such deviations. Others don't experience large falls or increases and hover more closely to their starting values. As time passes, the different random walks tend to be farther away from each other. This illustrates the non-stationary nature of random walks: the standard deviation calculated over an interval increases with time.

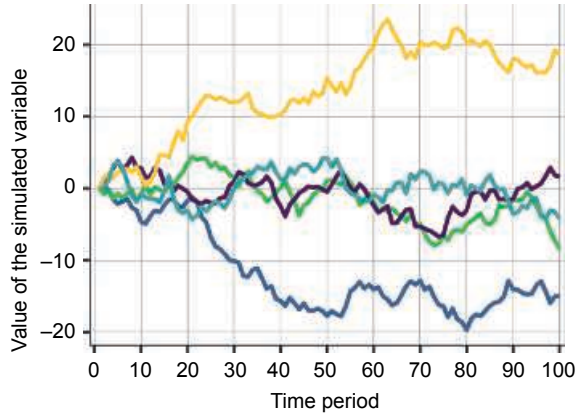


Figure 12.1 Five artificially generated random walk series

Source: Randomly generated.

Random walks may have trends, which are called drifts. A **random walk with drift** is non-stationary on two counts: both its expected value and its spread change with time. It turns out many real-life variables are well approximated by random walks, with or without drifts. These include the price of stocks and stock-market indices, exchange rates, and measures of technological progress. An important feature of random walk series is that they are impossible to predict (apart from their potential drift). As each step is completely random, independent of all previous steps, there is no way to use information from the past to predict the next steps. Another of their important features is that, after a change, they don't revert back to some value or trend line but continue their journey from that point.

In general, time series variables that have this random walk-like property are called variables with a **unit root**. In Under the Hood section 12.U1, we briefly discuss how we can detect unit root in a time series variable. The important message about such variables is that it's usually best to analyze their changes not their levels; we'll discuss this in more detail in Section 12.5.

Review Box 12.3 Non-stationarity in time series variables

Variables in time series data may be non-stationary in several ways. The most common forms of non-stationarity are:

- Trend: expected value increases or decreases with time.
- Seasonality: expected value is different in periodically recurring time periods.
- Random walk and other unit-root series: variance increases with time.

12.A1 CASE STUDY – Returns on a Company Stock and Market Returns

Question and data; preparation and exploration; visualization of time series

Our first case study asks how changes in the price of a company's stock are related to changes in market prices. Relative changes in prices are called returns: for example, an increase in price of 2% is a 2% return. Thus our question is how the returns on one company stock are related to market returns. This question is a frequent one asked in financial analysis. Answering it helps investors decide whether, and to what extent, the risks of investing in the company stock are related to the market risks. A company stock is riskier than the market if returns on the company stock tend to be even more positive when the market return is positive and even more negative when the market return is negative. The more so, the riskier the company stock. A core insight of finance is that investing in riskier stocks should be compensated by higher expected returns.

In this case study we examine how returns on the Microsoft company stock are related to the market returns. For market returns, we use returns on the Standard and Poor's 500 index (S&P500), which is a (weighted) average of 500 company stock prices listed on the New York Stock Exchange and Nasdaq.

When analyzing returns, we need to choose the time window for which the returns are calculated. That choice should be driven by the decision situation that will use the results of our analysis. In finance, portfolio managers often focus on monthly returns. Hence, we choose monthly returns to analyze. For comparison, we'll show results for daily returns, too; as a data exercise you are invited to analyze returns defined for other time windows.

