition	Monopolistic Competition	Oligopoly	Monopoly
Number of sellers	Many firms	Many firms	Few firms	Single firm
Barriers to entry	Very low	Low	High	Very high
Nature of substitute products	Very good substitutes	Good substitutes, but differentiated	Good substitutes or differentiated	No good substitutes
Nature of competition	Price only	Price, marketing, features	Price, marketing, features	Advertising
Pricing power	None	Some	Some to significant	Significant

Given the differences in cost structures, suitability of substitutes, and the degree of differentiation, simply classifying the primary characteristics of a market does not tell us the degree of pricing power for individual firms or the magnitude of the difference between market prices and the prices implied by perfect competition. What we would really like (especially for the regulation of markets) is to be able to measure the elasticity of firm demand directly, but that is difficult and subject to estimation error.

Consequently, regulators often use market shares (percentages of market sales) to measure the degree of monopoly or market power of firms in an industry. Often, mergers or acquisitions of companies in the same industry or market are not permitted by government authorities when they determine that the market share of the combined firms will be too high and, therefore, detrimental to the economy.

Market or industry **concentration measures** are often used as an indicator of market power. One concentration measure is the **N-firm concentration ratio**, which is calculated as the sum of the percentage market shares of the largest *N* firms in a market. While this measure is simple to calculate and understand, it does not directly measure market power or elasticity of demand.

One limitation of the N-firm concentration ratio is that it may be relatively insensitive to mergers of firms within an industry. This problem is reduced by using an alternative measure of market concentration, the **Herfindahl-Hirschman Index (HHI)**. The HHI is calculated as the sum of the squares of the market shares of the largest firms in the market. The following example illustrates this difference between the two measures and their calculation.

EXAMPLE: 4-firm concentration ratio and 4-firm HHI

Given the market shares of the following firms, calculate the 4-firm concentration ratio and the 4-firm HHI, both before and after a merger of Acme and Blake.

Firm	Market Share
Acme	25%
Blake	15%
Curtis	15%
Dent	10%
Erie	5%
Federal	5%

Answer:

Before the merger, the 4-firm concentration ratio for the market is 25 + 15 + 15 + 10 = 65%. After the merger, the Acme + Blake firm has 40% of the market, and the 4-firm concentration ratio is 40 + 15 + 10 + 5 = 70%. Although the 4-firm concentration ratio has only increased slightly, the market power of the largest firm in the industry has increased significantly, from 25% to 40%.

Before the merger, the 4-firm HHI is $0.25^2 + 0.15^2 + 0.15^2 + 0.10^2 = 0.1175$.

After the merger, the 4-firm HHI is $0.40^2 + 0.15^2 + 0.10^2 + 0.05^2 = 0.1950$, a significant increase.

A limitation that applies to both of our simple concentration measures is that barriers to entry are not considered in either case. Even a firm with high market share may not have much pricing power if barriers to entry are low and there is *potential competition*. With low barriers to entry, it may be the case that other firms stand ready to enter the market if firms currently in the market attempt to increase prices significantly. In this case, the elasticity of demand for existing firms may be high, even though they have relatively high market shares and industry concentration measures.

MODULE QUIZ 12.3

1. Which of the following is *most likely* an advantage of the Herfindahl-Hirschman Index (HHI) relative to the *N*-firm concentration ratio?
   A. The HHI is simpler to calculate.
   B. The HHI considers barriers to entry.
   C. The HHI is more sensitive to mergers.
2. A market characterized by low barriers to entry, good substitutes, limited pricing power, and marketing of product features is *best* characterized as:
   A. oligopoly.
   B. perfect competition.
   C. monopolistic competition.

KEY CONCEPTS

LOS 12.a

The breakeven quantity of production is the quantity for which price (P) = average total cost (ATC), and total revenue (TR) = total cost (TC).

A firm should shut down in the long run if P < ATC so that TR < TC. A firm should shut down in the short run (and the long run) if P < average variable cost (AVC) so that TR < total variable cost (TVC).

The long-run average total cost (LRATC) curve shows the minimum average total cost for each level of output, assuming that the plant size (scale of the firm) can be adjusted. A downward-sloping segment of an LRATC curve indicates economies of scale (increasing returns to scale). Over such a segment, increasing the scale of the firm reduces ATC. An upward-sloping segment of an LRATC curve indicates diseconomies of scale, where average unit costs will rise as the scale of the business (and long-run output) increases.

LOS 12.b

Perfect competition is characterized by the following:

- Many firms, each small relative to the market
- Very low barriers to entry into or exit from the industry
- Homogeneous products that are perfect substitutes; no advertising or branding
- No pricing power

Monopoly is characterized by the following:

- A single firm that comprises the whole market
- Very high barriers to entry into or exit from the industry
- Advertising used to compete with substitute products
- Significant pricing power

Monopolistic competition is characterized by the following:

- Many firms
- Low barriers to entry into or exit from the industry
- Differentiated products; heavy advertising and marketing expenditure
- Some pricing power

Oligopoly markets are characterized by the following:

- Few sellers
- High barriers to entry into or exit from the industry
- Products that may be homogeneous or differentiated by branding and advertising
- Firms that may have significant pricing power

LOS 12.c

Under monopolistic competition, firms face downward-sloping demand curves so that marginal revenue is less than price, and the price from the demand curve at the profit-maximizing quantity is the price at the optimal (profit-maximizing) level of output. Resources expended on product differentiation may increase ATC so that there are no economic profits at long-run equilibrium. There is no well-defined firm supply curve.

Monopolistic competition:

- Price > marginal revenue = marginal cost (in equilibrium)
- Zero economic profit in long-run equilibrium

LOS 12.d

Under oligopoly, the pricing strategy is not clear. Because firm decisions are interdependent, the optimal pricing and output strategy depends on assumptions made about other firms' cost structures and about competitors' responses to a firm's price changes.

Models of oligopoly pricing include the following:

- *Cournot*. Firms with a given MC that make pricing decisions simultaneously share the market equally in LR equilibrium.
- *Stackelberg*. When firm pricing decisions are sequential, the first firm to set its price (price leader) will maintain a larger share of the market, even in the long run.
- *Dominant firm*. A firm with a significantly lower cost of production will gain a disproportionate share of the market and essentially set the price that other competitors can charge.

In general, in the absence of collusion, the long-run oligopoly equilibrium is Nash equilibrium, one in which no competitor can unilaterally change price to increase its profits. There is no well-defined firm supply curve.

Oligopoly:

- Price > marginal revenue = marginal cost (in equilibrium)
- May have positive economic profit in long-run equilibrium, but may move toward zero economic profit over time

LOS 12.e

To identify the market structure in which a firm is operating, we can examine the number of firms in its industry, the prevalence of nonprice competition, and barriers to entry, then compare these to the characteristics that define each market structure. However, none of this provides us with the elasticity of firm demand.

Because direct estimation of the elasticity of firm demand is statistically difficult and imprecise, industry concentration measures are often used as a proxy for the degree of competition. A concentration ratio for N firms is calculated as the percentage of market sales accounted for by the N largest firms in the industry and is used as a simple measure of market structure and market power.

The Herfindahl-Hirschman Index measure of concentration is calculated as the sum of the squared market shares of the largest N firms in an industry and better reflects the effect of mergers on industry concentration.

Neither measure indicates market power directly. Both can be misleading when potential competition restricts pricing power.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 12.1

1. **A** In a purely competitive market, economic losses indicate that firms are overproducing, causing prices to fall below average total costs. This can occur in the short run. In the long run, however, market supply will decrease as firms exit the industry, and prices will rise to the point where economic profits are zero. (LOS 12.a)
2. **C** If price is greater than average variable cost, a firm will continue to operate in the short run because it is covering at least some of its fixed costs. (LOS 12.a)
3. **C** If a firm is generating sufficient revenue to cover its variable costs and part of its fixed costs, it should continue to operate in the short run. If average revenue is likely to remain below average total costs in the long run, the firm should shut down. (LOS 12.a)
4. **B** If minimum average total costs increase as plant size is increased, the firm is experiencing diseconomies of scale. (LOS 12.a)

Module Quiz 12.2

1. **B** The demand for products from firms competing in monopolistic competition is relatively elastic due to the availability of many close substitutes. If a firm increases its product price, it will lose customers to firms selling substitute products at lower prices. (LOS 12.c)
2. **B** Monopolistic competition is likely to result in a higher price and lower quantity of output compared to perfect competition. (LOS 12.c)
3. **C** The profit-maximizing output is the quantity at which marginal revenue equals marginal cost. In a price-searcher industry structure (i.e., any structure that is not perfect competition), price is greater than marginal revenue. (LOS 12.c)
4. **C** An oligopolistic industry has a great deal of interdependence among firms. One firm's pricing decisions or advertising activities will affect the other firms. (LOS 12.d)
5. **C** The kinked demand model assumes that each firm in a market believes that at some price, demand is more elastic for a price increase than for a price decrease. (LOS 12.d)
6. **A** The Nash equilibrium results when each nation pursues the strategy that is best, given the strategy that is pursued by the other nation.
   - Given that Germany complies with the agreement: France will get €8 billion if it complies, but €10 billion if it defaults. Therefore, France should default.

- Given that Germany defaults: France will get €2 billion if it complies, but €4 billion if it defaults. Therefore, France should default.

Because France is better off in either case by defaulting, France will default. Germany will follow the same logic and reach the same conclusion. (LOS 12.d)

Module Quiz 12.3

1. **C** Although the *N*-firm concentration ratio is simple to calculate, it can be relatively insensitive to mergers between companies with large market shares. Neither the HHI nor the *N*-firm concentration ratio consider barriers to entry. (LOS 12.e)

2. **C** These characteristics are associated with a market structure of monopolistic competition. Firms in perfect competition do not compete on product features. Oligopolistic markets have high barriers to entry. (LOS 12.e)

READING 13

UNDERSTANDING BUSINESS CYCLES

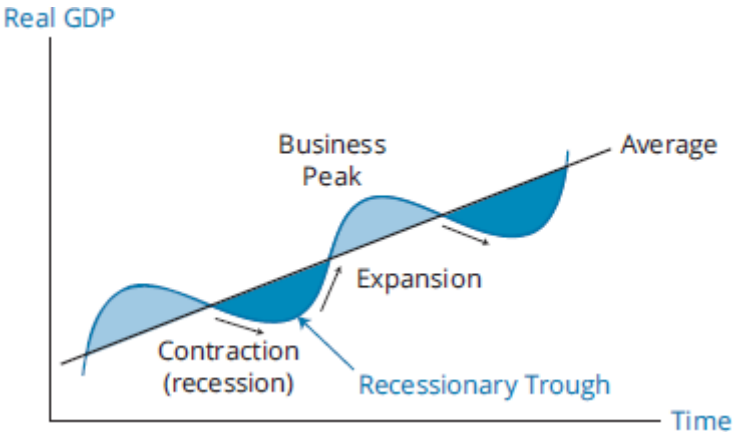
MODULE 13.1: BUSINESS CYCLES

Video covering this content is available online.

LOS 13.a: Describe the business cycle and its phases.

The **business cycle** is characterized by fluctuations in economic activity. The business cycle has four phases: **expansion** (real GDP is increasing), **peak** (real GDP stops increasing and begins decreasing), **contraction** or **recession** (real GDP is decreasing), and **trough** (real GDP stops decreasing and begins increasing). The phases are illustrated in Figure 13.1.

