CHAPTER

# Can you relate? Correlation and simple linear regression

9

## 9.1 INTRODUCTION

Making sense of data, as we have noted, is a required skill for any UX researcher. No matter what the data values are or how you got the data, you're the one who's responsible for making sense of the data. Your other team members may casually watch you conduct a usability test or two, but you are on the hook for interpreting the results. Perhaps just as importantly, everyone (including top brass) wants to know what your recommendations are based on your findings. Although rarely explicitly stated, you know what those execs are thinking: "You told me what you observed; now tell me what to do to increase my bottom line."

Of course, we've already discussed lots of ways to analyze data to gain meaning through statistical inference (hypothesis testing and confidence intervals) methods. But, what about using data to determine the *relationship* between quantities, in order to garner even more meaning? More specifically, how does the value of one variable change when the value of another variable changes? Furthermore—and here comes perhaps one of the most important concepts for not just UX, but all social and behavioral researchers—how can we *predict* the value of one variable from the knowledge of the value of another variable?

Enter the wonderful world of regression analysis.

## 9.2 CASE STUDY: DO RECRUITERS REALLY CARE ABOUT BOOLEAN AT BEHEMOTH.COM?

Let's return to our favorite employment Web site, Behemoth.com, where you were hired as usability researcher. As we mentioned in Chapter 3, Behemoth.com is aptly named, because it's one of the largest job search engines in the world, with over a million job openings at any time, over 1 million resumes in the database, and over 63 million job seekers per month. Behemoth has over 5000 employees spread around the world.

You'll also recall from Chapter 3 that one of the main sources of Behemoth's income is from employers and recruiters who (1) post jobs on the site and (2) buy access to its enormous database of resumes to search for good candidates to fill those jobs.

**Improving the User Experience through Practical Data Analytics.** http://dx.doi.org/10.1016/B978-0-12-800635-1.00009-4

You'll recall from Chapter 4 that you had to deliver the cold, hard facts about "Novix" to UX Director Hans Blitz: despite spending 80 million dollars on the new search, the current Behemoth search performed better than the new one in a head-to-head comparison usability test.

To put it mildly, Hans was shocked. But bad news doesn't get any better with age, so he quickly scheduled a meeting with Joey Vellucci, CEO, to break the news. You were glad you didn't have to be in the meeting, although a "fly-on-the-wall" perspective might have been interesting.

Strolling into work one morning, you spy Hans enjoying a cigarette in "smoker's alley," the only place where the few remaining Behemoth workers who smoke can indulge. It's behind the old mill, where 19th century workers brought in the raw materials for the looms, and where finished textiles were loaded onto trains. Hans sees you walking by and motions for you to come over for a chat.

"Well, Joey didn't go ballistic like I expected, but he's pretty annoyed and wants more answers pronto."

"Hmmm…what kind?"

"Well, first of all, he's curious about why our recruiting clients are complaining about our current search if our current search engine does better than Novix."

"Well, just because we did better than Novix doesn't mean that our current search is perfect."

"*You* tell him that," Hans offers with a chuckle. "He did raise some interesting points, though."

"Such as?" you ask.

"He looked at your report and was baffled that recruiters kept talking about the missing Boolean search; the big selling point that the Palo Alto guys kept hammering was that you don't need a Boolean with Novix, and that was supposed to be a *good* thing."

"Yeah," you admit. "It's a valid point. But what we found out is that a lot of the recruiters are really creatures of habit who have been using their same complex Boolean strings for a long time. You take that away from them and it's like taking a juicy T-Bone from a hungry dog. As a consequence, the lack of Boolean search capability may decrease adoption of the Novix search engine."

Hans' mood suddenly goes dark, and he glares at you with disdain: "Prove to me we spent 80 million getting rid of something our clients want!"


You're briefly taken aback, but decide to suppress the immediate urge to reply that UX should have been given the opportunity to run usability tests *before* the Novix purchase. You opt to utilize a more proactive approach: "Well, we could conduct an online usability study on our current search engine. We'll use an unmoderated study to get the sample sizes much higher than we would get with a standard usability test."

"OK," Hans says slowly and cautiously. "And?"

"After folks have used the search for a while, we could ask all the participants to rank their perception of usefulness with the different search fields, along with their likelihood to adopt the search engine. Then, I can calculate the correlation coefficient between the usefulness of the ability to perform a Boolean search and likelihood of adoption of the search engine, and perform some simple regression."

Hans hesitates: "I have no idea what you just said, but it sounds feasible. When can we have the results?"

"Well, it'll take some time to get the test together and screen for the right participants. After the results come in, we'll do the number crunching. Give me 2 weeks."

"Ok, but no more than 2 weeks right? Bad news doesn't get any better with time." He throws down his finished butt, stomps on it, turns abruptly on his heels and leaves in a huff.

Again, you're off to the races to get answers for Hans and to alleviate his Boolean angst. Realizing that you need higher sample sizes to make a convincing case for your results, you decide to go with an unmoderated online test of the current Behemoth search engine. The e-mail invite goes out to about 300 recruiters who are regularly searching for candidates. After answering some basic eligibility questions, they are carefully screened to disqualify any current Behemoth customers; you want newbies who've never used the engine before.

All the respondents are tasked with finding good candidates for the same three requisitions: (1) A Java Developer with at least 5 years experience, a bachelor's degree from MIT, a maximum salary requirement of $95,000 per year, willing to relocate, who is looking for a full-time position; (2) A Web Designer with skills using Axure, Photoshop, and Illustrator within 25 miles of San Diego, with an Active Confidential Security Clearance; and (3) A Business Analyst who has previously worked at Oracle, an average length of employment of no less than 1 year, with a resume dated no earlier than 2013, willing to travel up to 50% of the time.

After completing the tasks of finding candidates for the three positions, the respondents are asked about overall satisfaction with the search engine. In addition, they are specifically asked to rate their perception of usefulness for each of the fields in the search engine, on a scale of 1–5, where 1 = not at all useful and 5 = extremely useful. Table 9.1 shows the 15 specific search engine components respondents are asked to rate.

At the very end of the survey rating, you insert the moment of truth question: "Imagine that this search engine is available to you at no cost to find qualified candidates using the candidate databases you currently employ. Rate your likelihood of adopting this candidate search engine on a scale of 1–5, where 1 = not at all likely and 5 = extremely likely."

With everything in place, you launch the online study. After a week, you check into your online test tool. You're happy to find 233 responses. However, there were 36 incompletes, and another 17 who you have to disqualify for suspicious looking activity (mostly in the form of overly quick task completion times). You end up with 180 bona fide test responses. You download the Excel spreadsheet containing the rating scales of the search engine components.

Time to roll up your sleeves. The first thing you want to establish, of course, is the perceived value of Boolean search, and its correlation with likelihood of adoption.

## SIDEBAR: GET THOSE SAMPLE SIZES UP! UNMODERATED REMOTE USABILITY TESTING

Unmoderated remote usability testing is a low-cost and effective technique for collecting task-based usability data. As a consequence, the technique has experienced great growth in recent years by UX professionals.

Unmoderated remote usability tests allow you to:

- Collect data easily and efficiently.
- Find the right participants.
- Conduct studies on a limited budget.
- Test users in their natural environment.
- Test internationally without traveling.
- Validate or define lab-based research.
- Achieve statistical significance when comparing designs, identify areas of a Web site that need improvement, and conduct competitive benchmarking studies.

Those are the advantages, but how does it work? In a nutshell, the participant receives an e-mail invitation to participate in the study, and clicks on the link to begin. He/she is taken to a Web site where he/she is asked to complete a series of tasks. During task completion, data are being collected on a wide range of variables, like task completion rates, time to complete the tasks, pages visited, data entered, etc. Once finished, the participant is asked a series of questions regarding satisfaction on several different variables. In our case with Behemoth, the participant is asked to rate the usefulness of specific components of the search engine.

For an outstanding introduction and how-to guide on online usability studies—including planning, designing, launching, and data analysis—we heartily recommend *Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies*, by Bill Albert, Tom Tullis, and Donna Tedesco (Morgan Kaufmann, 2010). The book shows you how to use commercial tools, like User Zoom and Loop 11, but also offers discount approaches that can yield perfectly good results.

**Table 9.1** The 15 Search Engine Components

1. Ability to search by job title
2. Ability to search by years of experience
3. Ability to search by location
4. Ability to search by schools attended
5. Ability to search candidates by date of updated resume
6. Ability to search candidates by level of education
7. Ability to search by skills
8. Ability to search candidates by average length of employment at each company
9. Ability to search candidates by maximum salary
10. Ability to search candidates by job type he/she is looking for: full time, part time, temporary/contract, per diem, intern
11. Ability to search candidates by companies in which they have worked
12. Ability to search candidates by willingness to travel. (Expressed as “no travel ability required,” “up to 25%,” “up to 50%,” “up to 75%,” “up to 100%”)
13. Ability to search candidates by willingness to relocate
14. Ability to search candidates by security clearance. (Active Confidential, Inactive Confidential, Active Secret, Inactive Secret, Active Top Secret, Inactive Top Secret, Active Secret/SCI, Inactive Top Secret/SCI)
15. Ability to perform a Boolean search

## 9.3 THE CORRELATION COEFFICIENT

The "correlation coefficient" reflects the relationship between two variables. Specifically, it measures the strength of a straight-line relationship between two variables, and also tells you the direction of the relationship, if any. It is a numerical value that ranges between −1 and +1 and is typically denoted by "r":

$$-1 \leq r \leq +1$$

The absolute value of the correlation reflects the strength of the relationship. So, a correlation of −0.80 indicates a stronger relationship than +0.70.

There is an arithmetic formula to compute the correlation coefficient, but you should simply use Excel or SPSS to compute it for you. (We will show you how to do this later in this section.)

Let's first consider a small example before we dive into the Behemoth.com data set. Say that we have two variables: (1) Assessment of how sophisticated a Web design is, and (2) Amount of experience buying products online. We will assign the sophistication assessment as the output variable, "Y," and the amount of experience as the input variable, "X." It is common to use Y to notate the output variable and X to notate the input variable.

The correlation coefficient, "r," measures the relative degree to which a straight line fits the data. Let's consider the six "scatter plots" in Figure 9.1, and discuss them in light of the value of r.

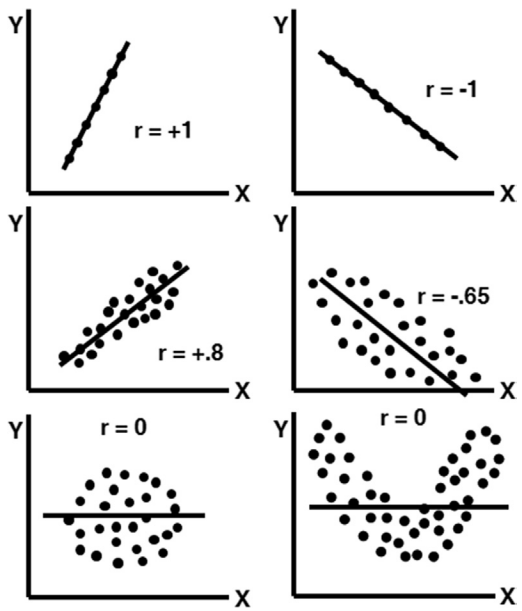



**FIGURE 9.1**

Scatter plots of Y and X data.