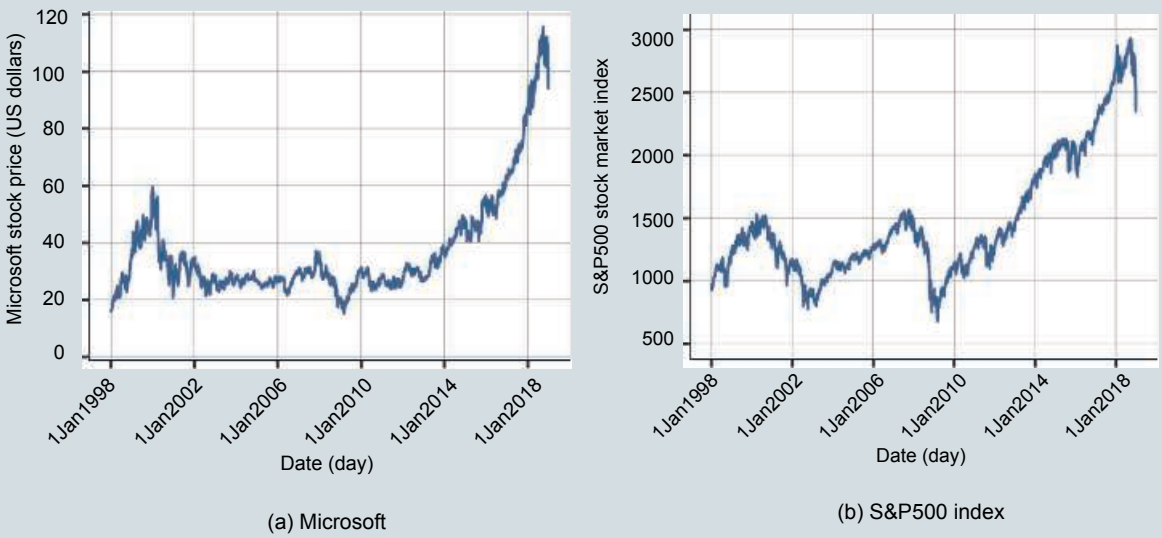


Figure 12.2 Stock prices, daily time series

Note: Daily closing price of the Microsoft company stock and the S&P500 stock market index.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018. N=5284.

We use the `stocks-sp500` dataset. It consists of daily data on the closing price of one company stock, Microsoft, and the S&P500. The data covers 21 years starting with December 31, 1997 and ending with December 31, 2018. This is time series data at daily frequency. The data has gaps as there are no observations for the days the markets are closed, such as weekends and holidays. The frequency of the Microsoft time series and the S&P500 time series is the same, including the gaps. We ignore those gaps for this case study and simply take returns between two consecutive days the markets were open. Both time series include 5284 days.

We'll analyze monthly returns, but let's start with visualizing the original time series of daily prices. Figure 12.2 shows the daily time series of the Microsoft stock price and the S&P500 index over this 21-year-long period.

There are ups and downs in each series, and the overall trend is positive in each. This is good to know, but our analysis requires monthly data as opposed to daily data. We now turn our attention to the monthly time series. We have defined the monthly time series of prices as the closing price on the last day of each month. Figure 12.3 shows the monthly time series of prices.

