CFA Institute®
CFA Program

QUANTITATIVE METHODS, ECONOMICS

CFA® Program Curriculum
2024 LEVEL 2 VOLUME 1

ISBN 978-1-953-33768-9 (paper)

CONTENTS

Economics

How to Use the CFA Program Curriculum

The CFA® Program exams measure your mastery of the core knowledge, skills, and abilities required to succeed as an investment professional. These core competencies are the basis for the Candidate Body of Knowledge (CBOK™). The CBOK consists of four components:

- A broad outline that lists the major CFA Program topic areas (www.cfainstitute.org/programs/cfa/curriculum/cbok)
- Topic area weights that indicate the relative exam weightings of the top-level topic areas (www.cfainstitute.org/programs/cfa/curriculum)
- Learning outcome statements (LOS) that advise candidates about the specific knowledge, skills, and abilities they should acquire from curriculum content covering a topic area: LOS are provided in candidate study sessions and at the beginning of each block of related content and the specific lesson that covers them. We encourage you to review the information about the LOS on our website (www.cfainstitute.org/programs/cfa/curriculum/study-sessions), including the descriptions of LOS "command words" on the candidate resources page at www.cfainstitute.org.
- The CFA Program curriculum that candidates receive upon exam registration

Therefore, the key to your success on the CFA exams is studying and understanding the CBOK. You can learn more about the CBOK on our website: www.cfainstitute.org/programs/cfa/curriculum/cbok.

The entire curriculum, including the practice questions, is the basis for all exam questions and is selected or developed specifically to teach the knowledge, skills, and abilities reflected in the CBOK.

ERRATA

The curriculum development process is rigorous and includes multiple rounds of reviews by content experts. Despite our efforts to produce a curriculum that is free of errors, there are instances where we must make corrections. Curriculum errata are periodically updated and posted by exam level and test date online on the Curriculum Errata webpage (www.cfainstitute.org/en/programs/submit-errata). If you believe you have found an error in the curriculum, you can submit your concerns through our curriculum errata reporting process found at the bottom of the Curriculum Errata webpage.

DESIGNING YOUR PERSONAL STUDY PROGRAM

An orderly, systematic approach to exam preparation is critical. You should dedicate a consistent block of time every week to reading and studying. Review the LOS both before and after you study curriculum content to ensure that you have mastered the

applicable content and can demonstrate the knowledge, skills, and abilities described by the LOS and the assigned reading. Use the LOS self-check to track your progress and highlight areas of weakness for later review.

Successful candidates report an average of more than 300 hours preparing for each exam. Your preparation time will vary based on your prior education and experience, and you will likely spend more time on some study sessions than on others.

CFA INSTITUTE LEARNING ECOSYSTEM (LES)

Your exam registration fee includes access to the CFA Program Learning Ecosystem (LES). This digital learning platform provides access, even offline, to all of the curriculum content and practice questions and is organized as a series of short online lessons with associated practice questions. This tool is your one-stop location for all study materials, including practice questions and mock exams, and the primary method by which CFA Institute delivers your curriculum experience. The LES offers candidates additional practice questions to test their knowledge, and some questions in the LES provide a unique interactive experience.

FEEDBACK

Please send any comments or feedback to info@cfainstitute.org, and we will review your suggestions carefully.

Quantitative Methods

LEARNING MODULE

1

Basics of Multiple Regression and Underlying Assumptions

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	describe the types of investment problems addressed by multiple linear regression and the regression process
☐	formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients
☐	explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions

INTRODUCTION

Multiple linear regression uses two or more independent variables to describe the variation of the dependent variable rather than just one independent variable, as in simple linear regression. It allows the analyst to estimate using more complex models with multiple explanatory variables and, if used correctly, may lead to better predictions, better portfolio construction, or better understanding of the drivers of security returns. If used incorrectly, however, multiple linear regression may yield spurious relationships, lead to poor predictions, and offer a poor understanding of relationships.

The analyst must first specify the model and make several decisions in this process, answering the following, among other questions: What is the dependent variable of interest? What independent variables are important? What form should the model take? What is the goal of the model—prediction or understanding of the relationship?

The analyst specifies the dependent and independent variables and then employs software to estimate the model and produce related statistics. The good news is that the software, such as shown in Exhibit 1, does the estimation, and our primary tasks are to focus on specifying the model and interpreting the output from this software, which are the main subjects of this content.

Exhibit 1: Examples of Regression Software

Software	Programs/Functions
Excel	Data Analysis > Regression
Python	scipy.stats.linregress
	statsmodels.lm
	sklearn.linear_model.LinearRegression
R	lm
SAS	PROC REG
	PROC GLM
STATA	regress

SUMMARY

- Multiple linear regression is used to model the linear relationship between one dependent variable and two or more independent variables.
- In practice, multiple regressions are used to explain relationships between financial variables, to test existing theories, or to make forecasts.
- The regression process covers several decisions the analyst must make, such as identifying the dependent and independent variables, selecting the appropriate regression model, testing if the assumptions behind linear regression are satisfied, examining goodness of fit, and making needed adjustments.
- A multiple regression model is represented by the following equation:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \ldots + b_kX_{ki} + \varepsilon_i,\ i = 1, 2, 3, \ldots, n,$$

  where Y is the dependent variable, Xs are the independent variables from 1 to k, and the model is estimated using n observations.
- Coefficient b_0 is the model's "intercept," representing the expected value of Y if all independent variables are zero.
- Parameters b_1 to b_k are the slope coefficients (or partial regression coefficients) for independent variables X_1 to X_k. Slope coefficient b_j describes the impact of independent variable X_j on Y, holding all the other independent variables constant.
- There are five main assumptions underlying multiple regression models that must be satisfied, including (1) linearity, (2) homoskedasticity, (3) independence of errors, (4) normality, and (5) independence of independent variables.
- Diagnostic plots can help detect whether these assumptions are satisfied. Scatterplots of dependent versus and independent variables are useful for detecting non-linear relationships, while residual plots are useful for detecting violations of homoskedasticity and independence of errors.

USES OF MULTIPLE LINEAR REGRESSION

2

☐ describe the types of investment problems addressed by multiple linear regression and the regression process

There are many investment problems in which the analyst needs to consider the impact of multiple factors on the subject of research rather than a single factor. In the complex world of investments, it is intuitive that explaining or forecasting a financial variable by a single factor may be insufficient. The complexity of financial and economic relations calls for models with multiple explanatory variables, subject to fundamental justification and various statistical tests.

Examples of how multiple regression may be used include the following:

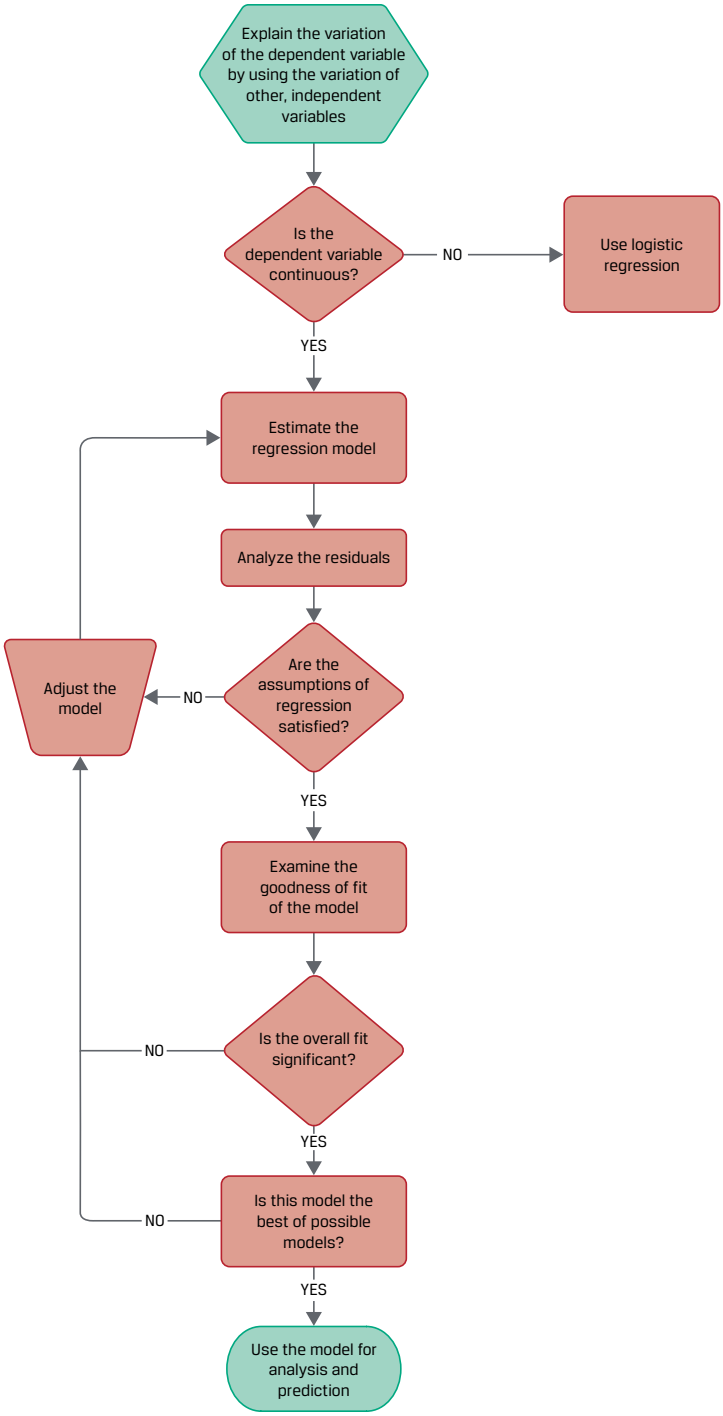
- A portfolio manager wants to understand how returns are influenced by a set of underlying factors; the size effect, the value effect, profitability, and investment aggressiveness. The goal is to estimate a Fama–French five-factor model that will provide an understanding of the factors that are important for driving a particular stock's excess returns.
- A financial adviser wants to identify whether certain variables, such as financial leverage, profitability, revenue growth, and changes in market share, can predict whether a company will face financial distress.
- An analyst wants to examine the effect of different dimensions of country risk, such as political stability, economic conditions, and environmental, social, and governance (ESG) considerations, on equity returns in that country.

Multiple regression can be used to identify relationships between variables, to test existing theories, or to forecast. We outline the general process of regression analysis in Exhibit 2. As you can see, there are many decisions that the analyst must make in this process.

For example, if the dependent variable is continuous, such as returns, the traditional regression model is typically the first step. If, however, the dependent variable is discrete—for example, an indicator variable such as whether a company is a takeover target or not a takeover target—then, as we shall see, the model may be estimated as a logistic regression.

In either case, the process of determining the best model follows a similar path. The model must first be specified, including independent variables that may be continuous, such as company financial features, or discrete (i.e., dummy variables), indicating membership in a class, such as an industry sector. Next, the regression model is estimated and analyzed to ensure it satisfies key underlying assumptions and meets the analyst's goodness-of-fit criteria. Once the model is tested and its out-of-sample performance is deemed acceptable, then it can be used for further identifying relationships between variables, for testing existing theories, or for forecasting.

Exhibit 2: The Regression Process



KNOWLEDGE CHECK

Assessment: Multiple Regression—Types of Investment Problems and Process

1. You are a junior analyst assisting in the development of various multiple regression models for your industry sector. Identify the action you should take to resolve each of the following issues:

Issue	Action
The dependent variable takes on a value of 1 if the company is a merger target and 0 otherwise.	
The analyst estimates a model with five independent variables, and none of these variables are significant explanatory variables.	
The residuals do not appear to be homoskedastic, thus violating a regression assumption.	
The regression assumptions are satisfied, the overall fit is significant, and the model is the best model of the possible models.	

Solution

Issue	Action
The dependent variable takes on a value of 1 if the company is a merger target and 0 otherwise.	Use logistic regression.
The analyst estimates a model with five independent variables, and none of these variables are significant explanatory variables.	Adjust the model and re-estimate.
The residuals do not appear to be homoskedastic, thus violating a regression assumption.	Adjust the model and re-estimate.
The regression assumptions are satisfied, the overall fit is significant, and the model is the best model of the possible models.	Use the model for analysis and prediction.

THE BASICS OF MULTIPLE REGRESSION 3

☐ formulate a multiple linear regression model, describe the relation between the dependent variable and several independent variables, and interpret estimated regression coefficients

The goal of simple regression is to explain the variation of the dependent variable, Y, using the variation of an independent variable, X. The goal of multiple regression is the same, to explain the variation of the dependent variable, Y, but using the variations in a set of independent variables, $X_1, X_2, \ldots, X_k$. Recall the variation of Y is

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

which we also refer to as the sum of squares total. The simple regression equation is

$$Y_i = b_0 + b_1X_i + \varepsilon_i,\ i=1, 2, 3, \ldots, n.$$

When we introduce additional independent variables to help explain the variation of the dependent variable, we have the multiple regression equation:

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \ldots + b_kX_{ki} + \varepsilon_i,\ i = 1, 2, 3, \ldots, n. \quad (1)$$

In this equation, the terms involving the k independent variables are the deterministic part of the model, whereas the error term, ε_i, is the stochastic or random part of the model. The model is estimated over n observations, where n must be larger than k.

It is important to note that a slope coefficient in a multiple regression, known as a **partial regression coefficient** or a *partial slope coefficient*, must be interpreted with care. A partial regression coefficient, b_j, describes the impact of that independent variable on the dependent variable, holding all the other independent variables constant. For example, in the multiple regression equation,

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \varepsilon_i,$$

the coefficient b_2 measures the change in Y for a one-unit change in X_2 assuming X_1 and X_3 are held constant. The estimated regression equation is

$$Y_i = \hat{b}_0 + \hat{b}_1X_{1i} + \hat{b}_2X_{2i} + \hat{b}_3X_{3i},$$

with ^ indicating estimated coefficients.

Consider an estimated regression equation in which the monthly excess returns of a bond index (RET) are regressed against the change in monthly government bond yields (BY) and the change in the investment-grade credit spreads (CS). The estimated regression, using 60 monthly observations, is

$$\text{RET} = 0.0023 - 5.0585\text{BY} - 2.1901\text{CS}.$$

We learn the following from this regression:

1. The bond index RET yields, on average, 0.0023% per month, or approximately 0.028% per year, if the changes in the government bond yields and investment-grade credit spreads are zero.
2. The change in the bond index return for a given one-unit change in the monthly government bond yield, BY, is –5.0585%, holding CS constant. This means that the bond index has an empirical duration of 5.0585.
3. If the investment-grade credit spreads, CS, increase by one unit, the bond index returns change by –2.1901%, holding BY constant.
4. For a month in which the change in the credit spreads is 0.001 and the change in the government bond yields is 0.005, the expected excess return on the bond index is

$$\text{RET} = 0.0023 - 5.0585(0.005) - 2.1901(0.001) = -0.0252, \text{ or } -2.52\%.$$

KNOWLEDGE CHECK

An institutional salesperson has just read the research report in which you estimated a regression of monthly excess returns on a portfolio, RETRF, against the Fama–French three factors:

- MKTRF, the market excess return;

- SMB, the difference in returns between small- and large-capitalization stocks; and
- HML, the difference in returns between value and growth stocks.

