2 MONITORING THE MACROECONOMIC ENVIRONMENT

CONTENTS

LEARNING OBJECTIVES

1. Understand the influence of the macroeconomic environment on business decisions
2. Define the inputs of a linear regression model
3. Interpret the output of a regression model
4. Predict the impact on business using linear regression

2.1 DEFINING THE MACROECONOMIC ENVIRONMENT

FedEx can make life easier by taking the stress out of getting your packages to remote locations "absolutely, positively on time." One might assume that all FedEx needs to do is to keep its planes and delivery systems well oiled. But there might be other factors, some surprisingly not in FedEx's control, that can affect its business functioning. One set of factors that can affect any business is categorized as economic factors, which can occur at the micro or macro level. Both micro- and macroeconomic factors tend to be outside the control of a particular business but nonetheless can significantly affect its revenue, sales, or demand for its products. Note that consumer demand (or just demand) is defined as the amount of goods that people are willing to buy at a given price.

Microeconomic factors are specific to a particular business and can include market size, distribution channel, suppliers, and competitors. *Macroeconomic factors*, on the other hand, tend to affect the entire economy. They can involve the current unemployment level, interest rates, taxes, gross domestic product (GDP), inflation, consumer confidence, and so on. Therefore, going back to our example of FedEx, it is important for FedEx to evaluate factors not only internal to it, such as price and profit, but also external economic factors such as the current employment rate. If more people are employed, they will have more money to spend on orders from online retailers; hence, FedEx will benefit. However, if unemployment increases, FedEx will need to include this change in its predictions.

2.1.1 Gross Domestic Product

Let's consider an important economic indicator of a country: its GDP. The GDP displays the financial health of a country because it includes the sum total of all goods and services produced in that country during a certain time period (e.g., a month or a quarter or a year). It can be calculated by summing up what everyone earned or what everyone spent. An increasing GDP indicates a growing economy that is likely to boost consumer spending, and businesses are likely to benefit.

2.1.2 Inflation

Similarly, *inflation* can affect many businesses. Inflation occurs in a country if money loses its value (i.e., the same amount of money is able to purchase fewer goods). It results in prices increasing and people being unable to purchase at the level at which they purchased previously. Inflation

may be caused either because the supply of products is not able to keep up with the demand or because the cost of products has risen significantly. Inflation increases costs to the business in terms of higher prices for supplies, higher rent, and more being paid for utilities. An extreme case of hyperinflation was witnessed by Zimbabwe in 2008 when the government started transacting in foreign currencies rather than the Zimbabwean dollar, which had been devalued significantly. However, a little bit of inflation is good for the economy since deflation, the opposite of inflation, can be quite harmful too. Deflation occurs when prices start falling. Falling prices motivate people to hold on to their savings because they think that future prices will fall further so their money will get them more. This causes a cycle of people holding on to their money, with the result that the economy—which relies on consumer spending—is thwarted and slows down. A slowing economy can result in a recession or, worst case, a depression. Japan, one of the prominent world economies, entered a period of slowing growth followed by deflation in the early 1990s.

2.1.3 Interest Rate

Another important economic indicator is the *interest rate*; this is the rate the lender levies on the borrower. The federal interest rate influences several other interest rates, such as bank lending rates to individuals, businesses, or other financial institutions. When interest rates are high, consumer spending decreases because there is a greater incentive to save rather than spend. Moreover, people may be less likely to borrow money to buy cars or homes. For businesses, debts can become more expensive because if the businesses borrow from a bank, they have to pay a higher interest rate. This makes them more cautious in borrowing more to grow. However, if interest rates are lower, businesses are motivated to borrow and invest in growth. Therefore, it is common to see central banks like the Federal Reserve increasing the interest rate if the economy seems to be growing too fast, or reducing the interest rate in order to facilitate economic growth. When interest rates are high, luxury product manufacturers are more likely to be affected compared to producers of staple products because the former rely more on people having access to cheap debt. People consider buying luxury products (e.g., handbags, shoes, clothing) when they feel they have disposable income that they don't need to spend on staple products like groceries, rent, and transportation.

2.1.4 Relevant Macroeconomic Environment

Examples abound of industries and companies that failed to consider the larger economic environment and incurred huge losses. In the early '80s, the petroleum industry invested more than $500 billion because it forecast that worldwide demand for oil would increase from 52 to 60 million barrels per day. Prices were expected to rise by 50%, so the investment made sense. But these predictions were based on the rosy outlook of rapid growth in industrialized economies. Reality did not match the optimistic outlook; a slowdown in GDP ensued in the industrialized economies, oil demand fell to 46 million barrels, and the petroleum industry—including drilling, refining, and shipping—was badly hit.

