Key Concepts

LO 21.1

The confidence interval for the regression coefficient, B_1, is calculated as:

$$b_1 \pm (t_c \times s_{b_1}), \text{ or } \left[b_1 - (t_c \times s_{b_1}) < B_1 < b_1 + (t_c \times s_{b_1})\right]$$

LO 21.2

The *p*-value is the smallest level of significance for which the null hypothesis can be rejected. Interpreting the *p*-value offers an alternative approach when testing for statistical significance.

LO 21.3

A *t*-test with n – 2 degrees of freedom is used to conduct hypothesis tests of the estimated regression parameters:

$$t = \frac{b_1 - B_1}{s_{b_1}}$$

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 $\hat{Y}_p = b_0 + b_1 X_p$.

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.

Qualitative independent variables (dummy variables) capture the effect of a binary independent variable:

- Slope coefficient is interpreted as the change in the dependent variable for the case when the dummy variable is one.
- Use one less dummy variable than the number of categories.

LO 21.4

Homoskedasticity refers to the condition of constant variance of the residuals. Heteroskedasticity refers to a violation of this assumption.

The effects of heteroskedasticity are as follows:

- The standard errors are usually unreliable estimates.
- The coefficient estimates (the b_j) aren't affected.
- If the standard errors are too small, but the coefficient estimates themselves are not affected, the *t*-statistics will be too large and the null hypothesis of no statistical significance is rejected too often. The opposite will be true if the standard errors are too large.

LO 21.5

The Gauss-Markov theorem says that if linear regression assumptions are true, then OLS estimators are the best linear unbiased estimators.

LO 21.6

The limitations of the Gauss-Markov theorem are that its conditions may not hold in practice and alternative estimators may be more efficient. Examples of alternative estimators include the weighted least squares estimator and the least absolute deviations estimator.

LO 21.7

In order to interpret *t*-statistics of regression coefficients when a sample size is small, we must assume the assumptions underlying linear regression hold. In practice, it is generally the case that sample sizes are large, meaning that *t*-statistics can be computed using homoskedasticity-only standard errors.

Concept Checkers

1. 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. $H_A: a_2 \neq 0$.
D. $H_A: a_2 > 0$.

Use the following information for Questions 2 through 4.

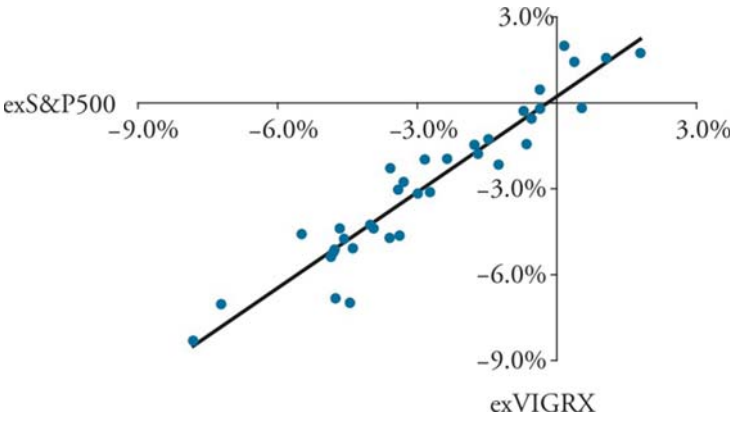
Bill Coldplay is analyzing the performance of the Vanguard Growth Index Fund (VIGRX) over the past three years. The fund employs a passive management investment approach designed to track the performance of the MSCI US Prime Market Growth index, a broadly diversified index of growth stocks of large U.S. companies.

Coldplay estimates a regression using excess monthly returns on VIGRX (exVIGRX) as the dependent variable and excess monthly returns on the S&P 500 index (exS&P) as the independent variable. The data are expressed in decimal terms (e.g., 0.03, not 3%).

$$exVIGRX_t = b_0 + b_1(exS\&P_t) + \varepsilon_t$$

A scatter plot of excess returns for both return series from June 2004 to May 2007 are shown in the following figure.

Analysis of Large Cap Growth Fund



Results from that analysis are presented in the following figures.

Coefficient	*Coefficient Estimate*	*Standard Error*
b_0	0.0023	0.0022
b_1	1.1163	0.0624

Source of Variation	*Sum of Squares*
Explained	0.0228
Residual	0.0024

2. The 90% confidence interval for b_0 is closest to:
   A. –0.0014 to +0.0060.
   B. –0.0006 to +0.0052.
   C. +0.0001 to +0.0045.
   D. –0.0006 to +0.0045.

3. Are the intercept term and the slope coefficient statistically significantly different from zero at the 5% significance level?

	Intercept term significant?	Slope coefficient significant?
A.	Yes	Yes
B.	Yes	No
C.	No	Yes
D.	No	No

4. Coldplay would like to test the following hypothesis: H_0: $B_1 \leq 1$ vs. H_A: $B_1 > 1$ at the 1% significance level. The calculated *t*-statistic and the appropriate conclusion are:

	Calculated *t*-statistic	Appropriate conclusion
A.	1.86	Reject H_0
B.	1.86	Fail to reject H_0
C.	2.44	Reject H_0
D.	2.44	Fail to reject H_0

5. 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. The 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.
   D. not justified, because the appropriate degrees of freedom used to calculate a confidence interval depends on the explained sum of squares.

Concept Checker Answers

1. **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 it is not equal to zero.

2. **A** Note that there are 36 monthly observations from June 2004 to May 2007, so n = 36. The critical two-tailed 10% *t*-value with 34 (n – 2 = 36 – 2 = 34) degrees of freedom is approximately 1.69. Therefore, the 90% confidence interval for b_0 (the intercept term) is 0.0023 +/– (0.0022)(1.69), or –0.0014 to +0.0060.

3. **C** The critical two-tailed 5% *t*-value with 34 degrees of freedom is approximately 2.03. The calculated *t*-statistics for the intercept term and slope coefficient are, respectively, 0.0023 / 0.0022 = 1.05 and 1.1163 / 0.0624 = 17.9. Therefore, the intercept term is not statistically different from zero at the 5% significance level, while the slope coefficient is.

4. **B** Notice that this is a one-tailed test. The critical one-tailed 1% *t*-value with 34 degrees of freedom is approximately 2.44. The calculated *t*-statistic for the slope coefficient is (1.1163 – 1) / 0.0624 = 1.86. Therefore, the slope coefficient is not statistically different from one at the 1% significance level, and Coldplay should fail to reject the null hypothesis.

5. **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.

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. This topic is also covered in:

Linear Regression with Multiple Regressors

Topic 22

Exam Focus

Multiple regression is, in many ways, simply an extension of regression with a single regressor. The coefficient of determination, t-statistics, and standard errors of the coefficients are interpreted in the same fashion. There are some differences, however; namely that the formulas for the coefficients and standard errors are more complicated. The slope coefficients are called partial slope coefficients because they measure the effect of changing one independent variable, assuming the others are held constant. For the exam, understand the implications of omitting relevant independent variables from the model, the adjustment to the coefficient of determination when adding additional variables, and the effect that heteroskedasticity and multicollinearity have on regression results.