All returns are stated in whole percentages (that is, 1 for 1%), and the estimated regression equation is

$$\text{RETRF} = 1.5324 + 0.5892\text{MKTRF} + -0.8719\text{SMB} + -0.0560\text{HML}.$$

Before this salesperson meets with her client firm, she asks you to do the following regarding your estimated regression model:

1. Interpret the intercept.

   Solution

   If the market excess return, SMB, and HML are each zero, then we expect a return on the portfolio of 1.534%.

2. Interpret each slope coefficient.

   Solution

   Each slope coefficient is interpreted assuming the other variables are held constant.
   • For MKTRF, if the market return increases by 1%, we expect the portfolio's return to increase by 0.5892%.
   • For SMB, if the size effect returns increase by 1%, we expect the portfolio's return to decrease by 0.8719%.
   • For HML, if the value effect returns increase by 1%, we expect the portfolio's return to decrease by 0.056%.

3. Calculate the predicted value of the portfolio's return if

   MKTRF = 1, SMB = 4, and HML = –2.

   Solution

   Given the expected values of the independent variables, the expected return on the portfolio is

   $$R = 1.534 + 0.5892(1) - 0.8719(4) - 0.0560(-2) = -1.2524.$$

ASSUMPTIONS UNDERLYING MULTIPLE LINEAR REGRESSION

4

☐ explain the assumptions underlying a multiple linear regression model and interpret residual plots indicating potential violations of these assumptions

Before we can conduct correct statistical inference on a multiple linear regression model estimated using ordinary least squares (OLS), we need to know whether the assumptions underlying that model are met. Suppose we have n observations on the dependent variable, Y, and the independent variables, $X_1, X_2, \ldots, X_k$, and we want to estimate the model

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{3i} + \ldots + b_kX_{ki} + \varepsilon_i,\ i = 1, 2, 3, \ldots, n.$$

In simple regression, we had four assumptions that needed to be satisfied so that we could make valid conclusions regarding the regression results. In multiple regression, we modify these slightly to reflect the additional independent variables:

1. Linearity: The relationship between the dependent variable and the independent variables is linear.
2. Homoskedasticity: The variance of the regression residuals is the same for all observations.
3. Independence of errors: The observations are independent of one another. This implies the regression residuals are uncorrelated across observations.
4. Normality: The regression residuals are normally distributed.
5. Independence of independent variables:

    5a. Independent variables are not random.

    5b. There is no exact linear relation between two or more of the independent variables or combinations of the independent variables.

The independence assumption is needed to enable the estimation of the coefficients. If there is an exact linear relationship between independent variables, the model cannot be estimated. In the more common case of approximate linear relationships, which may be indicated by significant pairwise correlations between the independent variables, the model can be estimated but its interpretation is problematic. In empirical work, the assumptions underlying multiple linear regression do not always hold. The statistical tools to detect violations and methods to mitigate their effects will be addressed later.

Regression software produces diagnostic plots, which are a useful tool for detecting potential violations of the assumptions underlying multiple linear regression. To illustrate the use of such plots, we first estimate a regression to analyze 10 years of monthly total excess returns of ABC stock using the Fama–French three-factor model. As noted previously, this model uses market excess return (MKTRF), size (SMB) and value (HML) as explanatory variables.

$$\text{ABC\_RETRF}_t = b_0 + b_1\text{MKTRF}_t + b_2\text{SMB}_t + b_3\text{HML}_t + \varepsilon_t$$

We start our analysis by generating a **scatterplot matrix** using software. This matrix is also referred to as a *pairs plot*.

CODE: SCATTERPLOT MATRIX

Using Python

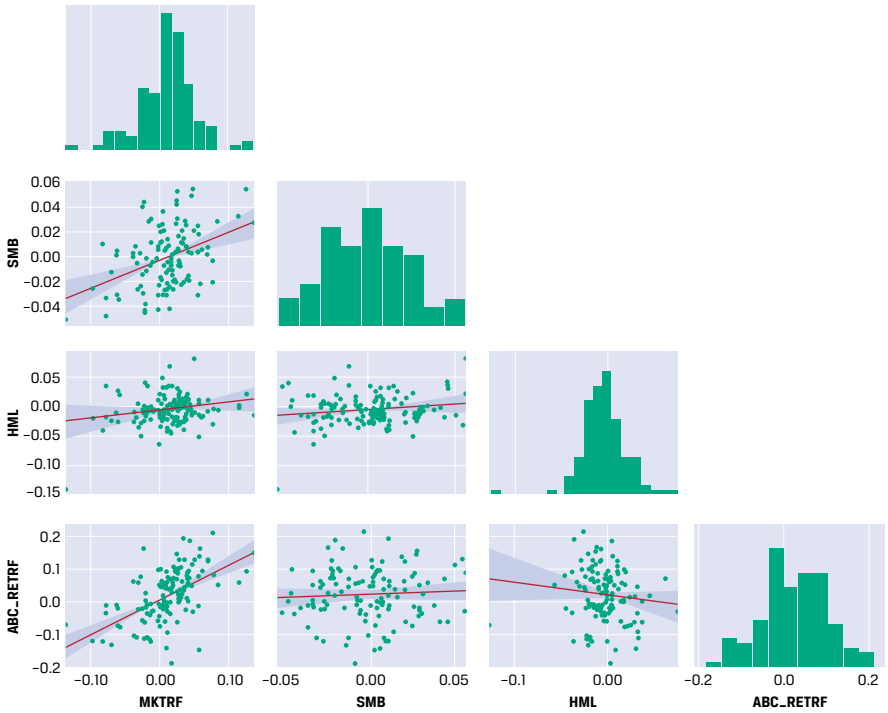
```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read_csv("ABC_FF.csv",parse_dates=True,index_col=0)

sns.pairplot(df)

plt.show()
```

Using R

```
df <- read.csv("data.csv")

pairs(df[c("ABC_RETRF","MKTRF","SMB","HML")])
```

The pairwise scatterplots for all variables are shown in Exhibit 3. For example, the bottom row shows the relationships for the following three pairs: ABC_RETRF and MKTRF, ABC_RETRF and SMB, and ABC_RETRF and HML. The simple regression line and corresponding 95% confidence interval for the variables in each pair are also shown, along with the histogram of each variable along the diagonal.

Exhibit 3: Scatterplot Matrix of ABC Returns and Fama–French Factors



You can see the following from the lower set of scatterplots between ABC_RET and the three independent variables:

- There is a positive relationship between ABC_RETF and the market factor, MKTRF.
- There seems to be no apparent relation between ABC_RETRF and the size factor, SMB. The reason is the scatterplot compares the two variables in isolation and does not show the "partial" correlation picked up by the regression, which explains why SMB is significant in the regression (see Exhibit 4) but not in the scatterplot.
- There is a negative relationship between ABC_RETF and the value factor, HML.

Looking at the scatterplots between the independent variables, SMB and HML have little or no correlation, as indicated by the relatively flat line for the SMB–HML pair. This is a desirable characteristic between explanatory variables.

An additional benefit of the scatterplot matrix is that all data points are displayed, so it can also be used to identify extreme values and outliers.

We now estimate the model of ABC's excess returns using software such as Microsoft Excel, Python, or R; results are shown in Exhibit 5. Focusing on the regression residuals, we look for clues to potential violations of the assumptions of multiple linear regression.