Factors in the economic environment that can affect a business vary. For a farmer, market size, price of fertilizers, or taxes might be more important than consumer confidence. On the other hand, the unemployment rate or interest rate may be more important for financial institutions.

Hence, keeping a close eye on the economic factors that are relevant to one's business is important. To avoid unpleasant shocks or avoid losing out on opportunities, firms must strive to incorporate these macroeconomic factors in their predictions. And it has become easier to do this over the years with the wide availability of different forms of data on economic indicators.

Businesses typically monitor the economic indicators that in the past have had the most impact on demand for their products or services. For Delta Air Lines, oil prices and GDP are likely to inform them whether to expect increased or decreased demand. Hence, changes in the macroeconomic environment can help a business predict future demand.

2.2 IMPACT OF MACROECONOMIC FACTORS ON BUSINESS OUTCOMES

All businesses are interested in improving their performance; they would like to sell more, earn more, and grow more in the next time period compared to the current time period. A time period could be a week, month, quarter, or year. In order to improve, businesses need to know what factors can help them perform better. In other words, they need to get better at predicting. From small manufacturers to large global suppliers, all need to be accurate in terms of predicting, for example, how many units of machinery they will be selling to their buyers in the next time period. Businesses need to understand both (1) how macroeconomic factors influence their sales and (2) how to predict future demand. For instance, if we considered only macroeconomic factors, would we predict that the Federal Reserve's interest rate will impact units ordered? One possibility could be that as the interest rate increases, businesses will be less likely to take out a loan to invest in further production, which could result in no change in the units produced or ordered. A second possibility is that buyers will take a more conservative approach and not order new units. This could result in the orders actually decreasing. Therefore, there is a good possibility that as the interest rate increases, units ordered decrease. Similarly, let's consider what happens when the GDP changes. GDP increases signal that the economy is doing well and that businesses are more likely to invest in growth strategies. This can result in units ordered increasing. Conversely, when GDP decreases, the economy shrinks, and orders may decrease.

Let's consider the influence of one more external variable of inflation. One possibility is that when inflation is very low, near zero, there is not much economic growth. In such a case, units ordered might actually decrease. As inflation increases a bit, orders might increase but might again decrease when inflation increases to a very high value. However, all of these are guesses without much data to back them up. If we had data about units ordered in past time periods, we could see how macroeconomic variables have impacted the number of units ordered. These data could be used to develop a model, which could then be used to make predictions for future time periods.

2.3 REGRESSION FOR PREDICTION

Analytic methods such as regression have commonly been used for prediction. Regression is considered the go-to method in a multitude of data analysis situations. Its popularity lies in the fact that its results are easy to interpret. This means one can find out which input variable has the most, the

least, or average influence on the output variable. The results also inform us whether the influence of a variable is positive or negative—for example, does an increasing unemployment rate increase or decrease demand for a product, and by how much? Regression allows decision-makers to clearly see the magnitude and direction of the impact of any input variable.

In the next section, we discuss linear regression as a way to predict the influence of macroeconomic factors on demand for products.

2.3.1 Linear Regression: Error-Based Learning

The simplest method of prediction is linear regression. It not only helps us develop a model that can explain the influence of macroeconomic variables on units ordered but also helps us make predictions about how many units are likely to be ordered in future time periods. We begin by understanding what is needed to run a linear regression.

2.3.2 The Inputs in a Linear Regression

In a regression, there are two types of variables: *predictors* and *outcomes*. Sometimes people refer to them as independent and dependent variables. The predictor (independent) variables are those that influence the outcome (dependent) variable. If a business wants to know the influence of interest rate on units ordered, then interest rate is the predictor variable, and units ordered is the outcome variable. This is because we are assuming that the units ordered depend on the interest rate and not the other way around.

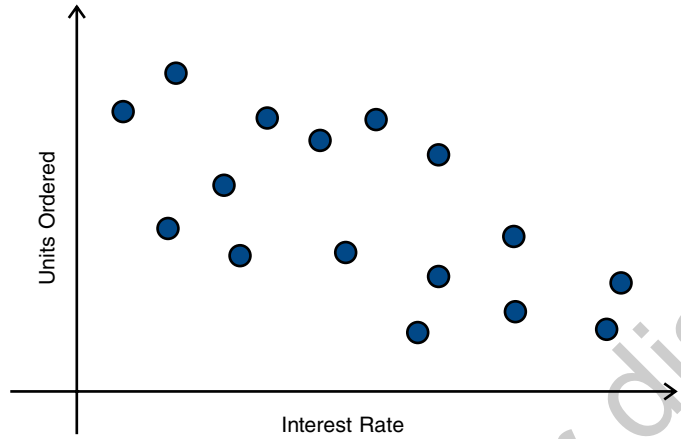
If the values that an outcome variable can take range over a continuum—for example, from 0 to 200 or 1 to 7—then we use a linear regression. Sometimes the outcome variable is not on a continuum. Instead, it appears as a categorical value, such as in predicting whether or not someone will buy a product. In this example, there are only two outcomes possible: yes (will buy) or no (will not buy). We cannot say yes has a higher numerical value than no, or vice versa. Therefore, if the outcome variable takes categorical values such as yes/no, high/low, or red/blue, then we use a logistic regression.