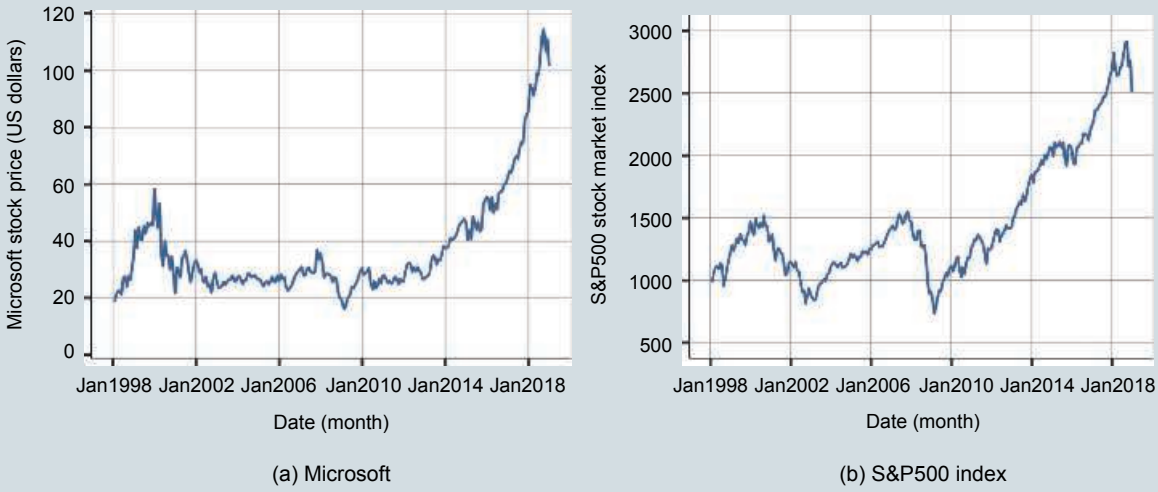


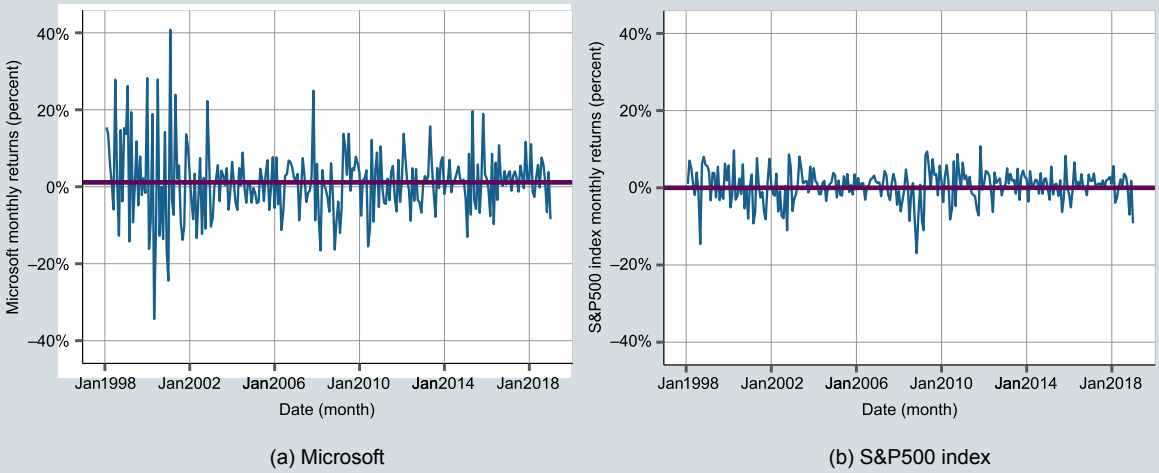
Figure 12.3 Stock prices, monthly time series

Note: Closing price on the last day of the month of the Microsoft company stock and the S&P500 stock market index.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018. N=253.

The monthly price time series look similar to the corresponding daily series, without the ups-and-downs within months. The trends are positive overall, mostly driven by the second half of the time period. In addition, both time series follow a random walk. The results of Phillips–Perron unit root tests are p-values of 0.99 (Microsoft) and 0.95 (S&P500), telling us not to reject the null hypothesis of a unit root (random walk) in either case (i.e., there may very well be a random walk in these series). As we will discuss later in Section 12.5, trends and random walks pose problems for regression. This is one reason why we only focus on returns instead of prices.

The question of our analysis is about returns: percent changes of prices. Monthly returns are the percent changes of prices between the last days of each month. With the 253 last-day-of-month

observations, we calculated 252 monthly returns. Figure 12.4 shows the monthly time series of the percent returns.



(a) Microsoft

(b) S&P500 index

Figure 12.4 Monthly returns time series

Note: Percent change of closing price between last days of the month of the Microsoft company stock and the S&P500 stock market index.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018, monthly frequency. N=252.

These figures show no trends. The Phillips–Perron test rejects its null hypothesis for both time series, with p-values less than 0.001. Thus the time series of monthly returns don't follow random walks, either. Monthly returns appear to fluctuate around zero, and the fluctuations tend to be larger for the return on the Microsoft stock than on the S&P500 index. In fact, the fluctuations are around the average return, which is small but positive for both, as displayed in Table 12.1.

Table 12.1 summarizes the most important statistics of the two monthly returns variables. The range is [-34.4,40.8] for Microsoft; it's less than half as wide, [-16.9,10.8], for the S&P500. We could see the same on Figure 12.4. The mean of the Microsoft monthly returns is 1.1%: it's 0.5% for the S&P500. The standard deviation is more than twice as large for returns on Microsoft: 9.1% versus 4.3%.

Table 12.1 Descriptive statistics on monthly returns

Variables	Min	Max	Mean	Std. dev.	N
Monthly returns on Microsoft (%)	−34.4	40.8	1.1	9.1	252
Monthly returns on the S&P500 (%)	−16.9	10.8	0.5	4.3	252

Note: Monthly percentage returns on the Microsoft stock and the S&P500 index.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018, monthly frequency N=252.

