2026

CFA®

Exam Prep

SchweserNotes™

Quantitative Methods and Economics

Level II Book 1

KAPLAN SCHWESER

Kaplan Schweser's Path to Success

Level II CFA® Exam

CFA®

Welcome

As the head of Advanced Designations at Kaplan Schweser, I am pleased to have the opportunity to help you prepare for the CFA® exam. Kaplan Schweser has decades of experience in delivering the most effective CFA exam prep products in the market and I know you will find them to be invaluable in your studies.

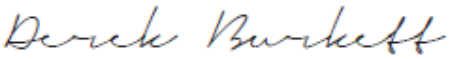
Our products are designed to be an integrated study solution across print and digital media to provide you the best learning experience, whether you are studying with a physical book, online, or on your mobile device.

Our core product, the SchweserNotes™, addresses all of the Topics, Learning Modules, and LOS in the CFA curriculum. Each reading in the SchweserNotes has been broken into smaller, bite-sized modules with Module Quizzes interspersed throughout to help you continually assess your comprehension. Online Topic Quizzes appear at the end of each Topic to help you assess your knowledge of the material before you move on to the next section.

All purchasers of the SchweserNotes receive online access to the Kaplan Schweser online platform (our learning management system or LMS) at www.Schweser.com. In the LMS, you will see a dashboard that tracks your overall progress and performance and also includes an Activity Feed, which provides structure and organization to the tasks required to prepare for the CFA exam. You also have access to the SchweserNotes, Module Quizzes, and Topic Quizzes content. I strongly encourage you to use the dashboard to track your progress and stay motivated.

Again, thank you for trusting Kaplan Schweser with your CFA exam preparation. We're here to help you throughout your journey to become a CFA charterholder.

Regards,



Derek Burkett, CFA, FRM, CAIA

Vice President (Advanced Designations)

Contact us for questions about your study package, upgrading your package, purchasing additional study materials, or for additional information:

888.325.5072 (U.S.) | +1 608.779.8327 (Int'l.)

staff@schweser.com | www.schweser.com/cfa

Book 1: Quantitative Methods and Economics

SchweserNotes™ 2026

Level II CFA®

SCHWESERNOTES™ 2026 LEVEL II CFA® BOOK 1: QUANTITATIVE METHODS AND ECONOMICS

Published in 2025 by Kaplan, Inc.

ISBN: 978-1-0788-5168-8

CONTENTS

ECONOMICS

Learning Outcome Statements (LOS)

1. Multiple Regression

The candidate should be able to:

a. describe the types of investment problems addressed by multiple linear regression and the regression process.
b. formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients.
c. explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.
d. evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.
e. formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests.
f. calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable.
g. describe how model misspecification affects the results of a regression analysis and how to avoid common forms of misspecification.
h. explain the types of heteroskedasticity and how it affects statistical inference.
i. explain serial correlation and how it affects statistical inference.
j. explain multicollinearity and how it affects regression analysis.
k. describe influence analysis and methods of detecting influential data points.
l. formulate and interpret a multiple regression model that includes qualitative independent variables.
m. formulate and interpret a logistic regression model.

2. Time-Series Analysis

The candidate should be able to:

a. calculate and evaluate the predicted trend value for a time series, modeled as either a linear trend or a log-linear trend, given the estimated trend coefficients.
b. describe factors that determine whether a linear or a log-linear trend should be used with a particular time series and evaluate limitations of trend models.
c. explain the requirement for a time series to be covariance stationary and describe the significance of a series that is not stationary.
d. describe the structure of an autoregressive (AR) model of order *p* and calculate one- and two-period-ahead forecasts given the estimated coefficients.
e. explain how autocorrelations of the residuals can be used to test whether the autoregressive model fits the time series.
f. explain mean reversion and calculate a mean-reverting level.
g. contrast in-sample and out-of-sample forecasts and compare the forecasting accuracy of different time-series models based on the root mean squared error criterion.
h. explain the instability of coefficients of time-series models.
i. describe characteristics of random walk processes and contrast them to covariance stationary processes.
j. describe implications of unit roots for time-series analysis, explain when unit roots are likely to occur and how to test for them, and demonstrate how a time series with a unit root can be transformed so it can be analyzed with an AR model.
k. describe the steps of the unit root test for nonstationarity and explain the relation of the test to autoregressive time-series models.
l. explain how to test and correct for seasonality in a time-series model and calculate and interpret a forecasted value using an AR model with a seasonal lag.
m. explain autoregressive conditional heteroskedasticity (ARCH) and describe how ARCH models can be applied to predict the variance of a time series.
n. explain how time-series variables should be analyzed for nonstationarity and/or cointegration before use in a linear regression.

o. determine an appropriate time-series model to analyze a given investment problem and justify that choice.