If a linear regression is examining the influence of one predictor on an outcome, it is a *univariate regression*. In our example, a univariate regression would be appropriate if we wanted to know only the influence of interest rate on units ordered. But it is more common to study the influence of several predictors on an outcome. Such a regression is referred to as a *multiple regression*. If we studied the influence of GDP, interest rate, and inflation on units ordered, then we would need to run a multiple regression. We will next discuss how a linear regression works and then use it for prediction.

2.3.3 Visualizing the Data With a Scatterplot

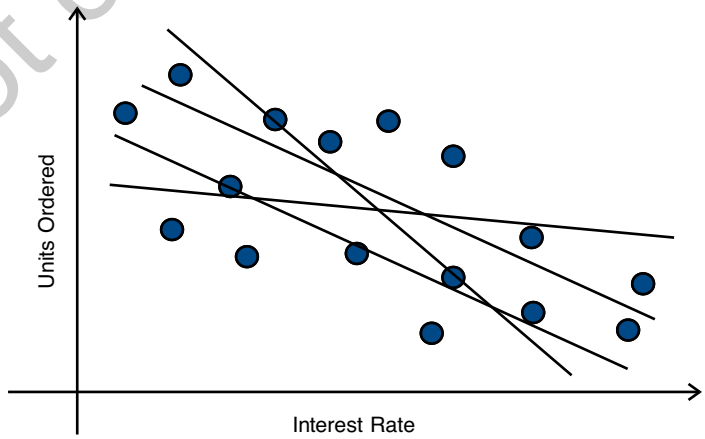
The first step usually is to visualize the data. For a linear regression, we can use a scatterplot to see how the outcome changes for each unit change in the predictor. For instance, we can plot the influence of interest on units ordered. The data could possibly look like the scatterplot in Figure 2.1.

FIGURE 2.1 ■ Visualizing the Data With a Scatterplot



Looking at the scatterplot, what can one infer about the relationship between interest rate and units ordered? Is it positive or negative? Do units ordered increase with increases in the interest rate? From the scatterplot, we can guess that the relationship looks negative: Units ordered decrease as the interest rate increases. But it is still a guess, and we want to be sure. Let's draw a line through the scatterplot. It is easier to interpret the data if they can be represented as a line rather than as several points on the scatterplot. If we can represent the scatterplot in the form of a line, the equation of the line can serve as a model to help us understand the relationship between units ordered and interest rate. For instance, if the line slopes downward, we will infer that units ordered decrease when the interest rate increases, or vice versa. Even more important, we can use the regression model to make future predictions about units ordered based on changes in the interest rate. However, after we proceed to draw a line, we realize that it is possible to draw several lines through the scatterplot (see Figure 2.2).

FIGURE 2.2 ■ Multiple Regression Lines



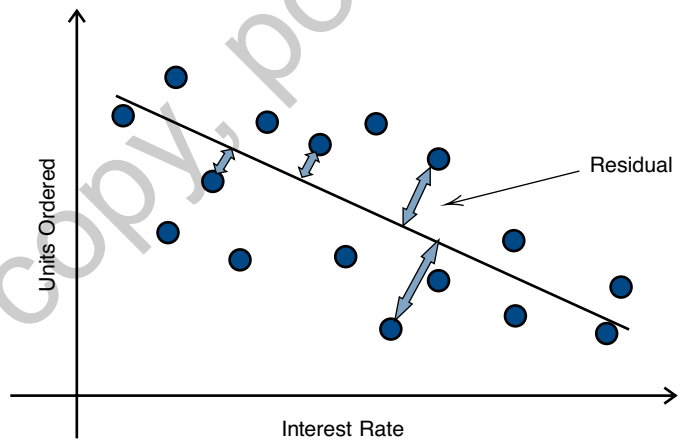
So, which of the many lines that can be drawn through the scatterplot is the one we want? We want the line that captures the predictive capabilities of the scatterplot as closely as possible (i.e., the line that best represents the relationship between the predictor and the outcome). We can use *ordinary least squares* (OLS) to find the line that best fits the data. Let's digress slightly to understand OLS.