The top left plot has all the data points exactly on a straight line, and the straight line has a positive slope. With real data, we would never see such an exact relationship; we present it only for illustration purposes. The value of r is +1. This value is the maximum value r can take on. And, the fact that the slope of the line is positive ensures that the sign of r is positive (i.e., +). Now, if you look at the top right plot, the relationship is also perfect—the data values are all right on the line. But, the line is downward sloping. The value of r, correspondingly, comes out −1. The "1" indicates that the fit to a straight line is perfect, while the "−" sign indicates, indeed, that the line is downward sloping.

This illustrates a useful property of r; while the numerical value of r tells us about how well the data fit a straight line (referred to as the [relative] *strength* of the relationship), its sign tells us the *direction* of the relationship, if any. A + value of r indicates a positive relationship—as one variable goes up, the other variable also goes up; as one variable goes down, the other variable also goes down. (You may also see this relationship described as a "positive" or "direct" correlation.)

A − [minus] value of r indicates an "inverse" relationship—as one variable goes up, the other goes down, and vice versa. (You may also see this relationship described as a "negative or "indirect" correlation.) A value of zero for r (or, as a practical matter, right near it) indicates that the two variables are unrelated linearly, and this is illustrated in the bottom left scatter plot; you can see that there is no indication at all of a relationship between the variables. The best-fitting line[1] is horizontal—indicating zero slope, the equivalent of no linear relationship. Whenever the best-fitting line has zero slope (i.e., is horizontal), the value of r is zero, and vice versa.

If we look at the middle left plot, the data values are not right on the line that goes through the values, but it will, intuitively, give you a pretty accurate value of Y from inputting the value of X into the equation of the line. The line might be, for example, $Y = 0.2 + 1.07X$, and when you plug in the value of X, the Y that comes out will be pretty close to the actual Y value for most all the data values. Without specific data, we cannot provide the exact value of r, but it might be in the neighborhood of +0.8; the line clearly has a positive slope. (The "r = 0.8" listed for this plot in Figure 9.1 is just a rough estimate by the authors when looking at the data.)

Let's compare this plot with the middle right plot. It should be clear to the reader that the fit to a straight line is not as good as the plot on the left—the data values are not as tightly clustered around the line as in the middle left plot—and also, the line best-fitting the data is downward sloping. Since the fit is less good, the value of r is lower, say, 0.65 (again, an estimate based on the authors' view at the data). Also, it is negative, reflecting the negative slope of the straight line best-fitting the data.

The final plot among the six is the bottom right plot. The purpose of that plot is to dramatically illustrate a key point that r is measuring the relative strength of a

[1] We will more precisely define "best-fitting line" later in the chapter.

*straight-line fit*. Clearly, there is a relationship, perhaps somewhat strong, between Y and X in the bottom right plot. But, it is a "U-shaped" curve. A straight line does not do a good job *at all* of fitting that data. The best-fitting line would be horizontal, or very close to it, and the value of r would, correspondingly, be equal to or very close to zero. After all, what does Y do as X increases? For the first half of the X's, Y goes down; for the second half of the X's, Y goes up. We might say that, *on average*, Y *does not do anything*—hence the zero slope of the best-fitting line and a value of r of zero.

Remember, if a line has a slope of zero (and is, thus, horizontal), it means that as X changes, Y does not change at all. After all, if a line is Y = 4 + 0*X, with its zero slope, no matter what X you plug in, the Y stays the same—indeed, Y is totally unaffected by X; that is why we noted earlier that a zero value for r indicates that there is no linear relationship at all between the two variables.

### 9.3.1 EXCEL

We will now describe how to find r using Excel. We will analyze the real-world Behemoth.com data a bit later. Let us use our limited data set to illustrate the finding of the correlation, r, using Excel. Suppose we have the five data points in Table 9.2 (which we can envision coming from respective 5-point Likert scales).

For the most part, when X is larger, Y is larger, so we would expect a positive value for r.

First we open Data Analysis in Excel and identify "Correlation." See the arrow in Figure 9.2. Then we click "OK." This gives us the Correlation dialog box, shown in Figure 9.3.

We enter the (input) data range (see vertical arrow in Figure 9.3), and, we ask the output to be on a page arbitrarily named "paul" (see horizontal arrow in Figure 9.3). After we click "OK," we get the answer—the correlation between the two variables—as shown in Figure 9.4.

Figure 9.4 tells us that the correlation, r, equals +0.895. We noted earlier that we anticipated a positive value, and, indeed, we do get a positive value. Notice that in Figure 9.4, the top right cell is empty. That is because the correlation coefficient between two variables is the same, regardless of which is the "Y" and which is the "X." In other words, the correlation between the Column 2 variable

**Table 9.2** Illustrative Data

| X | Y |
|---|---|
| 5 | 4 |
| 4 | 4 |
| 2 | 1 |
| 3 | 3 |
| 4 | 3 |

and the Column 1 variable (noted as 0.895144) is the same as the correlation between the Column 1 variable and the Column 2 variable. Also note that we have a value of 1.0 as the correlation between each variable and itself!! This is always the case and, of course, it makes perfect sense that a variable is perfectly correlated with itself.

Now we present finding r using SPSS.

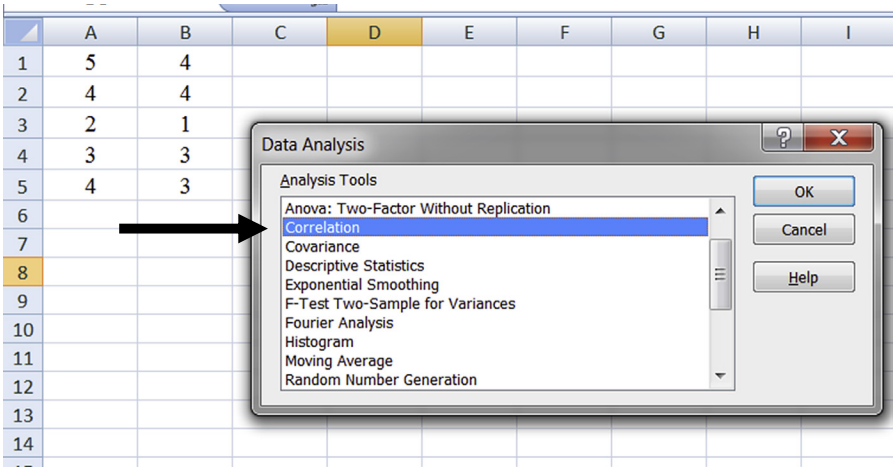



**FIGURE 9.2**

Data analysis with correlation highlighted; Excel with illustrative data.

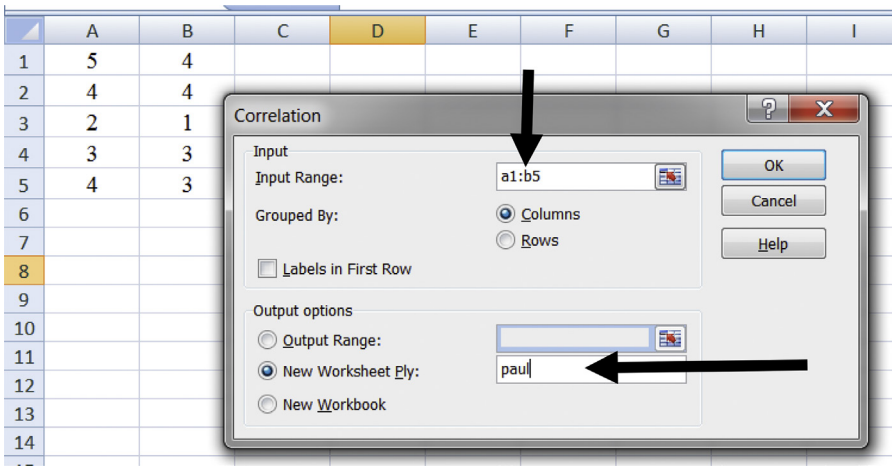



**FIGURE 9.3**

Correlation dialog box; Excel with illustrative data.

| | A | B | C |
|---|---|---|---|
| 1 | | *Column 1* | *Column 2* |
| 2 | Column 1 | 1 | |
| 3 | Column 2 | 0.895144 | 1 |
| 4 | | | |
| 5 | | | |

**FIGURE 9.4**

Correlation output; Excel with illustrative data.

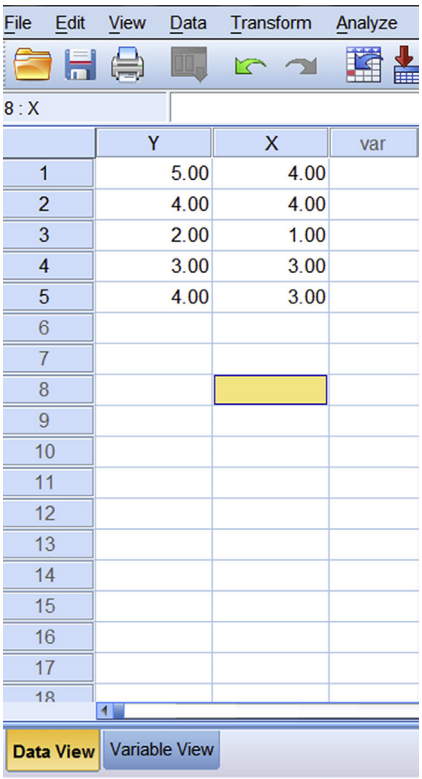



**FIGURE 9.5**

SPSS with illustrative input data.

### 9.3.2 SPSS

Figure 9.5 shows the same limited data set in SPSS. You can see in the figure that we used "Variable View" to change the variable names to Y and X. Of course, the Y values are what they are and the X values are what they are; however, it does not matter if the Y column is typed in to the left of the X column (as in Figure 9.5) or to the right of the X column.

## SIDEBAR: THE COEFFICIENT OF DETERMINATION (R-SQUARED)

Since this is the first example in which we found r for an actual data set (albeit, a small one!), we now introduce another interesting idea, one that helps interpret the actual value of r.

If we compute $r^2$, we get (0.895*0.895) = 0.801. The quantity, $r^2$, is called the "coefficient of determination," even though, often, it's referred to as simply the "$r^2$."

We can give a very useful interpretation to the 0.801. Based on the data, we estimate *that 80.1% of the variability in Y (i.e., the degree to which all the Y values are not the same) can be explained by the variability in X (i.e., the degree to which all the X values are not the same).* In loose terms, we might say that X is estimated to explain about 80.1% of Y, and if X were held constant, Y would vary only 19.9% as much as it varies now.