The most important conclusion of the exploratory analysis is that the time series of monthly returns don't follow trends or random walks. Monthly returns are higher, on average, for Microsoft than for the S&P500 index, and they vary more. That also means that prices themselves follow a positive trend (hence their average change is positive), which is steeper for Microsoft. Returns on

the Microsoft company stock are not only higher on average, but they also vary in a wider range, with a larger standard deviation. In a sense, that shows higher risk for the Microsoft company stock. However, recall that finance theory tells us that what matters is not simply how much returns vary, but how they vary in relation to market returns. To uncover that, we need to run a regression.

12.4 Time Series Regression

Regression with time series data is defined and estimated the same way as with other data. But we add something to our usual notation here: the **time index** of the variables, such as y_t and x_t. This additional notation serves two purposes. First, it reminds us that the regression is on time series data, which will allow for additional interpretation of its coefficients. Second, later we'll add observations from other time periods to the regression, and there it's essential to record the time period to know what's what. With this new notation, a linear regression with one explanatory variable on time series data is the following:

$$y_t^E = \alpha + \beta x_t \tag{12.6}$$

Instead of levels of y and x, we can regress the change in y on the change in x. A change in a variable is also called **first difference**, as it is a difference between time t and time $t-1$, the first preceding time period. (A second difference would be between t and $t-2$, and so on.) We use the Δ notation to denote a first difference:

$$\Delta y_t = y_t - y_{t-1} \tag{12.7}$$

A linear regression in differences is the following:

$$\Delta y_t^E = \alpha + \beta \Delta x_t \tag{12.8}$$

At a basic level, the coefficients have the same interpretation as before, except we can use the "when" word to refer to observations: α is the average left-hand-side variable when all right-hand-side variables are zero, and β shows the difference in the average left-hand-side variable for observations with different Δx_t.

But the regression coefficients have another, more informative, interpretation when the variables denote changes. Starting with the intercept: α is the average change in y when x doesn't change. The slope coefficient on Δx_t shows how much more y is expected to change when x changes by one more unit. Note the word "more" before how much y is expected to change in the interpretation of the slope. It is required there because we expect y to change anyway, by α, when x doesn't change. The slope shows how y is expected to change when x changes, in addition to α.

Besides first differences, we often have **relative changes** in regressions denoting how the value changed relative to its previous value. Often, we express such relative changes in percentages, which are then **percentage changes** so that a value of +1 means a 1 percent increase, and a value of –1 means a 1 percent decrease. With percentage differences in y or x, or both, the interpretation of the regression coefficients is a straightforward modification to first differences. A linear regression in percentage changes is the following:

$$pctchange(y_t)^E = \alpha + \beta\, pctchange(x_t) \tag{12.9}$$

where

$$pctchange(y_t) = 100\% \frac{y_t - y_{t-1}}{y_{t-1}} \tag{12.10}$$

Here α shows the expected percentage change in y when x doesn't change, and β shows how much more y is expected to change, in percentage points, when x increases by one more percent.

Finally, just as with cross-sectional regressions, we can approximate relative differences by log differences, which are here **log change**: first taking logs of the variables and then taking the first difference, for example, $\Delta \ln(y_t) = \ln(y_t) - \ln(y_{t-1})$. With time series data, we don't actually have to make use of the **log approximation** because we can directly compute relative, and percentage changes (in contrast with cross-sectional data, see Chapter 8, Section 8.2). Nevertheless, data analysts often use log differences in time series regressions – it is easy to carry out and is often a good approximation.

Review Box 12.4 Linear regression with changes in time series data

- With time series data, we often estimate regressions in changes.
- We use the Δ notation for changes:

$$\Delta x_t = x_t - x_{t-1} \tag{12.11}$$

- The regression in changes is

$$\Delta y_t^E = \alpha + \beta \Delta x_t \tag{12.12}$$

  - α : y is expected to change by α when x doesn't change.
  - β : y is expected to change by β more when x increases by one unit more.
- We often have variables in relative or percentage changes, or log differences that approximate such relative changes.