2.3.4 Ordinary Least Squares

Regression is called error-based learning because it attempts to ensure that the difference (or error) between the values of the outcome variable predicted by the regression model (the regression line in this case) and the actual values is as low as possible. What do we mean by this? If you look at Figure 2.3, it shows the scatterplot with the best-fitting line. That line predicts how many units will be ordered for each value of interest rate. However, the line does not go through each of the data points and, hence, is not an exact prediction. For some data points the line represents an underestimation of the units ordered, while for other data points it represents an overestimation.

These under- and overestimations of the data points by the regression line are called residuals. If we just add up the residuals to find the total error, then some of the over- and underestimations will cancel each other out. Therefore, to get an accurate estimate, we square the residuals. From the many possible lines, OLS finds the line that has the minimum value of these squared residuals. Hence, OLS is also called least-squares regression.

FIGURE 2.3 ■ OLS Regression Graphical Depiction With Residuals



2.3.5 Regression Model

The goal of regression analysis is to find the line that best captures the relationship between the predictor and the outcome variables. When we have just one outcome and one predictor variable, the best-fitting line can be expressed as an equation just like any line. In its simplest form, the equation of a line is $y = mx + b$, where m represents the slope and b represents the intercept.

For a linear regression line, the equation is

$$\text{Outcome} = \hat{\beta}_0 + \hat{\beta}_1 * \text{Predictor}$$

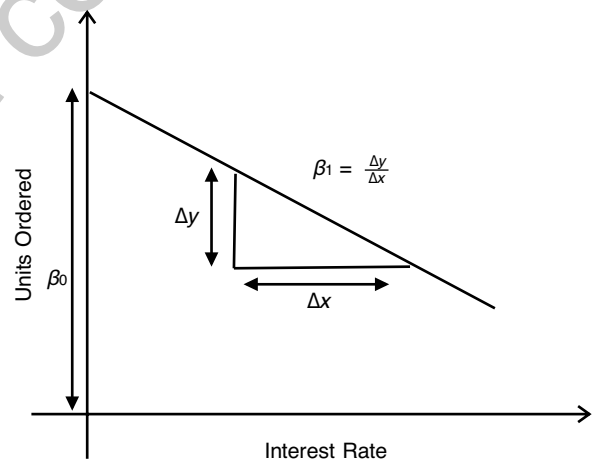
If we are interested in understanding the relationship between changes in the interest rate on units ordered, then the equation of the best-fitting line will be

$$\text{Units Ordered} = \hat{\beta}_0 + \hat{\beta}_1 * \text{Interest Rate}$$

The $\hat{\beta}$s (referred to as beta hats) are called the **coefficient estimates** because the regression equation is estimating the relationship between the predictor and outcome from the sample. While the population is each and every point in the world of the predictor and the outcome, the sample is a subgroup of the population data points that is available to us for regression analysis.

Look at Figure 2.4. It graphically explains different parts of the regression model. The predictor is on the x-axis, and the outcome is on the y-axis; therefore, they are generally represented as x and y variables, respectively, in the regression model. $\hat{\beta}_0$ is called the intercept and shows what the number of units ordered would be when there is no influence of any predictor variable. $\hat{\beta}_1$ shows the magnitude and direction of the influence of interest rate on units ordered. Magnitude means the degree or scale of influence (i.e., how much of an influence the predictor has on the outcome). It can also be defined as the rate of change in y for a unit change in x, aptly captured as Δ_y/Δ_x, where Δ is called delta. The nature of influence shows whether the influence is positive (outcome value increases with an increase in predictor values) or negative (outcome value decreases with an increase in predictor values). In the graph, we can see that the nature of influence is negative because as the interest rate increases, the number of units ordered decreases.

FIGURE 2.4 ■ Coefficient of Determination Depiction



Let's assume we ran a univariate regression and found that, as predicted by our regression model, the estimate of $\hat{\beta}_0$ was 4, and that it was –3 for $\hat{\beta}_1$. The $\hat{\beta}_1$ coefficient informs us that the direction of influence is negative—that is, when the interest rate increases, the number of units ordered decreases. It also informs us about the magnitude of influence: When the interest rate falls by 1 unit, the number of units ordered increases by 3 units. On the other hand, if $\hat{\beta}_1$ had been 3, we would have inferred that the influence of the predictor on the outcome is positive (i.e., when the interest rate increases by 1 unit, the number of units ordered increases by 3 units). Therefore, the beta coefficient indicates the magnitude and direction of influence of the predictor on the outcome. It is important to note that once we have the regression model, we can use it to predict the outcome variable for different possible values of the predictor.