Figure 13.1: Business Cycle



There are alternative ways to show business cycles. The curve in the figure illustrates the **classical cycle**, which is based on real GDP relative to a beginning value. The **growth cycle** refers to changes in the percentage difference between real GDP and its longer-term trend or potential value (shown as the "average" line in the figure). The **growth rate cycle** refers to changes in the annualized percentage growth rate from one period to the next and tends to show both peaks and troughs earlier than the other two measures. The growth rate cycle—which, like the growth cycle, shows GDP relative to a trend rate—is the preferred measure for economists and practitioners.

An expansion features growth in most sectors of the economy, with increasing employment, consumer spending, and business investment. As an expansion approaches

its peak, the rates of increase in spending, investment, and employment slow but remain positive, while inflation accelerates.

A contraction or recession is associated with declines in most sectors, with inflation typically decreasing. When a contraction reaches a trough and the economy begins a new expansion or **recovery**, economic growth becomes positive again and inflation is typically moderate, but employment growth might not start to increase until the expansion has taken hold convincingly.

Economists commonly consider two consecutive quarters of growth in real GDP to be the beginning of an expansion, and two consecutive quarters of declining real GDP to be the beginning of a contraction. Statistical agencies that date expansions and recessions, such as the National Bureau of Economic Research in the United States, look at a wider variety of additional economic data—especially unemployment, industrial production, and inflation—to identify turning points in the business cycle.

A key aspect of business cycles is that they recur, but not at regular intervals. Past business cycles have been as short as a year, or longer than a decade. The idea of a business cycle applies to economies that consist mainly of businesses. For economies that are mostly subsistence agriculture or dominated by state planning, fluctuations in activity are not really "business cycles" in the sense we are discussing here.

LOS 13.b: Describe credit cycles.

Credit cycles refer to cyclical fluctuations in interest rates and the availability of loans (credit). Typically, lenders are more willing to lend, and tend to offer lower interest rates, during economic expansions. Conversely, they are less willing to lend, and require higher interest rates, when the economy is slowing (contracting).

Credit cycles may amplify business cycles. Widely available or "loose" credit conditions during expansions can lead to "bubbles" (prices based on implausible expectations) in the markets for some assets, such as subprime mortgages in the period leading up to the financial crisis of 2007–2009. Some research suggests that expansions tend to be stronger, and contractions deeper and longer lasting, when they coincide with credit cycles. They do not always coincide, however, as historical data suggest credit cycles have been longer in duration than business cycles on average.

LOS 13.c: Describe how resource use, consumer and business activity, housing sector activity, and external trade sector activity vary over the business cycle and describe their measurement using economic indicators.

Business Cycles and Resource Use Fluctuation

Inventories are an important business cycle indicator. Firms try to keep enough inventory on hand to meet sales demand but do not want to keep too much of their capital tied up in inventory. As a result, the **inventory-sales ratio** in many industries tends toward a normal level in times of steady economic growth.

When an expansion is approaching its peak, sales growth begins to slow, and unsold inventories accumulate. This can be seen in an increase in the inventory-sales ratio above its normal level. Firms respond to this unplanned increase in inventory by reducing production, which is one of the causes of the subsequent contraction in the economy. An increase in inventories is counted in the GDP statistics as economic output whether the increase is planned or unplanned. An analyst who looks only at GDP growth, rather than the inventory-sales ratio, might see economic strength rather than the beginning of weakness.

The opposite occurs when a contraction reaches its trough. Having reduced their production levels to adjust for lower sales demand, firms find their inventories becoming depleted more quickly once sales growth begins to accelerate. This causes the inventory-sales ratio to decrease below its normal level. To meet the increase in demand, firms will increase output, and the inventory-sales ratio will increase back toward normal levels.

One of the ways that firms react to fluctuations in business activity is by adjusting their use of labor and physical capital. Adding and subtracting workers in lockstep with changes in economic growth would be costly for firms, in terms of both direct expenses and the damage it would do to employee morale and loyalty. Instead, firms typically begin by changing how they use their current workers, producing less or more output per hour or adjusting the hours they work by adding or removing overtime. Only when an expansion or contraction appears likely to persist will they hire or lay off workers.

Similarly, because it is costly to adjust production levels by frequently buying and selling plant and equipment, firms first adjust their production levels by using their existing physical capital more or less intensively. As an expansion persists, firms will increase their production capacity by investing more in plant and equipment. During contractions, however, firms will not necessarily sell plant and equipment outright. They can reduce their physical capacity by spending less on maintenance or by delaying the replacement of equipment that is nearing the end of its useful life.

Consumer Sector Activity

Consumer spending, the largest component of gross domestic product, depends on the level of consumers' current incomes and their expectations about their future incomes. As a result, consumer spending increases during expansions and decreases during contractions.

Consumer spending in some sectors is more sensitive to business cycle phases than spending in other sectors. Spending on **durable goods** is highly cyclical because they are often higher-value purchases. Consumers are more willing to purchase high-value durable goods (e.g., appliances, furniture, automobiles) during expansions, when incomes are increasing and economic confidence is high. During contractions (and sometimes extending into the early stages of expansions), consumers often postpone durable goods purchases until they are more confident about their employment status and prospects for income growth.

Consumer spending on **services** is also positively correlated with business cycle phases, but not to the same extent as durable goods spending. Services include some

discretionary spending, such as for travel, lodging, and restaurant meals, but they also include spending that is less discretionary, such as for telecommunications, health care, and insurance. Spending on **nondurable goods**, such as food at home or household products for everyday use, remains relatively stable over the business cycle.

Housing Sector Activity

Although the housing sector is a smaller part of the economy relative to overall consumer spending, cyclical swings in activity in the housing market can be large, so that the effect on overall economic activity is greater than it otherwise would be. Important determinants of the level of economic activity in the housing sector are as follows:

1. *Mortgage rates.* Low interest rates tend to increase home buying and construction, while high interest rates tend to reduce home buying and construction.
2. *Housing costs relative to income.* When incomes are cyclically high (low) relative to home costs, including mortgage financing costs, home buying and construction tend to increase (decrease). Housing activity can decrease even when incomes are rising if home prices are rising faster than incomes, as often occurs late in expansions.
3. *Speculative activity.* As we saw in the housing sector in 2007–08 in many economies, rising home prices can lead to purchases based on expectations of further gains. Higher prices led to more construction and eventually excess building. This resulted in falling prices that decreased or eliminated speculative demand and led to dramatic decreases in house prices and in housing activity overall.
4. *Demographic factors.* The proportion of the population in the 25- to 40-year-old segment is positively related to activity in the housing sector because these are the ages of greatest household formation. Strong population shifts from rural areas to cities, as may occur in newly industrializing economies, may require large increases in construction of new housing to accommodate those needs.

External Trade Sector Activity

The most important factors determining the level of a country's imports and exports are domestic GDP growth, GDP growth of trading partners, and currency exchange rates. Increasing growth of domestic GDP leads to increases in purchases of foreign goods (imports), while decreasing domestic GDP growth reduces imports. Exports depend on the growth rates of GDP of other economies (especially those of important trading partners). Increasing foreign incomes increase sales to foreigners (exports), and decreasing economic growth in foreign countries decreases domestic exports.

An increase in the value of a country's currency makes its goods more expensive to foreign buyers and foreign goods less expensive to domestic buyers, which tends to decrease exports and increase imports. A decrease in the value of a country's currency has the opposite effect (increasing exports and decreasing imports). Currencies affect import and export volumes over time in response to persistent trends in foreign exchange rates, rather than in response to short-term changes, which can be quite volatile.

Currency effects on imports and exports can differ in direction from GDP growth effects and respond to a complex set of variables. The effects of changes in GDP levels and growth rates are more direct and immediate.

Typical business cycle characteristics may be summarized as follows:

- *Trough:*
  - The GDP growth rate changes from negative to positive.
  - There is a high unemployment rate, and an increasing use of overtime and temporary workers.
  - Spending on consumer durable goods and housing may increase.
  - There is a moderate or decreasing inflation rate.
- *Expansion:*
  - The GDP growth rate increases.
  - The unemployment rate decreases as hiring accelerates.
  - There are investment increases in producers' equipment and home construction.
  - The inflation rate may increase.
  - Imports increase as domestic income growth accelerates.
- *Peak:*
  - The GDP growth rate decreases.
  - The unemployment rate decreases, but hiring slows.
  - Consumer spending and business investments grow at slower rates.
  - The inflation rate increases.
- *Contraction/recession:*
  - The GDP growth rate is negative.
  - Hours worked decrease; unemployment rate increases.
  - Consumer spending, home construction, and business investments decrease.
  - The inflation rate decreases with a lag.
  - Imports decrease as domestic income growth slows.

Economic Indicators

Economic indicators can be classified into three categories: **leading indicators** that have been known to change direction before peaks or troughs in the business cycle, **coincident indicators** that change direction at roughly the same time as peaks or troughs, and **lagging indicators** that tend not to change direction until after expansions or contractions are already underway.

MODULE QUIZ 13.1

1. In the early part of an economic expansion, inventory-sales ratios are *most likely* to:
   A. increase because sales are unexpectedly low.
   B. increase because businesses plan for expansion.
   C. decrease because of unexpected increases in sales.
2. The contraction phase of the business cycle is *least likely* accompanied by decreasing:
   A. unemployment.
   B. inflation pressure.

C. economic output.

3. Which economic sector would *most likely* correlate strongly and positively with credit cycles?
   A. Exports.
   B. Food retail.
   C. Construction.
4. An economic indicator that has turning points that tend to occur after the turning points in the business cycle is classified as a:
   A. lagging indicator.
   B. leading indicator.
   C. trailing indicator.
5. When they recognize the beginning of a recession, companies are *most likely* to adjust their stock of physical capital by:
   A. selling physical assets.
   B. deferring maintenance of equipment.
   C. canceling orders for new construction equipment.

KEY CONCEPTS

LOS 13.a

The business cycle has four phases:

1. *Expansion*. Real GDP is increasing (or GDP growth relative to trend is increasing).
2. *Peak*. Real GDP stops increasing/begins decreasing (GDP growth relative to trend peaks).
3. *Contraction*. Real GDP is decreasing (or GDP growth relative to trend is decreasing).
4. *Trough*. Real GDP stops decreasing/begins increasing (GDP growth relative to trend reaches a low).

Expansions feature increasing output, employment, consumption, investment, and inflation. Contractions are characterized by decreases in these indicators.

Business cycles are recurring, but they do not occur at regular intervals, can differ in strength or severity, and do not persist for specific lengths of time.

LOS 13.b

Credit cycles are cyclical fluctuations in interest rates and credit availability. Credit cycles may amplify business cycles and cause bubbles in the markets for some assets.

LOS 13.c

Inventory to sales ratios typically increase late in expansions when sales slow, and decrease near the end of contractions when sales begin to accelerate. Firms decrease or increase production to restore their inventory-sales ratios to their desired levels.

Because hiring and laying off employees have high costs, firms prefer to adjust their use of current employees. As a result, firms are slow to lay off employees early in contractions and slow to add employees early in expansions.

Firms use their physical capital more intensively during expansions, investing in new capacity only if they believe the expansion is likely to continue. They use physical capital less intensively during contractions, but they are more likely to reduce capacity by deferring maintenance and not replacing equipment than by selling their physical capital.