12.A2 CASE STUDY – Returns on a Company Stock and Market Returns

Time series regression with monthly returns

The main question of this case study is how returns on a company stock are related to market returns. To answer that question, we have estimated a simple regression with the percentage monthly returns on the Microsoft stock (*MSFT*) on the percentage monthly returns on the S&P500 index *(SP500)*:

$$pctchange(MSFT_t)^E = \alpha + \beta\, pctchange(SP500_t) \tag{12.13}$$

The result of this regression is presented in Table 12.2.

The intercept estimate shows that returns on the Microsoft stock tend to be 0.54 percent when the S&P500 index doesn't change. Its 95% confidence interval is [−0.34,1.44], which contains zero. This intercept estimate shows that, on average, the returns on Microsoft were 0.5% when the S&P500 didn't change during the 21-year period we examined. That can be interpreted as 0.5% extra returns on the Microsoft stock, compared to the market returns. But the confidence

interval is wide and contains zero, so we can't say with high confidence that the Microsoft stock has positive extra returns in the general pattern represented by our data of 1998–2018.

Table 12.2 Returns on Microsoft and market returns: Simple regression results

Variables	(1) Microsoft returns
S&P500 returns	1.26**
	(0.10)
Constant	0.54
	(0.45)
Observations	252
R-squared	0.36

Note: Dependent variable: monthly percentage returns on the Microsoft stock; explanatory variable: monthly percentage returns on the S&P500 index. Robust standard errors in parentheses.
** $p < 0.01$, * $p < 0.05$.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018, monthly frequency.

The slope estimate shows that returns on the Microsoft stock tend to be 1.26% higher when the returns on the S&P500 index are 1% higher. The 95% confidence interval is [1.06,1.46]. This interval doesn't include one, thus we can be more than 95% confident that the slope coefficient is larger than one in the general pattern represented by our data. Note that asterisks next to the coefficients show the p-values for testing whether the coefficient is zero. That's what statistical software test by default. Instead, the more interesting question here is whether the coefficient is larger than one. Even without carrying out that test in a formal way, the 95% CI helped answer this question.

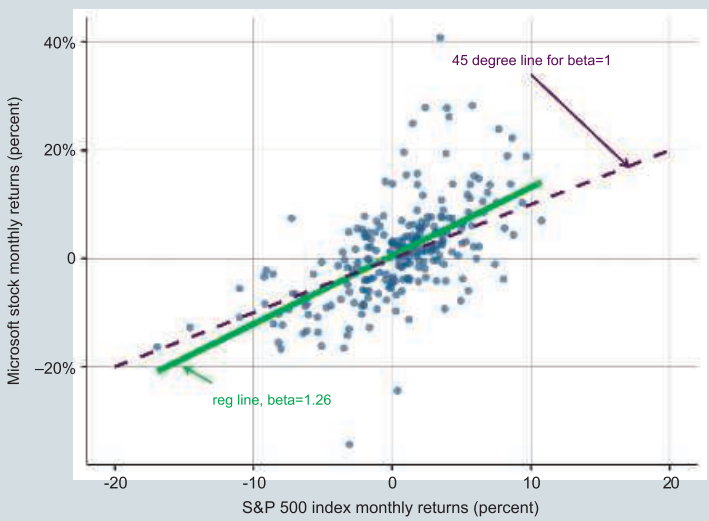


Figure 12.5 Returns on Microsoft and market returns: scatterplot and regression line

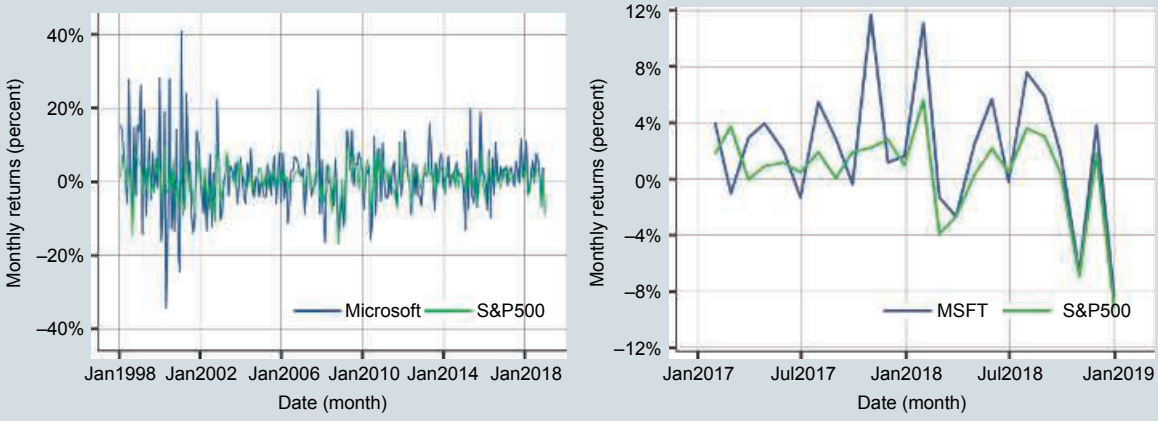
Note: Monthly percentage returns on the Microsoft stock and the S&P500 index. Scatterplot, regression line and the 45 degree line for comparison.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018, monthly frequency. N=252.