2.3.6 Multiple Regression

If we have several predictor variables that we think would influence the outcome variable, we run a multiple regression. An obvious question that comes to one's mind is why not run several univariate regressions and examine the influence of each predictor on the outcome variable. It is not a good idea to do so because then we are assuming that the predictors have an independent influence on the outcome. In reality many factors have a combined influence on the outcome variable. Therefore, when we have several predictor variables, we run a multiple regression.

For our example, we can run a multiple regression with the number of units ordered as the outcome variable and the interest rate, GDP, and inflation as the predictor variables. The multiple regression equation would be as follows:

$$\text{Units Ordered} = \beta_0 + \beta_1 * \text{Interest Rate} + \beta_2 * \text{GDP} + \beta_3 * \text{Inflation}$$

2.4 APPLICATION OF LINEAR REGRESSION FOR PREDICTION

Now that we understand how a regression works, we can apply it to a dataset and understand how macroeconomic factors can influence a business. Let's work with the example of medical equipment manufacturer MedDiagnostics, which manufactures and sells diagnostic imaging devices like X-ray systems and computed tomography (CT) scanners to hospitals and independent diagnostic centers across 40 different countries. (This example uses hypothetical data.) Therefore, it is in a business-to-business setting where it sells to other businesses rather than individuals. Beyond internal factors specific to the organization, macroeconomic factors can significantly affect units ordered. Therefore, MedDiagnostics wants to determine whether changes in macroeconomic factors such as GDP, interest rate, and inflation in a country will affect units ordered. MedDiagnostics has data from all 40 countries it serves from the last quarter (a three-month time period). The predictor variables are inflation, interest rate, and GDP for each country; the outcome variable is units ordered. Each of the variables is continuous (i.e., none of them is categorical in nature).

Since our aim is to predict the influence of macroeconomic variables on units ordered, we next partition the data into train and test sets.

2.4.1 Partitioning Data Into Train and Test Sets

The main aim of any predictive model is to be able to learn the relationship between the predictor variables and the outcome variable. The model is then used to predict values of the outcome variable for different values of the predictor variables that it has not seen previously, with as little error as possible. Therefore, the data sample from which the model learns the relationship is called the train dataset, and the data sample from which it predicts is called the test (or validation or hold-out) dataset. Such a procedure of testing on new datasets helps assess the predictive ability of the model. If we use all the data available to us to train the model to understand the relationship between the predictor and outcome, then we will not know how well the model will perform on a dataset it has not *seen* before. We might have *overfit* the model to the train set, and it may perform poorly on any new data sample. Such overfitting could occur because of certain characteristics specific to the train set that do not generalize to other sample datasets. Therefore, a common practice is to split the available data into two and use one part to train the model and the other part to test its predictive ability.

Following this procedure to test the model's predictive ability, we partition our data into train and test sets. Since we have the actual value of the outcome variable (i.e., the number of units ordered in a time period), we can easily compare them against the predicted values to understand how similar the linear regression model's predictions are to the observed/actual values. We specify that 70% of the data should be used as train data and 30% as test data; this spilt randomly picks up 70% of the data as train data.

Next, we run a linear multiple regression analysis on the train dataset.

```
## # A tibble: 4 × 5
##   term             estimate std.error statistic      p.value
##   <chr>               <dbl>     <dbl>     <dbl>        <dbl>
## 1 (Intercept)         5.13      0.834      6.15      0.00000235
## 2 inflation           0.102     0.375      0.272     0.788
## 3 interest_rate      -2.37      0.555     -4.27      0.000269
## 4 GDP_growth_rate     1.63      0.392      4.15      0.000358

## # A tibble: 1 × 2
##   r.squared  adj.r.squared
##   <dbl>              <dbl>
## 1 0.537              0.480
```

2.4.2 Interpreting the Regression Output

2.4.2.1 β Coefficient Estimates

When we look at the R output, it gives us information about the coefficient estimates of the intercept and the predictors of GDP, interest rate, and inflation. The first column labeled *term*

lists the names of the predictor variables beginning with the intercept. The second column labeled *estimate* provides the coefficient estimate of the predictors. The coefficient estimate value for inflation is 0.1. This is the beta coefficient that informs us of the direction and magnitude of influence of inflation on units ordered. The sign of the coefficient informs us that the direction of influence is positive (i.e., when inflation increases, units ordered increase). It also tells us about the magnitude of influence: When inflation increases by 1 unit, units ordered increase by 0.1 unit. Similarly, we can infer that when the interest rate increases by 1 unit, the number of units ordered decrease by 2.37 units because the beta coefficient is –2.37. The GDP variable captures the rate of change in the GDP. Hence, when the GDP increases by 1 unit, units ordered decrease by 1.63 units because the beta coefficient is 1.63.