In our example, Y = assessment of how sophisticated a specific design is, and X = amount of experience buying products online. So, in that context, an r of 0.895, and $r^2$ of 0.801, we would say that we estimate that about 80% of the variability in the respondents' opinions about how sophisticated the design is can be explained by how much experience a respondent has had buying online products. By the way, in this type of context, 80% would nearly always be considered a pretty high value!

We now pull down "Analyze" (we noted earlier in the book that "Analyze" is always how we begin a statistical analysis of any kind), highlight "Correlate," and go to the submenu item, "Bivariate." See arrows in Figure 9.6.

The term "bivariate" means correlation between *two* variables. The other choices are more complex, and are beyond the scope of this chapter.

After we click, we get the "Bivariate Correlations" dialog box, as shown in Figure 9.7. The word, "Correlations," is plural, since if your data set had, for example, three variables/columns (say, Y, X1, X2), the output would give you the correlation between each of the three pairs of variables; (Y, X1), (Y, X2), (X1, X2). Here, with only two variables, we will get, of course, only one correlation value (not counting the "1's"—the correlation of a variable with itself).

In Figure 9.7, we need to drag the Y and X over to the right-side box called "Variables." There is no need to change the defaults, including the check next to "Pearson" (see sidebar).

## SIDEBAR: KARL PEARSON

What we are finding is, technically, the *Pearson* correlation coefficient, named after the mathematician and biometrician (i.e., bio-statistician), Karl Pearson (1857–1936). He was born Carl Pearson, but changed his name purposely and officially to Karl, since he was a fervent fan of Karl Marx.

In 1911 he founded the world's first university statistics department at University College, London. In addition to his work on the correlation between variables, Dr. Pearson also headed up the work on the chi-square test we worked with in an earlier chapter. We noted that it was called the "Pearson chi-square test."

Another claim to fame, although he didn't know it at the time, was that when the 23-year-old (http://en.wikipedia.org/wiki/Albert_Einstein) Albert Einstein started a study group, the Olympia Academy, he suggested that the first book to be read was Karl Pearson's *The Grammar of Science*.

Dr Pearson had two daughters and a son. His son, Egon Pearson, became a prominent statistician in his own right, and succeeded his father as head of the Applied Statistics Department at University College. Egon Pearson, along with Jerzy Neyman, another prominent statistician, developed the basics of hypothesis testing as we know it today (improving ideas that had been earlier considered by Karl Pearson).

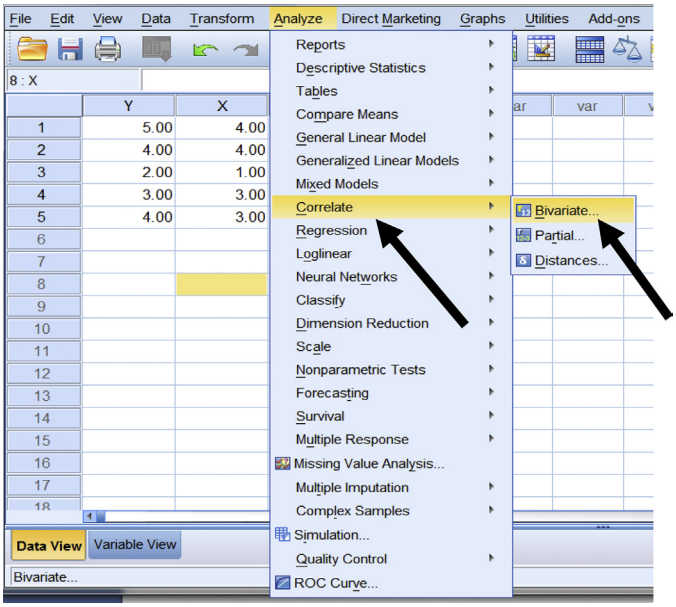



**FIGURE 9.6**

Asking SPSS to find a bivariate correlation; illustrative data.

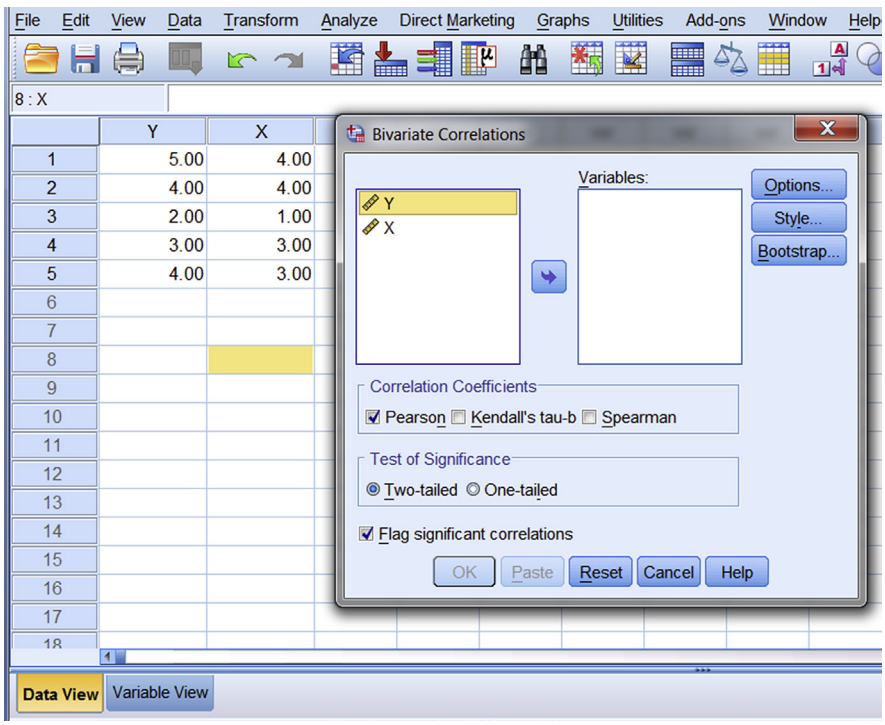



**FIGURE 9.7**

Bivariate Correlations dialog box; SPSS with illustrative data.

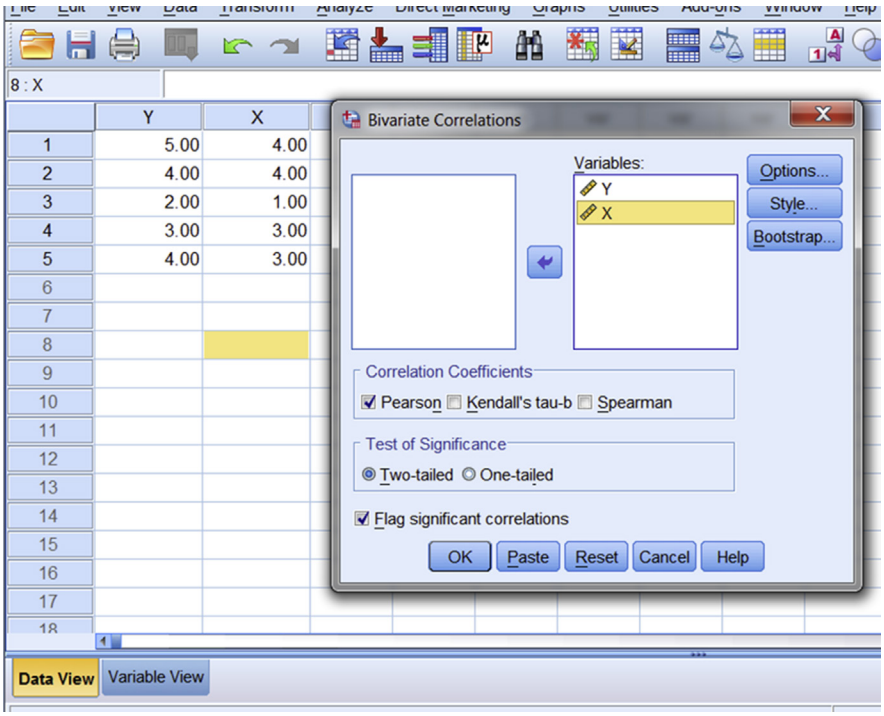



**FIGURE 9.8**

Variables brought over to Variables dialog box; SPSS with illustrative data.

In addition, we might as well see if the r, that we find, is statistically significant, and if so, have it flagged. (This is essentially testing our old friend the null hypothesis. H0 is that the true value r [often referred to as "ϱ"] is 0, vs. $\varrho \neq 0$. Refer to Chapter 1 for a brushup on hypothesis testing.) Again, these are the default options in the bottom portion of the "Bivariate Correlations" dialog box in Figure 9.7.

Figure 9.8 shows the variables brought over to the "Variables" dialog box.

We now click on OK, and get the output shown in Figure 9.9.

Of course, we get the same answer we got when doing the problem in Excel. However, the SPSS output provides even more value: it tells us that the 0.895 value of r is *statistically significant* (based on a significance level of 0.05), because (1) the asterisk on the 0.895 value tells us that, and, (2) the *p*-value = 0.04, which, indeed, is less than 0.05. (Recall that the *p*-value is always notated as "Sig." in SPSS.)

### 9.3.3 CORRELATION APPLICATION TO BEHEMOTH.COM

You'll recall that Hans Blitz told you to prove to him that Behemoth spent $80 million getting rid a feature that clients actually wanted.

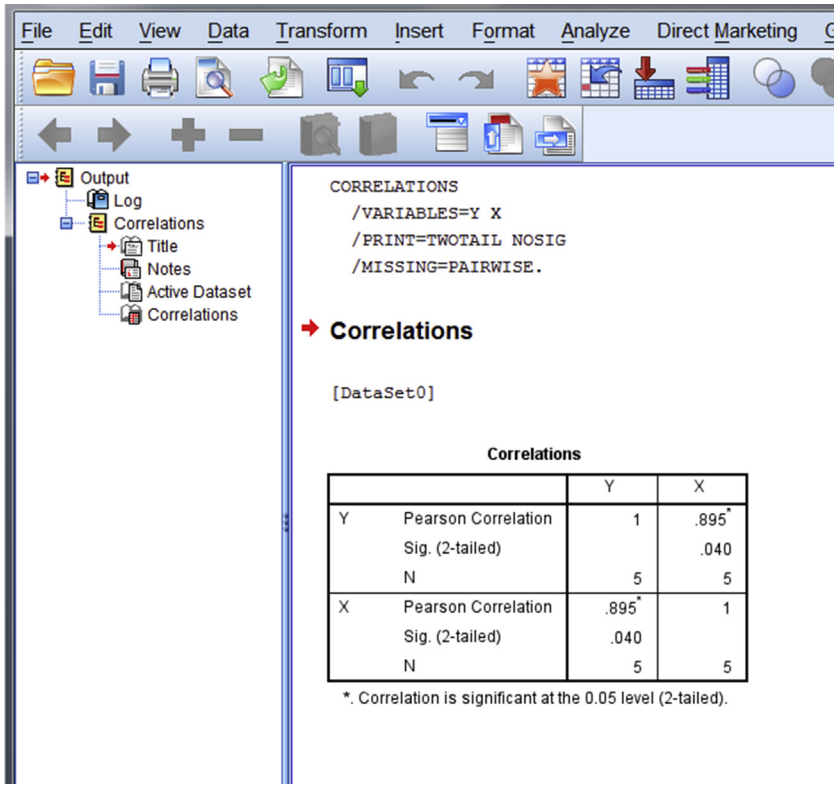



**FIGURE 9.9**

Correlation output; SPSS with illustrative data.