Finally, note that the R-squared is 0.36, which is a lot less than one. Thus, it's not that the Microsoft price increases by 1.26% whenever the S&P500 increases by 1%. That happens on average, but sometimes the Microsoft price change is even larger, it is sometimes smaller, and it is sometimes in the opposite direction. Apparently, there are a lot of ups and downs in the Microsoft stock price that are independent of how the market moves.

We can visualize this time series regression in two ways. The first one is the usual scatterplot with the regression line; Figure 12.5 shows them, together with the 45 degree line.

The regression line shows a strong positive association. The slope of the line is more than 45 degrees. The scatterplot also shows the wider range and larger standard deviation for returns on Microsoft than returns on the S&P500.

The second visualization of the regression makes use of the time series nature of the data; it overlays the two time series on each other. In principle, this second visualization can show directly whether, in what direction, and by how much the two time series move together. A positive slope coefficient would mean that the two series tend to have ups and downs in the same direction; a negative slope coefficient would mean that the ups and downs tend to be in opposite directions. Figure 12.6 shows the two time series together, both for the entire 1998–2018 time period and the last two years, 2017–2018. Note that these figures show two time series on the same graph, which raises the question of whether to plot them on the same scale, with a single *y* axis, or on two different scales, with a separate *y* axis for each (shown on the left and right side of the figure). Here we opted for a single *y* axis, because the magnitudes of the two time series are directly comparable: both are percentage changes. Moreover, the magnitudes are not only comparable, but our question is about comparing them directly.



(a) The entire time series, 1998–2018

(b) Two years only, 2017–2018

Figure 12.6 Stock and market returns over time

Note: Monthly percentage returns on the Microsoft stock and the S&P500 index.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018, monthly frequency. N=252.

The regression slope coefficient estimate is 1.26, so these figures should show that ups and downs tend to go together. But that is hard to see for the entire 1998–2018 time period, because it shows too many ups and downs on the same figure (252 of them). It's a lot more visible on the figure that zooms in on the last two years. That figure also shows that, while the ups and downs

tend to be in the same direction, that's not always the case. Moreover, when they move in the same direction, the Microsoft ups and downs tend to be larger, but that's not always the case, either. These facts are in line with a slope coefficient estimate that is greater than one, together with an R-squared that is a lot less than one.

Before concluding our case study, let's carry out some robustness checks and examine two more questions: whether the results are robust to how we measure changes, and whether we get similar results if we examine a different time series frequency. Table 12.3 shows the benchmark regression results, with monthly percent returns in column (1); column (2) shows the results when returns are monthly log changes, while columns (3) and (4) show the same regressions with returns calculated at daily frequency. Because the log changes are two orders of magnitude smaller than the corresponding percentage changes, and this may be important for the intercept estimates, we show four digits on this table.

Table 12.3 Returns on Microsoft and market returns: alternative measurements

Variables	(1) Monthly pct change	(2) Monthly log change	(3) Daily pct change	(4) Daily log change
S&P500 returns	1.2636**	1.2403**	1.1000**	1.0951**
	(0.1030)	(0.1003)	(0.0243)	(0.0236)
Constant	0.5396	0.0026	0.0266	0.0002
	(0.4529)	(0.0045)	(0.0202)	(0.0002)
Observations	252	252	5283	5283
R-squared	0.3573	0.3627	0.4492	0.4465

Note: Dependent variable: returns on the Microsoft stock; explanatory variable: returns on the S&P500 index. The returns are defined differently for the four regressions: monthly percentage changes for (1); monthly log changes for (2); daily percentage changes for (3); daily log changes for (4). Robust standard errors in parentheses.
$**$ $p<0.01$, $*p<0.05$.
Source: `stocks-sp500` dataset. December 31, 1997 to December 31, 2018.