Omitted Variable Bias

LO 22.1: Define and interpret omitted variable bias, and describe the methods for addressing this bias.

Omitting relevant factors from an ordinary least squares (OLS) regression can produce misleading or biased results. **Omitted variable bias** is present when two conditions are met: (1) the omitted variable is correlated with the movement of the independent variable in the model, and (2) the omitted variable is a determinant of the dependent variable. When relevant variables are absence from a linear regression model, the results will likely lead to incorrect conclusions as the OLS estimators may not accurately portray the actual data.

Omitted variable bias violates the assumptions of OLS regression when the omitted variable is in fact correlated with current independent (explanatory) variable(s). The reason for this violation is because omitted factors that partially describe the movement of the dependent variable will become part of the regression's error term since they are not properly identified within the model. If the omitted variable is correlated with the regression's slope coefficient, then the error term will also be correlated with the slope coefficient. Recall, that according to the assumptions of linear regression, the independent variable must be uncorrelated with the error term.

The issue of omitted variable bias occurs regardless of the size of the sample and will make OLS estimators inconsistent. The correlation between the omitted variable and the independent variable will determine the size of the bias (i.e., a larger correlation will lead to a larger bias) and the direction of the bias (i.e., whether the correlation is positive or negative). In addition, this bias can also have a dramatic effect on the test statistics used to determine whether the independent variables are statistically significant.

Testing for omitted variable bias would check to see if the two conditions addressed earlier are present. If a bias is found, it can be addressed by dividing data into groups and examining one factor at a time while holding other factors constant. However, in order to understand the full effects of all relevant independent variables on the dependent variable, we need to utilize multiple independent coefficients in our model. Multiple regression analysis is therefore used to eliminate omitted variable bias since it can estimate the effect of one independent variable on the dependent variable while holding all other variables constant.

Multiple Regression Basics

LO 22.2: Distinguish between single and multiple regression.

Multiple regression is regression analysis with more than one independent variable. It is used to quantify the influence of two or more independent variables on a dependent variable. For instance, simple (or univariate) linear regression explains the variation in stock returns in terms of the variation in systematic risk as measured by beta. With multiple regression, stock returns can be regressed against beta and against additional variables, such as firm size, equity, and industry classification, that might influence returns.

The general multiple linear regression model is:

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \ldots + B_kX_{ki} + \varepsilon_i$$

where:

Y_i = ith observation of the dependent variable Y, i = 1, 2, …, n
X_j = independent variables, j = 1, 2, …, k
X_{ji} = ith observation of the jth independent variable
B_0 = intercept term
B_j = slope coefficient for each of the independent variables
ε_i = error term for the ith observation
n = number of observations
k = number of independent variables

LO 22.5: Describe the OLS estimator in a multiple regression.

The multiple regression methodology estimates the intercept and slope coefficients such that the sum of the squared error terms, $\sum_{i=1}^{n}\varepsilon_i^2$, is minimized. The estimators of these coefficients are known as **ordinary least squares** (OLS) **estimators.** The OLS estimators are typically found with statistical software, but can also be computed using calculus or a trial-and-error method. The result of this procedure is the following regression equation:

$$\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki}$$

where the lowercase b_i's indicate an estimate for the corresponding regression coefficient

The residual, e_i, is the difference between the observed value, Y_i, and the predicted value from the regression, $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki})$$

LO 22.3: Interpret the slope coefficient in a multiple regression.

Let's illustrate multiple regression using research by Arnott and Asness (2003).[1] As part of their research, the authors test the hypothesis that future 10-year real earnings growth in the S&P 500 (EG10) can be explained by the trailing dividend payout ratio of the stocks in the index (PR) and the yield curve slope (YCS). YCS is calculated as the difference between the 10-year T-bond yield and the 3-month T-bill yield at the start of the period. All three variables are measured in percent.

Formulating the Multiple Regression Equation

The authors formulate the following regression equation using annual data (46 observations):

$$EG10 = B_0 + B_1PR + B_2YCS + \varepsilon$$

The results of this regression are shown in Figure 1.

Figure 1: Estimates for Regression of EG10 on PR and YCS

	Coefficient	*Standard Error*
Intercept	–11.6%	1.657%
PR	0.25	0.032
YCS	0.14	0.280

Interpreting the Multiple Regression Results

The interpretation of the estimated regression coefficients from a multiple regression is the same as in simple linear regression for the intercept term but significantly different for the slope coefficients:

- The **intercept term** is the value of the dependent variable when the independent variables are all equal to zero.
- Each slope coefficient is the estimated change in the dependent variable for a one-unit change in that independent variable, *holding the other independent variables constant.* That's why the slope coefficients in a multiple regression are sometimes called **partial slope coefficients.**

For example, in the real earnings growth example, we can make these interpretations:

- *Intercept term*: If the dividend payout ratio is zero and the slope of the yield curve is zero, we would expect the subsequent 10-year real earnings growth rate to be –11.6%.
- *PR coefficient*: If the payout ratio increases by 1%, we would expect the subsequent 10-year earnings growth rate to increase by 0.25%, *holding YCS constant.*
- *YCS coefficient*: If the yield curve slope increases by 1%, we would expect the subsequent 10-year earnings growth rate to increase by 0.14%, *holding PR constant.*

1. Arnott, Robert D., and Clifford S. Asness. 2003. "Surprise! Higher Dividends = Higher Earnings Growth." *Financial Analysts Journal*, vol. 59, no. 1 (January/February): 70–87.

Let's discuss the interpretation of the multiple regression slope coefficients in more detail. Suppose we run a regression of the dependent variable *Y* on a single independent variable *X*1 and get the following result:

$$Y = 2.0 + 4.5X1$$

The appropriate interpretation of the estimated slope coefficient is that if *X*1 increases by 1 unit, we would expect *Y* to increase by 4.5 units.

Now suppose we add a second independent variable *X*2 to the regression and get the following result:

$$Y = 1.0 + 2.5X1 + 6.0X2$$

Notice that the estimated slope coefficient for *X*1 changed from 4.5 to 2.5 when we added *X*2 to the regression. We would expect this to happen most of the time when a second variable is added to the regression, unless *X*2 is uncorrelated with *X*1, because if *X*1 increases by 1 unit, then we would expect *X*2 to change as well. The multiple regression equation captures this relationship between *X*1 and *X*2 when predicting *Y*.

Now the interpretation of the estimated slope coefficient for *X*1 is that if *X*1 increases by 1 unit, we would expect *Y* to increase by 2.5 units, *holding X2 constant*.