As a consequence, two of the columns of data you collected during the unmoderated usability test were the Likelihood of adoption (of the search engine) and the assessed usefulness of the Ability to perform a Boolean search, each rated on a 1–5 Likert scale, as mentioned earlier.

Therefore, we have 180 rows of data for these two variables (as we do for all of the data). Figure 9.10 displays the first 26 rows of data in Excel. Also, we should take note of the fact that the first row on the spreadsheet is taken up with the column labels.

As we noted, there are 180 data values in rows 2–181. When we open "Data Analysis" and then "Correlation," we have what is shown in Figure 9.11.

We show you Figure 9.11 specifically since it illustrates the use of a column label (i.e., title). Note how we entered the location of the input data: A1–B181, since there are 180 rows of data. We then check the box that asks whether we have a label in row 1. We do, and thus, we check that box; see arrow in Figure 9.11. Excel now knows to consider rows 2–181 as actual data rows.

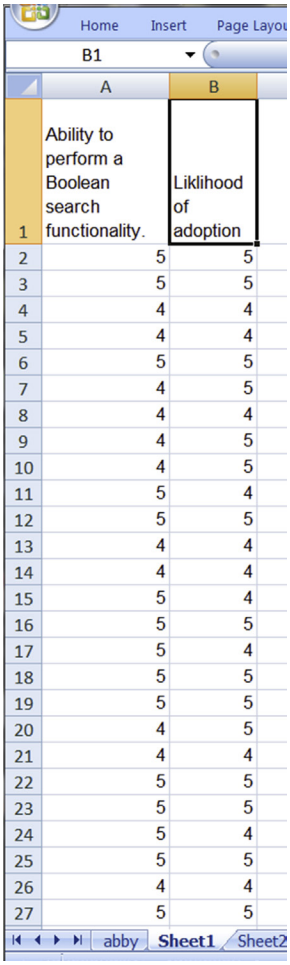



**FIGURE 9.10**

Behemoth.com data in Excel.

Our output is in Figure 9.12.

We can see that the correlation coefficient is +0.449 (the "+" sign is no surprise!). What this tells us is that a higher sense of usefulness of Boolean search capability is associated with an increased likelihood of adoption of the search engine. (If we performed this using SPSS, we would find that the *p*-value of r is 0.000 to 3 digits [actually, for those who care: 0.000000000247] and is thus highly statistically significant.)

Furthermore, we can say that the responder's opinion about the usefulness of the ability to perform a Boolean search, by itself, explains slightly over 20%

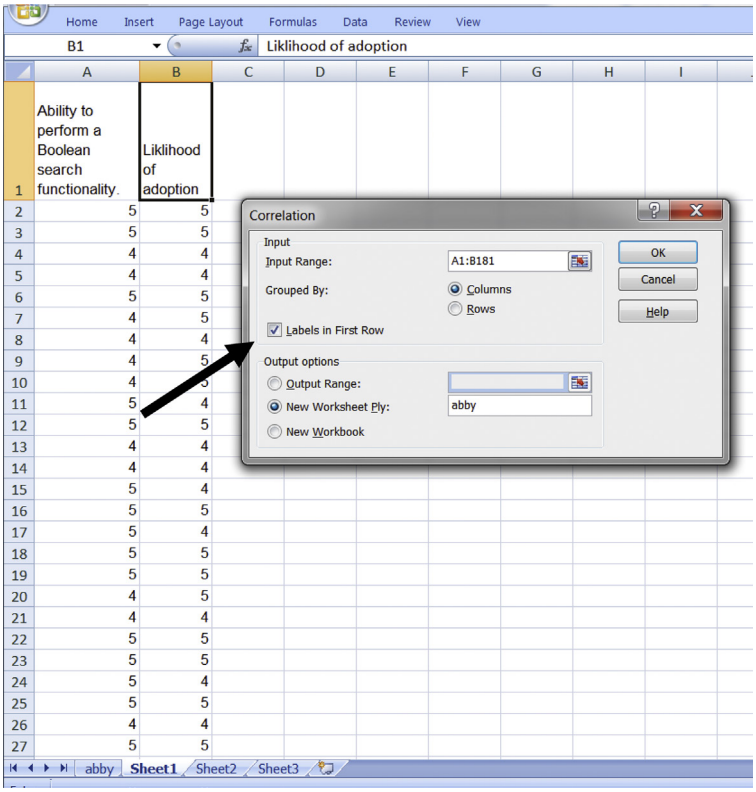



**FIGURE 9.11**

Correlation dialog box for Behemoth.com data; Excel.

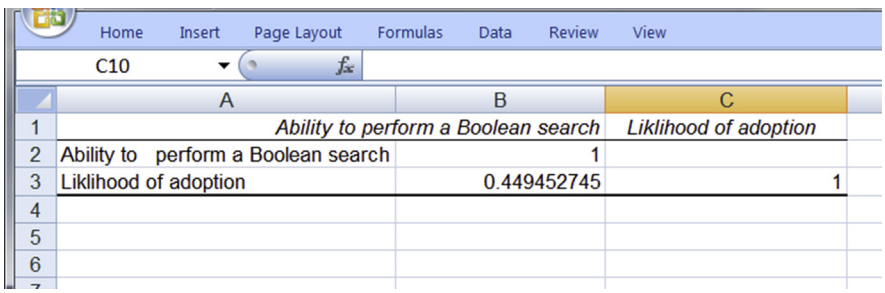



**FIGURE 9.12**

Correlation output for Behemoth.com data; Excel.

(0.449*0.449 = 0.202 or 20.2%) of the variability among the responders of their indicated likelihood to adopt the search engine.

The values of r and $r^2$ indicated the strength of the (linear) relationship between the two variables. This is certainly important, but we also need to determine what

the specific relationship is between the two variables. So, we now introduce Regression Analysis, which will determine the best fitting slope and intercept of this linear relationship based on the data. Using the scenario from this chapter, it will tell us, for example, how much an increased assessment of the usefulness of the ability to perform a Boolean search, say by one unit, will increase the likelihood of adoption of the search engine.

Onward!

## 9.4 LINEAR REGRESSION

The fundamental purpose of regression analysis and correlation analysis is to study the relationship between a "dependent variable" (which can be thought of as an *output* variable) and one or more "independent variables" (which can be thought of as *input* variables). In this chapter, we will have one independent variable—this form of regression is called "simple regression"; in the next chapter, we will have several input/independent variables (i.e., X's)—this will be called "multiple regression."

Let's return to the illustrative data set we used in the correlation section. This data set is shown in Table 9.2, but, for convenience, we repeat it in Table 9.3. We will illustrate the principles of regression analysis using this data set and then apply the methodology to the Behemoth.com data.

We traditionally refer to the "Y" as the *dependent variable*, and the "X" as the *independent variable*. In fact, you shall see that SPSS uses those terms.

Let us consider a straight-line (i.e., "linear") relationship between the two variables in Table 9.3. We usually start off considering a straight-line relationship first, unless the (Y, X) graph of the data points (similar to the graphs in Figure 9.1) strongly indicates that the relationship is clearly curved. The graph of the data in Table 9.3 is in Figure 9.13. The graph in Figure 9.13 is referred to as a "scatter diagram."

It is evident that a straight line fits the data pretty well, and that there is no meaningful indication of curvature. In Figure 9.14, we add a line that, intuitively, fits the data well.

Thus, we can pretty safely consider a straight-line relationship between X and Y, and not be concerned about more complex relationships. In fact, let us

**Table 9.3** Illustrative Data

| Y | X |
|---|---|
| 5 | 4 |
| 4 | 4 |
| 2 | 1 |
| 3 | 3 |
| 4 | 3 |

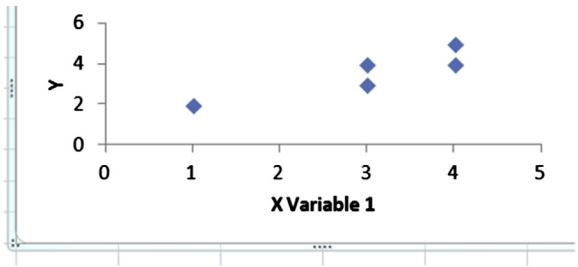



**FIGURE 9.13**

Scatter diagram.

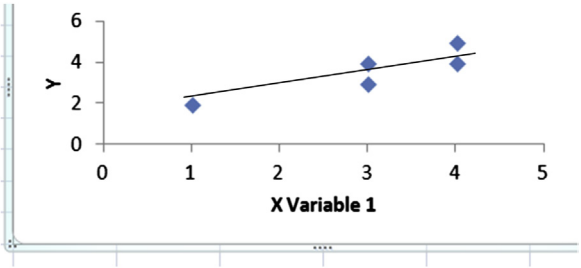



**FIGURE 9.14**

Adding to the scatter diagram a line that, intuitively, fits well.

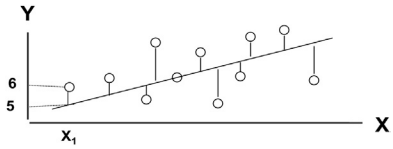



**FIGURE 9.15**

Illustration of the least-squares definition of "best."

determine the "best-fitting" line to the data (which, surely, will be close to the line drawn "by eye" by the authors in Figure 9.14, but likely will not be identical to it).

**But hold on!** We cannot find the "best-fitting line" without deciding how to define "best-fitting." Well, in about 99.99% of the cases you would encounter, "best-fitting" is defined as the *least-squares line*. This is the line that minimizes the sum (for all the data values) of the squared differences between the actual Y value and the predicted Y value from using the line. Put another way, it is the line that best fits the actual scores and minimizes the error in prediction, using this criterion. This is illustrated in Figure 9.15.