The third column provides the **standard error** of the beta coefficient estimates. Intuitively, the standard error captures how precise the estimate of the coefficient is. A smaller value of standard error shows a more precise estimate than a larger value.

2.4.2.2 Significance Testing

The fourth column called the *statistic* reports the t statistic, and the fifth column is the p value. The t statistic is estimated by dividing the coefficient by its standard error. Let's understand the t statistic and p value together because they are related. While the regression provides us with a beta coefficient associated with the predictor, how do we know whether this predictor truly has any effect on the number of units ordered? For this we use the t statistic and p value together to find out whether the predictor has a significant influence on the outcome (i.e., we perform *significance testing*). A p value of less than 0.05 suggests a significant influence, and informs us that the chances of finding the relationship between predictor and outcome in the data when there is no such true relationship in the population is only 5%. Keep in mind that $p < 0.05$ is an arbitrary threshold. One can set a more conservative threshold of 0.01. For the predictor of inflation, we find that the t statistic is 0.27 and the p value is 0.79. Therefore, if we had to interpret the result, we would say the following: Inflation does not have a significant influence on units ordered. However, the results indicate that both the interest rate and GDP have a significant influence on units ordered.

2.4.2.3 Coefficient of Determination

We can find the regression model that best fits the data, but how well it truly explains the relation between predictor and outcome variables is a different question. For this we evaluate the goodness of fit of the regression model using R^2, the *coefficient of determination*. R^2 informs us how much of the variation in the outcome variable can be explained by the predictor variables that we have used. R^2 can vary between 0 and 1. For example, if we get a value of 0.6403, it means that about 64% of the variation in the outcome variable can be explained by the predictors included in our regression model, while 36% is not explained.

Adding more predictor variables increases R^2 but makes the model less generalizable and leads to the problem of overfitting. When we add many predictors so that we can increase the predictive power of our regression model, we may get a model that perfectly fits a specific data sample but would do poorly with other data samples. An overfit model captures not only the

relationship we are studying but all other quirks or noise of this specific dataset. Other datasets will have other types of noise, which this regression model will fare poorly in capturing. So a parsimonious model in which we have fewer predictors is better because it is more generalizable. Generalizability indicates how well the regression model will perform on other datasets that are studying the same relationship. In order to adjust for the number of predictors included in the regression model, we can look at adjusted R^2. In our case the following output shows that R^2 is 0.537 and adjusted R^2 is 0.48, which indicates some adjustment because of the three predictor variables that we have included.

```
## # A tibble: 1 × 2
##   r.squared   adj.r.squared
##     <dbl>             <dbl>
## 1   0.537             0.480
```

2.4.3 Performing Prediction With Linear Regression

Once we have interpreted the results of the train data, the next task is to assess how well our regression model predicts units ordered in the test data. This is an important step because in the absence of test data, we will have no way of finding out if we have overfit the model and how well our model will do with new data points.

```
## # A tibble: 1 × 3
##   .metric  .estimator   .estimate
##   <chr>    <chr>            <dbl>
## 1 rmse     standard          1.13

## # A tibble: 1 × 3
##   .metric  .estimator   .estimate
##   <chr>    <chr>            <dbl>
## 1 rsq      standard         0.773

## # A tibble: 1 × 3
##   .metric  .estimator   .estimate
##   <chr>    <chr>            <dbl>
## 1 mae      standard         0.881
```

The output includes `rmse`, `rsq`, and `mae`. Let's understand what each one represents. `rmse` stands for root mean square error and helps capture how close the values estimated by the regression model are to the actual (i.e., observed) values of the outcome variable. Hence, `rmse` is calculated on the test dataset and not the train dataset because we are determining

the predictive ability of our regression model. The difference between the actual and estimated values are called residuals. To prevent residual values from canceling each other out, the residuals are squared and added together to find out how close the regression model's estimates are to the actual outcome variable values. The squared residuals are divided by the sample size (i.e., the number of observations) of the test data to obtain the average value of the residuals. `rmse` measures how spread out these residuals are (i.e., `rmse` is the square root of the average value of the residuals). Lower values of `rmse` are considered better and indicate data points close to the regression line (in the univariate case). Higher `rmse` values indicate a poorly fitting regression line with points spread away from the line. A common question concerns what is a good `rmse` value. To determine this, remember that `rmse` values depend on the value of the outcome variable. If the outcome variable values range from 1 to 200, then a `rmse` value of 0.75 is pretty low; however, if the outcome value ranges from 0 to 1, then 0.75 is quite a high `rmse` value. In our case, since units ordered range from 2 to 12, a value of 1.13 is not bad.

`rsq` or *R*-squared is the correlation between actual (i.e., observed) values of orders and the values predicted by the regression model. The value is 0.773.