Consumer spending fluctuates with the business cycle. Durable goods spending is highly sensitive to business cycles, and spending on services is somewhat sensitive, but spending on nondurable goods is relatively less sensitive to business cycles.

The level of activity in the housing sector is affected by mortgage rates, demographic changes, the ratio of income to housing prices, and investment or speculative demand for homes resulting from recent price trends.

Domestic imports tend to rise with increases in GDP growth and domestic currency appreciation, while increases in foreign incomes and domestic currency depreciation tend to increase domestic export volumes.

Leading indicators have turning points that tend to precede those of the business cycle.

Coincident indicators have turning points that tend to coincide with those of the business cycle.

Lagging indicators have turning points that tend to occur after those of the business cycle.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 13.1

1. **C** Early in an expansion, inventory-sales ratios typically decrease below their normal levels as accelerating sales draw down inventories of produced goods. (LOS 13.c)

2. **A** An economic contraction is likely to feature increasing unemployment (i.e., decreasing employment), along with declining economic output and decreasing inflation pressure. (LOS 13.a)

3. **C** Credit cycles are associated with interest rates and the availability of credit, which is important in the financing of construction and the purchase of property. (LOS 13.b)

4. **A** Lagging indicators have turning points that occur after business cycle turning points. (LOS 13.c)

5. **B** Physical capital adjustments to downturns typically are made through aging of equipment plus postponing maintenance. (LOS 13.c)

READING 14

FISCAL POLICY

MODULE 14.1: FISCAL POLICY OBJECTIVES

Video covering this content is available online.

LOS 14.a: Compare monetary and fiscal policy.

Fiscal policy refers to a government's use of spending and taxation to influence economic activity. The budget is said to be *balanced* when tax revenues equal government expenditures. A **budget surplus** occurs when government tax revenues exceed expenditures, and a **budget deficit** occurs when government expenditures exceed tax revenues. An increase in the deficit (or a decrease in a surplus) is considered expansionary in that it tends to increase GDP. A decrease in a deficit (or increase in a surplus) is considered contractionary in that it tends to decrease GDP.

Monetary policy refers to the central bank's actions that affect the quantity of money and credit in an economy to influence economic activity. Monetary policy is said to be expansionary (or accommodative or easy) when the central bank increases the quantity of money and credit in an economy. Conversely, when the central bank is reducing the quantity of money and credit in an economy, monetary policy is said to be contractionary (or restrictive or tight).

Policymakers use both monetary and fiscal policies with the goals of maintaining stable prices and producing positive economic growth. Fiscal policy can also be used as a tool for redistribution of income and wealth.

LOS 14.b: Describe roles and objectives of fiscal policy as well as arguments as to whether the size of a national debt relative to GDP matters.

Objectives of fiscal policy may include the following:

- Influencing the level of economic activity and aggregate demand
- Redistributing wealth and income among segments of the population
- Allocating resources among economic agents and sectors in the economy

In general, decreased taxes and increased government spending both *increase* a budget deficit, overall demand, economic growth, and employment; increased taxes and decreased government spending *decrease* these. Budget deficits are increased in

response to recessions, and budget deficits are decreased to slow growth when inflation is too high.

Keynesian economists believe that fiscal policy, through its effect on aggregate demand, can have a strong effect on economic growth when the economy is operating at less than full employment. Monetarists believe that the effect of fiscal stimulus is only temporary, and that monetary policy should be used to increase or decrease inflationary pressures over time. Monetarists do not believe that monetary policy should be used in an attempt to influence aggregate demand to counter cyclical movements in the economy.

Discretionary fiscal policy refers to the spending and taxing decisions of a national government that are intended to stabilize the economy. In contrast, **automatic stabilizers** are built-in fiscal devices triggered by the state of the economy. For example, during a recession, tax receipts will fall, and government expenditures on unemployment insurance payments will increase. Both of these tend to increase budget deficits and are expansionary. Similarly, during boom times, higher tax revenues coupled with lower outflows for social programs tend to decrease budget deficits and are contractionary.

When a government runs fiscal deficits, it incurs debt that needs to be repaid as well as ongoing interest expense. Total deficits, annual deficits, and interest expense can all be evaluated relative to annual GDP. When these ratios increase beyond certain levels, it may be a cause for concern, and the solvency of the country may be questioned.

A country's **debt ratio** is the ratio of aggregate debt to GDP. Because taxes are linked to GDP, when an economy grows in real terms, so will tax revenues. If the real interest rate on the government's debt is higher than the real growth rate of the economy, then the debt ratio will increase over time (keeping tax rates constant). Similarly, if the real interest rate on the government's debt is lower than real growth in GDP, the debt ratio will decrease (i.e., improve) over time.

Arguments *for* being concerned with the size of a fiscal deficit are as follows:

- Higher deficits lead to higher future taxes. Higher future taxes will lead to disincentives to work and entrepreneurship. This leads to lower long-term economic growth.
- If markets lose confidence in the government, investors may not be willing to refinance the debt. This can lead to the government defaulting (if debt is in a foreign currency) or having to simply print money (if the debt is in local currency). Printing money would ultimately lead to higher inflation.
- Increased government borrowing will tend to increase interest rates, and firms may reduce their borrowing and investment spending as a result, thus decreasing the impact on aggregate demand of deficit spending. This is referred to as the **crowding-out effect** because government borrowing is taking the place of private-sector borrowing.

Arguments *against* being concerned with the size of a fiscal deficit are as follows:

- If the debt is primarily held by domestic citizens, the scale of the problem is overstated.
- If the debt is used to finance productive capital investment, future economic gains will be sufficient to repay the debt.
- Fiscal deficits may prompt needed tax reform.
- Deficits would not matter if **Ricardian equivalence** holds, which means private-sector savings in anticipation of future tax liabilities just offset the government deficit.
- If the economy is operating at less than full capacity, deficits do not divert capital away from productive uses and can aid in increasing GDP and employment.

MODULE QUIZ 14.1

1. Both monetary and fiscal policy are used to:
   A. balance the budget.
   B. achieve economic targets.
   C. redistribute income and wealth.
2. Roles and objectives of fiscal policy *most likely* include:
   A. controlling the money supply to limit inflation.
   B. adjusting tax rates to influence aggregate demand.
   C. using government spending to control interest rates.

MODULE 14.2: FISCAL POLICY TOOLS AND IMPLEMENTATION

Video covering this content is available online.

LOS 14.c: Describe tools of fiscal policy, including their advantages and disadvantages.

Fiscal policy tools include spending tools and revenue tools.

Spending Tools

Transfer payments, also known as entitlement programs, redistribute wealth, taxing some and making payments to others. Examples include government-run retirement income plans (such as Social Security in the United States) and unemployment insurance benefits. Transfer payments are not included in GDP computations.

Current spending refers to government purchases of goods and services on an ongoing and routine basis.

Capital spending refers to government spending on infrastructure, such as roads, schools, bridges, and hospitals. Capital spending is expected to boost future productivity of the economy.

Spending tools are commonly justified as means of pursuing the following goals:

- Provide services such as national defense that benefit all the residents in a country.
- Invest in infrastructure to enhance economic growth.

- Support the country's growth and unemployment targets by directly affecting aggregate demand.
- Provide a minimum standard of living.
- Subsidize investment in research and development for certain high-risk ventures consistent with future economic growth or other goals (e.g., green technology).

Revenue Tools

Direct taxes are levied on income or wealth. These include income taxes, taxes on income for national insurance, wealth taxes, estate taxes, corporate taxes, capital gains taxes, and Social Security taxes. Some progressive taxes (such as income taxes and wealth taxes) collect revenue for wealth and income redistributing.

Indirect taxes are levied on goods and services. These include sales taxes, value-added taxes (VATs), and excise taxes. Indirect taxes can be used to reduce consumption of some goods and services (e.g., alcohol, tobacco, gambling).

Desirable attributes of tax policy are as follows:

- *Simplicity* to use and enforce
- *Efficiency*, defined here as minimizing interference with market forces and not acting as a deterrent to working
- *Fairness* is quite subjective, but two of the commonly held beliefs are **horizontal equality** (people in similar situations should pay similar taxes) and **vertical equality** (richer people should pay more in taxes)
- *Sufficiency*, in that taxes should generate enough revenues to meet the spending needs of the government

Advantages of fiscal policy tools include:

- Social policies (e.g., discouraging tobacco use) can be implemented quickly via indirect taxes.
- Quick implementation of indirect taxes also means that government revenues can be increased without significant additional costs.

Disadvantages of fiscal policy tools include:

- Direct taxes and transfer payments take time to implement, delaying the impact of fiscal policy.
- Capital spending also takes a long time to implement; the economy may have recovered by the time its impact is felt.

Announcing a change in fiscal policy may have significant effects on expectations. For example, an announcement of a future increase in taxes may immediately reduce current consumption, rapidly producing the desired goal of reducing aggregate demand.

Not all fiscal policy tools affect economic activity equally. Spending tools are most effective in increasing aggregate demand. Tax reductions are somewhat less effective, as people may not spend the entire amount of the tax savings. Tax reductions for those with low incomes will be more effective in increasing aggregate demand, as those with

lower incomes tend to spend a larger proportion of income on consumption; that is, they save a smaller proportion of income and have a higher **marginal propensity to consume (MPC)**.

Fiscal Multiplier

Changes in government spending have magnified effects on aggregate demand because those whose incomes increase from increased government spending will, in turn, increase their spending, which increases the incomes and spending of others. The magnitude of the *multiplier effect* depends on the tax rate and on the marginal propensity to consume.

To understand the calculation of the multiplier effect, consider an increase in government spending of $100 when the MPC is 80% and the tax rate is 25%. The increase in spending increases incomes by $100, but $25 (100 × 0.25) of that will be paid in taxes. **Disposable income** is equal to income after taxes, so disposable income increases by $100 × (1 − 0.25) = $75. With an MPC of 80%, additional spending by those who receive the original $100 increase is $75 × 0.8 = $60.

This additional spending will increase others' incomes by $60 and disposable incomes by $60 × 0.75 = $45, from which they will spend $45 × 0.8 = $36.

Because each iteration of this process reduces the amount of additional spending, the effect reaches a limit. The **fiscal multiplier** determines the potential increase in aggregate demand resulting from an increase in government spending:

$$\text{fiscal multiplier} = \frac{1}{1 - \text{MPC}(1 - t)}$$

Here, with a tax rate of 25% and an MPC of 80%, the fiscal multiplier is 1 / [1 − 0.8(1 − 0.25)] = 2.5, and the increase of $100 in government spending has the potential to increase aggregate demand by $250.

The fiscal multiplier is inversely related to the tax rate (higher tax rate decreases the multiplier) and is directly related to the marginal propensity to consume (higher MPC increases the multiplier).

Balanced Budget Multiplier

To balance the budget, the government could increase taxes by $100 to just offset a $100 increase in spending. Changes in taxes also have a magnified effect on aggregate demand. An increase in taxes will decrease disposable income and consumption expenditures, thereby decreasing aggregate demand. The initial decrease in spending from a tax increase of $100 is 100 × MPC = 100 × 0.8 = $80; beyond that, the multiplier effect is the same as we described for a direct increase in government spending, and the overall decrease in aggregate demand for a $100 tax increase is 100(MPC) × fiscal multiplier—or, for our example, 100(0.8)(2.5) = $200.