3. Machine Learning

The candidate should be able to:

a. describe supervised machine learning, unsupervised machine learning, and deep learning.
b. describe overfitting and identify methods of addressing it.
c. describe supervised machine learning algorithms—including penalized regression, support vector machine, k-nearest neighbor, classification and regression tree, ensemble learning, and random forest—and determine the problems for which they are best suited.
d. describe unsupervised machine learning algorithms—including principal components analysis, k-means clustering, and hierarchical clustering—and determine the problems for which they are best suited.

4. Big Data Projects

The candidate should be able to:

a. identify and explain steps in a data analysis project.
b. describe objectives, steps, and examples of preparing and wrangling data.
c. evaluate the fit of a machine learning algorithm.
d. describe objectives, methods, and examples of data exploration.
e. describe methods for extracting, selecting and engineering features from textual data.
f. describe objectives, steps, and techniques in model training.
g. describe preparing, wrangling, and exploring text-based data for financial forecasting.

5. Currency Exchange Rates: Understanding Equilibrium Value

The candidate should be able to:

a. calculate and interpret the bid–offer spread on a spot or forward currency quotation and describe the factors that affect the bid–offer spread.
b. identify a triangular arbitrage opportunity and calculate its profit, given the bid–offer quotations for three currencies.
c. explain spot and forward rates and calculate the forward premium/discount for a given currency.
d. calculate the mark-to-market value of a forward contract.
e. explain international parity conditions (covered and uncovered interest rate parity, forward rate parity, purchasing power parity, and the international Fisher effect).
f. describe relations among the international parity conditions.
g. evaluate the use of the current spot rate, the forward rate, purchasing power parity, and uncovered interest parity to forecast future spot exchange rates.
h. explain approaches to assessing the long-run fair value of an exchange rate.
i. describe the carry trade and its relation to uncovered interest rate parity and calculate the profit from a carry trade.
j. explain how flows in the balance of payment accounts affect currency exchange rates.
k. explain the potential effects of monetary and fiscal policy on exchange rates.
l. describe objectives of central bank or government intervention and capital controls and describe the effectiveness of intervention and capital controls.
m. describe warning signs of a currency crisis.

6. Economic Growth

The candidate should be able to:

a. compare factors favoring and limiting economic growth in developed and developing economies.
b. describe the relation between the long-run rate of stock market appreciation and the sustainable growth rate of the economy.
c. explain why potential GDP and its growth rate matter for equity and fixed income investors.
d. contrast capital deepening investment and technological progress and explain how each affects economic growth and labor productivity.
e. demonstrate forecasting potential GDP based on growth accounting relations.
f. explain how natural resources affect economic growth and evaluate the argument that limited availability of natural resources constrains economic growth.

g. explain how demographics, immigration, and labor force participation affect the rate and sustainability of economic growth.
h. explain how investment in physical capital, human capital, and technological development affects economic growth.
i. compare classical growth theory, neoclassical growth theory, and endogenous growth theory.
j. explain and evaluate convergence hypotheses.
k. describe the economic rationale for governments to provide incentives to private investment in technology and knowledge.
l. describe the expected impact of removing trade barriers on capital investment and profits, employment and wages, and growth in the economies involved.

WELCOME TO THE 2026 LEVEL II SCHWESERNOTES™

Thank you for trusting Kaplan Schweser to help you reach your goals. We are pleased that you have chosen us to assist you in preparing for the Level II CFA Exam. In this introduction, I want to explain the resources included with these SchweserNotes, suggest how you can best use Schweser materials to prepare, and direct you towards other educational resources you will find helpful as you study for the exam.

Besides the SchweserNotes themselves, there are many educational resources available at Schweser.com. Log in using the individual username and password that you received when you purchased your SchweserNotes.