Exhibit 4: ABC Returns Explained Using Fama–French Three-Factor Model

Regression Statistics

Multiple *R*	0.6238
R-Squared	0.3891
Adjusted *R*-Squared	0.3733
Standard Error	0.0628
Observations	120

ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	3	0.2914	0.0971	24.6278	0.0000
Residual	116	0.4575	0.0039		
Total	119	0.7489			

	Coefficient	Standard error	*t*-Stat.	*P*-value	Lower 95%	Upper 95%
Intercept	0.0052	0.0061	0.8435	0.4007	–0.0070	0.0173
MKTRF	1.2889	0.1538	8.3791	0.0000	0.9842	1.5935
SMB	–0.5841	0.2664	–2.1922	0.0304	–1.1118	–0.0564
HML	–-0.6810	0.2231	–3.0523	0.0028	–1.1229	–0.2391

Exhibit 5: ABC Returns Explained Using Fama–French Three-Factor Model

Regression Statistics

Multiple *R*	0.6238
R-Squared	0.3891
Adjusted *R*-Squared	0.3733
Standard Error	0.0628
Observations	120

ANOVA

	Df	*SS*	*MS*	*F*	*Significance F*
Regression	3	0.2914	0.0971	24.6278	0.0000
Residual	116	0.4575	0.0039		
Total	119	0.7489			

	Coefficient	Standard error	*t*-Stat.	*P*-value	Lower 95%	Upper 95%
Intercept	0.0052	0.0061	0.8435	0.4007	–0.0070	0.0173
MKTRF	1.2889	0.1538	8.3791	0.0000	0.9842	1.5935
SMB	–0.5841	0.2664	–2.1922	0.0304	–1.1118	–0.0564
HML	–-0.6810	0.2231	–3.0523	0.0028	–1.1229	–0.2391

CODE: REGRESSION

Using Python

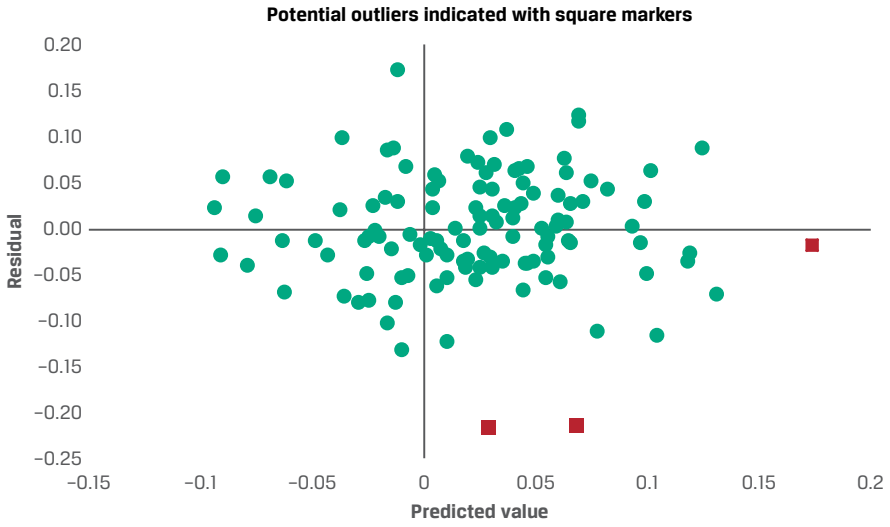
```
import pandas as pd
from statsmodels.formula.api import ols
df = pd.read_csv("data.csv")
model = ols('ABC_RETRF ~ MKTRF+SMB+HML',data=df).fit()
print(model.summary())
```

Using R

```
df <- read.csv("data.csv")
model <- lm('ABC_RETRF~ MKTRF+SMB+HML',data=df)
print(summary(model))
```

We start by looking at a scatterplot of residuals against the dependent variable, as shown in Exhibit 6. We can use this scatterplot to uncover potential assumption violations and to help identify outliers in our data.

Exhibit 6: Residuals vs. Predicted Value of Dependent Variable

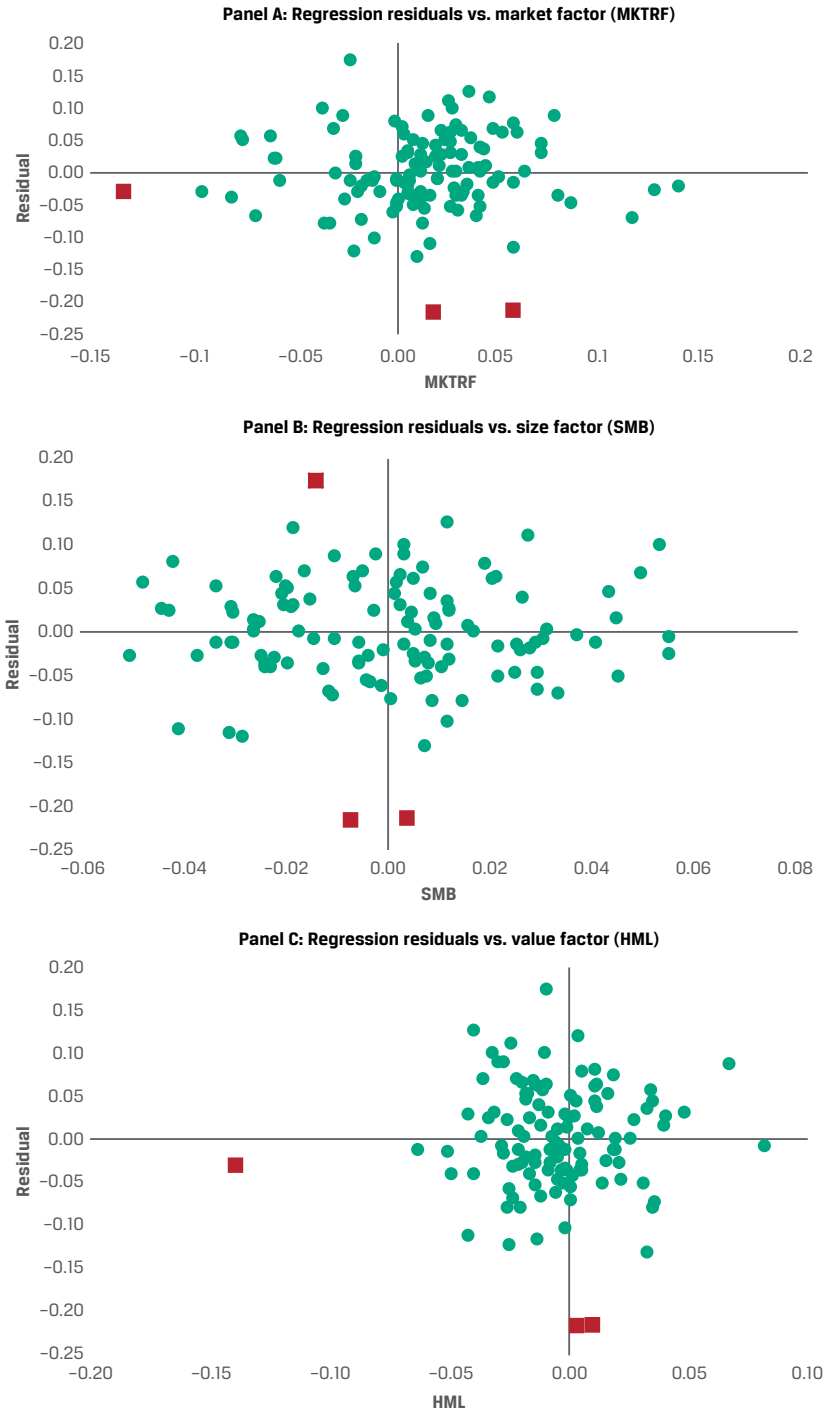


Potential outliers indicated with square markers

As indicated by the line centered near residual value 0.00, a visual inspection of Exhibit 6 does not reveal any directional relationship between the residuals and the predicted values from the regression model. This outcome is good, because we want residuals to behave in an independent manner compared to what the model predicts, and suggests the regression's errors have a constant variance and are uncorrelated with each other, thereby satisfying several of the underlying assumptions of multiple linear regression. Notably, we detect three residuals (square markers) that may be outliers, Months 7, 25, and 95. This information can be used to check for shocks from factors not considered in the model that may have occurred at these points in time.