Combining the total increase in aggregate demand from a $100 increase in government spending with the total decrease in aggregate demand from a $100 tax increase shows

that the net effect on aggregate demand of both is an increase of $250 – $200 = $50, so we can say that the balanced budget multiplier is positive.

If, instead of a $100 increase in taxes, we increased taxes by 100 / MPC = 100 / 0.8 = $125 and increased government spending by $100, the net effect on aggregate demand would be zero.

Ricardian Equivalence

Increases in the current deficit mean greater taxes in the future. To maintain their preferred pattern of consumption over time, taxpayers may increase current savings (reduce current consumption) to offset the expected cost of higher future taxes. If taxpayers reduce current consumption and increase current saving by just enough to repay the principal and interest on the debt the government issued to fund the increased deficit, there is no effect on aggregate demand. This is known as Ricardian equivalence, after economist David Ricardo. If taxpayers underestimate their future liability for servicing and repaying the debt so that aggregate demand is increased by equal spending and tax increases, Ricardian equivalence does not hold. Whether it does is an open question.

LOS 14.d: Explain the implementation of fiscal policy and difficulties of implementation as well as whether a fiscal policy is expansionary or contractionary.

Fiscal policy is implemented through changes in taxes and spending. This is called **discretionary fiscal policy** (as opposed to automatic stabilizers, discussed previously). Discretionary fiscal policy would be designed to be expansionary when the economy is operating below full employment. Fiscal policy aims to stabilize aggregate demand. During recessions, actions can be taken to increase government spending or decrease taxes. Either change tends to strengthen the economy by increasing aggregate demand, putting more money in the hands of corporations and consumers to invest and spend. During inflationary economic booms, actions can be taken to decrease government spending or increase taxes. Either change tends to slow the economy by decreasing aggregate demand, taking money out of the hands of corporations and consumers and causing both investment and consumption spending to decrease.

Discretionary fiscal policy is not an exact science. First, economic forecasts might be wrong, leading to incorrect policy decisions. Second, complications arise in practice that delay both the implementation of discretionary fiscal policy and the impact of policy changes on the economy. The lag between recessionary or inflationary conditions in the economy and the impact on the economy of fiscal policy changes can be divided into three types:

1. **Recognition lag.** Discretionary fiscal policy decisions are made by a political process. The state of the economy is complex, and it may take policymakers time to recognize the nature and extent of the economic problems.
2. **Action lag.** Governments take time to discuss, vote on, and enact fiscal policy changes.

3. **Impact lag.** Time elapses between the enactment of fiscal policy changes and when the impact of the changes on the economy actually takes place. It takes time for corporations and individuals to act on the fiscal policy changes, and fiscal multiplier effects occur only over time as well.

These lags can actually make fiscal policy counterproductive. For example, if the economy is in a recession phase, fiscal stimulus may be deemed appropriate. However, by the time fiscal stimulus is implemented and has its full impact, the economy may already be on a path to a recovery driven by the private sector.

Additional macroeconomic issues may hinder the usefulness of fiscal policy:

- *Misreading economic statistics*. The full employment level for an economy is not precisely measurable. If the government relies on expansionary fiscal policy mistakenly at a time when the economy is already at full capacity, it will simply drive inflation higher.
- *Crowding-out effect*. Expansionary fiscal policy may crowd out private investment. Greater government borrowing tends to increase interest rates, which decreases private investments (some investments will be uneconomic with higher borrowing costs). This crowding-out effect can reduce the impact of expansionary fiscal policy on aggregate demand. Opinions vary on the magnitude of this effect.
- *Supply shortages*. If economic activity is slow due to resource constraints (low availability of labor or other resources) and not due to low demand, expansionary fiscal policy will fail to achieve its objective and will probably lead to higher inflation.
- *Limits to deficits*. There is a limit to expansionary fiscal policy. If the markets perceive that the deficit is already too high as a proportion of GDP, funding the deficit will be problematic. This could lead to higher interest rates and actually make the situation worse.
- *Multiple targets*. If the economy has high unemployment coupled with high inflation, fiscal policy cannot address both problems simultaneously.

To determine if fiscal policy is expansionary or contractionary, economists often focus on *changes* in the budget surplus or deficit. An increase (decrease) in surplus is indicative of a contractionary (expansionary) fiscal policy. Similarly, an increase (decrease) in deficit is indicative of an expansionary (contractionary) fiscal policy.

PROFESSOR'S NOTE

For the exam, an increase (decrease) in a revenue item (e.g., sales tax) should be considered contractionary (expansionary), and an increase (decrease) in a spending item (e.g., construction of highways) should be considered expansionary (contractionary).

A government's intended fiscal policy is not necessarily obvious from just examining the current deficit. Consider an economy that is in recession so that transfer payments are increased and tax revenue is decreased, leading to a deficit. This does not necessarily indicate that fiscal policy is expansionary as, at least to some extent, the deficit is a natural outcome of the recession without any explicit action of the

government. Economists often use a measure called the **structural budget deficit** (or **cyclically adjusted budget deficit**) to gauge fiscal policy. This is the deficit that would occur based on current policies if the economy were at full employment.

MODULE QUIZ 14.2

1. A government enacts a program to subsidize farmers with an expansive spending program of $10 billion. At the same time, the government enacts a $10 billion tax increase over the same period. Which of the following statements *best* describes the impact on aggregate demand?
   A. Lower growth, because the tax increase will have a greater effect.
   B. No effect, because the tax and spending effects just offset each other.
   C. Higher growth, because the spending increase will have a greater effect.
2. A government reduces spending by $50 million. The tax rate is 30%, and consumers exhibit a marginal propensity to consume of 80%. The change in aggregate demand caused by the change in government spending is *closest* to:
   A. −$66 million.
   B. −$114 million.
   C. −$250 million.
3. The size of a national debt is *most likely* to be a concern for policymakers if:
   A. Ricardian equivalence holds.
   B. a crowding-out effect occurs.
   C. debt is used to finance capital growth.
4. A government is concerned about the timing of fiscal policy changes and is considering requiring the compilation and reporting of economic statistics weekly, rather than quarterly. The new reporting frequency is intended to decrease the:
   A. action lag.
   B. impact lag.
   C. recognition lag.
5. Fiscal policy is *most likely* to be expansionary if tax rates:
   A. and government spending both decrease.
   B. decrease and government spending increases.
   C. increase and government spending decreases.

KEY CONCEPTS

LOS 14.a

Fiscal policy is a government's use of taxation and spending to influence the economy. Monetary policy deals with determining the quantity of money supplied by the central bank. Both policies aim to achieve economic growth with price level stability, although governments use fiscal policy for social and political reasons as well.

LOS 14.b

Objectives of fiscal policy can include (1) influencing the level of economic activity, (2) redistributing wealth or income, and (3) allocating resources among industries.

There are arguments for being concerned with the size of a fiscal deficit:

- Higher future taxes lead to disincentives to work, negatively affecting long-term economic growth.

- Fiscal deficits may not be financed by the market when debt levels are high.
- There is a crowding-out effect as government borrowing increases interest rates and decreases the private-sector investment.

There are arguments against being concerned with the size of a fiscal deficit:

- Debt may be financed by domestic citizens.
- Deficits for capital spending can boost the productive capacity of the economy.
- Fiscal deficits may prompt needed tax reform.
- Ricardian equivalence may prevail: private savings rise in anticipation of the need to repay principal on government debt.
- When the economy is operating below full employment, deficits do not crowd out private investment.

LOS 14.c

Fiscal policy tools include spending tools and revenue tools. Spending tools include transfer payments, current spending (goods and services used by government), and capital spending (investment projects funded by government). Revenue tools include direct and indirect taxation.

An advantage of fiscal policy is that indirect taxes can be used to quickly implement social policies and can also be used to quickly raise revenues at a low cost.

Disadvantages of fiscal policy include time lags for implementing changes in direct taxes and time lags for capital spending changes to have an impact.

LOS 14.d

Fiscal policy is implemented by government changes in taxing and spending policies. Delays in realizing the effects of fiscal policy changes limit their usefulness. Delays can be caused by the following:

- *Recognition lag*. Policymakers may not immediately recognize when fiscal policy changes are needed.
- *Action lag*. Governments take time to enact needed fiscal policy changes.
- *Impact lag*. Fiscal policy changes take time to affect economic activity.

In general, fiscal policies that increase the government deficit are often considered expansionary, and policies that decrease the deficit are considered contractionary. However, because a deficit will naturally increase during a recession and decrease during an expansion, the impact of a current deficit is often judged relative to a structural (cyclically adjusted) deficit amount.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 14.1

1. **B** Both monetary and fiscal policies primarily strive to achieve economic targets such as inflation and GDP growth. Balancing the budget is not a goal for monetary policy and is a potential outcome of fiscal policy. Fiscal policy (but not monetary policy) may secondarily be used as a tool to redistribute income and wealth. (LOS 14.a)

2. **B** Influencing the level of aggregate demand through taxation and government spending is an objective of fiscal policy. Controlling inflation and interest rates are typical objectives of monetary policy. (LOS 14.b)

Module Quiz 14.2

1. **C** The amount of the spending program exactly offsets the amount of the tax increase, leaving the budget unaffected. The multiplier for government spending is greater than the multiplier for a tax increase. Therefore, the balanced budget multiplier is positive. All of the government spending enters the economy as increased expenditure, whereas spending is reduced by only a portion of the tax increase. (LOS 14.c)

2. **B** fiscal multiplier = 1 / [1 − MPC (1 − T)] = 1 / [1 − 0.80(1 − 0.3)] = 2.27

   change in government spending = −$50 million

   change in aggregate demand = −(50 × 2.27) = −$113.64 million

   (LOS 14.c)

3. **B** Crowding out refers to the possibility that government borrowing causes interest rates to increase and private investment to decrease. If government debt is financing the growth of productive capital, this should increase future economic growth and tax receipts to repay the debt. Ricardian equivalence is the theory that if government debt increases, private citizens will increase savings in anticipation of higher future taxes, and it is an argument against being concerned about the size of government debt and budget deficits. (LOS 14.d)

4. **C** More frequent and current economic data would make it easier for authorities to monitor the economy and to recognize problems. The reduction in the time between economic reports should reduce the recognition lag. (LOS 14.d)

5. **B** Increases in government spending and decreases in taxes are expansionary fiscal policy. Decreases in spending and increases in taxes are contractionary fiscal policy. (LOS 14.d)

READING 15

MONETARY POLICY

MODULE 15.1: CENTRAL BANK OBJECTIVES AND TOOLS

Video covering this content is available online.

LOS 15.a: Describe the roles and objectives of central banks.