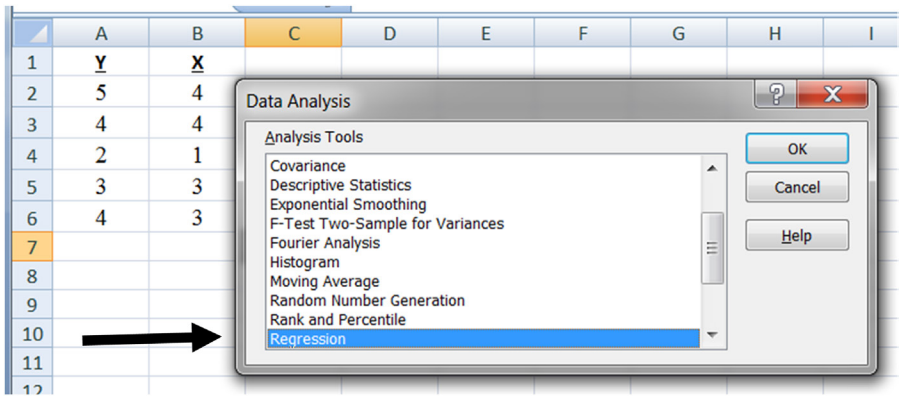



**FIGURE 9.16**

Regression command within data analysis; Excel with illustrative data.

To calculate this line, the vertical differences from the dots to the line are first squared and summed. For example, at $X_1$, the line predicts 5, but the actual data value equals 6, for a difference of 1. It can be proved that the least-squares line is unique. In other words, there cannot be a tie for which line is the least-squares line. Perhaps more importantly, Excel and SPSS will find it for us.

First, let's begin by providing our notation[2] for the formula for a simple regression least-squares line:

$$Yc = a + b * X$$

where:

- **"Yc"** is the predicted ("computed") value of Y based on a known value of X,
- **"b"** represents the slope of the line,
- **"a"** represents the intercept, or the point at which the line crosses the Y-axis (sometimes called the "Y-intercept").

There are some very tedious mathematical formulas you can use to calculate the slope and intercept of the least-squares line (sometimes called the "regression line"), but both Excel and SPSS will save you lots of time and headaches. Let's start with Excel.

### 9.4.1 EXCEL

To do a regression analysis in Excel (and what we have been doing for all statistical analyses in Excel) we first open "Data Analysis." Then, we scroll down to "Regression." See Figure 9.16, with the arrow pointing to the command.

[2]There is no standard notation for this least-squares line. If you looked at 10 statistics/predictive analytics/data analysis texts, you might see five or six different notations for the slope and intercept.

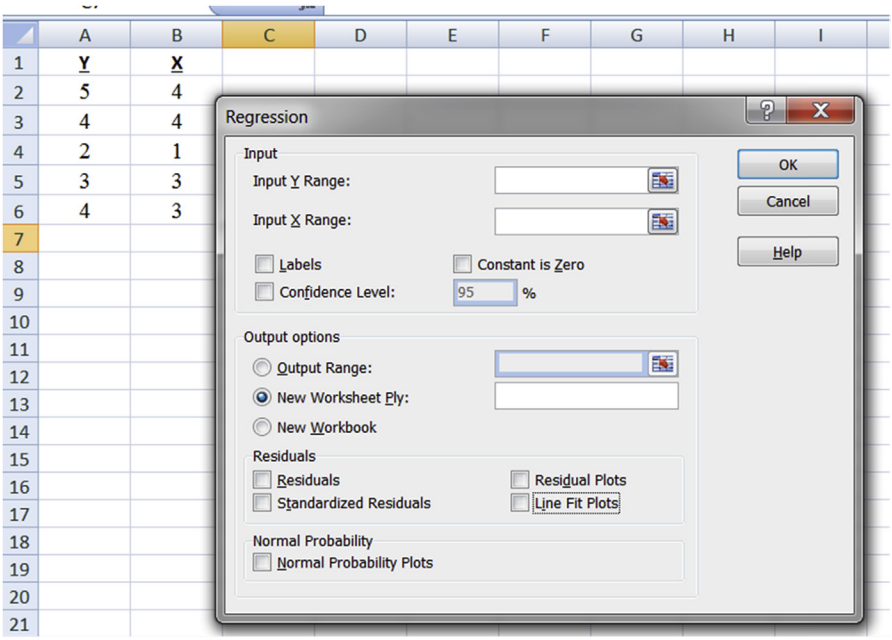



**FIGURE 9.17**

Regression dialog box; Excel with illustrative data.

We click on "Regression," and get the dialog box shown in Figure 9.17.

We enter the location of the Y variable data and the location of the X variable data, and enter an arbitrary name of a new worksheet: "JARED." That brings us to Figure 9.18.

Note that we checked "Labels" and listed the data as (a1:a6) and (b1:b6), even though there are no data values in row 1.

We now click "OK," and find the output in Figure 9.19.

There is a lot to digest in Figure 9.19. But let's take it step-by-step, and you'll be fine.

First, we note the least-square (best-fitting) line by examining the circled column in the bottom left of the figure; the intercept is 1.1 and the slope is 0.833. (The intercept is labeled "intercept" and the slope is labeled by "X" [row 18 in Figure 9.19], which is standard notation in all statistical software. Since there can be more than one X, the label is to indicate which X the slope pertains to.) It is understood that the value (in this case: 0.833) is the slope of the X listed. Our line, thus, is:

$$Yc = 1.1 + 0.833 * X.$$

The slope of 0.833 means that for each unit increase in X (i.e., X goes up by 1), we predict that Y goes up by 0.833. If X is 0, then our prediction for Y is 1.1, since that's our intercept.

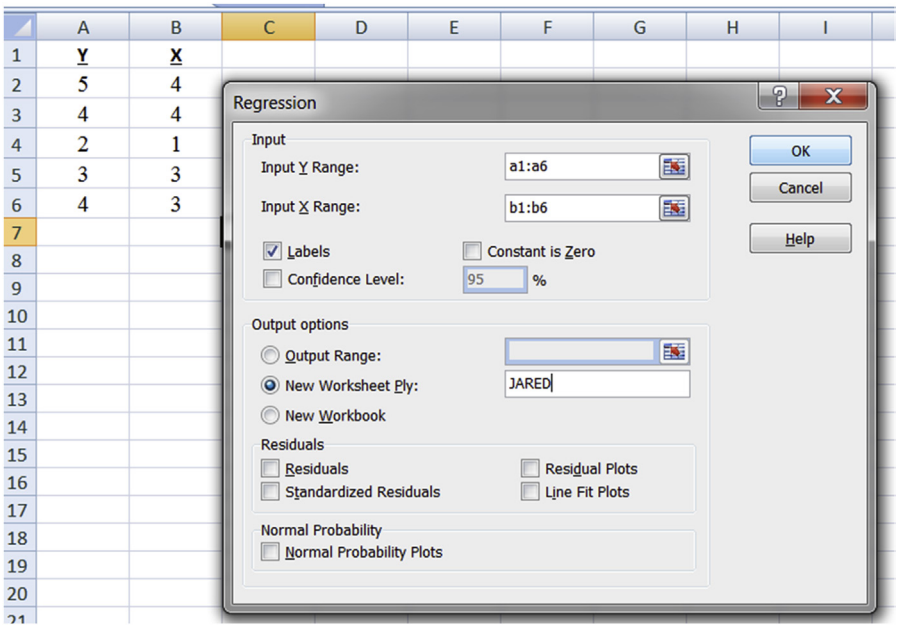
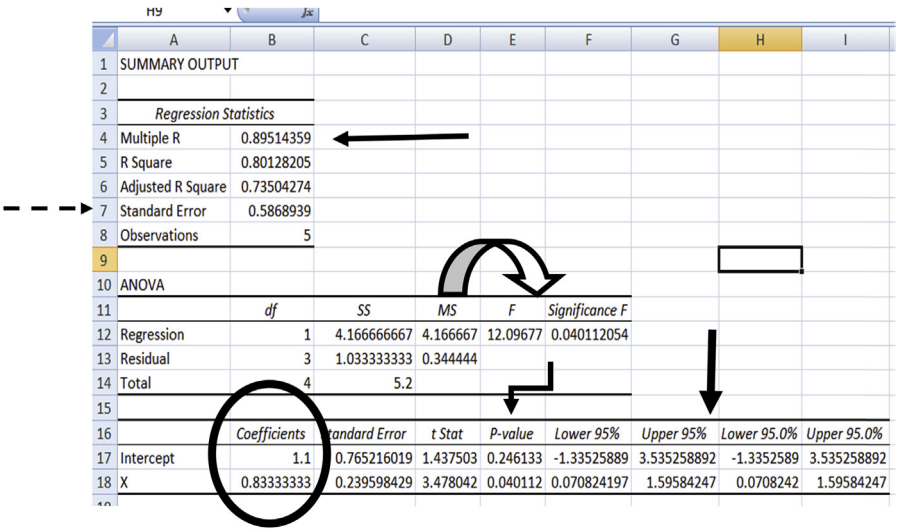



**FIGURE 9.18**

Filling in the Regression dialog box; Excel with illustrative data.



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | *Regression Statistics* | | | | | | | | |
| 4 | Multiple R | 0.89514359 | | | | | | | |
| 5 | R Square | 0.80128205 | | | | | | | |
| 6 | Adjusted R Square | 0.73504274 | | | | | | | |
| 7 | Standard Error | 0.5868939 | | | | | | | |
| 8 | Observations | 5 | | | | | | | |
| 9 | | | | | | | | | |
| 10 | ANOVA | | | | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| 12 | Regression | 1 | 4.166666667 | 4.166667 | 12.09677 | 0.040112054 | | | |
| 13 | Residual | 3 | 1.033333333 | 0.344444 | | | | | |
| 14 | Total | 4 | 5.2 | | | | | | |
| 15 | | | | | | | | | |
| 16 | | *Coefficients* | *tandard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| 17 | Intercept | 1.1 | 0.765216019 | 1.437503 | 0.246133 | -1.33525889 | 3.535258892 | -1.3352589 | 3.535258892 |
| 18 | X | 0.83333333 | 0.239598429 | 3.478042 | 0.040112 | 0.070824197 | 1.59584247 | 0.0708242 | 1.59584247 |

**FIGURE 9.19**

Regression analysis output; Excel with illustrative data.

OK, here's where things really get interesting. If we have a value of X, we can insert it into the equation for the line, and compute Yc, the value of Y that is predicted for the value of X we input. For example, if X = 3, we predict that Y is

$$1.1 + 0.833\,(3) = 3.599$$

But wait, there's more! Check out the correlation coefficient, which is 0.895 (see solid horizontal arrow in Figure 9.19,[3] labeled "Multiple R"). This is a reasonably high value (and, of course, is the same value we found when we did a correlation analysis with these same data earlier in the chapter). Loosely, but pragmatically interpreted, it means we should expect, for the most part, the predicted value of Y and the actual value of Y to be reasonably close to one another. If we examine the data set, we see that the average of the (two) Y values when X = 3 is 3.5, which, indeed, is close to the predicted value, Yc, of 3.599.

If we look right below "Multiple R," we see "R Square," which equals 0.801. As earlier, this indicates that a bit over 80% of the variability in Y (i.e., how come Y is not always the same!!) is due to the fact that X is not always the same. Indeed, if X were always the same, the variability in Y would be only about 20% as much as it is now.