Exhibit 7 presents plots of the regression residuals versus each of the three factors in Panels A, B, and C. A visual inspection does not indicate any directional relationship between the residuals and the explanatory variables, suggesting there is no violation of a multiple linear regression assumption. Importantly, the three potential outliers detected in the residual versus predicted value plot are also apparent in Exhibit 7, as indicated by the square markers.

Exhibit 7: Regression Residuals vs. Factors (Independent Variables)



CODE: RESIDUAL ANALYSIS

Using Python

```
import pandas as pd
```

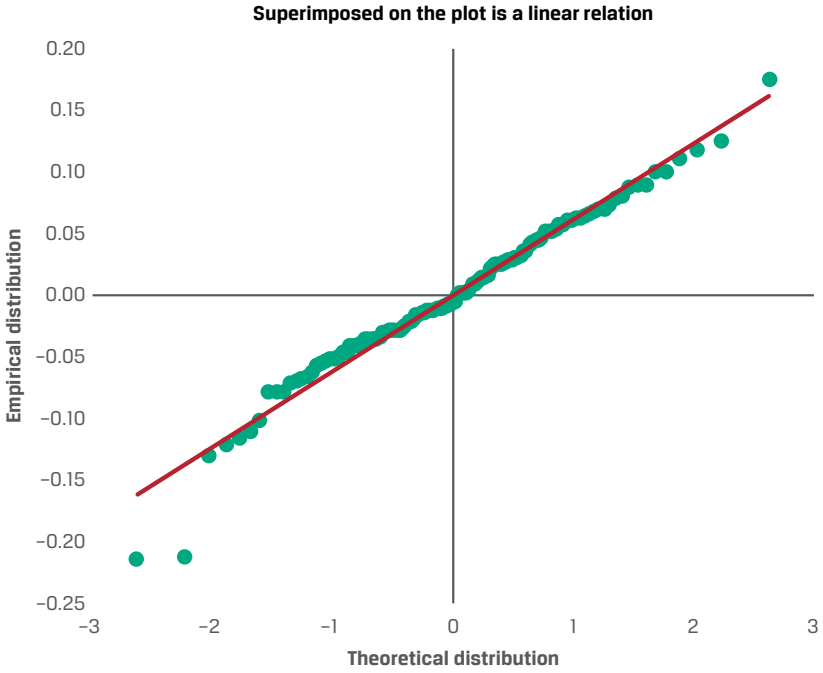
```
import matplotlib.pyplot as plt

import statsmodels.api as sm

import numpy as np

df = pd.read_csv("data.csv",parse_dates=True,index_col=0)

model = ols('ABC_RETRF ~ MKTRF+SMB+HML',data=df).fit()

fig = sm.graphics.plot_partregress_grid(model)

fig.tight_layout(pad=1.0)

plt.show()

fig = sm.graphics.plot_ccpr_grid(model)

fig.tight_layout(pad=1.0)

plt.show()
```

Using R

```
library(ggplot2)

library(gridExtra)

df <- read.csv("data.csv")

model <- lm('ABC_RETRF~ MKTRF+SMB+HML',data=df)

df$res <- model$residuals

g1 <- ggplot(df,aes(y=res, x=MKTRF))+geom_point()+
xlab("MKTRF")+ylab("Residuals")

g2 <- ggplot(df,aes(y=res, x=SMB))+geom_point()+ xlab("SMB")+
ylab("Residuals")

g3 <- ggplot(df,aes(y=res, x=HML))+geom_point()+ xlab("HML")+
ylab("Residuals")

grid.arrange(g1,g2,g3,nrow=3)
```

Finally, in Exhibit 8 we present a **normal Q-Q plot**. A normal Q-Q plot, or simply a Q-Q plot, is used to visualize the distribution of a variable by comparing it to a normal distribution. In the case of regression, we can use a Q-Q plot to compare the model's standardized residuals to a theoretical standard normal distribution. If the residuals are normally distributed, they should align along the diagonal. Recall that 5% of observations that are normally distributed should fall below −1.65 standard deviations, so the 5th percentile residual observation should appear at −1.65 standard deviations.

Exhibit 8: Normal Q-Q Plot of Regression Residuals



Superimposed on the plot is a linear relation

However, after –2 standard deviations, observations 25 and 95 fall well below the theoretical standard normal distribution range, while Observation 7, lying above the diagonal line around +2.5 standard deviations, is somewhat above the theoretical range. This evidence again suggests these three residual observations are potential outliers. However, setting them aside, the normal Q-Q plot does provide ample evidence that the regression residuals overall are distributed consistently with the normal distribution. Thus, we can conclude that the regression model error term is close to being normally distributed.

KNOWLEDGE CHECK

You are analyzing price changes of a cryptocurrency (CRYPTO) using the price changes for gold (GOLD) and a technology stock index (TECH), based on five years of monthly observations. You also run several diagnostic charts of your regression results. In a meeting with your research director, she asks you to do the following:

1. Identify any assumptions that may be violated if we examine the correlation between GOLD and TECH and find a significant pairwise correlation.

   Solution

   This result may indicate an approximate linear relation between GOLD and TECH, which would be a violation of the independence of independent variables, and should be explored further.

2. Describe the purpose of a plot of the regression residuals versus the predicted value of CRYPTO.

Solution

This plot is useful for examining whether there is any clustering or pattern that may suggest the residuals are not homoskedastic and whether there are any potential outliers.

3. Describe the purpose of a plot of the regression residuals versus GOLD.

Solution

This plot is useful for examining whether there are any extreme values of the independent variables that may influence the estimated regression parameters and whether there is any relationship between the residuals and an independent variable, which suggests the model is misspecified.

4. Describe the purpose of a normal Q-Q plot of residuals.

Solution

The normal Q-Q plot is useful for exploring whether the residuals are normally distributed, a key assumption of linear regression.

5. A pairwise scatterplot is used to detect whether:

   A. there is a linear relationship between the dependent and independent variables.

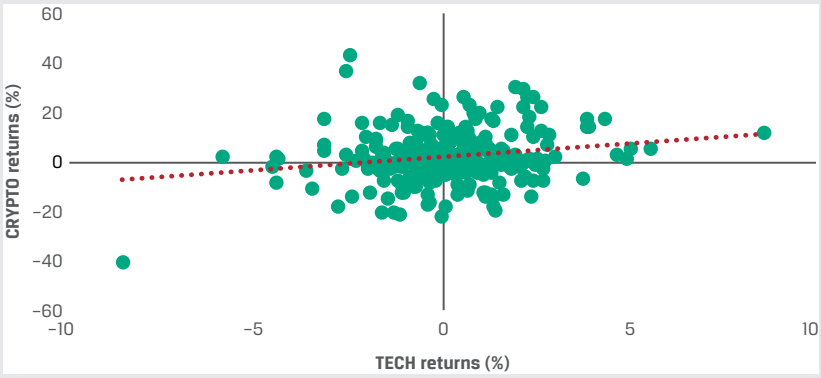
   B. the residual terms exhibit heteroskedasticity.

   C. the residual terms are normally distributed.

Solution

A is correct. The pairwise scatterplot is useful for visualizing the relationships between the dependent and explanatory variables.

6. Interpret this scatterplot showing price changes for the cryptocurrency (CRYPTO) and the tech index (TECH):



Solution

Based on the plot, there appears to be a positive relationship between CRYPTO and TECH, which may be significant. Several potential outliers are also apparent.