LO 22.4: Describe homoskedasticity and heteroskedasticity in a multiple regression.

In multiple regression, homoskedasticity and heteroskedasticity are just extensions of their definitions discussed in the previous topic. **Homoskedasticity** refers to the condition that the variance of the error term is constant for all independent variables, *X*, from i = 1 to n: $\text{Var}(\varepsilon_i \mid X_i) = \sigma^2$. **Heteroskedasticity** means that the dispersion of the error terms varies over the sample. It may take the form of conditional heteroskedasticity, which says that the variance is a function of the independent variables.

Measures of Fit

LO 22.6: Calculate and interpret measures of fit in multiple regression.

The **standard error of the regression** (SER) measures the uncertainty about the accuracy of the predicted values of the dependent variable, $\hat{Y}_i = b_0 + b_1 X_i$. Graphically, the relationship is stronger when the actual x,y data points lie closer to the regression line (i.e., the e_i are smaller).

Formally, SER is the standard deviation of the predicted values for the dependent variable about the regression line. Equivalently, it is the standard deviation of the error terms in the regression. SER is sometimes specified as s_e.

Recall that regression minimizes the sum of the squared vertical distances between the predicted value and actual value for each observation (i.e., prediction errors). Also, recall that the sum of the squared prediction errors, $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$, is called the sum of squared residuals, SSR (not to be confused with SER). If the relationship between the variables in the regression is very strong (actual values are close to the line), the prediction errors, and the SSR, will be small. Thus, as shown in the following equations, the standard error of the regression is a function of the SSR:

$$SER = \sqrt{s_e^2} = \sqrt{\frac{SSR}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}\left[Y_i - (b_0 + b_i X_i)\right]^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}}$$

where:

n = number of observations

k = number of independent variables

$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ = SSR = the sum of squared residuals

$\hat{Y}_i = b_0 + b_i X_i$ = a point on the regression line corresponding to a value of X_i. It is the expected (predicted) value of Y, given the estimated relation between X and Y.

Similar to the standard deviation for a single variable, SER measures the degree of variability of the actual Y-values relative to the estimated Y-values. The SER gauges the "fit" of the regression line. *The smaller the standard error, the better the fit.*

Coefficient of Determination, R^2

The multiple coefficient of determination, R^2, can be used to test the overall effectiveness of the entire set of independent variables in explaining the dependent variable. Its interpretation is similar to that for simple linear regression: the percentage of variation in the dependent variable that is *collectively* explained by all of the independent variables. For example, an R^2 of 0.63 indicates that the model, as a whole, explains 63% of the variation in the dependent variable.

R^2 is calculated the same way as in simple linear regression.

$$R^2 = \frac{\text{total variation} - \text{unexplained variation}}{\text{total variation}} = \frac{TSS - SSR}{TSS} = \frac{\text{explained variation}}{\text{total variation}} = \frac{ESS}{TSS}$$

Adjusted R^2

Unfortunately, R^2 by itself *may not be a reliable measure of the explanatory power of the multiple regression model*. This is because R^2 almost always increases as independent variables are added to the model, even if the marginal contribution of the new variables is not statistically significant. Consequently, a relatively high R^2 may reflect the impact of a large set of independent variables rather than how well the set explains the dependent variable. This problem is often referred to as overestimating the regression.

To overcome the problem of overestimating the impact of additional variables on the explanatory power of a regression model, many researchers recommend adjusting R^2 for the number of independent variables. The *adjusted* R^2 value is expressed as:

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

R_a^2 is less than or equal to R^2. So while adding a new independent variable to the model will increase R^2, it may either *increase or decrease* the R_a^2. If the new variable has only a small effect on R^2, the value of R_a^2 may decrease. In addition, R_a^2 may be less than zero if the R^2 is low enough.

Example: Calculating R^2 and adjusted R^2

An analyst runs a regression of monthly value-stock returns on five independent variables over 60 months. The total sum of squares for the regression is 460, and the sum of squared errors is 170. **Calculate** the R^2 and adjusted R^2.

Answer:

$$R^2 = \frac{460 - 170}{460} = 0.630 = 63.0\%$$

$$R_a^2 = 1 - \left[\left(\frac{60-1}{60-5-1}\right) \times (1 - 0.63)\right] = 0.596 = 59.6\%$$

The R^2 of 63% suggests that the five independent variables together explain 63% of the variation in monthly value-stock returns.

Example: Interpreting adjusted R^2

Suppose the analyst now adds four more independent variables to the regression, and the R^2 increases to 65.0%. **Identify** which model the analyst would most likely prefer.

Answer:

With nine independent variables, even though the R^2 has increased from 63% to 65%, the adjusted R^2 has decreased from 59.6% to 58.7%:

$$R_a^2 = 1 - \left[\left(\frac{60-1}{60-9-1}\right) \times (1-0.65)\right] = 0.587 = 58.7\%$$

The analyst would prefer the first model because the adjusted R^2 is higher and the model has five independent variables as opposed to nine.

Assumptions of Multiple Regression

LO 22.7: Explain the assumptions of the multiple linear regression model.

As with simple linear regression, most of the assumptions made with the multiple regression pertain to ε, the model's error term:

- A linear relationship exists between the dependent and independent variables. In other words, the model in LO 22.2 correctly describes the relationship.
- The independent variables are not random, and there is no exact linear relation between any two or more independent variables.
- The expected value of the error term, conditional on the independent variables, is zero [i.e., $E(\varepsilon | X_1, X_2, ...X_k) = 0$].
- The variance of the error terms is constant for all observations [i.e., $E(\varepsilon_i^2) = \sigma_\varepsilon^2$].
- The error term for one observation is not correlated with that of another observation [i.e., $E(\varepsilon_i \varepsilon_j) = 0, j \neq i$].
- The error term is normally distributed.

Multicollinearity

LO 22.8: Explain the concept of imperfect and perfect multicollinearity and their implications.

Multicollinearity refers to the condition when two or more of the independent variables, or linear combinations of the independent variables, in a multiple regression are highly correlated with each other. This condition distorts the standard error of the regression and the coefficient standard errors, leading to problems when conducting *t*-tests for statistical significance of parameters.

The degree of correlation will determine the difference between perfect and imperfect multicollinearity. If one of the independent variables is a perfect linear combination of the other independent variables, then the model is said to exhibit **perfect multicollinearity**. In this case, it will not be possible to find the OLS estimators necessary for the regression results.

An important consideration when performing multiple regression with dummy variables is the choice of the number of dummy variables to include in the model. Whenever we want to distinguish between n classes, we must use $n - 1$ dummy variables. Otherwise, the regression assumption of no exact linear relationship between independent variables would be violated. In general, if every observation is linked to only one class, all dummy variables are included as regressors, and an intercept term exists, then the regression will exhibit perfect multicollinearity. This problem is known as the **dummy variable trap**. As mentioned, this issue can be avoided by excluding one of the dummy variables from the regression equation (i.e., $n - 1$ dummy variables). With this approach, the intercept term will represent the omitted class.

