CFA Institute®
CFA Program

QUANTITATIVE METHODS

CFA® Program Curriculum
2025 • LEVEL II • VOLUME 1

ISBN 978-1-961409-21-7 (paper)

ISBN 978-1-961409-32-3 (ebook)

August 2024

CONTENTS

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/cbok)
- Topic area weights that indicate the relative exam weightings of the top-level topic areas (www.cfainstitute.org/en/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 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/-/media/documents/support/programs/cfa-and-cipm-los-command-words.ashx.
- The CFA Program curriculum that candidates receive access to 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 curriculum, including the practice questions, is the basis for all exam questions. The curriculum is selected or developed specifically to provide candidates with the knowledge, skills, and abilities reflected in the CBOK.

CFA INSTITUTE LEARNING ECOSYSTEM (LES)

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

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 you can demonstrate the knowledge, skills, and abilities described by the LOS and the assigned reading. Use the LOS as a 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 topics than on others.

ERRATA

The curriculum development process is rigorous and involves multiple rounds of reviews by content experts. Despite our efforts to produce a curriculum that is free of errors, in some instances, we must make corrections. Curriculum errata are periodically updated and posted by exam level and test date 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.

OTHER FEEDBACK

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

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

1

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

LEARNING MODULE OVERVIEW

- 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:

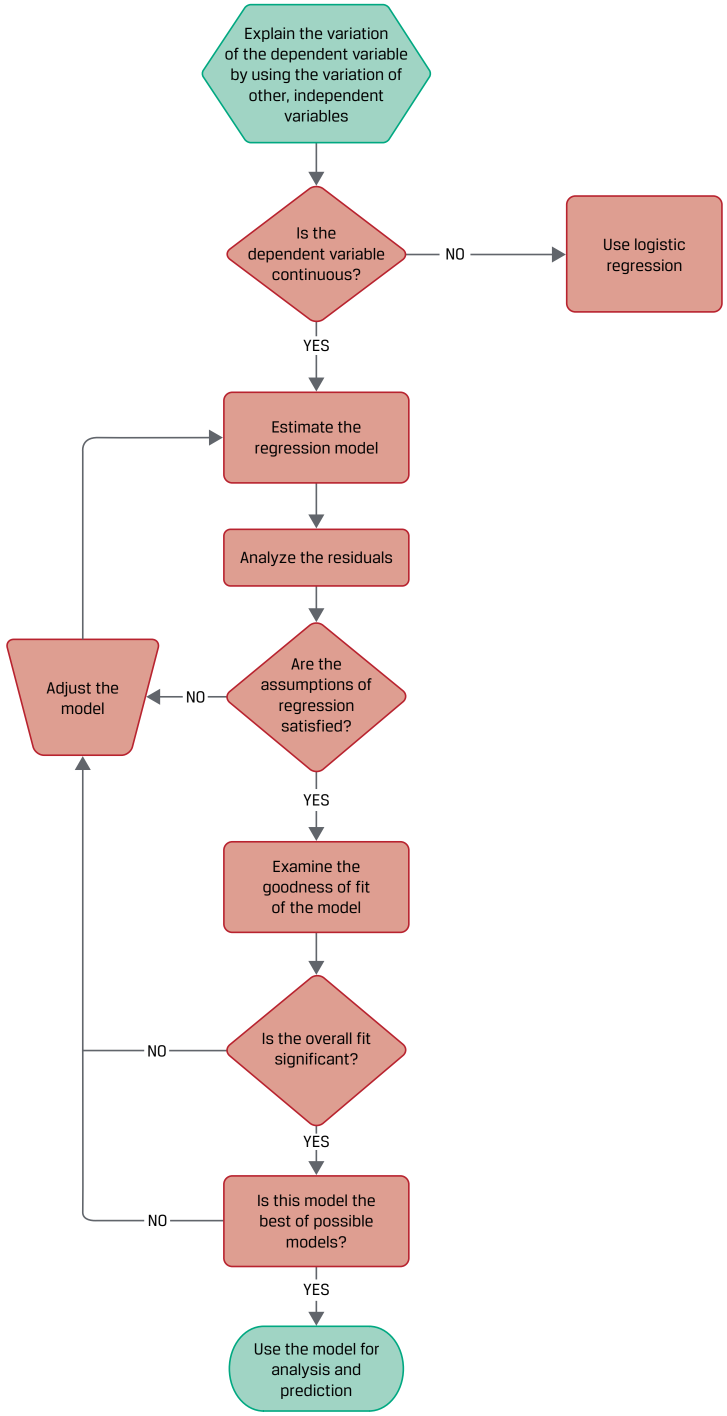
- A portfolio manager wants to understand how returns are influenced by a set of underlying factors; the size effect, the value effect, profitability, and investment aggressiveness. The goal is to estimate a Fama–French five-factor model that will provide an understanding of the factors that are important 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.