7. A normal Q-Q plot is used to detect whether:

   A. there is a linear relationship between the dependent and independent variables.

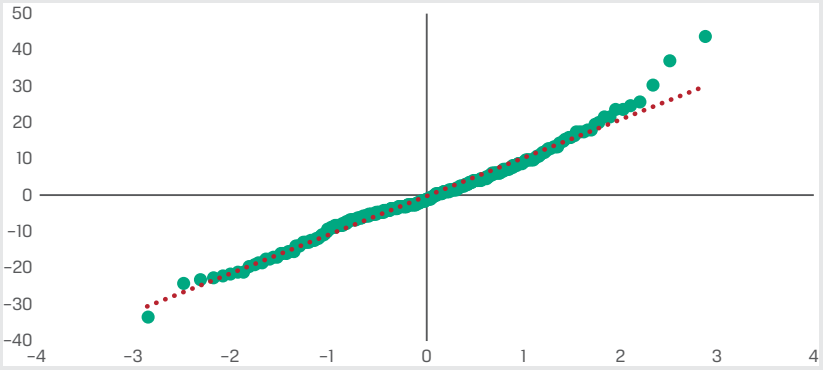
B. the regression residual terms exhibit heteroskedasticity.

C. the regression residual terms are normally distributed.

Solution

C is correct. The normal Q-Q plot is useful for exploring whether the residuals are normally distributed.

8. Interpret this normal Q-Q plot from our regression of CRYPTO price changes:



Solution

Based on the plot, the residuals are not normally distributed, as indicated by the deviation of residuals from the diagonal evident past ±2 standard deviations, and several potential outliers are also apparent. This normal Q-Q plot suggests the distribution of residuals is "fat-tailed." Note that fat-tailed distributions of residuals are a commonly observed feature of financial data time series.

PRACTICE PROBLEMS

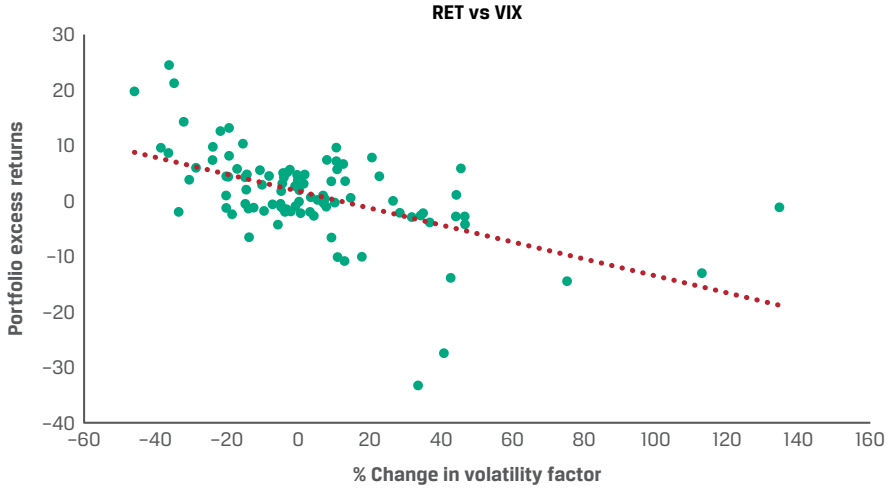
The following information relates to questions 1-5

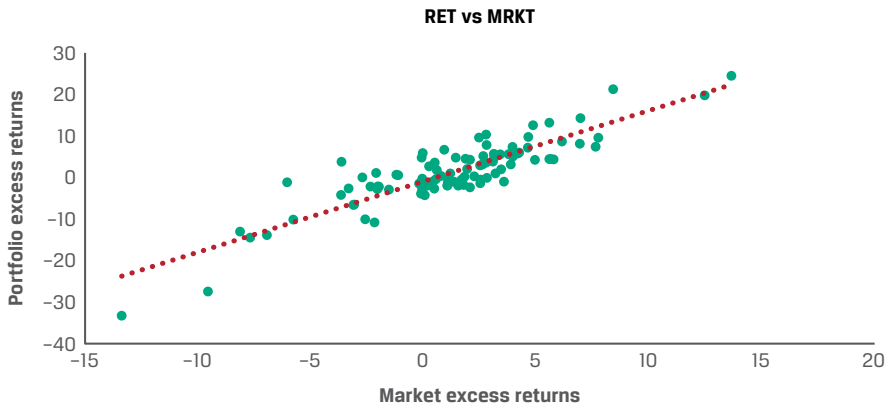
You are a junior analyst at an asset management firm. Your supervisor asks you to analyze the return drivers for one of the firm's portfolios. She asks you to construct a regression model of the portfolio's monthly excess returns (RET) against three factors: the market excess return (MRKT), a value factor (HML), and the monthly percentage change in a volatility index (VIX).

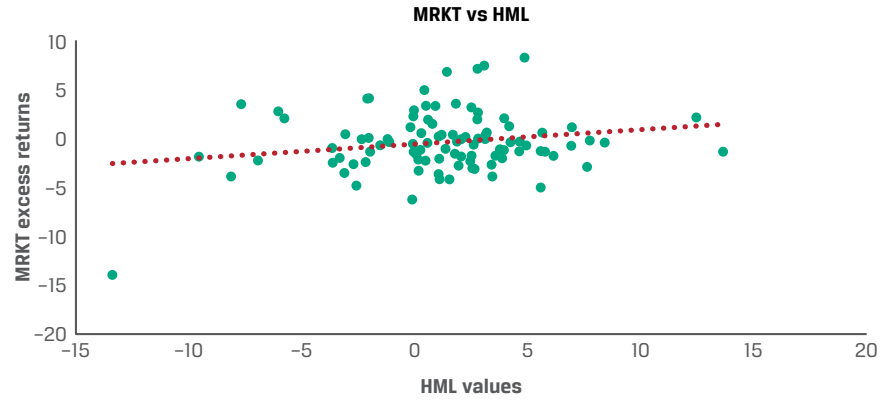
You collect the data and run the regression, and the resulting model is

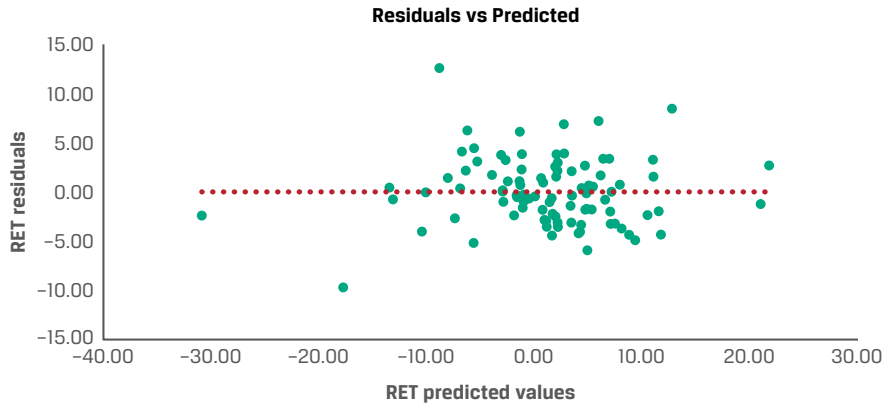
$Y_{RET} = -0.999 + 1.817X_{MRKT} + 0.489X_{HML} + 0.037X_{VIX}$.

You then create some diagnostic charts to help determine the model fit.









1. Determine the type of regression model you should use.

A. Logistic regression

B. Simple linear regression

C. Multiple linear regression

2. Determine which one of the following statements about the coefficient of the volatility factor (VIX) is true.

A. A 1.0% increase in X_{VIX} would result in a –0.962% decrease in Y_{RET}.

B. A 0.037% increase in X_{VIX} would result in a 1.0% increase in Y_{RET}.

C. A 1.0% increase in X_{VIX}, holding all the other independent variables constant, would result in a 0.037% increase in Y_{RET}.