In addition to the least-squares line and the correlation coefficient (and its square, $r^2$, the coefficient of determination), there are a few other noteworthy values in the output of Figure 9.19.

If you look at the bottom right of the output (see vertical arrow), you see a 95% confidence interval for each of the coefficients (i.e., intercept and slope[4]). Let's take them one by one.

Our best estimate of the intercept is 1.1; however, a 95% confidence interval for the true value of the intercept is −1.34 to 3.54. However, we can see that the intercept is not significant, since its *p*-value is 0.24 (see the bent arrow in Figure 9.19). Therefore, we cannot rule out that its true value equals zero. Quite often, however, the intercept is not a quantity that, by itself, is of great value to us.

Now let's look at the confidence interval for the slope. Keep in mind that the slope is crucially important; whether it's zero or not directly indicates whether the variables are actually related. Here, we get a value for the slope of 0.833. The 95% confidence interval of the true slope is 0.071 to 1.596. Its *p*-value (0.040) is below the traditional 0.05 benchmark value. Therefore, at significance level equal to 0.05, the slope is statistically significant.

---

[3] The reader will note that the correlation is labeled "Multiple R." This is simply reflecting oversimplification (sloth?) on Excel's part. Excel did not want to bother writing *simple R* when there is only one X, and *multiple R* when there is more than one X, and decided to just write *multiple R* no matter how many X's there are. We obviously weren't involved in the usability testing. ☺

[4] The reader may note that the confidence intervals for the intercept and for the slope are each written twice! This, again, is simply reflecting laziness on Excel's part. You can specify a confidence level other than 95%, and if you do, Excel gives that confidence interval to you, but also, automatically, gives you the confidence interval for 95%. If you do not specify another confidence level (and one virtually never does so), Excel gives you the confidence interval for the 95% default and then gives you the automatic one for 95%.

So, we now can formally conclude that the two variables are, indeed, linearly related.

### SIDEBAR: EXCEL'S WEIRD LABEL FOR THE *P*-VALUE FOR THE F-STATISTIC

We want to add that the middle section of the output, the ANOVA table (you saw ANOVA tables in several earlier chapters), gives you a *p*-value also, relative to the F-statistic. You can see the F-statistic value of 12.097 (see curved arrow in Figure 9.19); its *p*-value is just to the right of it and equals 0.040. But, wait a moment!!!! *This value is exactly the same as the* p-*value for the slope!!*

For reasons unknown to the authors, Excel calls the *p*-value for the F-statistic "Significance F," but we assure you that this is the *p*-value (and should be called *p*-value!!). Any time we are running a simple regression (recall: this means there is only one X variable), the *F-statistic will have the same* p-*value as the* p-*value for the slope (t-test),* and provide exactly the same information content. In fact, in writing up a report on the results of a simple regression, you would not want to separately discuss the two *p*-values, since it would be a redundancy. In the next chapter, Chapter 10, the *p*-value for the F-statistic and that for the slope will have different values and will mean different things.

There is one final thing that we wish to impart about the output in Figure 9.19, and that is the "Standard Error," as listed in row 7 in the top section of the output (see dashed horizontal arrow in Figure 9.19). Its value equals 0.587, and its notation is often: Sy.x. This is a key value for finding a confidence interval for a prediction, often a very important thing to find. In essence, this is the standard deviation estimate of the error of a prediction if we had the correct regression line. However, we do not have the exact correct regression line (finding which, in theory, would require infinite data!!). However, if the sample size is reasonably large (say, at least 25), and we are predicting for a value of X that is near the mean of our data, we can, as an approximation, use the standard error value as if it were the overall standard deviation of the prediction. With this caveat, the formula for a 95% confidence interval for a prediction is

$$\text{Yc} \pm \text{TINV}\,(0.05,\, n\text{–}2)\ *\ \text{Sy.x},$$

where "$n$" is the sample size (in this example, $n = 5$) and TINV is an Excel command that provides a value from the t-distribution. The first value (i.e., 0.05) reflects wanting 95% confidence—it would be 0.01 for 99% confidence, 0.10 for 90% confidence, etc.; the second value, ($n$–2), is a degrees-of-freedom number—you really don't need to know the details/derivation of why that value is what it is—it is easy to determine, since you know the value of $n$, the sample size, and hence, you obviously know the value of ($n$–2). For our earlier example, where we predicted a value of Yc to be 3.599, a 95% confidence interval for what the value will actually come out *for an individual person* is:

$$3.599 \pm \text{TINV}\,(0.05,\, 3)\ *\ (0.587)$$

$$3.599 \pm (3.182)\ *\ (0.587)$$

$$3.599 \pm 1.865$$

or

$$(1.734 \text{ to } 5),$$

with the realization that we cannot get a value that exceeds 5.

If we determine the actual value for the confidence interval, using the relatively complex formula, we would get 1.339 to 5, a bit wider, but not that different, even though, after all, $n$ is only 5. In a real application, in which $n$ is not so small (such as in the Behemoth.com data set, in which $n = 180$), the difference from the theoretically true confidence interval will be very much smaller, and virtually for sure, the difference will be immaterial. Indeed, the difference is not that big *even with our sample size of only 5*!

This confidence interval for what will actually occur in an individual case when X = 3 is wider that you might like, but that is because, as we have noted, it is based on only five data values. If we had pretty much the same results for $n = 30$, the interval would be much more precise, around 2.45 to 4.75.

### 9.4.2 SPSS

Figure 9.20 shows the same sample data in SPSS. We have already gone into "Variable View" to label the columns Y and X.

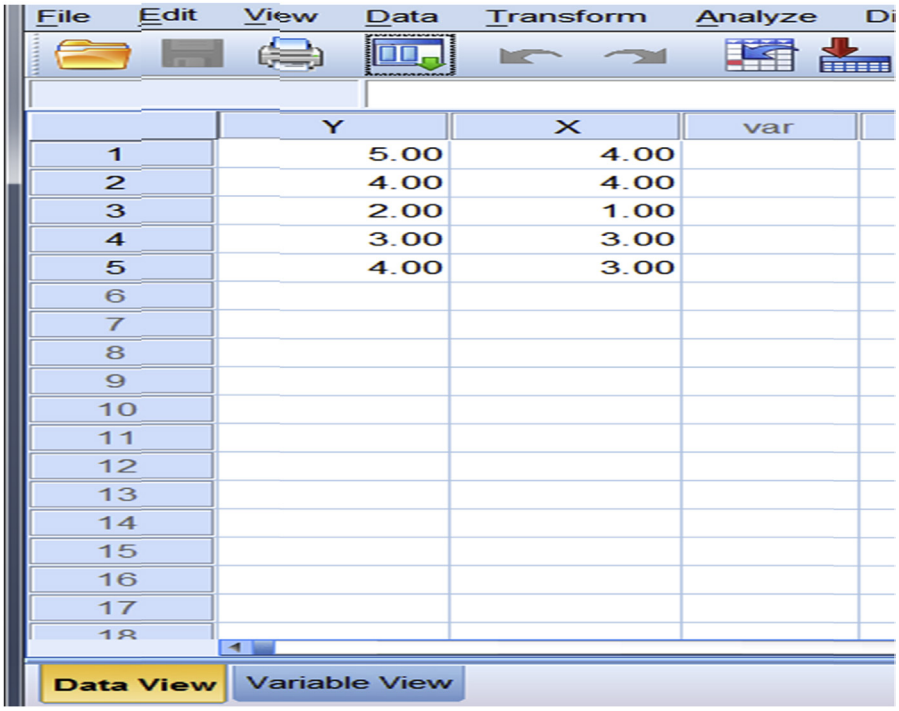



**FIGURE 9.20**

SPSS template for illustrative data regression analysis.

We now pull down "Analyze," and go to "Regression," and then "Linear," as shown in Figure 9.21 (see arrows).

After we click/let go of "Regression/Linear," we get the dialog box shown in Figure 9.22.

We now drag "Y" over to the "Dependent" rectangle, and X to the "Independent" rectangle. (We weren't kidding you when we said that you needed to become familiar with the terms "dependent variable" and "independent variable"!) This is shown in Figure 9.23 (see arrows).

We are now ready to click "OK" and obtain our output. The output is shown in Figure 9.24.

The output has exactly the same values that the Excel output had, although it does not have the confidence intervals for the coefficients, but they're easy to determine. See Sidebar coming up. The coefficients are circled; notice that SPSS calls the intercept the "constant." That's fine, and we point it out only to illustrate how each

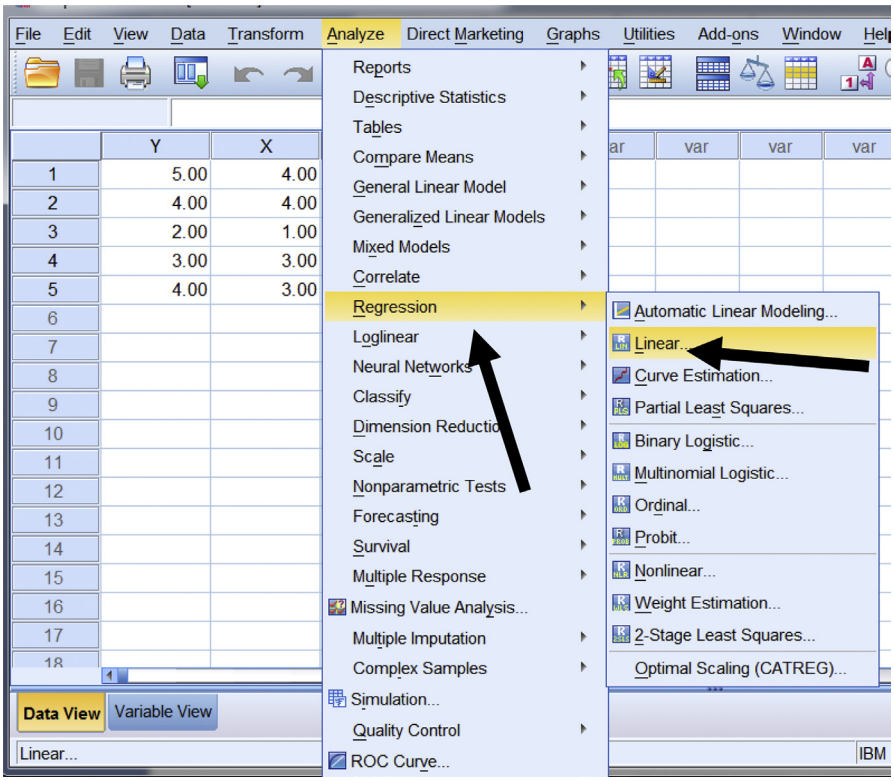



**FIGURE 9.21**

Accessing linear regression in SPSS; illustrative data.