Imperfect multicollinearity arises when two or more independent variables are highly correlated, but less than perfectly correlated. When conducting regression analysis, we need to be cognizant of imperfect multicollinearity since OLS estimators will be computed, but the resulting coefficients may be improperly estimated. In general, when using the term multicollinearity, we are referring to the *imperfect case*, since this regression assumption violation requires detecting and correcting.

Effect of Multicollinearity on Regression Analysis

As a result of multicollinearity, there is a *greater probability that we will incorrectly conclude that a variable is not statistically significant* (e.g., a Type II error). Multicollinearity is likely to be present to some extent in most economic models. The issue is whether the multicollinearity has a significant effect on the regression results.

Detecting Multicollinearity

The most common way to detect multicollinearity is the situation where t-tests indicate that none of the individual coefficients is significantly different than zero, while the R^2 is high. This suggests that the variables together explain much of the variation in the dependent variable, but the individual independent variables do not. The only way this can happen is when the independent variables are highly correlated with each other, so while their common source of variation is explaining the dependent variable, the high degree of correlation also "washes out" the individual effects.

High correlation among independent variables is sometimes suggested as a sign of multicollinearity. In fact, as a general rule of thumb: If the absolute value of the sample correlation between any two independent variables in the regression is greater than 0.7, multicollinearity is a potential problem. However, this only works if there are exactly two independent variables. If there are more than two independent variables, while individual variables may not be highly correlated, linear combinations might be, leading to multicollinearity. High correlation among the independent variables suggests the possibility of multicollinearity, but low correlation among the independent variables *does not necessarily* indicate multicollinearity is *not* present.

Example: Detecting multicollinearity

Bob Watson runs a regression of mutual fund returns on average P/B, average P/E, and average market capitalization, with the following results:

Variable	*Coefficient*	*p-Value*
Average P/B	3.52	0.15
Average P/E	2.78	0.21
Market Cap	4.03	0.11
R^2	89.6%	

Determine whether or not multicollinearity is a problem in this regression.

Answer:

The R^2 is high, which suggests that the three variables as a group do an excellent job of explaining the variation in mutual fund returns. However, none of the independent variables individually is statistically significant to any reasonable degree, since the *p*-values are larger than 10%. This is a classic indication of multicollinearity.

Correcting Multicollinearity

The most common method to correct for multicollinearity is to omit one or more of the correlated independent variables. Unfortunately, it is not always an easy task to identify the variable(s) that are the source of the multicollinearity. There are statistical procedures that may help in this effort, like stepwise regression, which systematically remove variables from the regression until multicollinearity is minimized.

Key Concepts

LO 22.1

Omitted variable bias is present when two conditions are met: (1) the omitted variable is correlated with the movement of the independent variable in the model, and (2) the omitted variable is a determinant of the dependent variable.

LO 22.2

The multiple regression equation specifies a dependent variable as a linear function of two or more independent variables:

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \ldots + B_kX_{ki} + \varepsilon_i$$

The intercept term is the value of the dependent variable when the independent variables are equal to zero. Each slope coefficient is the estimated change in the dependent variable for a one-unit change in that independent variable, holding the other independent variables constant.

LO 22.3

In a multivariate regression, each slope coefficient is interpreted as a partial slope coefficient in that it measures the effect on the dependent variable from a change in the associated independent variable holding other things constant.

LO 22.4

Homoskedasticity means that the variance of error terms is constant for all independent variables, while heteroskedasticity means that the variance of error terms varies over the sample. Heteroskedasticity may take the form of conditional heteroskedasticity, which says that the variance is a function of the independent variables.

LO 22.5

Multiple regression estimates the intercept and slope coefficients such that the sum of the squared error terms is minimized. The estimators of these coefficients are known as ordinary least squares (OLS) estimators. The OLS estimators are typically found with statistical software.

LO 22.6

The standard error of the regression is the standard deviation of the predicted values for the dependent variable about the regression line:

$$SER = \sqrt{\frac{SSR}{n-k-1}}$$

The coefficient of determination, R^2, is the percentage of the variation in Y that is explained by the set of independent variables.

- R^2 increases as the number of independent variables increases—this can be a problem.
- The adjusted R^2 adjusts the R^2 for the number of independent variables.
- $R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$

LO 22.7

Assumptions of multiple regression mostly pertain to the error term, ε_i.

- A linear relationship exists between the dependent and independent variables.
- The independent variables are not random, and there is no exact linear relation between any two or more independent variables.
- The expected value of the error term is zero.
- The variance of the error terms is constant.
- The error for one observation is not correlated with that of another observation.
- The error term is normally distributed.

LO 22.8

Perfect multicollinearity exists when one of the independent variables is a perfect linear combination of the other independent variable. Imperfect multicollinearity arises when two or more independent variables are highly correlated, but less than perfectly correlated.

Concept Checkers

Use the following table for Question 1.

Source	*Sum of Squares (SS)*
Explained	1,025
Residual	925

1. The total sum of squares (TSS) is closest to:
   A. 100.
   B. 1.108.
   C. 1,950.
   D. 0.9024.

Use the following information to answer Questions 2 and 3.

Multiple regression was used to explain stock returns using the following variables:

Dependent variable:

RET = annual stock returns (%)

Independent variables:

MKT = market capitalization = market capitalization / $1.0 million

IND = industry quartile ranking (IND = 4 is the highest ranking)

FORT = Fortune 500 firm, where {FORT = 1 if the stock is that of a Fortune 500 firm, FORT = 0 if not a Fortune 500 stock}

The regression results are presented in the tables below.

	Coefficient	*Standard Error*	*t-Statistic*	*p-Value*
Intercept	0.5220	1.2100	0.430	0.681
Market capitalization	0.0460	0.0150	3.090	0.021
Industry ranking	0.7102	0.2725	2.610	0.040
Fortune 500	0.9000	0.5281	1.700	0.139

2. Based on the results in the table, which of the following most accurately represents the regression equation?
   A. 0.43 + 3.09(MKT) + 2.61(IND) + 1.70(FORT).
   B. 0.681 + 0.021(MKT) + 0.04(IND) + 0.139(FORT).
   C. 0.522 + 0.0460(MKT) + 0.7102(IND) + 0.9(FORT).
   D. 1.21 + 0.015(MKT) + 0.2725(IND) + 0.5281(FORT).

3. The expected amount of the stock return attributable to it being a Fortune 500 stock is closest to:
   A. 0.522.
   B. 0.046.
   C. 0.710.
   D. 0.900.

4. Which of the following situations is not possible from the results of a multiple regression analysis with more than 50 observations?