There are several key **roles of central banks**:

1. *Sole supplier of currency*. Central banks have the sole authority to supply money. Traditionally, such money was backed by gold; the central bank stood ready to convert the money into a prespecified quantity of gold. Later on, the gold backing was removed, and money supplied by the central bank was deemed **legal tender** by law. Money not backed by any tangible value is termed **fiat money**. As long as fiat money holds its value over time and is acceptable for transactions, it can continue to serve as a medium of exchange.
2. *Banker to the government and other banks*. Central banks provide banking services to the government and other banks in the economy.
3. *Regulator and supervisor of payments system*. In many countries, central banks may regulate the banking system by imposing standards of risk-taking allowed and reserve requirements of banks under its jurisdiction. Central banks also oversee the payments system to ensure smooth clearing operations domestically and in conjunction with other central banks for international transactions.
4. *Lender of last resort*. Central banks' ability to print money allows them to supply money to banks that are experiencing shortages. This government backing tends to prevent runs on banks (i.e., large-scale withdrawals) by assuring depositors that their funds are secure.
5. *Holder of gold and foreign exchange reserves*. Central banks are often the repositories of the nation's gold and reserves of foreign currencies.
6. *Conductor of monetary policy*. Central banks control or influence the quantity of money supplied in an economy and growth of money supply over time.

The primary objective of a central bank is to *control inflation* so as to promote price stability. High inflation is not conducive to a stable economic environment. High inflation leads to **menu costs** (i.e., cost to businesses of constantly having to change their prices) and **shoe leather costs** (i.e., costs to individuals of making frequent trips

to the bank so as to minimize their holdings of cash that are depreciating in value due to inflation).

In addition to price stability, some central banks have other stated goals:

- Stability in exchange rates with foreign currencies
- Full employment
- Sustainable positive economic growth
- Moderate long-term interest rates

The target inflation rate in most developed countries is a range around 2% to 3%. A target of zero inflation is not used because that increases the risk of deflation, which can be disruptive for an economy.

While most developed countries have an explicit target inflation rate, the U.S. Fed and the Bank of Japan do not. In the United States, this is because the Fed has the additional goals of maximum employment and moderate long-term interest rates. In Japan, it is because deflation, rather than inflation, has been a persistent problem in recent years.

Some developed countries, and several developing countries, choose a target level for the exchange rate of their currency with that of another country, primarily the U.S. dollar. This is referred to as **pegging** their exchange rate with the dollar. If their currency appreciates (i.e., becomes relatively more valuable), they can sell their domestic currency reserves for dollars to reduce the exchange rate. While such actions may be effective in the short run, for stability of the exchange rate over time, the monetary authorities in the pegging country must manage interest rates and economic activity to achieve their goal. This can lead to increased volatility of their money supply and interest rates. The pegging country essentially commits to a policy intended to make its inflation rate equal to the inflation rate of the country to which it pegs its currency.

LOS 15.b: Describe tools used to implement monetary policy tools and the monetary transmission mechanism, and explain the relationships between monetary policy and economic growth, inflation, interest, and exchange rates.

Monetary policy is implemented using the **monetary policy tools** of the central bank. The three main policy tools of central banks are as follows:

1. *Policy rate*. In the United States, banks can borrow funds from the Fed if they have temporary shortfalls in reserves. The rate at which banks can borrow reserves from the Fed is called the *discount rate*. For the European Central Bank (ECB), it is called the *refinancing rate*.

   One way to lend money to banks is through a *repurchase agreement*. The central bank purchases securities from banks that, in turn, agree to repurchase the securities at a higher price in the future. The percentage difference between the purchase price and the repurchase price is effectively the rate at which the central bank is lending to member banks. The Bank of England uses this method, and its policy rate is called the *two-week repo (repurchase) rate*. A lower rate reduces banks' cost of funds,

encourages lending, and tends to decrease interest rates overall. A higher policy rate has the opposite effect, decreasing lending and increasing interest rates.

In the United States, the *federal funds rate* is the rate that banks charge each other on overnight loans of reserves. The Fed sets a target for this market-determined rate and uses open market operations to influence it toward the target rate.

2. *Reserve requirements*. By increasing the reserve requirement (the percentage of deposits that banks are required to retain as reserves), the central bank effectively decreases the funds that are available for lending and thereby decreases the money supply, which will tend to increase interest rates. A decrease in the reserve requirement will increase the funds available for lending and the money supply, which will tend to decrease interest rates. This tool only works well to increase the money supply if banks are willing to lend and customers are willing to borrow.
3. *Open market operations*. Buying and selling of securities by the central bank is referred to as open market operations. When the central bank buys securities, cash replaces securities in investor accounts, banks have excess reserves, more funds are available for lending, the money supply increases, and interest rates decrease. Sales of securities by the central bank have the opposite effect, reducing cash in investor accounts, excess reserves, funds available for lending, and the money supply, which will tend to cause interest rates to increase. In the United States, open market operations are the Fed's most commonly used tool and are important in achieving the federal funds target rate.

Monetary Transmission Mechanism

The **monetary transmission mechanism** refers to the ways in which a change in monetary policy, specifically the central bank's policy rate, affects the price level and inflation. A change in the policy rates that the monetary authorities control directly is transmitted to prices through four channels: other short-term rates, asset values, currency exchange rates, and expectations.

We can examine the transmission mechanism in more detail by considering the effects of a change to a contractionary monetary policy implemented through an increase in the policy rate:

- Banks' *short-term lending rates will increase* in line with the increase in the policy rate. The higher rates will decrease aggregate demand as consumers reduce credit purchases and businesses cut back on investment in new projects.
- Bond prices, equity prices, and *asset prices in general will decrease* as the discount rates applied to future expected cash flows are increased. This may have a wealth effect because a decrease in the value of households' assets may increase the savings rate and decrease consumption.
- Both consumers and businesses may decrease their expenditures because their *expectations for future economic growth decrease*.
- The increase in interest rates may attract foreign investment in debt securities, leading to an *appreciation of the domestic currency relative to foreign currencies*. An

appreciation of the domestic currency increases the foreign currency prices of exports and can reduce demand for the country's export goods.

Taken together, these effects act to decrease aggregate demand and put downward pressure on the price level. A decrease in the policy rate would affect the price level through the same channels, but in the opposite direction.

Monetary Policy Relation With Economic Growth, Inflation, Interest, and Exchange Rates

If money neutrality holds, changes in monetary policy and the policy rate will have no effect on real output. In the short run, however, changes in monetary policy can affect real economic growth as well as interest rates, inflation, and foreign exchange rates. The effects of a change to a more expansionary monetary policy may include any or all of the following:

- The central bank buys securities, which increases bank reserves.
- The interbank lending rate decreases as banks are more willing to lend each other reserves.
- Other short-term rates decrease as the increase in the supply of loanable funds decreases the equilibrium rate for loans.
- Longer-term interest rates also decrease.
- The decrease in real interest rates causes the currency to depreciate in the foreign exchange market.
- The decrease in long-term interest rates increases business investments in plant and equipment.
- Lower interest rates cause consumers to increase their purchases of houses, autos, and durable goods.
- Depreciation of the currency increases foreign demand for domestic goods.
- The increases in consumption, investment, and net exports all increase aggregate demand.
- The increase in aggregate demand increases inflation, employment, and real GDP.

MODULE QUIZ 15.1

1. A central bank's policy goals *least likely* include:
   A. price stability.
   B. minimizing long-term interest rates.
   C. maximizing the sustainable growth rate of the economy.
2. A country that targets a stable exchange rate with another country's currency *least likely*:
   A. accepts the inflation rate of the other country.
   B. will sell its currency if its foreign exchange value rises.
   C. must also match the money supply growth rate of the other country.
3. If a country's inflation rate is below the central bank's target rate, the central bank is *most likely* to:
   A. sell government securities.
   B. increase the reserve requirement.

C. decrease the overnight lending rate.

4. A central bank conducts monetary policy primarily by altering the:
   A. policy rate.
   B. inflation rate.
   C. long-term interest rate.
5. Purchases of securities in the open market by the monetary authorities are *least likely* to increase:
   A. excess reserves.
   B. cash in investor accounts.
   C. the interbank lending rate.
6. An increase in the policy rate will *most likely* lead to an increase in:
   A. business investments in fixed assets.
   B. consumer spending on durable goods.
   C. the foreign exchange value of the domestic currency.

MODULE 15.2: MONETARY POLICY EFFECTS AND LIMITATIONS

Video covering this content is available online.

LOS 15.c: Describe qualities of effective central banks; contrast their use of inflation, interest rate, and exchange rate targeting in expansionary or contractionary monetary policy; and describe the limitations of monetary policy.

For a central bank to succeed in its inflation targeting policies, it should have three essential qualities:

1. *Independence*. For a central bank to be effective in achieving its goals, it should be free from political interference. Reducing the money supply to reduce inflation can also be expected to decrease economic growth and employment. The political party in power has an incentive to boost economic activity and reduce unemployment before elections. For this reason, politicians may interfere with the central bank's activities, compromising its ability to manage inflation. Independence should be thought of in relative terms (degrees of independence) rather than absolute terms. Even in the case of relatively independent central banks, the heads of the banks may be appointed by politicians.

   Independence can be evaluated based on both **operational independence** and **target independence**. Operational independence means that the central bank is allowed to independently determine the policy rate. Target independence means the central bank also defines how inflation is computed, sets the target inflation level, and determines the horizon over which the target is to be achieved. The ECB has both target and operational independence, while most other central banks have only operational independence.
2. *Credibility*. To be effective, central banks should follow through on their stated intentions. If a government with large debts, instead of a central bank, sets an inflation target, the target would not be credible because the government has an incentive to allow inflation to exceed the target level. On the other hand, a credible central bank's targets can become self-fulfilling prophecies. If the market believes

that a central bank is serious about achieving a target inflation rate of 3%, wages and other nominal contracts will be based on 3% inflation, and actual inflation will then be close to that level.

3. *Transparency*. Transparency on the part of central banks aids their credibility. Transparency means that central banks periodically disclose the state of the economic environment by issuing **inflation reports**. Transparent central banks periodically report their views on the economic indicators and other factors they consider in their interest rate setting policy. When a central bank makes clear the economic indicators it uses in establishing monetary policy and how they will be used, it not only gains credibility, but also makes policy changes easier to anticipate and implement.

Inflation, Interest Rate, and Exchange Rate Targeting by Central Banks

Central banks have used various economic variables and indicators over the years to make monetary policy decisions. In the past, some have used **interest rate targeting**, increasing the money supply when specific interest rates rose above the target band and decreasing the money supply (or the rate of money supply growth) when interest rates fell below the target band. Currently, **inflation targeting** is the most widely used tool for making monetary policy decisions, and it is the method required by law in some countries. Central banks that currently use inflation targeting include the United Kingdom, Brazil, Canada, Australia, Mexico, and the ECB.

The **neutral interest rate** of an economy is the growth rate of the money supply that neither increases nor decreases the economic growth rate:

$$\text{neutral interest rate} = \text{real trend rate of economic growth} + \text{inflation target}$$

When the policy rate is above the neutral rate, monetary policy is said to be contractionary; when the policy rate is below the neutral rate, monetary policy is said to be expansionary.

The most common inflation rate target is 2%, with a permitted deviation of ±1% so the target band is 1% to 3%. The reason the inflation target is not 0% is that variations around that rate would allow for negative inflation (i.e., deflation), which is considered disruptive to the smooth functioning of an economy. Central banks are not necessarily targeting current inflation, which is the result of prior policy and events, but inflation in the range of two years in the future.