SchweserNotes™

These notes consist of five volumes that include complete coverage of all 10 Topic areas and all the Learning Outcome Statements (LOS). Examples and Module Quizzes (multiple-choice questions) are provided along the way to help you master the material and check your progress. At the end of each major topic area, you can take an online Topic Quiz for that subject. Topic Quiz questions are created to be exam-like in format and difficulty, to help you evaluate how well your study of each topic has prepared you for the actual exam. Finally, there are three online Checkpoint Exams, each covering multiple topics to evaluate your retention.

Practice Questions

Studies have shown that to retain what you learn, it is essential that you quiz yourself often. For this purpose we offer SchweserPro™ QBank, which contains thousands of Level II practice questions and explanations. Questions are available for each module and topic. Build your own quizzes by specifying the topics and the number of questions. SchweserPro™ QBank is an important learning aid for achieving the depth of proficiency needed at Level II. It should not, however, be considered a replacement for rehearsing with “exam-type” questions as found in our Schweser Mock Exams.

Mock Exams

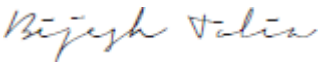
Kaplan Schweser offers six full-length mock exams: Mock Exams 1 through 6 each contain complete 88-question tests, with answer explanations. These are important tools for gaining the speed and skills you will need to pass the exam. You can use our Performance Tracker to monitor how you are performing compared to other Kaplan Schweser Level II candidates.

How to Succeed

The Level II CFA exam is a formidable challenge (42 readings and 371 Learning Outcome Statements), so you must devote considerable time and effort to be adequately prepared. There is no shortcut! You must learn the material, know the terminology and techniques, understand the concepts (candidates used to rote memorization should change their approach), and be able to answer 88 questions quickly and mostly-correctly. Fifteen hours per week for 25 weeks is a useful estimate of the study time required on average, but different candidates will need more or less time, depending on their individual backgrounds and experience.

There is no way around it; CFA Institute will test you in a way that will reveal how well you know the curriculum. You should begin early and stick to your study plan. Read the SchweserNotes and complete the Module Quizzes. Prepare for and attend a live class, an online class, or a study group each week. Take quizzes often using SchweserPro QBank, and go back to review previous topics regularly. At the end of each topic area, take the online Topic Quiz to check your progress. You should try to finish reading the curriculum at least four weeks before the exam so that you have sufficient time for Mock Exams and for further review of those topics that you have not yet mastered.

Best regards,



Dr. Bijesh Tolia, CFA, CA
Level II Manager

Kaplan Schweser

READING 1

MULTIPLE REGRESSION

EXAM FOCUS

Multiple linear regression models explain the variation in a dependent variable using more than one independent variables. You should know how to use an *F*-test to test the effectiveness of nested models. Become familiar with the effects that heteroskedasticity, serial correlation, and multicollinearity have on regression results, and be able to identify common model misspecifications. Finally, understand the role of influential observations in the estimated model, and the use of logistic regression models.

MODULE 1.1: BASICS OF MULTIPLE REGRESSION AND UNDERLYING ASSUMPTIONS

LOS 1.a: Describe the types of investment problems addressed by multiple linear regression and the regression process.

Given the complexities of financial and economic relations, a simple one-factor linear regression model is usually inadequate. **Multiple regression** models allow for consideration of multiple underlying influences (independent variables) on the dependent variable.

We can use multiple regression models to:

1. **Identify relationships between variables:** For example, an analyst may perform exploratory analysis of the factors that influence returns on small-cap stocks. Or, the analyst may wish to determine if the three-factor Fama-French model (market, size, and style) actually adequately explains cross-sectional returns for a sample time period.
2. **Forecast variables:** For example, an analyst may seek to forecast cash flows for a company, or to predict the probability of company default.
3. **Test existing theories:** For example, analysts may want to assess if corporate debt issuers with high levels of intangibles on their balance sheet (in addition to other known factors) explain credit risk premiums for those issuers.

Warm-Up: Multiple Regression Basics

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, $i = 1, 2, \ldots, n$
X_j = independent variables, $j = 1, 2, \ldots, k$
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

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 result of this process is the following regression equation:

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1X_{1i} + \hat{b}_2X_{2i} + \ldots + \hat{b}_kX_{ki}$$

where the "^" indicates an estimate for the corresponding regression coefficient

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

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{b}_0 + \hat{b}_1X_{1i} + \hat{b}_2X_{2i} + \ldots + \hat{b}_kX_{ki}\right)$$

The Level I curriculum covered simple linear regression and the t-test for statistical significance of the slope coefficient. For Level II, in order to interpret regression results, we can alternatively use the ***p*-value** to evaluate the null hypothesis that a slope coefficient is equal to zero.