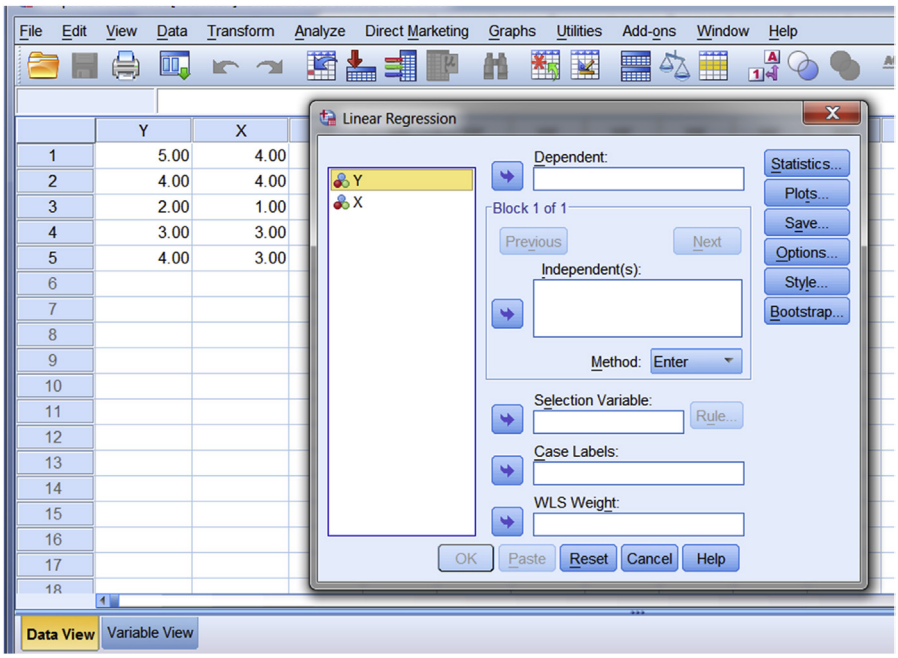



**FIGURE 9.22**

The Linear Regression dialog box; SPSS with illustrative data.

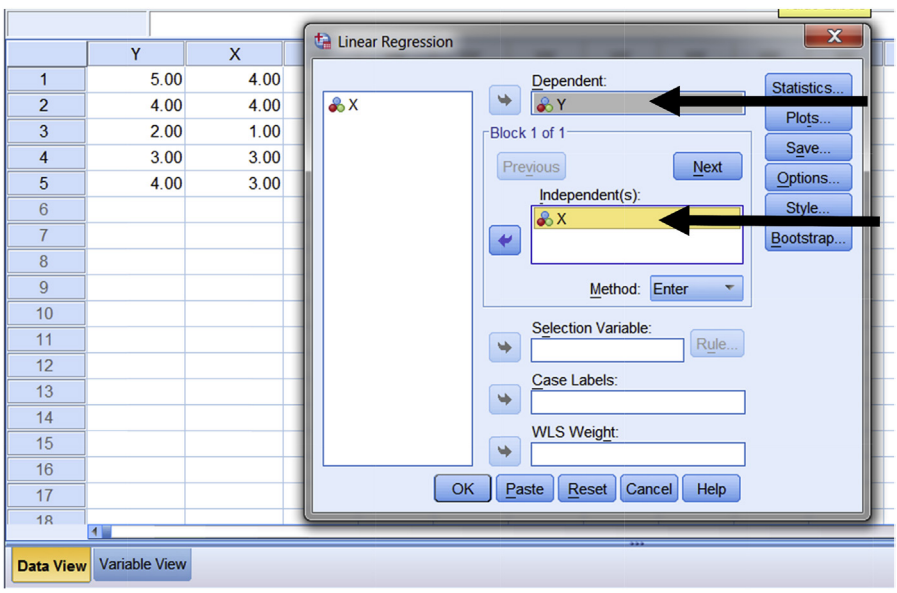



**FIGURE 9.23**

Getting ready to receive the linear regression output; SPSS with illustrative data.

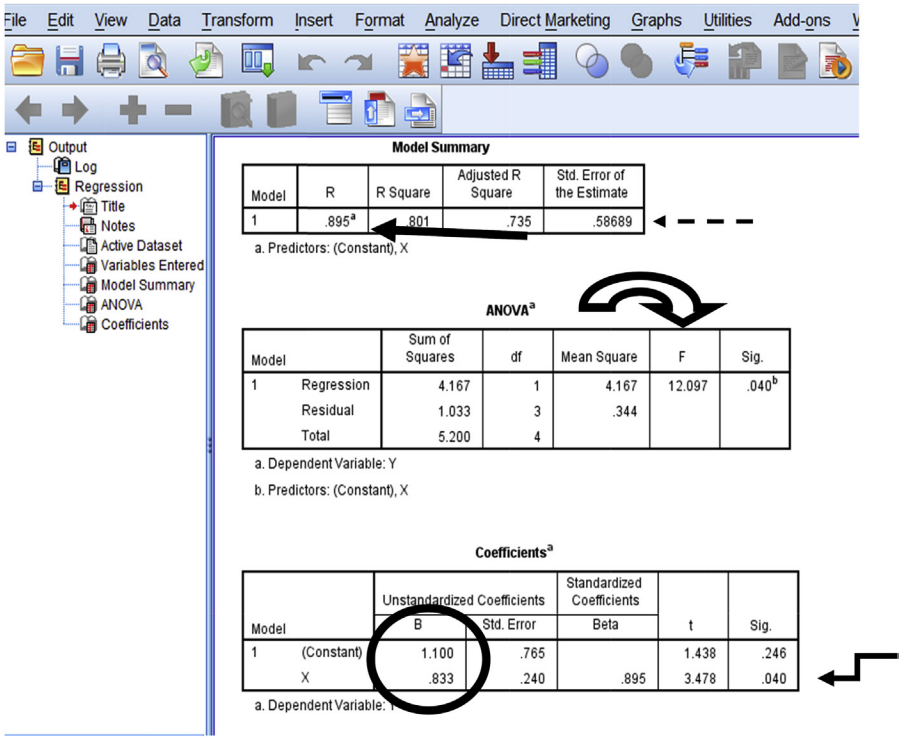



**FIGURE 9.24**

Linear regression output; SPSS with illustrative data.

statistical software package often uses slightly different (but still "sensible") names for selected quantities. We have also added the same types of arrows that are in Figure 9.19, so you will easily see the correspondence between the Excel and SPSS outputs.

## SIDEBAR: CONFIDENCE INTERVALS FOR INTERCEPT AND SLOPE IN SPSS

If you want the confidence intervals for the intercept ("constant") and slope to be displayed in SPSS, follow these instructions:

Go back to Figure 9.23, and click "Statistics," as in Figure 9.25 (see horizontal arrow), and then click "Confidence intervals" under the "Regression Coefficients" section (see vertical arrow in Figure 9.25).

Finally, click "Continue" (see dashed arrow in Figure 9.25), which takes you back to Figure 9.23.

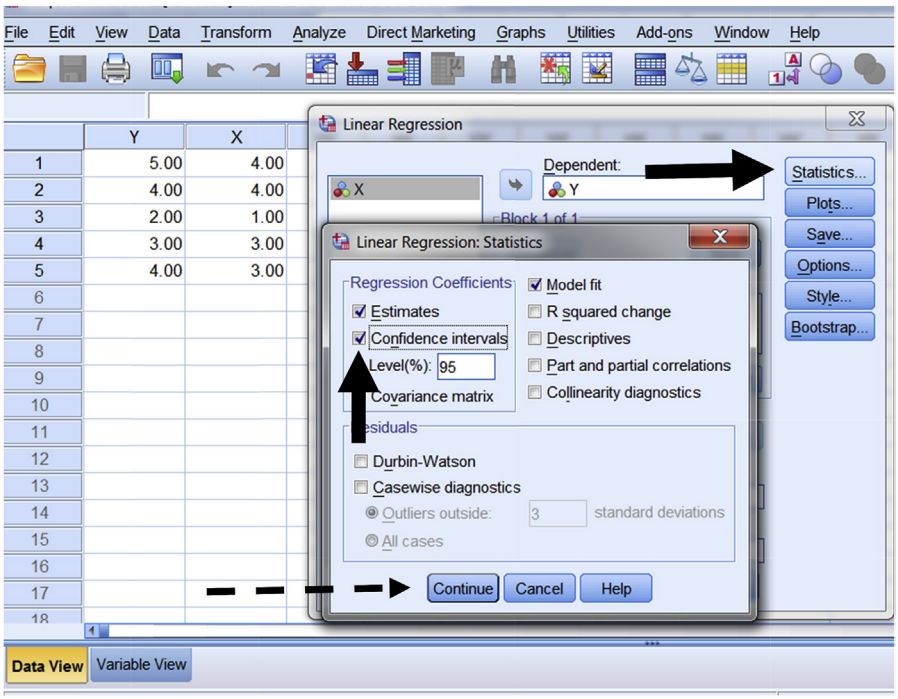



**FIGURE 9.25**

Extra command in SPSS to obtain confidence interval for intercept/constant and slope; illustrative data.

Now the output looks like Figure 9.26, the same as in Figure 9.24, except for the added confidence intervals for the intercept and slope, as circled in Figure 9.26. The values, of course, are the same values we obtained in the Excel output.

## 9.5 LINEAR REGRESSION ANALYSIS OF BEHEMOTH.COM DATA

OK, we're finally ready to apply our linear regression technique to the 180 data points of the Behemoth.com data (see Figure 9.10). Applying the same steps we used for the sample data to the Behemoth data, we get the results shown in Figure 9.27.

We can see that, of course, the correlation is the same as it was in the correlation analysis section: 0.449.