Some countries, especially developing countries, use **exchange rate targeting**. That is, they target a foreign exchange rate between their currency and another (often the U.S. dollar), rather than targeting inflation. As an example, consider a country that has targeted an exchange rate for its currency versus the U.S. dollar. If the foreign exchange value of the domestic currency falls relative to the U.S. dollar, the monetary authority must use foreign reserves to purchase their domestic currency (which will reduce money supply growth and increase interest rates) to reach the target exchange rate. Conversely, an increase in the foreign exchange value of the domestic currency above the target rate will require sale of the domestic currency in currency markets to reduce

its value (increasing the domestic money supply and decreasing interest rates) to move toward the target exchange rate. One result of exchange rate targeting may be greater volatility of the money supply because domestic monetary policy must adapt to the necessity of maintaining a stable foreign exchange rate.

Over the short term, the targeting country can purchase or sell its currency in the foreign exchange markets to influence the exchange rate. There are limits, however, on how much influence currency purchases or sales can have on exchange rates over time. For example, a country may run out of foreign reserves with which to purchase its currency when the exchange value of its currency is still below the target exchange rate.

The net effect of exchange rate targeting is that the targeting country will have the same inflation rate as the country with the targeted currency, and the targeting country will need to follow monetary policy and accept interest rates that are consistent with this goal regardless of domestic economic circumstances.

Limitations of Monetary Policy

This transmission mechanism for monetary policy previously described does not always produce the intended results. In particular, long-term rates may not rise and fall with short-term rates because of the effect of monetary policy changes on expected inflation.

If individuals and businesses believe that a decrease in the money supply intended to reduce inflation will be successful, they will expect lower future inflation rates. Because long-term bond yields include a premium for expected inflation, long-term rates could fall (tending to increase economic growth), even while the central bank has increased short-term rates to slow economic activity. Conversely, increasing the money supply to stimulate economic activity could lead to an increase in expected inflation rates and long-term bond yields, even as short-term rates fall.

From a different perspective, monetary tightening may be viewed as too extreme—increasing the probability of a recession, making long-term bonds more attractive, and reducing long-term interest rates. If money supply growth is seen as inflationary, higher expected future asset prices will make long-term bonds relatively less attractive and will increase long-term interest rates. Bond market participants that act in this way have been called **bond market vigilantes**. When the central bank's policy is credible and investors believe that the inflation target rate will be maintained over time, this effect on long-term rates will be small.

Another situation in which the transmission mechanism may not perform as expected is if demand for money becomes very elastic and individuals willingly hold more money, even without a decrease in short-term rates. Such a situation is called a **liquidity trap**. Increasing the growth of the money supply will not decrease short-term rates under these conditions because individuals hold the money in cash balances instead of investing in interest-bearing securities. If an economy is experiencing deflation even though monetary policy has been expansionary, liquidity trap conditions may be present.

Compared to inflation, deflation is more difficult for central banks to reverse. In a deflationary environment, monetary policy needs to be expansionary. However, the central bank is limited to reducing the nominal policy rate to zero. Once it reaches zero, the central bank has a limited ability to further stimulate the economy.

Another reason standard tools for increasing the money supply might not increase economic activity is that even with increasing excess reserves, banks may not be willing to lend. When what has become known as the *credit bubble* collapsed in 2008, banks around the world lost equity capital and desired to rebuild it. For this reason, they decreased their lending even as money supplies were increased, and short-term rates fell. With short-term rates near zero, economic growth still poor, and a threat of deflation, central banks began a policy termed **quantitative easing**.

In the United Kingdom, quantitative easing entailed large purchases of British government bonds in the maturity range of three to five years. The intent was to reduce interest rates to encourage borrowing and to generate excess reserves in the banking system to encourage lending. Uncertainty about the economy's future caused banks to behave quite conservatively and willingly hold more excess reserves, rather than make loans.

In the United States, billions of dollars were made available for the Fed to buy assets other than short-term Treasury securities. Large amounts of mortgage securities were purchased from banks to encourage bank lending and to reduce mortgage rates in an attempt to revive the housing market, which had collapsed. When this program did not have the desired effect, a second round of quantitative easing (QE2) was initiated. The Fed purchased long-term Treasury bonds in large quantities (hundreds of billions of dollars) with the goal of bringing down longer-term interest rates and generating excess reserves to increase lending and economic growth. The Fed has also purchased securities with credit risk as part of its quantitative easing, improving banks' balance sheets but perhaps shifting risk from the private sector to the public sector.

Monetary Policy in Developing Economies

Developing countries face problems in successfully implementing monetary policy. Without a liquid market in their government debt, interest rate information may be distorted and open market operations difficult to implement. In a very rapidly developing economy, it may be quite difficult to determine the neutral rate of interest for policy purposes. Rapid financial innovation may change the demand to hold monetary aggregates. Central banks may lack credibility because of past failure(s) to maintain inflation rates in a target band and might not be given independence by the political authority.

LOS 15.d: Explain the interaction of monetary and fiscal policy.

Monetary policy and fiscal policy may each be either expansionary or contractionary, so there are four possible scenarios:

1. **Expansionary fiscal and monetary policy.** In this case, the impact will be highly expansionary, taken together. Interest rates will usually be lower (due to monetary

policy), and the private and public sectors will both expand.

2. **Contractionary fiscal and monetary policy.** In this case, aggregate demand and GDP would be lower, and interest rates would be higher due to tight monetary policy. Both the private and public sectors would contract.
3. **Expansionary fiscal policy and contractionary monetary policy.** In this case, aggregate demand will likely be higher (due to fiscal policy), while interest rates will be higher (due to increased government borrowing and tight monetary policy). Government spending as a proportion of GDP will increase.
4. **Contractionary fiscal policy and expansionary monetary policy.** In this case, interest rates will fall from decreased government borrowing and from the expansion of the money supply, increasing both private consumption and output. Government spending as a proportion of GDP will decrease due to contractionary fiscal policy. The private sector would grow as a result of lower interest rates.

Not surprisingly, the fiscal multipliers for different types of fiscal stimulus differ, and the effects of expansionary fiscal policy are greater when it is combined with expansionary monetary policy. The fiscal multiplier for direct government spending increases has been much higher than the fiscal multiplier for increases in transfers to individuals or tax reductions for workers. Transfer payments to the poor have the greatest relative impact, followed by tax cuts for workers, and broader-based transfers to individuals (not targeted). For all types of fiscal stimulus, the impact is greater when the fiscal actions are combined with expansionary monetary policy. This may reflect the impact of greater inflation, falling real interest rates, and the resulting increase in business investments.

MODULE QUIZ 15.2

1. Qualities of effective central banks include:
   A. credibility and verifiability.
   B. comparability and relevance.
   C. independence and transparency.

2. Monetary policy is likely to be *least* responsive to domestic economic conditions if policymakers employ:
   A. inflation targeting.
   B. interest rate targeting.
   C. exchange rate targeting.

3. Monetary policy is *most likely* to fail to achieve its objectives when the economy is:
   A. growing rapidly.
   B. experiencing deflation.
   C. experiencing disinflation.

KEY CONCEPTS

LOS 15.a

Central bank roles include supplying currency, acting as a banker to the government and to other banks, regulating and supervising the payments system, acting as a lender

of last resort, holding the nation's gold and foreign currency reserves, and conducting monetary policy.

Central banks have the objective of controlling inflation, and some have additional goals of maintaining currency stability, full employment, positive sustainable economic growth, or moderate interest rates.

LOS 15.b

Policy tools available to central banks include the policy rate, reserve requirements, and open market operations. The policy rate is called the discount rate in the United States, the refinancing rate by the ECB, and the two-week repo rate in the United Kingdom.

Decreasing the policy rate, decreasing reserve requirements, and making open market purchases of securities are all expansionary. Increasing the policy rate, increasing reserve requirements, and making open market sales of securities are all contractionary.

The transmission mechanism for changes in the central bank's policy rate through to prices and inflation includes one or more of the following:

- Short-term bank lending rates
- Asset prices
- Expectations for economic activity and future policy rate changes
- Exchange rates with foreign currencies

A contractionary monetary policy (increase in policy rate) will tend to decrease economic growth, increase market interest rates, decrease inflation, and lead to appreciation of the domestic currency in foreign exchange markets. An expansionary monetary policy (decrease in policy rate) will have opposite effects, tending to increase economic growth, decrease market interest rates, increase inflation, and reduce the value of the currency in foreign exchange markets.

LOS 15.c

Effective central banks exhibit independence, credibility, and transparency.

- *Independence*. The central bank is free from political interference.
- *Credibility*. The central bank follows through on its stated policy intentions.
- *Transparency*. The central bank makes it clear what economic indicators it uses and reports on the state of those indicators.

Most central banks set target inflation rates, typically 2%–3%, rather than targeting interest rates, as was once common. When inflation is expected to rise above (fall below) the target band, the money supply is decreased (increased) to reduce (increase) economic activity.

Developing economies sometimes target a stable exchange rate for their currency relative to that of a developed economy, selling their currency when its value rises above the target rate and buying their currency with foreign reserves when the rate falls below the target. The developing country must follow a monetary policy that

supports the target exchange rate and essentially commits to having the same inflation rate as the developed country.

Reasons that monetary policy may not work as intended:

- Monetary policy changes may affect inflation expectations to such an extent that long-term interest rates move opposite to short-term interest rates.
- Individuals may be willing to hold greater cash balances without a change in short-term rates (liquidity trap).
- Banks may be unwilling to lend greater amounts, even when they have increased excess reserves.
- Short-term rates cannot be reduced below zero.
- Developing economies face unique challenges in using monetary policy due to undeveloped financial markets, rapid financial innovation, and lack of credibility of the monetary authority.

LOS 15.d

Interaction of monetary and fiscal policies:

Monetary Policy	Fiscal Policy	Interest Rates	Output	Private Sector Spending	Public Sector Spending
Tight	Tight	Higher	Lower	Lower	Lower
Easy	Easy	Lower	Higher	Higher	Higher
Tight	Easy	Higher	Higher	Lower	Higher
Easy	Tight	Lower	Varies	Higher	Lower

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 15.1

1. **B** Central bank goals often include the following: maximum employment, which is interpreted as the maximum sustainable growth rate of the economy; stable prices; and *moderate* (not minimum) long-term interest rates. (LOS 15.a)

2. **C** The money supply growth rate may need to be adjusted to keep the exchange rate within acceptable bounds, but it is not necessarily the same as that of the other country. The other two statements are true. (LOS 15.a)

3. **C** Decreasing the overnight lending rate would add reserves to the banking system, which would encourage bank lending, expand the money supply, reduce interest rates, and allow GDP growth and the rate of inflation to increase. Selling government securities or increasing the reserve requirement would have the opposite effect, reducing the money supply and decreasing the inflation rate. (LOS 15.b)

4. **A** The primary method by which a central bank conducts monetary policy is through changes in the target short-term rate or policy rate. (LOS 15.b)

5. **C** Open market purchases by monetary authorities decrease the interbank lending rate by increasing excess reserves that banks can lend to one another—and therefore, increasing their willingness to lend. (LOS 15.b)

6. **C** An increase in the policy rate is likely to increase longer-term interest rates, causing decreases in consumption spending on durable goods and business investments in plant and equipment. The increase in rates, however, makes investment in the domestic economy more attractive to foreign investors, increasing demand for the domestic currency and causing the currency to appreciate. (LOS 15.b)

Module Quiz 15.2

1. **C** The three qualities of effective central banks are independence, credibility, and transparency. (LOS 15.c)

2. **C** Exchange rate targeting requires monetary policy to be consistent with the goal of a stable exchange rate with the targeted currency, regardless of domestic economic conditions. (LOS 15.c)

3. **B** Monetary policy has a limited ability to act effectively against deflation because the policy rate cannot be reduced below zero, and demand for money may be highly elastic (liquidity trap). (LOS 15.c)

READING 16

INTRODUCTION TO GEOPOLITICS

MODULE 16.1: GEOPOLITICS

Video covering this content is available online.