The *p*-value is the smallest level of significance for which the null hypothesis can be rejected. We test the significance of coefficients by comparing the *p*-value to the chosen significance level:

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

LOS 1.b: Formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients.

Let's illustrate multiple regression using research by Arnott and Asness (2003).[1] As part of their research, the authors tested 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 formulated the following regression equation using annual data (46 observations):

$$EG10 = b_0 + b_1 PR + b_2 YCS + \varepsilon$$

The results of this regression are shown in Figure 1.1.

Figure 1.1: 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

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 somewhat 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 1-unit change in that independent variable, *holding the other independent variables constant*. For this reason, the slope coefficients in a multiple regression are sometimes called **partial slope coefficients**.

For example, regarding the real earnings growth model, 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.*

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:

$$Y = 2.0 + 4.5\,X_1$$

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:

$$Y = 1.0 + 2.5\,X_1 + 6.0\,X_2$$

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 expect this to happen when a second variable is included.

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* X_2 *constant*.

LOS 1.c: Explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions.

Assumptions underlying a multiple regression model include:

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

Residual plots allow analysts to get a preliminary indication of violation of regression assumptions. We will discuss formal statistical tests for the detection of violation of (the last three) regression assumptions later.

A normal quantile-quantile plot (usually called simply a **Q-Q plot**), is used to compare a variable's distribution to that of a normal distribution. We can employ a Q-Q plot to evaluate the standardized residuals of a regression model: the residuals should lie along a diagonal if they follow a normal distribution. Recall that 5% of normally distributed observations should be below −1.65 standard deviations.

EXAMPLE: Office rent model

An analyst wants to model the determinants of rents for office properties in a large city in the United States. Using a sample of 191 observations, she has estimated the following model:

$$rent_i = b_0 + b_1\,age_i + b_2\,distance_i + b_3\,restaurant_i + \varepsilon_i$$

where:

rent = monthly rent per square feet ($)
age = age of the property (in years)
distance = distance from the nearest metro station (in miles)
restaurant = number of lunch locations within walking distance

Regression Output:

Coefficients	Estimate	Std. Error
(Intercept)	44.67	2.01
Age	−0.31	0.05
Distance	−0.01	0.001
Restaurant	1.29	0.29