3 THE BASICS OF MULTIPLE REGRESSION

☐ 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

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

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

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

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 1.5324%.

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

   $$\text{MKTRF} = 1, \text{SMB} = 4, \text{and HML} = -2.$$

   Solution

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

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

4 ASSUMPTIONS UNDERLYING MULTIPLE LINEAR REGRESSION

☐ 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()
```

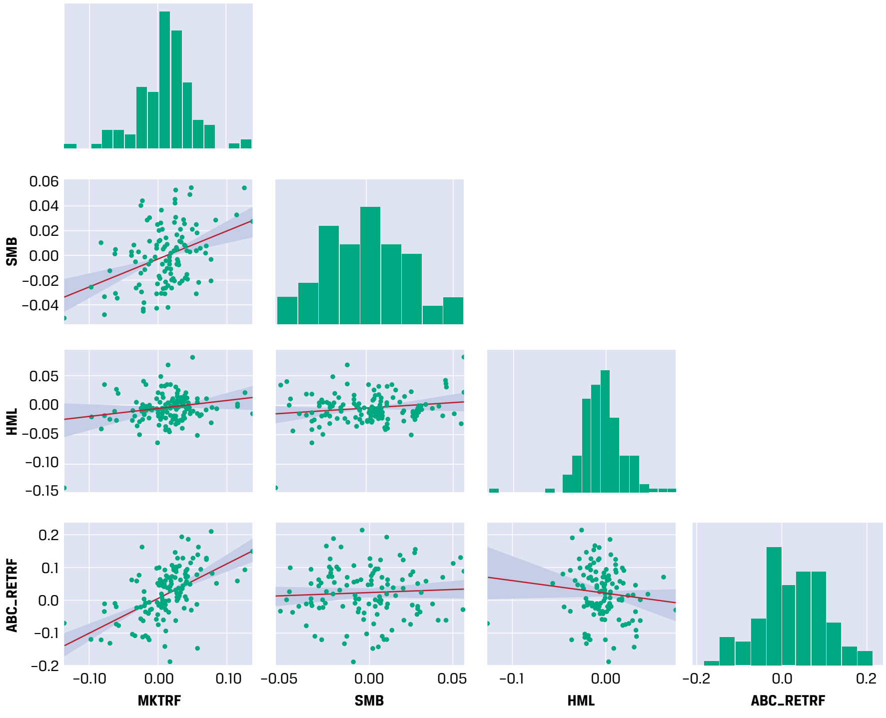
Using R

```
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 4. 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