	R^2	Adjusted R^2
A.	71%	69%
B.	83%	86%
C.	54%	12%
D.	10%	–2%

5. Assumptions underlying a multiple regression are most likely to include:
   A. The expected value of the error term is $0.00 < i < 1.00$.
   B. Linear and non-linear relationships exist between the dependent and independent variables.
   C. The error for one observation is not correlated with that of another observation.
   D. The variance of the error terms is not constant for all observations.

Concept Checker Answers

1. **C** TSS = 1,025 + 925 = 1,950

2. **C** The coefficients column contains the regression parameters.

3. **D** The regression equation is 0.522 + 0.0460(MKT) + 0.7102(IND) + 0.9(FORT). The coefficient on FORT is the amount of the return attributable to the stock of a Fortune 500 firm.

4. **B** Adjusted R^2 must be less than or equal to R^2. Also, if R^2 is low enough and the number of independent variables is large, adjusted R^2 may be negative.

5. **C** Assumptions underlying a multiple regression include: the error for one observation is not correlated with that of another observation; the expected value of the error term is zero; a linear relationship exists between the dependent and independent variables; the variance of the error terms is constant.

The following is a review of the Quantitative Analysis principles designed to address the learning objectives set forth by GARP®. This topic is also covered in:

Hypothesis Tests and Confidence Intervals in Multiple Regression

Topic 23

Exam Focus

This topic addresses methods for dealing with uncertainty in a multiple regression model. Hypothesis tests and confidence intervals for single- and multiple-regression coefficients will be discussed. For the exam, you should know how to use a *t*-test to assess the significance of the individual regression parameters and an *F*-test to assess the effectiveness of the model as a whole in explaining the dependent variable. Also, be able to identify the common model misspecifications. Focus on interpretation of the regression equation and the test statistics. Remember that most of the test and descriptive statistics discussed (e.g., *t*-stat, *F*-stat, and R^2) are provided in the output of statistical software. Hence, application and interpretation of these measurements are more likely than actual computations on the exam.

LO 23.1: Construct, apply, and interpret hypothesis tests and confidence intervals for a single coefficient in a multiple regression.

Hypothesis Testing of Regression Coefficients

As with simple linear regression, the magnitude of the coefficients in a multiple regression tells us nothing about the importance of the independent variable in explaining the dependent variable. Thus, we must conduct hypothesis testing on the estimated slope coefficients to determine if the independent variables make a significant contribution to explaining the variation in the dependent variable.

The *t*-statistic used to test the significance of the individual coefficients in a multiple regression is calculated using the same formula that is used with simple linear regression:

$$t = \frac{b_j - B_j}{s_{b_j}} = \frac{\text{estimated regression coefficient} - \text{hypothesized value}}{\text{coefficient standard error of } b_j}$$

The *t*-statistic has $n - k - 1$ degrees of freedom.

Professor's Note: An easy way to remember the number of degrees of freedom for this test is to recognize that "k" is the number of regression coefficients in the regression, and the "1" is for the intercept term. Therefore, the degrees of freedom is the number of observations minus k minus 1.

Determining Statistical Significance

The most common hypothesis test done on the regression coefficients is to test statistical significance, which means testing the null hypothesis that the coefficient is zero versus the alternative that it is not:

"testing statistical significance" $\Rightarrow H_0: b_j = 0$ versus $H_A: b_j \neq 0$

Example: Testing the statistical significance of a regression coefficient

Consider again, from the previous topic, the hypothesis that future 10-year real earnings growth in the S&P 500 (EG10) can be explained by the trailing dividend payout ratio of the stocks in the index (PR) and the yield curve slope (YCS). **Test** the statistical significance of the independent variable PR in the real earnings growth example at the 10% significance level. Assume that the number of observations is 46. The results of the regression are reproduced in the following figure.

Coefficient and Standard Error Estimates for Regression of EG10 on PR and YCS

	Coefficient	*Standard Error*
Intercept	–11.6%	1.657%
PR	0.25	0.032
YCS	0.14	0.280

Answer:

We are testing the following hypothesis:

$H_0: PR = 0$ versus $H_A: PR \neq 0$

The 10% two-tailed critical *t*-value with 46 – 2 – 1 = 43 degrees of freedom is approximately 1.68. We should reject the null hypothesis if the *t*-statistic is greater than 1.68 or less than –1.68.

The *t*-statistic is:

$$t = \frac{0.25}{0.032} = 7.8$$

Therefore, because the *t*-statistic of 7.8 is greater than the upper critical *t*-value of 1.68, we can reject the null hypothesis and conclude that the PR regression coefficient is statistically significantly different from zero at the 10% significance level.

Interpreting *p*-Values

The *p*-value is the smallest level of significance for which the null hypothesis can be rejected. An alternative method of doing hypothesis testing of the coefficients is to compare the *p*-value to the significance level:

- If the *p*-value is less than significance level, the null hypothesis can be rejected.
- If the *p*-value is greater than the significance level, the null hypothesis cannot be rejected.

Example: Interpreting *p*-values

Given the following regression results, **determine** which regression parameters for the independent variables are statistically significantly different from zero at the 1% significance level, assuming the sample size is 60.

Variable	*Coefficient*	*Standard Error*	*t-Statistic*	*p-Value*
Intercept	0.40	0.40	1.0	0.3215
X1	8.20	2.05	4.0	0.0002
X2	0.40	0.18	2.2	0.0319
X3	–1.80	0.56	–3.2	0.0022

Answer:

The independent variable is statistically significant if the *p*-value is less than 1%, or 0.01. Therefore *X*1 and *X*3 are statistically significantly different from zero.

Figure 1 shows the results of the *t*-tests for each of the regression coefficients of our 10-year earnings growth example, including the *p*-values.

Figure 1: Regression Results for Regression of EG10 on PR and YCS

	Coefficient	*Standard Error*	*t-statistic*	*p-value*
Intercept	–11.6%	1.657%	–7.0	< 0.0001
PR	0.25	0.032	7.8	< 0.0001
YCS	0.14	0.280	0.5	0.62

As we determined in a previous example, we can reject the null hypothesis and conclude that PR is statistically significant. We can also draw the same conclusion for the intercept term because –7.0 is less than the lower critical value of –1.68 (because it is a two-tailed test). However, we fail to reject the null hypothesis for YCS, so we cannot conclude that YCS has a statistically significant effect on the dependent variable, EG10, when PR is also included in the model. The *p*-values tell us exactly the same thing (as they always will): the

intercept term and PR are statistically significant at the 10% level because their *p*-values are less than 0.10, while YCS is not statistically significant because its *p*-value is greater than 0.10.

Other Tests of the Regression Coefficients

You should also be prepared to formulate one- and two-tailed tests in which the null hypothesis is that the coefficient is equal to some value other than zero, or that it is greater than or less than some value.

Example: Testing regression coefficients (two-tail test)

Using the data from Figure 1, **test** the null hypothesis that PR is equal to 0.20 versus the alternative that it is not equal to 0.20 using a 5% significance level.

