Fundamentals of Correlation and Regression

Reagan Brown

Introduction

1

Correlation and regression and their place in the universe.

Introduction

Correlation and regression are statistical tools that are used to assess the relationship between variables and to predict scores on a variable given scores on other variables. To predict and associate. Very useful functions. All very cool. And, I think you'll be surprised at how simple it all is. Things will get complicated in time; even then, the advanced concepts are just a series of additions, none too complicated, to a rather simple foundation.

Predictive Versus Explanatory Research

Let's talk about uses for correlation and regression. Correlation and regression are tools, nothing more. As with ANOVA, *t* tests, and chi squares, they are tools we use to analyze data. *Why* we analyze these data is another story. Psychological research can serve two general purposes: prediction and explanation. Predictive research is conducted to see how well a variable (or set of variables) predicts another variable. If we determine that the relationship between these variables is strong enough for applied purposes, then predictive research is also concerned with establishing the means for making these predictions in the future.

Explanatory research is concerned with causal issues. "Explanation is probably the ultimate goal of scientific inquiry, not only because it satisfies the need to understand phenomena, but also because it is the key for creating the requisite conditions for the achievement of specific objectives" (Pedhazur, 1997, p. 241). Stated differently, understanding causality is important because if we understand *how* something occurs, we have the means to change *what* occurs. That's powerful stuff.

Thus, with explanatory research we seek to understand why something is occurring. Why do children succeed or fail in school? Why do people feel

satisfied or dissatisfied with their job? Why do some people continually speak in the form of questions? It should be obvious that explanation is more difficult than mere prediction. With prediction we don't care why something is happening. All we want to do is predict it. Understanding why something is occurring may help to predict it, but it's not necessary. Explanation requires more than simply finding variables related to the dependent variable – it requires the identification of the variables that actually cause the phenomenon. Many variables, although not actually causing the phenomenon, will predict simply because they are related to causal variables. Many variables predict, but only a subset of these variables are the actual causes.

So, explanatory research is more difficult than predictive research. What may be surprising is that the analytical tools used for predictive and explanatory research are sometimes the same. That's right, correlation and regression can be used for both types of research.

The Role of Theory

You might ask, how then is predictive research different from explanatory research, aside from their end goals? The answer is they can involve different analytic tools, but there are some other important differences. Foremost among these is the role of theory. Theory need not play any role at all in predictive research. It's possible to go completely theory-free and have successful predictive research. Just try a bunch of variables and see what works. Because it doesn't matter *why* something predicts, we don't have to possess a good reason for trying and using a variable if it predicts. That said, predictive research based on a sound theory is more likely to succeed than theory-free predictive research.

The situation is completely different for explanatory research. A sound theoretical basis is essential for explanatory research. Because explanatory research is all about why different outcomes occur, we must include all of the relevant variables in our analysis. No throwing a bunch of variables in the experiment just to see what works. A set of variables, chosen with little regard to any previous work, will not likely include the actual cause. (Also, including too many irrelevant variables can corrupt our analysis in other ways.) Furthermore, there is no way to fix explanatory research that was incorrectly conceived. "Sound thinking within a theoretical frame of reference and a clear understanding of the analytic methods used are probably the best safeguards against drawing unwarranted, illogical, or nonsensical conclusions" (Pedhazur, 1997, p. 242). I don't know about you, but I don't want to draw unwarranted, illogical, or nonsensical conclusions.

The following hypothetical study illustrates the differences between predictive and explanatory research. In this study, researchers measured the number of classical music CDs, books, computers, and desks in the houses of parents of newborns. Ten years later they measured the mathematical intelligence of these children. An analysis revealed that the combined number of classical music CDs and desks strongly correlated with mathematical intelligence.

The first issue to address is: Is this sound predictive research? Yes, the number of classical CDs and desks are strongly related to mathematical intelligence and can be used to predict math IQ scores with excellent accuracy. (Just a reminder, this study is not real. I had a lot of fun making it up.)

A second question is: Is this sound explanatory research? No, and it's not even close. These variables were chosen simply because they corre-

lated with the dependent variable, not because there was a logical reason for them to affect math ability. To think that the possession of these items is the cause of mathematical intelligence for these children is to make the classic mistake of equating a strong relationship with a causal relationship. If you're still not convinced, ask yourself this: Would supplying classical music and furniture to households of newborns that didn't have those items raise the math scores of children living in those households? The cause of a given variable is also the means for changing the status of people on that variable.

Let us close this section by stating that correlation and regression analysis are statistical tools that can be used for both predictive and explanatory research. A sound theoretical foundation is helpful for the former, essential for the latter.

Research Designs

No chapter that even mentions causal, or explanatory, research would be complete without a short discussion of research design. Statistics are fun and all, but it is the research design (and associated methodology) that allows us to draw, or prevents us from drawing, clear conclusions about causality.

The three basic research designs are: the true experiment, the quasi experiment, and the non experiment (also called a correlational study, but that's a terrible name). These three designs differ in two aspects: how subjects are assigned to conditions (through a random or non random process) and whether the independent variables are manipulated by the experimenter. Some variables can be manipulated, like type of reinforcement schedule, and some can't, like height or SAT score.