3. Identify the regression assumption that may be violated based on Chart 1, RET vs. VIX.

A. Independence of errors

B. Independence of independent variables

C. Linearity between dependent variable and explanatory variables

4. Identify which chart, among Charts 2, 3, and 4, is *most likely* to be used to assess homoskedasticity.

A. Chart 2

B. Chart 3

C. Chart 4

5. Identify which chart, among Charts 2, 3, and 4, is *most likely* to be used to assess independence of independent variables.

A. Chart 2

B. Chart 3

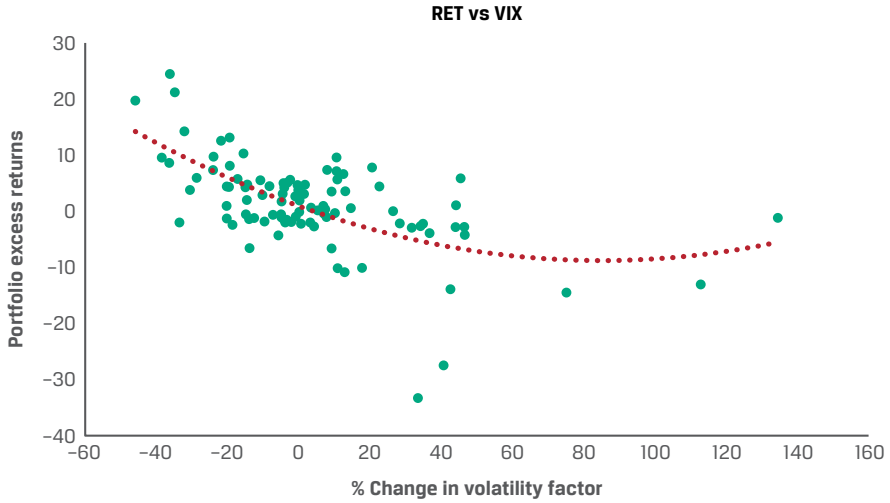
C. Chart 4

SOLUTIONS

1. C is correct. You should use a multiple linear regression model since the dependent variable is continuous (not discrete) and there is more than one explanatory variable. If the dependent variable were discrete, then the model should be estimated as a logistic regression.

2. C is correct. The coefficient of the volatility factor (X_{VIX}) is 0.037. It should be interpreted to mean that holding all the other independent variables constant, a 1% increase (decrease) would result in a 0.037% increase (decrease) in the monthly portfolio excess return (Y_{RET}).

3. C is correct. Chart 1 is a scatterplot of RET versus VIX. Linearity between the dependent variable and the independent variables is an assumption underlying multiple linear regression. As shown in the following Revised Chart 1, the relationship appears to be more curved (i.e., quadratic) than linear.



4. C is correct. To assess homoskedasticity, we must evaluate whether the variance of the regression residuals is constant for all observations. Chart 4 is a scatterplot of the regression residuals versus the predicted values, so it is very useful for visually assessing the consistency of the variance of the residuals across the observations. Any clusters of high and/or low values of the residuals may indicate a violation of the homoskedasticity assumption.

5. B is correct. Chart 3 is a scatterplot comparing the values of two of the independent variables, MRKT and HML. This chart would most likely be used to assess the independence of these explanatory variables.

LEARNING MODULE

2

Evaluating Regression Model Fit and Interpreting Model Results

LEARNING OUTCOMES

Mastery	*The candidate should be able to:*
☐	evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit
☐	formulate hypotheses on the significance of two or more coefficients in a multiple regression model and interpret the results of the joint hypothesis tests
☐	calculate and interpret a predicted value for the dependent variable, given the estimated regression model and assumed values for the independent variable

SUMMARY

- In multiple regression, adjusted R^2 is used as a measure of model goodness of fit since it does not automatically increase as independent variables are added to the model. Rather, it adjusts for the degrees of freedom by incorporating the number of independent variables.
- Adjusted R^2 will increase (decrease) if a variable is added to the model that has a coefficient with an absolute value of its t-statistic greater (less) than 1.0.
- Akaike's information criterion (AIC) and Schwarz's Bayesian information criteria (BIC) are also used to evaluate model fit and select the "best" model among a group with the same dependent variable. AIC is preferred if the purpose is prediction, BIC is preferred if goodness of fit is the goal, and lower values of both measures are better.
- Hypothesis tests of a single coefficient in a multiple regression, using t-tests, are identical to those in simple regression.

- The joint *F*-test is used to jointly test a subset of variables in a multiple regression, where the "restricted" model is based on a narrower set of independent variables nested in the broader "unrestricted" model. The null hypothesis is that the slope coefficients of all independent variables outside the restricted model are zero.
- The general linear *F*-test is an extension of the joint *F*-test, where the null hypothesis is that the slope coefficients on all independent variables in the unrestricted model are equal to zero.
- Predicting the value of the dependent variable using an estimated multiple regression model is similar to that in simple regression. First, sum, for each independent variable, the estimated slope coefficient multiplied by the assumed value of that variable, and then add the estimated intercept coefficient.
- In multiple regression, the confidence interval around the forecasted value of the dependent variable reflects both model error and sampling error (from forecasting the independent variables); the larger the sampling error, the larger is the standard error of the forecast of *Y* and the wider is the confidence interval.

1 GOODNESS OF FIT

☐ evaluate how well a multiple regression model explains the dependent variable by analyzing ANOVA table results and measures of goodness of fit

In the simple regression model, the **coefficient of determination**, also known as *R*-squared or R^2, is a measure of the goodness of fit of an estimated regression to the data. R^2 can also be defined in multiple regression as the ratio of the variation of the dependent variable explained by the independent variables (sum of squares regression) to the total variation of the dependent variable (sum of squares total).

$$R^2 = \frac{\text{Sum of squares regression}}{\text{Sum of squares total}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2},$$

where n is the number of observations in the regression, Y_i is an observation on the dependent variable, $\hat{Y}_i$ is the predicted value of the dependent variable based on the independent variables, and $\overline{Y}$is the mean of the dependent variable.

In multiple linear regression, however, R^2 is less appropriate as a measure of a model's goodness of fit. This is because as independent variables are added to the model, R^2 will increase or will stay the same, but it will not decrease. Problems with using R^2 in multiple regression include the following:

- The R^2 cannot provide information on whether the coefficients are statistically significant.
- The R^2 cannot provide information on whether there are biases in the estimated coefficients and predictions.
- The R^2 cannot tell whether the model fit is good. A good model may have a low R^2, as in many asset-pricing models, and a bad model may have a high R^2 due to overfitting and biases in the model.

Overfitting of a regression model is a situation in which the model is too complex, meaning there may be too many independent variables relative to the number of observations in the sample. A result of overfitting is that the coefficients on the independent variables may not represent true relationships with the dependent variable.

An alternative measure of goodness of fit is the **adjusted R^2** ($\overline{R}^2$), which is typically part of the multiple regression output produced by most statistical software packages. A benefit of using the adjusted R^2 is that it does not automatically increase when another independent variable is added to a regression. This is because it adjusts for the degrees of freedom as follows, where k is the number of independent variables:

$$\overline{R}^2 = 1 - \left(\frac{\text{Sum of squares error}/(n-k-1)}{\text{Sum of squares total}/(n-1)}\right). \quad (1)$$

Mathematically, the relation between R^2 and $\overline{R}^2$ is

$$\overline{R}^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)(1-R^2)\right]. \quad (2)$$

Note that if $k \geq 1$, then R^2 is strictly greater than adjusted R^2. Further, the adjusted R^2 may be negative, whereas the R^2 has a minimum of zero.