Exhibit 1: Plot of Residual vs. Predicted Values

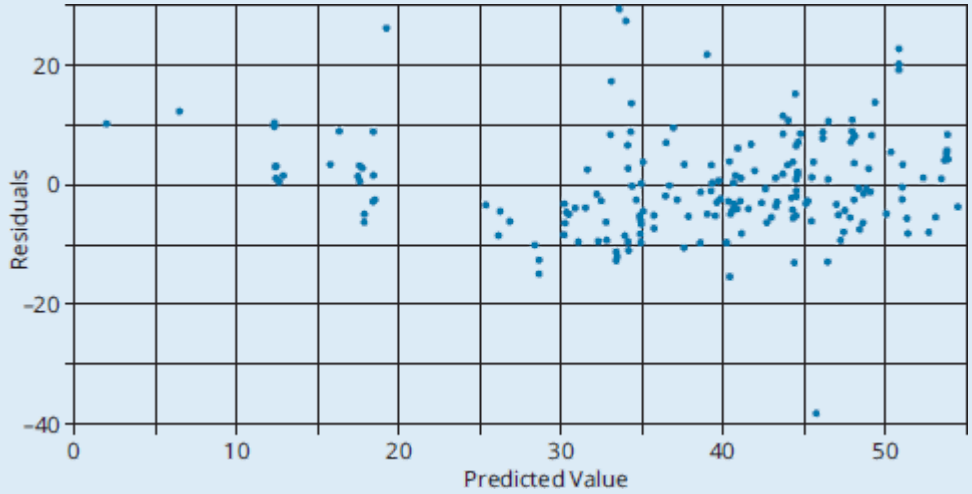


Exhibit 2: Plots of the Residuals vs. the Independent Variables

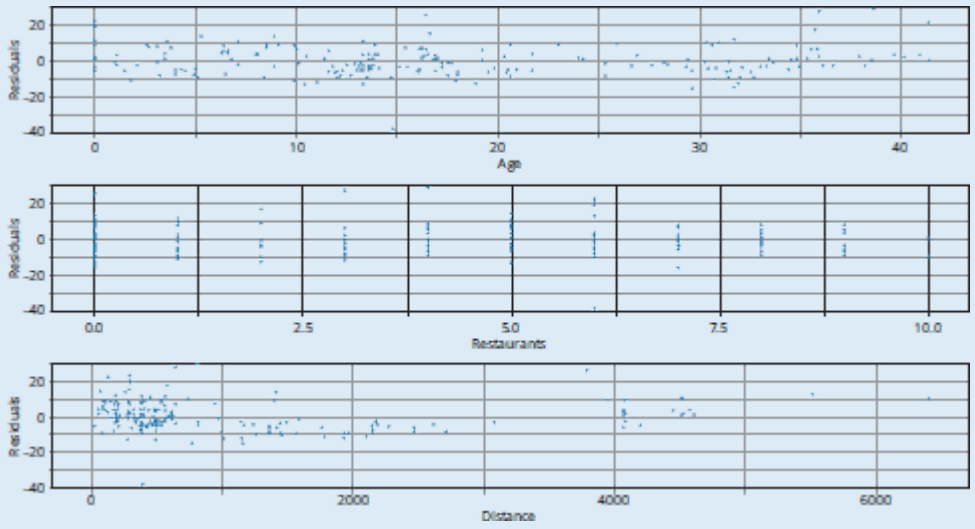
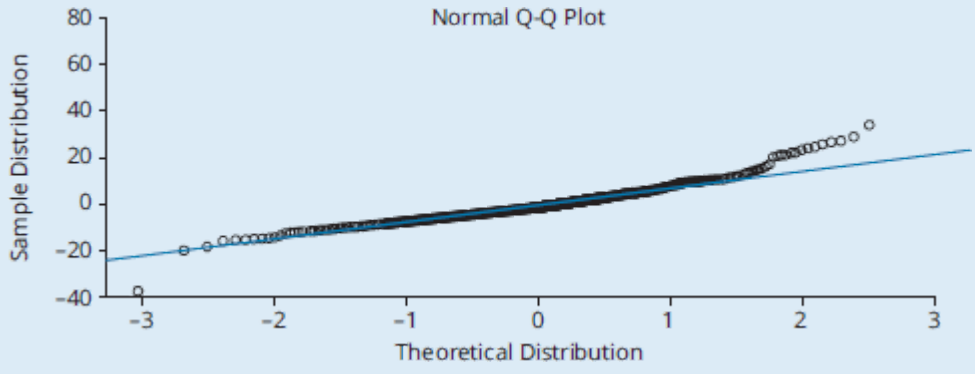


Exhibit 3: Normal Q-Q Plot of Residuals



1. Interpret the plot of regression residuals versus the predicted value of the dependent variable as shown in Exhibit 1.
2. Interpret the plot of regression residuals versus the independent variables as shown in Exhibit 2.
3. Interpret the Normal Q-Q plot of regression residuals as shown in Exhibit 3.

Answer:

1. Exhibit 1 does not indicate a systematic pattern or directional relationship between the predicted value of the dependent variable and the model residuals as indicated by the horizontal line centered on 0. This conforms to the requirement that residuals are independent of the predicted value of the dependent variable. Visually, it appears that the residuals have a constant variance and are uncorrelated with each other.
2. Exhibit 2 does not indicate a directional relationship between the residuals and any of the independent variables. The residuals are scattered around the horizontal line 0 across different values of the independent variables. This is desirable as it conforms to the requirement that the residuals are unrelated to the independent variables.
3. For a standard normal distribution, only 5% of the observations should be beyond −1.65 standard deviations of 0. We see that a few observations are beyond −2 standard deviations in Exhibit 3, with one outlier beyond −3 standard deviations. Similarly, it appears that there are a higher-than-expected number of observations beyond +2 standard deviations. Exhibit 3 also shows skewness in the right tail where the observations are skewed above the line of theoretical distribution. In summary, the distribution of the residuals deviates from a normal distribution as it has fatter tails and right skewness.

MODULE QUIZ 1.1

1. Which of the following investment problems is *least likely* to be addressed by using a multiple regression model?
   A. Prediction of the likelihood of monetary tightening by the central bank using macroeconomic variables.
   B. Uncovering a systematic pattern in the value of a currency using historical daily price data for that currency.
   C. Determining if the five-factor Fama-French model can be improved by adding an earnings momentum factor.