LOS 16.a: Describe geopolitics from a cooperation versus competition perspective.

Geopolitics refers to interactions among nations, including the actions of **state actors** (national governments) and **nonstate actors** (corporations, nongovernment organizations, and individuals).

Geopolitics also refers to the study of how geography affects interactions among nations and their citizens. For example, firms located in coastal countries naturally tend to be the dominant participants in international shipping.

One way to examine geopolitics is through analysis of the extent to which individual countries cooperate with one another. Potential areas for cooperation include diplomatic and military matters, and economic and cultural interactions. In terms of economics, areas of cooperation include freedom of movement across borders for goods, services, and capital; agreements to harmonize tariffs; international standardization of rules; and transfers of information and technology.

While a country that engages with other countries on these matters may be considered cooperative and one that does not may be considered noncooperative, the extent of cooperation actually varies along a spectrum. A country might be more cooperative on some issues and less cooperative on others, and its degree of cooperation can change over time or with the outcomes of the country's domestic politics. A country's current decision makers and the length of its political cycle are factors to consider when analyzing geopolitics.

A country will typically cooperate with other countries when doing so advances its national interests. For example, a country may cooperate with its neighbors in a military alliance if doing so will further its interests in protecting its citizens from foreign invaders.

We can analyze a country's national interests as a hierarchy, with its top priorities being those that ensure its survival. A country's **geophysical resource endowment** may influence its priorities. For example, a country that has mineral resources but lacks arable land needs to trade minerals for food; therefore, it has an interest in cooperating with other countries to keep international trade lanes open.

Nonstate actors often have interests in cooperating across borders. Individuals and firms seek to direct their resources to their highest-valued uses, and some of those uses may be in other countries. To facilitate the flow of resources, state and nonstate actors may cooperate on **standardization** of regulations and processes. One key example of standardization among countries is International Financial Reporting Standards for firms presenting their accounting data to the public, which we will examine in the Financial Statement Analysis topic area.

Cultural factors, such as historical emigration patterns or a shared language, can be another influence on a country's level of cooperation. Among these cultural factors are a country's formal and informal **institutions**, such as laws, public and private organizations, or distinct customs and habits. Strong and stable institutions can make cooperation easier for state and nonstate actors. For example, countries that produce and export large amounts of cultural content tend to be those with legal and ethical institutions that protect intellectual property. Cultural exchange is one means through which a country may exercise **soft power**, which is the ability to influence other countries without using or threatening force.

LOS 16.b: Describe geopolitics and its relationship with globalization.

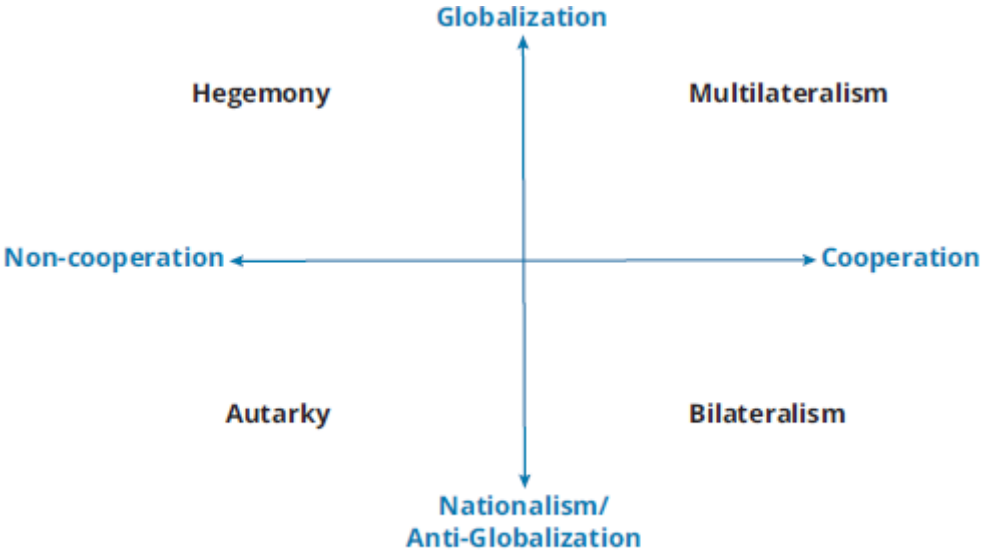
Globalization refers to the long-term trend toward worldwide integration of economic activity and cultures. Data from the World Bank suggest economic openness, as measured by international trade as a percentage of total output, increased steadily from about 25% in the early 1970s to about 60% before the 2008 financial crisis, and has remained near that level since then. We may contrast globalization with **nationalism**—which, in this context, refers to a nation pursuing its own economic interests independently of, or in competition with, the economic interests of other countries.

PROFESSOR'S NOTE
Debate about what the word *nationalism* means is beyond the scope of the CFA curriculum. We use it here only in the sense of opposition or resistance to globalization.

As we did with cooperation versus noncooperation, we can think of countries' actions along a spectrum from globalization to nationalism. In general, countries that are closer to the globalization end of the spectrum are those that more actively import and export goods and services, permit freer movement of capital across borders and exchange of currencies, and are more open to cultural interaction.

In Figure 16.1, we draw each spectrum as an axis. This creates four quadrants, each of which we can associate with a type of behavior by countries. While individual countries rarely fit neatly into one of these categories, this gives us a general framework within which we can describe geopolitical actions.

Figure 16.1: Archetypes of Globalization and Cooperation



Source: Reproduced from Level I CFA Curriculum learning module, "Introduction to Geopolitics," with permission from CFA Institute.

Characteristics we may associate with each of these categories are as follows:

- **Autarky** (noncooperation and nationalism) refers to a goal of national self-reliance, including producing most or all necessary goods and services domestically. Autarky is often associated with a state-dominated society in general, with attributes such as government control of industry and media.
- **Hegemony** (noncooperation and globalization) refers to countries that are open to globalization but have the size and scale to influence other countries without necessarily cooperating.
- **Bilateralism** (cooperation and nationalism) refers to cooperation between two countries. A country that engages in bilateralism may have many such relationships with other countries while tending not to involve itself in multicountry arrangements.
- **Multilateralism** (cooperation and globalization) refers to countries that engage extensively in international trade and other forms of cooperation with many other countries. Some countries may exhibit **regionalism**, cooperating multilaterally with nearby countries but less so with the world at large.

Some of the nonstate actors within a country may be more oriented toward globalization than their governments. Businesses may look outside their home country for opportunities to increase profits, reduce costs, and sell to new markets. Investors may seek higher returns or diversification by investing outside their home country. Nonstate actors might buy and sell foreign securities (**portfolio investment flows**) or own physical production capacity in other countries (**foreign direct investment**).

LOS 16.c: Describe functions and objectives of the international organizations that facilitate trade, including the World Bank, the International Monetary Fund,

and the World Trade Organization.

Perhaps the best way to understand the roles of the organizations designed to facilitate trade is to examine their own statements.

The following is according to the **International Monetary Fund (IMF)** (more available at www.IMF.org):

Article I of the Articles of Agreement sets out the IMF's main goals:

- Promoting international monetary cooperation
- Facilitating the expansion and balanced growth of international trade
- Promoting exchange stability
- Assisting in the establishment of a multilateral system of payments
- Making resources available (with adequate safeguards) to members experiencing balance of payments difficulties

The following is according to the **World Bank** (www.worldbank.org):

> The World Bank is a vital source of financial and technical assistance to developing countries around the world. Our mission is to fight poverty with passion and professionalism for lasting results and to help people help themselves and their environment by providing resources, sharing knowledge, building capacity, and forging partnerships in the public and private sectors.
>
> We are not a bank in the common sense; we are made up of two unique development institutions owned by 187 member countries: the International Bank for Reconstruction and Development (IBRD) and the International Development Association (IDA).
>
> Each institution plays a different but collaborative role in advancing the vision of inclusive and sustainable globalization. The IBRD aims to reduce poverty in middle-income and creditworthy poorer countries, while IDA focuses on the world's poorest countries.
>
> […] Together, we provide low-interest loans, interest-free credits, and grants to developing countries for a wide array of purposes that include investments in education, health, public administration, infrastructure, financial and private sector development, agriculture, and environmental and natural resource management.

The following is according to the **World Trade Organization (WTO)** (more available at www.WTO.org):

> The World Trade Organization (WTO) is the only international organization dealing with the global rules of trade between nations. Its main function is to ensure that trade flows as smoothly, predictably, and freely as possible.
>
> […] Trade friction is channeled into the WTO's dispute settlement process where the focus is on interpreting agreements and commitments, and how to ensure that countries' trade policies conform with them. That way, the risk of disputes spilling over into political or military conflict is reduced.
>
> […] At the heart of the system—known as the multilateral trading system—are the WTO's agreements, negotiated and signed by a large majority of the world's trading nations, and ratified in their parliaments. These agreements are the legal ground-rules for international commerce. Essentially, they are contracts, guaranteeing member countries important trade rights. They also bind governments to keep their trade policies within agreed limits to everybody's benefit.

LOS 16.d: Describe geopolitical risk.

Geopolitical risk is the possibility of events that interrupt peaceful international relations. We can classify geopolitical risk into three types:

1. **Event risk** refers to events about which we know the timing but not the outcome, such as national elections.
2. **Exogenous risk** refers to unanticipated events, such as outbreaks of war or rebellion.
3. **Thematic risk** refers to known factors that have effects over long periods, such as human migration patterns or cyber risks.

Geopolitical risk affects investment values by increasing or decreasing the risk premium investors require to hold assets in a country or region. To forecast the effect on investments of a geopolitical risk, we need to consider its probability (*likelihood*), the magnitude of its effects on investment outcomes (*impact*), and how quickly investment values would reflect these effects (*velocity*).

We can use our framework of cooperation and globalization to help estimate the **likelihood of geopolitical risk**. Countries that are more cooperative and globalized tend to have less likelihood of some geopolitical risks, such as armed conflict, but may have greater likelihood of other risks, such as the supply chain disruptions that followed the COVID-19 pandemic in 2020 and 2021.

To analyze the **velocity of geopolitical risk**, we can classify risks as high velocity (short term), medium velocity, or low velocity (long term). Exogenous risks often have high-velocity effects on financial markets and investment values. **Black swan risk** is a term for the risk of low-likelihood exogenous events that have substantial short-term effects. Investors with longer time horizons typically do not need to react to these kinds of events, but investors with shorter horizons might find it necessary to react.

Medium-velocity risks can potentially damage specific companies or industries by increasing their costs or disrupting their production processes, while low-velocity risks tend to affect them in the "environmental, social, and governance" realm. Analyzing these kinds of risk is important for investors with long time horizons.

LOS 16.e: Describe tools of geopolitics and their impact on regions and economies.

We can consider **tools of geopolitics**, the means by which (primarily) state actors advance their interests in the world, as falling into three broad categories of national security, economic, and financial.