Answer:

We are testing the following hypothesis:

H_0: PR = 0.20 versus H_A: PR ≠ 0.20

The 5% two-tailed critical *t*-value with 46 – 2 – 1 = 43 degrees of freedom is approximately 2.02. We should reject the null hypothesis if the *t*-statistic is greater than 2.02 or less than –2.02.

The *t*-statistic is:

$$t = \frac{0.25 - 0.20}{0.032} = 1.56$$

Therefore, because the *t*-statistic of 1.56 is between the upper and lower critical *t*-values of –2.02 and 2.02, we cannot reject the null hypothesis and must conclude that the PR regression coefficient is not statistically significantly different from 0.20 at the 5% significance level.

Example: Testing regression coefficients (one-tail test)

Using the data from Figure 1, test the null hypothesis that the intercept term is greater than or equal to –10.0% versus the alternative that it is less than –10.0% using a 1% significance level.

Answer:

We are testing the following hypothesis:

H_0: Intercept $\geq$ –10.0% versus H_A: Intercept $<$ –10.0%

The 1% one-tailed critical *t*-value with 46 – 2 – 1 = 43 degrees of freedom is approximately 2.42. We should reject the null hypothesis if the *t*-statistic is less than –2.42.

The *t*-statistic is:

$$t = \frac{-11.6\% - (-10.0\%)}{1.657\%} = -0.96$$

Therefore, because the *t*-statistic of –0.96 is not less than –2.42, we cannot reject the null hypothesis.

Confidence Intervals for a Regression Coefficient

The confidence interval for a regression coefficient in multiple regression is calculated and interpreted the same way as it is in simple linear regression. For example, a 95% confidence interval is constructed as follows:

$$b_j \pm (t_c \times s_{b_j})$$

or

estimated regression coefficient ± (critical *t*-value)(coefficient standard error)

The critical *t*-value is a two-tailed value with n – k – 1 degrees of freedom and a 5% significance level, where *n* is the number of observations and *k* is the number of independent variables.

Example: Calculating a confidence interval for a regression coefficient

Calculate the 90% confidence interval for the estimated coefficient for the independent variable PR in the real earnings growth example.

Answer:

The critical *t*-value is 1.68, the same as we used in testing the statistical significance at the 10% significance level (which is the same thing as a 90% confidence level). The estimated slope coefficient is 0.25 and the standard error is 0.032. The 90% confidence interval is:

$$0.25 \pm (1.68)(0.032) = 0.25 \pm 0.054 = 0.196 \text{ to } 0.304$$

Professor's Note: Notice that because zero is not contained in the 90% confidence interval, we can conclude that the PR coefficient is statistically significant at the 10% level. Constructing a confidence interval and conducting a t-test with a null hypothesis of "equal to zero" will always result in the same conclusion regarding the statistical significance of the regression coefficient.

Predicting the Dependent Variable

We can use the regression equation to make predictions about the dependent variable *based on forecasted values of the independent variables*. The process is similar to forecasting with simple linear regression, only now we need predicted values for more than one independent variable. The predicted value of dependent variable *Y* is:

$$\hat{Y}_i = b_0 + b_1\hat{X}_{1i} + b_2\hat{X}_{2i} + \ldots + b_k\hat{X}_{ki}$$

where:

$\hat{Y}_i$ = the predicted value of the dependent variable

b_j = the estimated slope coefficient for the *j*th independent variable

$\hat{X}_{ji}$ = the forecast of the *j*th independent variable, j = 1, 2, …, *k*

Professor's Note: The prediction of the dependent variable uses the estimated intercept and all of the estimated slope coefficients, regardless of whether the estimated coefficients are statistically significantly different from zero. For example, suppose you estimate the following regression equation: $\hat{Y} = 6 + 2X_1 + 4X_2$*, and you determine that only the first independent variable* (X_1) *is statistically significant (i.e., you rejected the null that* $B_1 = 0$*). To predict Y given forecasts of* $X_1 = 0.6$ *and* $X_2 = 0.8$*, you would use the complete model:* $\hat{Y} = 6 + (2 \times 0.6) + (4 \times 0.8) = 10.4$*. Alternatively, you could drop X2 and reestimate the model using just X1, but remember that the coefficient on X1 will likely change.*

Example: Calculating a predicted value for the dependent variable

An analyst would like to use the estimated regression equation from the previous example to calculate the predicted 10-year real earnings growth for the S&P 500, assuming the payout ratio of the index is 50%. He observes that the slope of the yield curve is currently 4%.

Answer:

$$\widehat{EG10} = -11.6\% + 0.25(50\%) + 0.14(4\%) = 1.46\%$$

The model predicts a 1.46% real earnings growth rate for the S&P 500, assuming a 50% payout ratio, when the slope of the yield curve is 4%.

Joint Hypothesis Testing

LO 23.2: Construct, apply, and interpret joint hypothesis tests and confidence intervals for multiple coefficients in a multiple regression.

LO 23.3: Interpret the F-statistic.

LO 23.5: Interpret confidence sets for multiple coefficients.

A joint hypothesis tests two or more coefficients at the same time. For example, we could develop a null hypothesis for a linear regression model with three independent variables that sets two of these coefficients equal to zero: H_0: $b_1 = 0$ and $b_2 = 0$ versus the alternative hypothesis that one of them is not equal to zero. That is, if just one of the equalities in this null hypothesis does not hold, we can reject the entire null hypothesis. Using a joint hypothesis test is preferred in certain scenarios since testing coefficients individually leads to a greater chance of rejecting the null hypothesis. For example, instead of comparing one t-statistic to its corresponding critical value in a joint hypothesis test, we are testing two t-statistics. Thus, we have an additional opportunity to reject the null. A robust method for applying joint hypothesis testing, especially when independent variables are correlated, is known as the *F*-statistic.

The *F*-Statistic

An *F*-test assesses how well the set of independent variables, as a group, explains the variation in the dependent variable. That is, the *F*-statistic is used to test whether *at least one* of the independent variables explains a significant portion of the variation of the dependent variable.

For example, if there are four independent variables in the model, the hypotheses are structured as:

$H_0: B_1 = B_2 = B_3 = B_4 = 0$ versus H_A: at least one $B_j \neq 0$

The *F*-statistic, *which is always a one-tailed test*, is calculated as:

$$\frac{ESS/k}{SSR/(n-k-1)}$$

where:
ESS = explained sum of squares
SSR = sum of squared residuals

Professor's Note: The explained sum of squares and the sum of squared residuals are found in an analysis of variance (ANOVA) table. We will analyze an ANOVA table from a multiple regression shortly.