2. A multiple regression model with two explanatory variables was fitted as follows: $Y = 2.30 + 5.02 X_1 - 4.55 X_2$. Which of the following is the *least appropriate* interpretation of this model?
   A. The forecasted value of Y is 2.30 when both X_1 and X_2 are equal to zero.
   B. A 1% increase in X_1 would lead to a 7.32% increase in Y.
   C. A 1% increase in X_2 would lead to a 4.55% decrease in Y.

3. Which of the following *least accurately* represents an assumption of the multiple regression model?
   A. The relationship between the X variables is linear.

B. The variance of the error is constant.
C. The residual distribution is normal.

MODULE 1.2: EVALUATING REGRESSION MODEL FIT AND INTERPRETING MODEL RESULTS

LOS 1.d: Evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit.

ANOVA Tables

By decomposing the total variation in the dependent variable into the explained and unexplained components, we can evaluate the quality of model fit. **Analysis of variance (ANOVA)** is a statistical procedure that provides this information.

The results of the ANOVA procedure are presented in an ANOVA table, which accompanies a multiple regression output. A generic ANOVA table is presented in Figure 1.2.

Figure 1.2: ANOVA Table

Source	df (Degrees of Freedom)	SS (Sum of Squares)	MS (Mean Square = SS/df)
Regression	k	RSS	MSR
Error	n − k − 1	SSE	MSE
Total	n − 1	SST	

The first column indicates the "source" of the variation: regression is the explained component of the variation, while error is the unexplained component.

Coefficient of Determination, R^2

R^2 evaluates the overall effectiveness of the entire set of independent variables in explaining the dependent variable. It is 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.

$$R^2 = \frac{\text{total variation} - \text{unexplained variation}}{\text{total variation}}$$

$$= \frac{SST - SSE}{SST} = \frac{\text{explained variation}}{\text{total variation}} = \frac{RSS}{SST}$$

Adjusted R^2

Unfortunately, R^2 almost always increases as more 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 efficiently the set explains the dependent variable. This problem is often referred to as overestimating or **overfitting** the regression model.

To overcome the problem of overfitting (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 will always be 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 (i.e., the absolute value of the coefficient's t-statistic is less than 1), the value of R_a^2 will decrease.

EXAMPLE: Calculating R^2 and adjusted R^2

An analyst runs a regression of monthly 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 stock returns. The R_a^2 is, as expected, a somewhat lower value.

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 is most likely to 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 will prefer the first model because the adjusted R^2 is higher, and the model has five independent variables as opposed to nine.

While the adjusted R^2 penalizes overfitting, it does not indicate the quality of model fit, nor does it indicate statistical significance of the slope coefficients. We can formally evaluate the overall model fit using an *F*-test (discussed later).

For evaluating a regression model, regression output may include the Akaike's information criterion (AIC) and the Schwarz's Bayesian information criteria (BIC). Both AIC and BIC evaluate the quality of model fit *among competing models for the same dependent variable*. Lower values indicate a better model under either criteria.

AIC is used if the goal is to have a better forecast, while BIC is used if the goal is a better goodness of fit.

These metrics can be calculated as follows:

$$\text{AIC} = n \times \ln\left(\frac{\text{SSE}}{n}\right) + 2(k+1)$$

$$\text{BIC} = n \times \ln\left(\frac{\text{SSE}}{n}\right) + \ln(n) \times (k+1)$$

where:
k = number of independent variables

The variable *k* is a penalty parameter in both criteria: higher values of *k* result in higher values of the criteria. Because ln(n) is greater than 2 for even small sample sizes, the BIC metric imposes a higher penalty for overfitting.

EXAMPLE: Goodness of fit for the rent model

Continuing our example on rental price per square foot, the following shows the results when a single factor, two factors, and all three factors are used in the model.

Independent Variable(s)	K	SSR	SSE	R^2	R^2-Adj	AIC	BIC
Age	1	3,318.9	32,627.3	9.23%	8.75%	985.9	992.4
Age + Distance	2	20,946.1	15,000.2	58.27%	57.8%	839.4	849.2
Age + Distance + Restaurants	3	22,395.6	13,550.5	62.30%	61.7%	822.0	835.0

A. Which model is the most appropriate for use in generating forecasts?

B. Which model has a better goodness of fit?

Answer:

A. The model with all three independent variables has the lowest AIC, and hence is the most appropriate model for generating forecasts.