National security tools may include armed conflict, espionage, or bilateral or multilateral agreements designed to reinforce or prevent armed conflict. We can say a national security tool is *active* if a country is currently using it or *threatened* if a country is not currently using it but appears likely to do so. Armed conflict affects

regions and economies by destroying productive capital and causing migration away from areas of conflict.

Economic tools can be cooperative or noncooperative. Examples of cooperative economic tools include free trade areas, common markets, and economic and monetary unions (each of which we describe in our reading on international trade and capital flows). Examples of noncooperative economic tools include domestic content requirements, voluntary export restraints, and nationalization (i.e., the state taking control) of companies or industries.

Financial tools include foreign investment and the exchange of currencies. We can view countries as using these tools cooperatively if they allow foreign investment and the free exchange of currencies, or noncooperatively when they restrict these activities. **Sanctions**, or restrictions on a specific geopolitical actor's financial interests, are a financial tool that state actors may use alongside national security tools.

LOS 16.f: Describe the impact of geopolitical risk on investments.

Because analyzing geopolitical risks requires effort, time, and resources, investors should consider whether the **impact of geopolitical risk** is likely to be high or low, and focus their analysis on risks that could have a high impact. With regard to those risks, investors should determine whether they are likely to have *discrete impacts* on a company or industry or *broad impacts* on a country, a region, or the world. Business cycles can affect the impact of geopolitical risk, in that these risks may have greater impacts on investment values when an economy is in recession than they would have during an expansion.

Investors can use qualitative or quantitative **scenario analysis** to gauge the potential effects of geopolitical risks on their portfolios. To help identify geopolitical risks over time, investors may identify **signposts**, or data that can signal when the likelihood of an event is increasing or decreasing, such as volatility indicators in financial markets.

MODULE QUIZ 16.1

1. A state actor that is generally cooperative with other countries and primarily nationalist in pursuing its objectives is *most accurately* said to exhibit:
   A. autarky.
   B. hegemony.
   C. bilateralism.
2. Which of the following tools of geopolitics is *best* described as a noncooperative economic tool?
   A. Voluntary export restraints.
   B. Regional free trade agreements.
   C. Restrictions on conversion of currencies.
3. When investing for a long time horizon, a portfolio manager should *most likely* devote resources to analyzing:
   A. event risks.
   B. thematic risks.

C. exogenous risks.

4. Which international organization is primarily concerned with providing economic assistance to developing countries?
   A. World Bank.
   B. World Trade Organization.
   C. International Monetary Fund.

KEY CONCEPTS

LOS 16.a

Geopolitics refers to interactions among nations. On various issues ranging from diplomacy and military force to economic or cultural openness, countries lie along a spectrum from cooperative to noncooperative.

LOS 16.b

Globalization refers to integration of economic activity and cultures among countries, and can be contrasted with nationalism, which refers to a country pursuing its own interests independently of other countries. Analysts should view geopolitical actions as being on a spectrum from nationalism to globalization.

We may describe geopolitics and its relationship with globalization using the following four broad categories: autarky (noncooperation and nationalism), hegemony (noncooperation and globalization), bilateralism (cooperation and nationalism), and multilateralism (cooperation and globalization).

LOS 16.c

The International Monetary Fund facilitates trade by promoting international monetary cooperation and exchange rate stability, assists in setting up international payments systems, and makes resources available to member countries with balance of payments problems.

The World Bank provides low-interest loans, interest-free credits, and grants to developing countries for many specific purposes. It also provides resources and knowledge and helps form private/public partnerships with the overall goal of fighting poverty.

The World Trade Organization has the goal of ensuring that trade flows freely and works smoothly. Its main focus is on instituting, interpreting, and enforcing numerous multilateral trade agreements that detail global trade policies for a large majority of the world's trading nations.

LOS 16.d

Categories of geopolitical risk are event risk (when the timing is known), exogenous risk (unanticipated events), and thematic risk (known factors that have long-term effects).

LOS 16.e

Tools of geopolitics include national security tools, economic tools, and financial tools.

National security tools may include armed conflict, espionage, or bilateral or multilateral national security agreements.

Cooperative economic tools include free trade areas, common markets, and economic and monetary unions. Noncooperative economic tools include domestic content requirements, voluntary export restraints, and nationalization.

Financial tools include foreign investment, exchange of currencies, and sanctions.

LOS 16.f

Investors should analyze the likelihood of a geopolitical risk, the impact on investment values of an event if it occurs, and the velocity with which it would affect investment values.

ANSWER KEY FOR MODULE QUIZZES

Module Quiz 16.1

1. **C** Bilateralism is characterized by nationalism (as contrasted with globalization) and cooperation. Both autarky and hegemony are characterized by noncooperation. (LOS 16.b)

2. **A** Voluntary export restraints (exporting less of a good than the global market demands) are an example of a noncooperative economic tool. Restrictions on the exchange of currencies are a financial tool. Free trade agreements are a cooperative economic tool. (LOS 16.e)

3. **B** Thematic risks are those that have effects over the long term. Event risks and exogenous risks are more likely to have high-velocity impacts on investment values, but they are less of a focus for investors with longer time horizons. (LOS 16.d)

4. **A** The World Bank provides technical and financial assistance to economically developing countries. The World Trade Organization is primarily concerned with settling disputes among countries concerning international trade. The International Monetary Fund promotes international trade and exchange rate stability and assists member countries that experience balance of payments trouble. (LOS 16.c)

READING 17

INTERNATIONAL TRADE

MODULE 17.1: INTERNATIONAL TRADE

Video covering this content is available online.

LOS 17.a: Describe the benefits and costs of international trade.

Historically, economic models of trade have focused on the gains that result when countries with lower relative cost of (a comparative advantage in) the production of a good specialize in producing that good and export it, importing goods for which other countries have a lower relative cost or production. This increases the total output of goods and the wealth of both countries. Comparative advantage results from differences in technology and resource endowments across countries.

Newer models of trade emphasize gains from economies of scale that reduce costs of export goods, an increased variety of goods produced, decreasing costs and improved quality from additional competition, and more efficient allocation of productive resources.

Free trade can also benefit consumers by reducing the pricing power of domestic monopolies. Countries that produce the same good can also operate in market of monopolistic competition with differentiated products. Consider the global market for automobiles. Several countries export one type of autos and import other types. Consumers benefit from the greater variety offered, as well as the reduced costs from specialization.

While domestic consumers of imported goods and domestic producers of exported goods both gain, international trade imposes costs as well. The most cited costs of free trade are the loss of domestic jobs in an importing industry and increased economic inequality. Consider a country with a comparative advantage in the production of steel that exports to another country with a higher relative cost of steel production. When free trade is permitted, consumers of steel in the importing country gain and workers and producers of steel lose. In the exporting country, increased demand for steel production may increase prices to domestic consumers, while workers and companies in the steel industry will gain, at least in the short run.

In a situation where labor costs are lower in the exporting country, free trade can decrease wages and employment in the domestic economy. Consider textile production in a country with a relatively high cost of labor. For a labor-intensive industry such as some textile production, importing textiles from a lower-cost country will result in job and wage losses in the domestic industry, possibly increasing income inequality.

When we speak of the gains from trade in general, it is based on economic analysis suggesting that, overall, the gains from trade are greater than the losses—especially in the long run. That is, the gainers could, at least in theory, compensate the losers, with net gains shared by both importing and exporting countries. While free trade surely imposes costs on the workers and companies in an industry facing competition from imports in the short run, it is argued that in the long run—after workers receive additional training and find work in other industries—the short-run costs of free trade are mitigated to some significant degree, or even reversed.

LOS 17.b: Compare types of trade restrictions, such as tariffs, quotas, and export subsidies, and their economic implications.

There are many reasons (at least stated reasons) given by governments that impose trade restrictions. Some have support among economists as conceivably valid in terms of increasing a country's welfare, while others get little or no support from economic theory. Some of the reasons for trade restrictions that have support from economists are as follows:

- *Infant industry.* Protection from foreign competition is given to new industries to give them an opportunity to grow to an internationally competitive scale and get up the learning curve in terms of efficient production methods.
- *National security.* Even if imports are cheaper, it may be in the country's best interest to protect producers of goods crucial to the country's national defense so that those goods are available domestically in the event of conflict.

Other arguments for trade restrictions that have little support in theory are as follows:

- *Protecting domestic jobs.* While some jobs are certainly lost, and some groups and regions are negatively affected by free trade, other jobs (in export industries or growing domestic goods and services industries) will be created, and prices for domestic consumers will be less without import restrictions.
- *Protecting domestic industries*. Industry firms often use political influence to get protection from foreign competition—usually to the detriment of consumers, who pay higher prices.

Other arguments include retaliation for foreign trade restrictions, government collection of tariffs (like taxes on imported goods), countering the effects of government subsidies paid to foreign producers, and preventing foreign exports at less than their cost of production (*dumping*).

Types of trade restrictions include the following:

- **Tariffs.** These are taxes on imported goods collected by the government.
- **Quotas.** These are limits on the amount of imports allowed over some period.
- **Export subsidies.** These are government payments to firms that export goods.
- **Minimum domestic content.** This is the requirement that some percentage of product content must be from the domestic country.

- **Voluntary export restraint.** A country voluntarily restricts the amount of a good that can be exported, often in the hope of avoiding tariffs or quotas imposed by its trading partners.

Economic Implications of Trade Restrictions

We will now examine the effects of the primary types of trade restrictions, tariffs, and subsidies.

A **tariff** placed on an imported good increases its domestic price, decreases the quantity imported, and increases the quantity supplied domestically. Domestic producers gain, foreign exporters lose, and the domestic government gains by the amount of the tariff revenues.

A **quota** restricts the quantity of a good imported to the quota amount. Domestic producers gain, and domestic consumers lose from an increase in the domestic price. The right to export a specific quantity to the domestic country is granted by the domestic government, which may or may not charge for the import licenses to foreign countries. If the import licenses are sold, the domestic government gains the revenue.

We illustrate the overall welfare effects of quotas and tariffs for a small country in Figure 17.1. We define a quota that is equivalent to a given tariff as a quota that will result in the same decrease in the quantity of a good imported as the tariff. Defined this way, a tariff and an equivalent quota both increase the domestic price from P_{world}, the price that prevails with no trade restriction, to $P_{\text{protection}}$.

At P_{world}, before any restriction, the domestic quantity supplied is QS_1, and the domestic quantity demanded is QD_1, with the difference equal to the quantity imported, $\text{QD}_1 - \text{QS}_1$. Placing a tariff on imports increases the domestic price to $P_{\text{protection}}$, increases the domestic quantity supplied to QS_2, and decreases the domestic quantity demanded to QD_2. The difference is the new quantity imported. An equivalent quota will have the same effect, decreasing the quantity imported to $\text{QD}_2 - \text{QS}_2$.

The entire shaded area in Figure 17.1 represents the loss of consumer surplus in the domestic economy. The portion with vertical lines, the area to the left of the domestic supply curve between $P_{\text{protection}}$ and P_{world}, represents the gain in the producer surplus of domestic producers. The portion with horizontal lines, the area bounded by $\text{QD}_2 - \text{QS}_2$ and $\text{P}_{\text{protection}} - \text{P}_{\text{world}}$, represents the gain to the domestic government from tariff revenue. The two remaining triangular areas are the deadweight loss from the restriction on free trade.