To determine whether at least one of the coefficients is statistically significant, the calculated *F*-statistic is compared with the **one-tailed** critical *F*-value, F_c, at the appropriate level of significance. The degrees of freedom for the numerator and denominator are:

$df_{numerator} = k$

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

where:
n = number of observations
k = number of independent variables

The decision rule for the *F*-test is:

Decision rule: reject H_0 if F (test-statistic) > F_c (critical value)

Rejection of the null hypothesis at a stated level of significance indicates that at least one of the coefficients is significantly different than zero, which is interpreted to mean that at least one of the independent variables in the regression model makes a significant contribution to the explanation of the dependent variable.

Professor's Note: It may have occurred to you that an easier way to test all of the coefficients simultaneously is to just conduct all of the individual t-tests and see how many of them you can reject. This is the wrong approach, however, because if you set the significance level for each t-test at 5%, for example, the significance level from testing them all simultaneously is NOT 5%, but rather some higher percentage. Just remember to use the F-test on the exam if you are asked to test all of the coefficients simultaneously.

Example: Calculating and interpreting the *F*-statistic

An analyst runs a regression of monthly value-stock returns on five independent variables over 60 months. The total sum of squares is 460, and the sum of squared residuals is 170. Test the null hypothesis at the 5% significance level (95% confidence) that all five of the independent variables are equal to zero.

Answer:

The null and alternative hypotheses are:

$$H_0: B_1 = B_2 = B_3 = B_4 = B_5 = 0 \text{ versus } H_A: \text{at least one } B_j \neq 0$$

$$ESS = TSS - SSR = 460 - 170 = 290$$

$$F = \frac{58.0}{3.15} = 18.41$$

The critical *F*-value for 5 and 54 degrees of freedom at a 5% significance level is approximately 2.40. Remember, it's a **one-tailed** test, so we use the 5% *F*-table! Therefore, we can reject the null hypothesis and conclude that at least one of the five independent variables is significantly different than zero.

Professor's Note: When testing the hypothesis that all the regression coefficients are simultaneously equal to zero, the F-test is always a one-tailed test, despite the fact that it looks like it should be a two-tailed test because there is an equal sign in the null hypothesis.

Interpreting Regression Results

Just as in simple linear regression, the variability of the dependent variable or **total sum of squares** (TSS) can be broken down into **explained sum of squares** (ESS) and **sum of squared residuals** (SSR). As shown previously, the coefficient of determination is:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum e_i^2}{\sum(Y_i - \bar{Y})^2}$$

Regression results usually provide R^2 and a host of other measures. However, it is useful to know how to compute R^2 from other parts of the results. Figure 2 is an ANOVA table of the results of a regression of hedge fund returns on lockup period and years of experience of the manager. In the ANOVA table, the value of 90 represents TSS, the ESS equals 84.057, and the SSR is 5.943. Although the output results provide the value $R^2 = 0.934$, it can also be computed using TSS, ESS, and SSR like so:

$$R^2 = \frac{84.057}{90} = 1 - \frac{5.943}{90} = 0.934$$

The **coefficient of multiple correlation** is simply the square root of *R*-squared. In the case of a multiple regression, the coefficient of multiple correlation is always positive.

Figure 2: ANOVA Table

R-squared	0.934					
Adj *R*-squared	0.890					
Standard error	1.407					
Observations	6					
	Degrees of Freedom	*SS*	*MS*	*F*		
Explained	2	84.057	42.029	21.217		
Residual	3	5.943	1.981			
Total	5	90				
Variables	*Coeff*	*Std Error*	*t-stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	–4.4511	3.299	–1.349	0.270	–14.950	6.048
Lockup	2.057	0.337	6.103	0.009	0.984	3.130
Experience	2.008	0.754	2.664	0.076	–0.391	4.407

The results in Figure 2 produce the following equation:

$$\hat{Y}_i = -4.451 + 2.057 \times X_{1i} + 2.008 \times X_{2i}$$

This equation tells us that holding other variables constant, increasing the lockup period will increase the expected return of a hedge fund by 2.057%. Also, holding other variables constant, increasing the manager's experience one year will increase the expected return of a hedge fund by 2.008%. A hedge fund with an inexperienced manager and no lockup period will earn a negative return of –4.451%.

The ANOVA table outputs the standard errors, *t*-statistics, probability values (p-values), and confidence intervals for the estimated coefficients. These can be used in a hypothesis test for each coefficient. For example, for the independent variable experience (b_2), the output indicates that the standard error is $se(b_2) = 0.754$, which yields a *t*-statistic of: 2.008 / 0.754 = 2.664. The critical *t*-value at a 5% level of significance is $t_{0.025} = 3.182$. Thus, a hypothesis stating that the number of years of experience is not related to returns could not be rejected. In other words, the result is to not reject the null hypothesis that $B_2 = 0$. This is also seen with the provided confidence interval. Upper and lower limits of the confidence interval can be found in the ANOVA results.

$$[b_2 - t_{\alpha/2} \times se(b_2)] < B_2 < [b_2 + t_{\alpha/2} \times se(b_2)]$$

$$(2.008 - 3.182 \times 0.754) < B_2 < (2.008 + 3.182 \times 0.754)$$

$$-0.391 < B_2 < 4.407$$

Since the confidence interval contains the value zero, then the null hypothesis: H_0: $B_2 = 0$ cannot be rejected in a two-tailed test at the 5% level of significance. Figure 2 provides a third way of performing a hypothesis test by providing a p-value. The **p-value** indicates the

minimum level of significance at which the two-tailed hypothesis test can be rejected. In this case, the p-value is 0.076 (i.e., 7.6%), which is greater than 5%.

The statistics for b_1 indicate that a null hypothesis can be rejected at a 5% level using a two-tailed test. The *t*-statistic is 6.103, and the confidence interval is 0.984 to 3.13. The p-value of 0.9% is less than 5%.

The statistics in the ANOVA table also allow for the testing of the joint hypothesis that both slope coefficients equal zero.

H_0: $B_1 = B_2 = 0$

H_A: $B_1 \neq 0$ or $B_2 \neq 0$

The test statistic in this case is the ***F*-statistic** where the degrees of freedom are indicated by two numbers: the number of slope coefficients (2) and the sample size minus the number of slope coefficients minus one (6 – 2 – 1 = 3). The *F*-statistic given the hedge fund data can be calculated as follows:

$$F = \frac{ESS/df}{SSR/df} = \frac{84.057/2}{5.943/3} = \frac{42.029}{1.981} = 21.217$$

The critical *F*-statistic at a 5% significance level is $F_{0.05} = 9.55$. Since the value from the regression results is greater than that value: $F = 21.217 > 9.55$, a researcher would reject the null hypothesis: H_0: $B_1 = B_2 = 0$. It should be noted that rejecting the null hypothesis indicates one or both of the coefficients are significant.