`mae` denotes mean absolute error and captures the same meaning as `rmse`. One concern with `rmse` is that when large residuals are squared, they can dominate its value (i.e., large residuals get weighed more in `rmse`). Hence, to address this concern, we can calculate the absolute value of the residuals, sum them, and then divide by the sample size to obtain `mae`. The `mae` in our test dataset is 0.881.

2.5 A FEW THINGS TO REMEMBER

Let's go over a few things we need to keep in mind while using linear regression.

- When there is more than one predictor variable, don't run many univariate regressions; always run a multiple regression.
- There is no harm in standardizing the data; it helps in comparing the interpretations across different predictor variables that may be using different scales or units.
- However, standardizing might make interpretations a bit difficult. Keep in mind the parsimony-versus-fit argument when adding or removing predictor variables: Adding predictors can improve R^2 but reduces the generalizability of the regression model.
- There are times when running a linear regression might give inaccurate results. One reason could be the choice of predictor variables that are highly correlated with each other. This results in a problem called multicollinearity. We discuss the use of penalized regression to address multicollinearity in Chapter 10. Similarly, data collected across time may suffer from an issue of dependence called autocorrelation.

Knowing which external variables can impact a business can help a firm formulate strategies and plan for the future. As we saw in our analysis of the impact of macroeconomic factors on units ordered, knowing the magnitude and nature of impact of certain variables allows us to focus on the ones that exert a significant influence and to pay less attention to those that may seem to be important but do not have a significant influence.

In sum, regression is a sound go-to method for conducting several types of analysis such as interpretation and prediction. The code to analyze the MedDiagnostics dataset is available in the next section in R.

2.6 IMPLEMENTATION USING R: PREDICTING UNITS ORDERED FOR MEDDIAGNOSTICS

MedDiagnostics has data from 40 different countries. The data are in a comma-separated values (.csv) file that has columns showing the predictor variables of inflation, interest rate, and GDP and the outcome variable of units ordered. Each of the rows captures a different country. Each of the variables is continuous (i.e., none of them is categorical in nature).

We will perform the linear regression analysis using R software.[1]

After loading the tidymodels package,[2] let's import the dataset.

```
library(tidymodels)
macro <- read.csv(url
("http://data.mishra.us/files/chapter_macroeconomic/macro_environ.csv"))
```

Next, we partition the data into train and test sets in order to avoid *overfitting*. Since we have the actual value of the outcome variables (i.e., the number of units ordered in a time period), we can easily compare them with the predicted values to understand how close our linear regression model's predictions are.

We use the `initial_split` function. This function randomly splits a dataset into train data and test data. However, it requires us to specify what percentage of the total data should be randomly assigned to the train set versus the test set. Using `prop = 0.7`, we specify that 70% of the data should be used as train data and 30% as test data. Since the `initial_split` function randomly splits data, we use R's `set.seed` function to create reproducible results. You may choose to set different `set.seed` values, but given the dataset size coefficient, estimates from your analysis will be slightly different.

```
set.seed(123)
datasplit <- initial_split(macro, prop =0.7,
                                  strata = orders)
trainData <- training(datasplit)
testData  <- testing(datasplit)
```

2.6.1 Multiple Regression on the Train Dataset

Next we run a linear multiple regression analysis on the train dataset. `linear_reg(engine = "lm")` defines an ordinary least-squares regression model. orders~ means variable order is the outcome variable, and `.` means that all variables other than order are predictor variables. It is equivalent to writing `order ~ inflation + interest_rate + GDP_growth_rate`. Using the `tidy()` function, we print the results of the linear regression analysis.

```
lin_model <- linear_reg(engine = "lm")

linear_fit <-
  lin_model %>%
  fit(orders ~ ., data = trainData)

tidy(linear_fit)

## # A tibble: 4 × 5
##   term             estimate  std.error  statistic  p.value
##   <chr>            <dbl>     <dbl>      <dbl>          <dbl>
## 1 (Intercept)       5.13     0.834       6.15      0.00000235
## 2 inflation         0.102    0.375       0.272     0.788
## 3 interest_rate    -2.37     0.555      -4.27      0.000269
## 4 GDP_growth_rate   1.63     0.392       4.15      0.000358
```

The R output gives us information about the coefficient estimates of the intercept and the predictors of GDP, interest rate, and inflation. The β coefficient estimate value for inflation is 0.1 (i.e., when inflation increases by 1 unit, units ordered increase by 0.1 units). Similarly, we can infer that when interest rate increases by 1 unit, units ordered decrease by –2.37 units because the β coefficient is –2.37. The GDP variable captures the rate of change in GDP. Hence, when GDP increases by 1 unit, units ordered decrease by 1.63 units because the beta coefficient is 1.63.