The following are two key observations about $\overline{R}^2$ when adding a new variable to a regression:

- If the coefficient's t-statistic > |1.0|, then $\overline{R}^2$ increases.
- If the coefficient's t-statistic < |1.0|, then $\overline{R}^2$ decreases.

Note that a t-statistic with an absolute value of 1.0 does not indicate the independent variable is different from zero at typical levels of significance, 5% and 1%. So, adjusted R^2 does not set a very high bar for the statistic to increase.

Consider the regression output provided in Exhibit 1, which shows the results from the regression of portfolio returns on the returns for five hypothetical fundamental factors, which we shall call Factors 1 through 5. The goal of this regression is to identify the factors that best explain the returns on the portfolio.

Exhibit 1: Regression of Portfolio Excess Returns on Five Factors

Multiple R	0.7845
R-Squared	0.6155
Adjusted R-Squared	0.5718
Standard Error	0.0113
Log-Likelihood	-74.054
Observations	50

ANOVA Table

Source	Degrees of freedom	Sum of squares	Mean squares	*F*-statistic	Significance of *F*-statistic
Regression	5	90.6234	18.1247	14.0853	< 0.0000
Residual	44	56.6182	1.2868		
Total	49	147.2416			

					95% confidence interval	
	Coefficient	**Standard error**	***t*-Statistic**	***P*-value**	**Lower bound**	**Upper bound**
Intercept	2.1876	0.1767	−12.3787	0.0000	−2.5437	−1.8314
Factor 1	1.5992	0.2168	7.3756	0.0000	1.1622	2.0361
Factor 2	0.1923	0.7406	0.2596	0.7964	−1.3002	1.6847
Factor 3	−0.7126	0.5854	−1.2172	0.2300	−1.8925	0.4673
Factor 4	3.3376	1.3493	2.4736	0.0173	0.6182	6.0570
Factor 5	−2.6832	8.3919	−0.3197	0.7507	−19.5959	14.2295

CODE: REGRESSION STATISTICS

Using Microsoft Excel

Let depvar be the range of cells for the dependent variable, and let indvar be the range of cells for the independent variables.

=LINEST(depvar,indvar,TRUE,TRUE) or Data Analysis > Regression

Using Python

Let df be the data frame containing the data.

```
import statsmodels.api as sm

from statsmodels.stats.anova import anova_lm

from statsmodels.formula.api import ols

formula='Portfolio ~ Factor1+Factor2+Factor3+Factor4+Factor5'

results=ols(formula,df).fit()

print(results.summary())
```

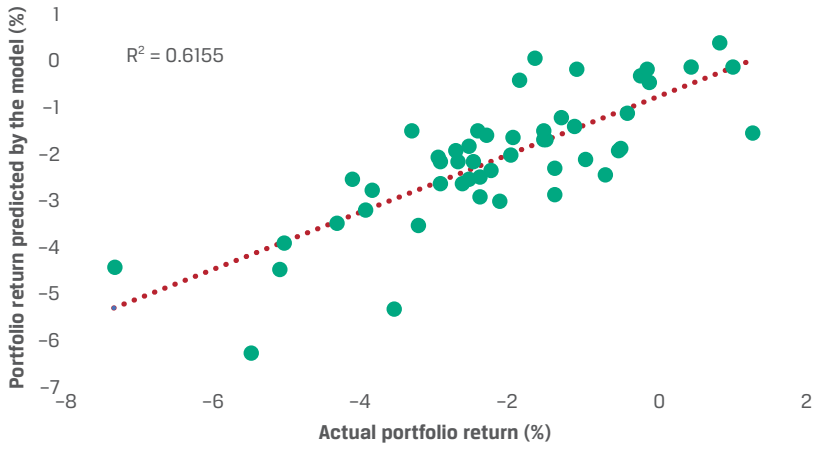
Using R

Let df be the data frame containing the data.

```
model11 <- lm(df$Portfolio ~

df$Factor1+df$Factor2+df$Factor3+df$Factor4+df$Factor5)

anova(model11)

summary(model11)
```

We see in Exhibit 1 that R^2 is 0.6155, or 61.55%, and can we visualize this relationship using the graph in Exhibit 2.

Exhibit 2: Predicted vs. Actual Portfolio Excess Returns Based on Regression of Returns on a Model with Five Factors



We can use the **analysis of variance (ANOVA)** table in Exhibit 1 to describe the model fit. We know from simple regression that the R^2 is the ratio of the sum of squares regression to the sum of squares total. We confirm this as

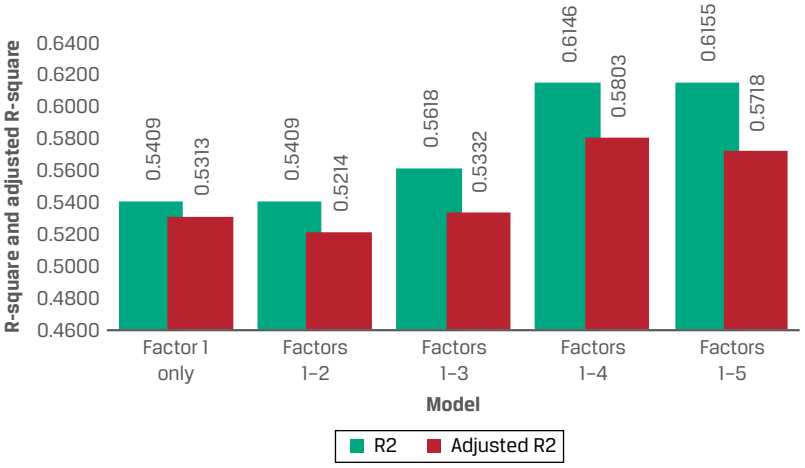
$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{90.6234}{147.2416} = 0.6155$$

and the adjusted R^2 (using Equation 3) as

$$\overline{R}^2 = 1 - \left[\left(\frac{50-1}{50-5-1}\right)\left(1 - \frac{90.6234}{147.2416}\right)\right] = 0.5718.$$

The effect of successively adding each factor to the model is shown in Exhibit 3. The regression of the portfolio returns starts with the returns of Factor 1, then in the next model adds Factor 2, and so on, until all five are included in the full model. Note that with each added variable, the R^2 either stays the same or increases. However, while the adjusted R^2 increases when Factors 3 and 4 are added, it declines when Factors 2 and 5 are added to those models, respectively. This illustrates the relationship between the $|t$-statistic$|$ of the added variable and adjusted R^2.

Exhibit 3: R^2 and Adjusted R^2 for Models Adding Factors to Explain Excess Returns



Importantly, the following should be noted:

- Unlike in simple regression, there is no neat interpretation of the adjusted R^2 in a multiple regression setting in terms of percentage of the dependent variable's variation explained.
- The adjusted R^2 does not address whether the regression coefficients are significant or the predictions are biased; this requires examining residual plots and other statistics.
- R^2 and adjusted R^2 are not generally suitable for testing the significance of the model's fit; for this, we explore the ANOVA further, calculating the *F*-statistic and other goodness-of-fit metrics.

KNOWLEDGE CHECK

You are a junior portfolio manager (PM) reviewing your firm's research on diversified manufacturers. You are considering Model 1, a cross-sectional regression of return on assets (ROA) for a sample of 26 diversified manufacturing companies on capital expenditures scaled by beginning year PPE (CAPEX):

$$\text{Model 1: ROA}_i = b_0 + b_{CAPEX}\text{CAPEX}_i + \varepsilon_i.$$

Multiple *R*	0.9380
R-Squared	0.8799
Adjusted *R*-Squared	0.8749
Standard Error	1.5274
Log-Likelihood	–46.842
Observations	26