Comparing columns (1) and (2) reveals that using log differences instead of percentage changes makes little difference in terms of the slope coefficient: it's 1.24 instead of 1.26, and the standard error estimates are similar, too. At the same time, the intercept estimates are different. Our benchmark results, repeated in column (1), suggest excess returns of 0.5% (with a very wide confidence interval). The intercept with log changes, in column (2), suggests half as large excess returns, of 0.26 percent (corresponding to a log change of 0.0026). While the point estimates are different, the confidence intervals are wide, so the conclusion about excess returns that we can infer from both regressions is very uncertain.

When estimating the regression using daily returns, we get a smaller slope coefficient, 1.1 instead of 1.26 (columns 3 and 1). The 95% confidence interval doesn't contain one, even with the smaller point estimate in column (3), so we can be quite confident that the Microsoft stock is risky in the general patterns represented by the data in terms of daily returns as well as in terms of monthly returns. Naturally, the intercept estimate is a lot smaller for daily returns, and its confidence interval, too, contains zero. Finally, when using log changes for measuring returns, we get very similar results for the slope and also for the intercept, using daily frequency data (columns 3 and 4). These findings are very similar to what we have seen for monthly returns.

Taken together, the results of these robustness checks suggest that defining returns in terms of percent changes or log changes makes little difference for our main question, the slope coefficient in the regression. It matters somewhat for the intercept estimates, but those are very imprecisely estimated anyway. At the same time, choosing the time series frequency is important for the slope coefficient. In our case, a higher time series frequency led to a smaller coefficient estimate, although it, too, is above one, and its 95% CI is entirely above one, too.

This concludes our case study. Our question was how returns on the Microsoft company stock move together with market returns, measured by returns on the S&P500 index. Using data on monthly percentage returns from 21 years, we have estimated a slope coefficient of 1.26 (95% CI [1.04,1.48]). This means that returns on the Microsoft stock tend to be larger than average by 1.26% in months when returns on the S&P500 is larger than average by 1%. This estimate is larger than one, suggesting that the Microsoft stock is risky: its price tends to move in the same direction as the market, only even more so. Along the way we also estimated that the returns on Microsoft tend to be larger than the market returns, but that conclusion, based on the intercept estimate, is less certain to hold in the general pattern.

Note that what we estimated is very close to what is known as the **"beta" of an asset** in finance. That beta is the **slope coefficient** of a regression almost like ours, except that returns there are measured above the risk-free rate. That may make a difference, because the risk-free rate tends to change through time, too. In any case, the finance "beta" is the slope coefficient in a simple regression of time series data. Hence its name. Similarly, the **intercept coefficient** of the relevant regression is known as "alpha" in finance, measuring extra return, also called excess return. We, as all data analysts, denote the intercept and slope coefficients of a simple regression by the Greek letters α and β, regardless of the content of the variables. But it's good to remember that the finance usage of these Greek letters is more specific.

This case study illustrated some of the important questions we need to address for time series regressions. The first task was to determine the time series frequency for the regression. We opted for monthly frequency because that conforms to the decision frequency of many investors. Our results with daily frequency showed that this decision mattered for the regression estimates. The second task was to define returns. We chose percentage changes, a quite natural choice, because percentage returns are the unit of interest for investments. Our additional results have shown that an alternative measure, log changes, yields very similar results, at least in terms of the slope coefficient estimates. It turns out that using changes in the regression made sure that the time series don't have trends and are not random walks, even though the original time series of prices had both trend and followed a random walk. And that was important to get good estimates, a topic that we'll discuss in the next section.

12.5 Trends, Seasonality, Random Walks in a Regression

Trends, seasonality, and random walks can present serious threats to uncovering meaningful patterns in time series data. Fortunately, we know how to detect them. In this section we also discuss what to do about them.

Consider a simple time series regression in levels and not changes, $y_t^E = \alpha + \beta x_t$. If both y and x have a positive trend, the slope coefficient β will be positive whether the two variables are related or