The third column provides the standard error of the beta coefficient estimates. The fourth column is called the *statistic*, and the fifth column is the p value. For the predictor of inflation, we find that the t statistic is 0.27 and the p is 0.79. Therefore, inflation does not have a statistically significant influence on units ordered. However, the results indicate that both interest rate and GDP have a significant influence on units ordered.

Using `glance()`, we can print R^2 and other measures of model performance. For instance, an F test of overall significance (defined with `statistic` and p `value`) shows whether our linear

regression model with three predictor variables provides a better fit with the train data than an intercept-only model with no predictor variables.

```
glance(linear_fit)
## # A tibble: 1 × 12
##       r.squared adj.r.squared sigma statistic p.value  df    logLik AIC   BIC
##          <dbl>       <dbl>     <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl>
## 1        0.537       0.480     1.97  9.30      0.000292 3     -56.6  123.  130.
## #...  3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

If you want to print traditional linear regression summary tables, you can use the following code. The output of this code is essentially a combined version of what `tidy(linear_fit)` and `glance(linear_fit)` have created.

```
linear_fit %>%
 pluck("fit") %>%
 summary()

##
## Call:
## stats::lm(formula = orders ~ ., data = data)
##
## Residuals:
##  Min                          1Q     Median        3Q       Max
## -3.14574                  -1.09321  -0.08497   1.11301   3.04204
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.1301    0.8339     6.152    2.35e-06  ***
## inflation        0.1019    0.3745     0.272    0.787806
## interest_rate   -2.3670    0.5549    -4.265    0.000269  ***
## GDP_growth_      1.6283    0.3921     4.153    0.000358  ***
  rate
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.972 on 24 degrees of freedom
## Multiple R-squared: 0.5375, Adjusted R-squared: 0.4796
## F-statistic: 9.296 on 3 and 24 DF, p-value: 0.0002924
```

2.6.2 Performing Prediction With Linear Regression

We can assess how well our regression model predicts units ordered in the test data. The `predict()` function helps us predict values of orders based on the variables of `inflation`, `interest_rate`, and `GDP_growth_rate` in the test data. We then combine the prediction made by the regression model (.pred) and the true values of the variable orders. This helps us to estimate key metrics such as `rmse`, `rsq`, and `mae` that inform us about the performance of our model on test data.

```
order_pred <- predict(linear_fit, new_data = testData)

prediction_comparison <- testData %>%
    bind_cols(order_pred)

prediction_comparison %>%
   rmse(truth = orders, estimate = .pred)

## # A tibble: 1 × 3
##   .metric  .estimator  .estimate
##   <chr>    <chr>           <dbl>
## 1 rmse     standard        1.13

prediction_comparison %>%
    rsq(truth = orders, estimate = .pred)

## # A tibble: 1 × 3
##   .metric  .estimator  .estimate
##   <chr>    <chr>           <dbl>
## 1 rsq      standard        0.773

prediction_comparison %>%
    mae(truth = orders, estimate = .pred)

## # A tibble: 1 × 3
##   .metric  .estimator  .estimate
##   <chr>    <chr>           <dbl>
## 1 mae      standard        0.881
```

2.7 UNDERSTANDING THE CHAPTER

Students are encouraged to answer the following questions. They can then review the chapter to check their answers.

1. If a scatterplot is a good visualization tool, why is it difficult to use the scatterplot to infer the relationship between the predictor and outcome variables?
2. The CEO of a company would like to know which factors influence sales. They found that the following multiple regression had the best fit. The model is as follows:

$$Sales = \hat{\beta}_0 + \hat{\beta}_1 Price + \hat{\beta}_2 Spokesperson + \hat{\beta}_3 Advertisement$$

Sales refers to the sales amount, *price* is the price of the product, *spokesperson* refers to whether the products were purchased before or after the current spokesperson was employed, and *advertisement* is the amount spent on the advertisement. The following table summarizes the multiple linear regression model output.

Variable	Estimate	*p* value
Intercept	3.583	0.718
Price	–1.367	0.037
Spokesperson	0.782	0.812
Advertisement	1.306	0.047

   a. Using the provided information, would you interpret that price had a significant influence on sales? Was the influence of price on sales direct or inverse?
   b. Among all the predictor variables, which variable had the biggest influence on sales? What value from the preceding table would your answer be based on?
   c. If the coefficient of determination R^2 was 76.72 when these three predictor variables were used, what would it indicate?
   d. Using the model output, can we write the regression equation? If yes, what will it look like?

3. When predictor variables in a multiple linear regression are highly correlated, does it cause autocorrelation or multicollinearity? Why?