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

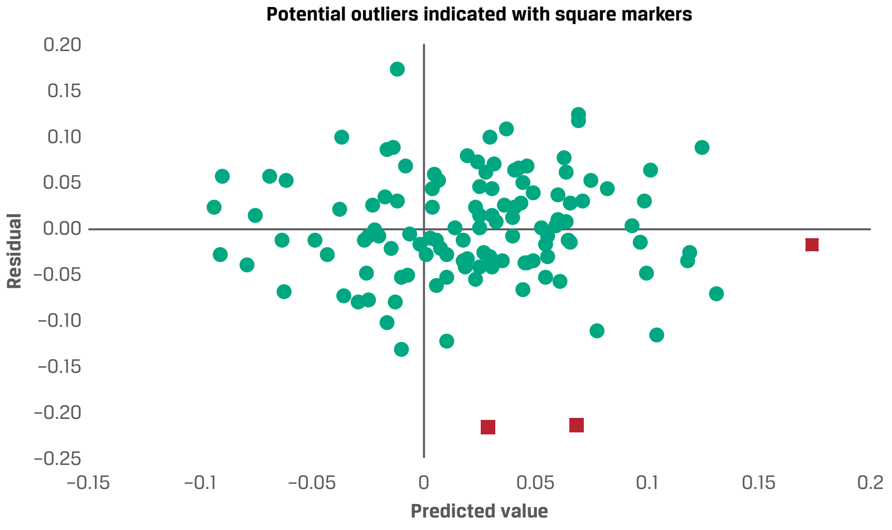
```
df <- read.csv("data.csv")

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 5. We can use this scatterplot to uncover potential assumption violations and to help identify outliers in our data.

Exhibit 5: Residuals vs. Predicted Value of Dependent Variable



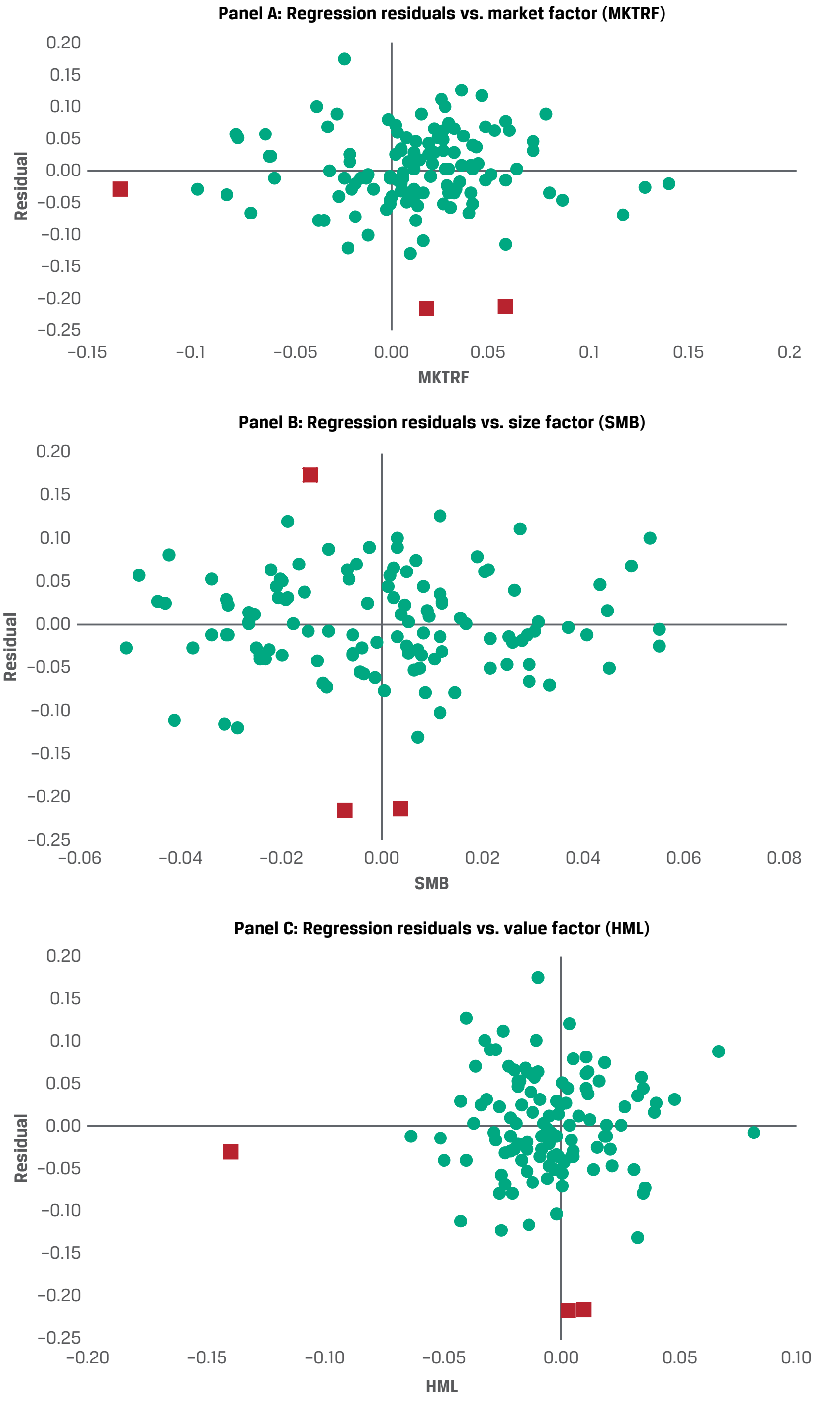
Potential outliers indicated with square markers

As indicated by the line centered near residual value 0.00, a visual inspection of Exhibit 5 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 6 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 6, as indicated by the square markers.

Exhibit 6: Regression Residuals vs. Factors (Independent Variables)



CODE: RESIDUAL ANALYSIS

Using Python