Specification Bias

Specification bias refers to how the slope coefficient and other statistics for a given independent variable are usually different in a simple regression when compared to those of the same variable when included in a multiple regression. To illustrate this point, the following three OLS results correspond to a two-variable regression using only the indicated independent variable and the results for a three-variable:

$$\hat{Y}_i = 1 + 2 \times (\text{lockup})_i$$
$$t = 3.742$$

$$\hat{Y}_i = 11.714 + 1.714 \times (\text{experience})_i$$
$$t = 2.386$$

$$\hat{Y}_i = -4.451 + 2.057 \times (\text{lockup})_i + 2.008 \times (\text{experience})_i$$
$$t = 6.103 \qquad t = 2.664$$

Specification bias is indicated by the extent to which the coefficient for each independent variable is different when compared across equations (e.g., for lockup, the slope is 2 in the two-variable equation, and the slope is 2.057 in the multivariate regression). This is because in the two-variable regression, the slope coefficient includes the effect of the included independent variable in the equation and, to some extent, the indirect effect of the excluded

variable(s). In this case, the bias for the coefficient on the lockup coefficient was not large because the experience variable was not significant as indicated in its two-variable regression ($t = 2.386 < t_{0.025} = 2.78$) and was not significant in the multivariable regression either.

R^2 and Adjusted R^2

LO 23.7: Interpret the R^2 and adjusted R^2 in a multiple regression.

To further analyze the importance of an added variable to a regression, we can compute an adjusted coefficient of determination, or **adjusted R^2**. The reason adjusted R^2 is important is because, mathematically speaking, the coefficient of determination, R^2, must go up if a variable with any explanatory power is added to the regression, even if the marginal contribution of the new variables is not statistically significant. Consequently, a relatively high R^2 may reflect the impact of a large set of independent variables rather than how well the set explains the dependent variable. This problem is often referred to as overestimating the regression.

When computing both the R^2 and the adjusted R^2, there are a few pitfalls to acknowledge, which could lead to invalid conclusions.

1. If adding an additional independent variable to the regression improves the R^2, this variable is not necessary statistically significant.

2. The R^2 measure may be spurious, meaning that the independent variables may show a high R^2; however, they are not the exact cause of the movement in the dependent variable.

3. If the R^2 is high, we cannot assume that we have found all relevant independent variables. Omitted variables may still exist, which would improve the regression results further.

4. The R^2 measure does not provide evidence that the most or least appropriate independent variables have been selected. Many factors go into finding the most robust regression model, including omitted variable analysis, economic theory, and the quality of data being used to generate the model.

Restricted vs. Unrestricted Least Squares Models

A restricted least squares regression imposes a value on one or more coefficients with the goal of analyzing if the restriction is significant. To explain this concept, it is useful to note that there is an implied restriction in each of the two variable regressions:

$$\hat{Y}_i = b_0 + b_{lockup} \times (lockup)_i$$
$$\hat{Y}_i = b_0 + b_{experience} \times (experience)_i$$

In essence, each of the two-variable regressions is a restricted regression where the coefficient on the omitted variable is restricted to zero. To help illustrate the concept, the more elaborate subscripts have been used in these expressions. Using the indicated notation, the first specification that only includes "lockup" is restricting $b_{experience}$ to 0. In the unrestricted

multivariable regression, both b_{lockup} and $b_{experience}$ are allowed to assume the values that minimize the SSR. The R^2 from the restricted regression is called a **restricted R^2** or R_r^2. For comparison, the **unrestricted R^2** from the specification that includes both independent variables is given the notation R_{ur}^2, and both are included in an *F*-statistic that can test if the restriction is significant or not:

$$F = \frac{(R_{ur}^2 - R_r^2)/m}{(1 - R_{ur}^2)/(n - k_{ur} - 1)}$$

The symbol "*m*" refers to the number of restrictions, which in the example discussed would be equal to one. This *F*-stat is known as the **homoskedasticity-only *F*-statistic** since it can only be derived from R^2 when the error terms display homoskedasticity. An alternative formula for computing this *F*-stat is to use the sum of squared residuals in place of the R^2:

$$F = \frac{(SSR_{ur} - SSR_r)/m}{SSR_{ur}/(n - k_{ur} - 1)}$$

In the event that the error terms are not homoskedastic, a hetroskedasticity-robust *F*-stat would be applied. This statistic is used more frequently in practice; however, as the sample size, *n*, increases, these two types of *F*-statistics will converge.

LO 23.4: Interpret tests of a single restriction involving multiple coefficients.

With the *F*-statistic, we constructed a null hypothesis that tested multiple coefficients being equal to zero. However, what if we wanted to test whether one coefficient was equal to another such that: $H_0: b_1 = b_2$? The alternative hypothesis in this scenario would be that the two are not equal to each other. Hypothesis tests of single restrictions involving multiple coefficients requires the use of statistical software packages, but we will examine the methodology of two different approaches.

The first approach is to directly test the restriction stated in the null. Some statistical packages can test this restriction and output a corresponding *F*-stat. This is the easier of the two methods; however, a second method will need to be applied if your statistical package cannot directly test the restriction.

The second approach transforms the regression and uses the null hypothesis as an assumption to simplify the regression model. For example, in a regression with two independent variables: $Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + \varepsilon_i$, we can add and subtract B_2X_{1i} to ultimately transform the regression to: $B_0 + (B_1 - B_2)X_{1i} + B_2(X_{1i} + X_{2i}) + \varepsilon_i$. One of the coefficients will drop out in this equation when assuming that the null hypothesis of $B_1 = B_2$ is valid. We can remove the second term from our regression equation so that: $B_0 + B_2(X_{1i} + X_{2i}) + \varepsilon_i$. We observe that the null hypothesis test changes from a single restriction involving multiple coefficients to a single restriction on just one coefficient.

Professor's Note: Remember that this process is typically done with statistical software packages, so on the exam, you would simply be asked to describe and/or interpret these tests.

Model Misspecification

LO 23.6: Identify examples of omitted variable bias in multiple regressions.

Recall from the previous topic that omitting relevant factors from a regression can produce misleading or biased results. Similar to simple linear regression, omitted variable bias in multiple regressions will result if the following two conditions occur:

- The omitted variable is a determinant of the dependent variable.
- The omitted variable is correlated with *at least* one of the independent variables.

As an example of omitted variable bias, consider a regression in which we're trying to predict monthly returns on portfolios of stocks (R) using three independent variables: portfolio beta (B), the natural log of market capitalization (lnM), and the natural log of the price-to-book ratio ln(PB). The correct specification of this model is as follows:

$$R = b_0 + b_1B + b_2\ln M + b_3\ln PB + \varepsilon$$

Now suppose we did not include lnM in the regression model:

$$R = a_0 + a_1B + a_2\ln PB + \varepsilon$$

If lnM is correlated with any of the remaining independent variables (B or lnPB), then the error term is also correlated with the same independent variables, and the resulting regression coefficients are biased and inconsistent. That means our hypothesis tests and predictions using the model will be unreliable.