The best-fitting (i.e., least-squares) line is (see circle in Figure 9.27):

$$Yc = 2.456 + 0.460 * X$$

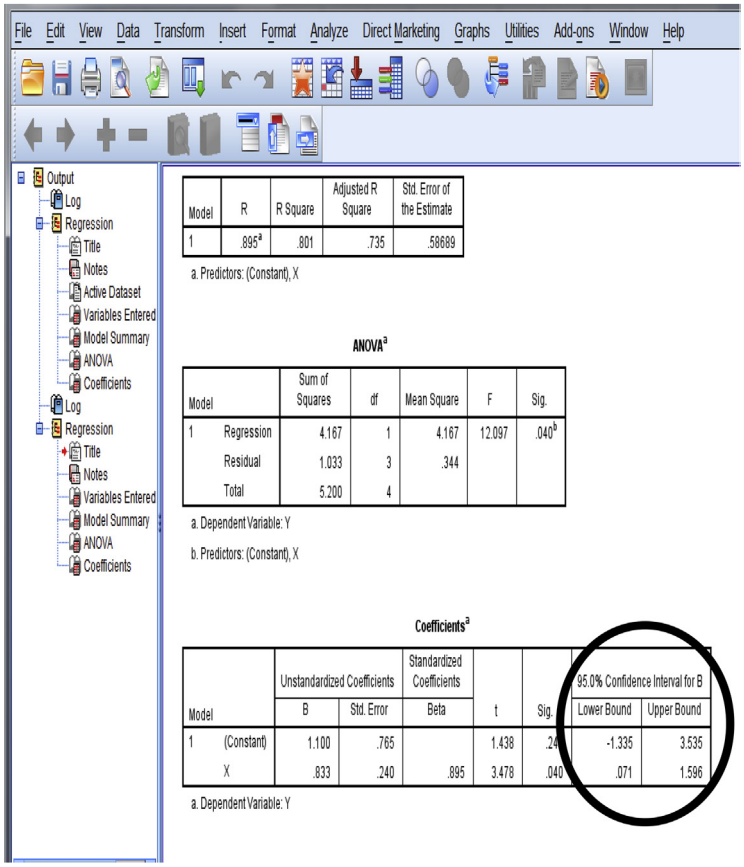
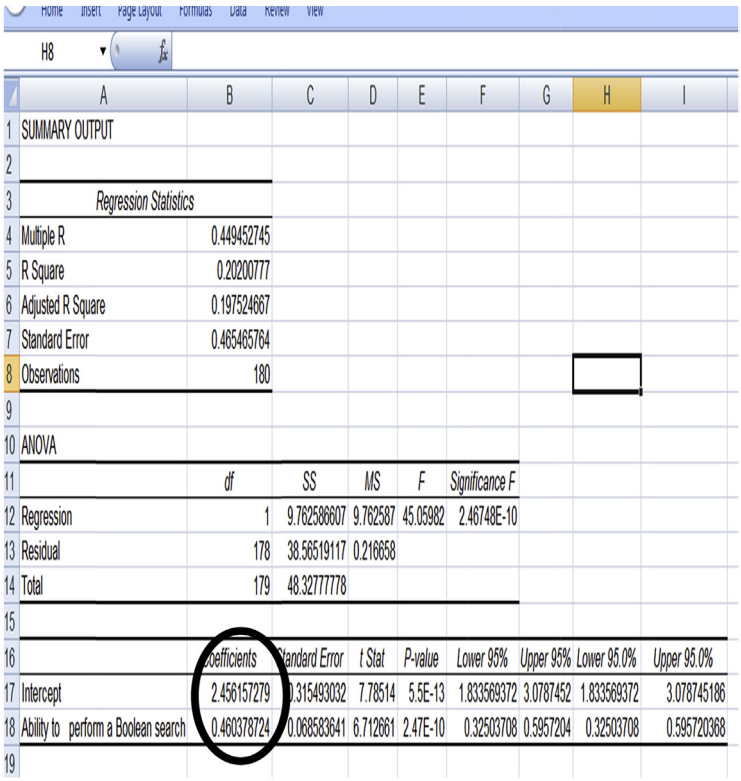



**FIGURE 9.26**

Augmented SPSS output; illustrative data.

So, for each unit increase in the perceived usefulness of the ability to do a Boolean search, there is a corresponding increase of 0.46 (on the 5-point Likert scale) of "Likelihood of Adoption" of the search engine. A person who gives a 5 for usefulness of the ability to do a Boolean search has a predicted value of about 4.76 on the 5-point scale of likelihood to adopt the search engine.

A confidence interval for this value, using the methodology described in the previous section, is 3.85 to 5.00. A person who gives a 4 for the usefulness of doing the Boolean search has a predicted value of 4.30 for likelihood of adoption of the search engine. A confidence interval for this value is 3.39 to 5.00. We would, of course, prefer narrower intervals, but we must remember that we are using only one X variable in the prediction process, and while the $r^2$ is about 20%, which is often very good



| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | | | |
| 2 | | | | | | | | | |
| 3 | *Regression Statistics* | | | | | | | | |
| 4 | Multiple R | 0.449452745 | | | | | | | |
| 5 | R Square | 0.20200777 | | | | | | | |
| 6 | Adjusted R Square | 0.197524667 | | | | | | | |
| 7 | Standard Error | 0.465465764 | | | | | | | |
| 8 | Observations | 180 | | | | | | | |
| 9 | | | | | | | | | |
| 10 | ANOVA | | | | | | | | |
| 11 | | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| 12 | Regression | 1 | 9.762586607 | 9.762587 | 45.05982 | 2.46748E-10 | | | |
| 13 | Residual | 178 | 38.56519117 | 0.216658 | | | | | |
| 14 | Total | 179 | 48.32777778 | | | | | | |
| 15 | | | | | | | | | |
| 16 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 95.0%* | *Upper 95.0%* |
| 17 | Intercept | 2.456157279 | 0.315493032 | 7.78514 | 5.5E-13 | 1.833569372 | 3.0787452 | 1.833569372 | 3.078745186 |
| 18 | Ability to perform a Boolean search | 0.460378724 | 0.068583641 | 6.712661 | 2.47E-10 | 0.32503708 | 0.5957204 | 0.32503708 | 0.595720368 |
| 19 | | | | | | | | | |

**FIGURE 9.27**

Linear regression output for Behemoth.com data; Excel.

(i.e., high) for just one X variable, there is still 80% of the variability ($1-r^2 = 0.8$) unaccounted for.

The *p*-value of the slope is 0.000; thus, the slope is highly significant; remember that the value "2.47E-10" actually means 0.000000000247 as the notation indicates that we should move the decimal point 10 places to the left. It is, of course, the same *p*-value we found much earlier when we analyzed the correlation. This very low *p*-value indicates that there is virtually no doubt that there is a positive linear relationship between the *usefulness of the Ability to do a Boolean search, and the Likelihood of Adoption of the search engine. Furthermore, the r-square value of 0.202 means we estimate that the former, by itself, explains more than 20% of the responder's choice for the Likelihood of Adoption of the search engine query.*

The results are the same, of course, in SPSS. See Figure 9.28.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .449[a] | .202 | .198 | .46547 |

a. Predictors: (Constant), VAR00001

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 9.763 | 1 | 9.763 | 45.060 | .000[b] |
| | Residual | 38.565 | 178 | .217 | | |
| | Total | 48.328 | 179 | | | |

a. Dependent Variable: VAR00002

b. Predictors: (Constant), VAR00001

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 2.456 | .315 | | 7.785 | .000 |
| | VAR00001 | .460 | .069 | .449 | 6.713 | .000 |

a. Dependent Variable: VAR00002

**FIGURE 9.28**

Linear Regression output for Behemoth.com data; SPSS.

## 9.6 MEANWHILE, BACK AT BEHEMOTH

The statistical analysis complements your qualitative findings perfectly. The recruiters complained mightily that the Novix search engine had eliminated Boolean searching; now you find that a higher sense of usefulness of Boolean search capability is associated with an increased likelihood of adoption of the search engine that contains a Boolean capability—with statistical significance! Recruiters want Boolean, and they're more likely to use your engine if you have it. Killing Boolean was a mistake—and a costly one.

You send the results to Hans in an e-mail, and within a minute, he pops his head in your cube.

"You sure 'bout all this stats stuff?"

"Yes, but I can walk you through the details if you like."

"Nah. But do me a favor; put it all in a nice, sharp Powerpoint. Max 5 pages. Bullets for everything. Keep the stats jargon to a minimum, but give enough details so Joey will think you know what you're talking about."

"Sure, how's early afternoon tomorrow?"

"Perfect, I've got my running 1pm meeting with him tomorrow. Maybe the Chianti he likes during lunch will defuse the inevitable explosion!"

## 9.7 SUMMARY

In this chapter, we have introduced correlation and regression analysis. Both of these techniques deal with the relationship between a "dependent variable" or output variable that we label "Y," and an "independent variable" or input variable that we label "X."

The correlation, r, is a dimensionless quantity that ranges between −1 and 1, and indicates the strength and direction of a linear relationship between the two variables; the (hypothesis) test of its significance is also discussed. We also note that the coefficient of determination, $r^2$, has a direct interpretation as the proportion of variability in Y explained by X (in a linear relationship).

We consider example scatter diagrams (graphs of the X, Y points) and discuss how they correspond with the respective values of r. We also demonstrate in both Excel and SPSS how to obtain the correlation.

Regression analysis quantifies the linear relationship between Y and X, by providing a least-squares line from which we can input a value of X and obtain a predicted (best estimate) value of Y, using the line's corresponding slope and intercept. We note how to perform a regression analysis in both Excel and SPSS, and discuss various confidence intervals of interest, as well as hypothesis testing to decide if we should conclude that there truly is a linear relationship between Y and X "beyond a reasonable doubt." In each case—correlation and regression—our illustrations use a small data set that is easier for the reader to follow, and then we apply the technique to the prototype real-world data from Behemoth.com.

## 9.8 ADDENDUM: A QUICK DISCUSSION OF SOME ASSUMPTIONS IMPLICIT IN INTERPRETING THE RESULTS

When we perform "statistical inference" (more or less, for us, confidence intervals, and hypothesis testing) in a correlation or regression analysis, there are three theoretical assumptions we are technically making.

One assumption, called "normality," says that if we hold X constant at any (and every) value, and were to look at many values of Y at that X value, the Y values would form a normal distribution.

A second assumption, called "constant variability" (or often by the ugly word, "homoscedasticity," which is said to mean "constant variability" in Greek [and sometimes, it is spelled with the first "c" being a "k"]), says that the normal curves for each X have the same variability (which as we might recall from Chapter 1,

means equally tall and thin/short and fat curves) for all values of X. For this to be exactly true, it is often a bit dubious.

However, these two assumptions are referred to as "robust." Essentially, this means that if the two assumptions are "moderately violated," it does not materially affect the results of the analysis. In the world of user experience data, it is unlikely that any assumption violations are sufficiently large to affect the results materially. There are ways to test these assumptions, but they are well beyond the scope of this chapter.

The third assumption, called "independence," is that the data points are independent. This is a more critical assumption (because it is not robust), but is usually the easiest to avoid violating. If each respondent provides one row of data and there is no connection between the respondents/data points, the assumption is generally satisfied fully.

Overall, the majority of people who perform correlation and regression analyses do not worry much about these assumptions, and in the vast majority of cases, there is no problem with concluding an accurate interpretation of the results. Still, if the results arrived at seem to very much belie common sense, perhaps somebody familiar with these assumptions should be called upon for consultation.

## 9.9 EXERCISE

1. Consider the Excel data in the file "Chapter 9.Exercise 1," which has 402 data points on Y (column A) and X (column B).
   - **a.** Run a correlation analysis. Is the correlation significant at $\alpha = 0.05$? What percent of the variability in Y is explained by the linear relationship with X?
   - **b.** Run a regression analysis. What is the least-squares line? What do you predict Y to be when X = 4?
   - **c.** Repeat parts (a) and (b) using SPSS and the data in the file named "Chapter 9..Exercise 1.data." The output is in a file named "Chapter 9..Exercise 1.output."

   The answers are in a Word file named, "Chapter 9.Exercise 1.ANSWERS."