```
import pandas as pd
```

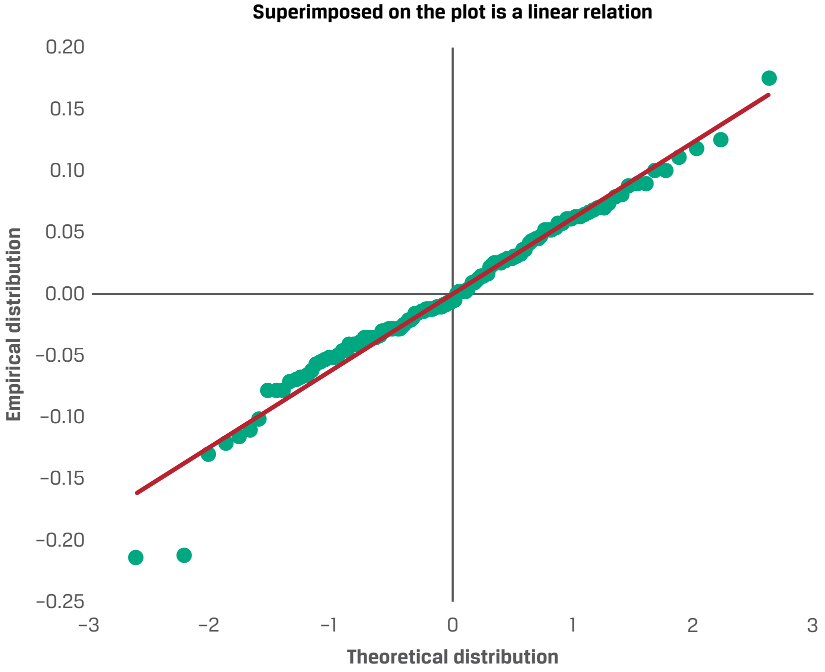
```
import matplotlib.pyplot as plt
import statsmodels.api as sm
import numpy as np
df = pd.read_csv(“data.csv”,parse_dates=True,index_col=0)
model = ols('ABC_RETRF ~ MKTRF+SMB+HML',data=df).fit()
fig = sm.graphics.plot_partregress_grid(model)
fig.tight_layout(pad=1.0)
plt.show()
fig = sm.graphics.plot_ccpr_grid(model)
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 7 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 7: 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.