Put these two factors together, and we get our three basic types of experimental designs (Chart

CHART 1 Experimental Design Characteristics

	Random Assignment	Manipulation
True Experiment	✔	✔
Quasi Experiment	✘	✔
Non Experiment	✘	✘

1). The true experiment has random assignment to groups and a manipulated independent variable. Due to the random assignment, the groups likely begin the study equal on all relevant variables, meaning that after the manipulation has occurred the differences observed between the groups on the dependent variable are the result of the experimenter's manipulations (i.e., the independent variable). The great advantage of this design is that, if done correctly, causal claims are clear and easy to substantiate. There are some disadvantages to this design, but let's not concern ourselves with those.

In the quasi experiment, people are not randomly assigned to groups, but there is a manipulated independent variable. Aside from the lack of random assignment, the quasi experiment is like the true experiment. However, that one difference makes all of the difference. The non random assignment to groups is a fundamental weakness. Only random assignment offers any assurance that the groups start out equal. And if the groups start out unequal, there is no way to know if the observed differences on the dependent variable are due to the manipulated variable or to pre-existing differences. To summarize, there are an infinite number of possible causes for the differences observed on the dependent variable, of which the independent variable is but one. At least, however, the manipulated variable is a good candidate for the cause. So there's that. You may be asking, "If there are so many problems that result from not randomly assigning people to groups, why would anyone ever fail to randomly assign?" The answer

is sometimes we are simply unable to randomly assign people to groups. The groups are pre-existing (i.e., they were formed before the study) and unalterable. An example would be the effect of two different teaching techniques on classes of introductory psychology students. The students picked the class (including instructor, dates, times and locations); it is not possible for the researcher to assign them, randomly or otherwise, to one class or the other. That's the real world, and sometimes it constrains our research.

In the third design, the non experiment, people are not randomly assigned to groups; there is also no manipulation. In fact, there are often not even groups. A classic example of this type of design is a study designed to determine what causes success in school. The dependent variable is scholastic achievement, and the independent variable is any number of things (IQ, SES, various personality traits). You will note that all of these various independent variable are continuous variables – there are no groups. And of course, nothing is manipulated; the people in the study bring their own IQ status (or SES or what have you) with them. As with the quasi experimental design, there are an infinite number of possible causes for the differences observed on the dependent variable. However, because nothing was manipulated in the non experiment, there isn't even a good candidate for causality. Every possible cause must be evaluated in light of theory and previous research. It's an enormous chore. So why would anyone use this design? Well, some variables can't be manipulated for ethical reasons (e.g., the effects of smoking on human health) or practical reasons (e.g., height). Conducting research on topics where the independent variable can't be eliminated requires researchers to make the best of a bad hand (to use a poker metaphor).

Terminology: Independent Variable or Predictor?

I've used the term independent variable in the previous section without defining it. Independent variable has a twofold definition. An independent variable is a variable that is manipulated by the researcher; it is also a presumed cause of the dependent variable. Another oft-used term, similar to independent variable, is predictor. A predictor differs from independent variable on both parts of the previous definition. A predictor is not manipulated by the researcher, and causal claims are not being made with it.

So there we have it, independent variable and predictor, two terms describing the variable that that starts the study. Independent variables are manipulated and causal. Predictors are not manipulated and are non causal. That seems simple enough, but what of research conducted within a non experimental design where the variable is not manipulated but is thought to be the cause of the dependent variable? This situation is rather confusing, but the causality issue is likely the more relevant factor, making this variable an independent variable.

One last issue, when the term predictor is used, the variable analogous to dependent variable is referred to as the criterion. So it's independent and dependent variables when causality is an issue and predictor and criterion variables when it is not.

One Last Thing Before We Proceed

I didn't invent any of this stuff. In this book I am merely explaining concepts and principles that long ago entered into the body of foundational statistics knowledge. The origins of regression analysis date to work by Gauss and Legendre some two hundred years ago. Everything important about correlation was described over a century ago by

two other researchers. Correlation and regression analysis are not new concepts – they are classics.

Basic Statistics

2

Variability. Sampling Error. Standardized Scores.

It doesn't get any better than this.

Introduction

Before we can discuss correlation and regression, there are a few basic statistics we need to discuss. These should be very familiar concepts, so we won't spend much time on them.

Populations and Samples

There are two terms that are important to understand in the world of statistics. Just to scare you a little bit, failure to understand these terms may mean that you use the wrong equation because, you guessed it, sometimes there are different equations for samples versus populations.

Population: Everyone relevant to a study. If your study is about people in general, then your population consists of every person on the planet. If your study is about students in an art history class being taught a certain way at a certain place, then your population is everyone in that class. Aside from studies with narrowly defined populations, we never measure the entire population. Sometimes researchers like to pretend that they have measured a population just because their sample is big, but they're just pretending.

Sample: A subset of the population. If there are ten million in the population, and you measure all but one, you've measured a sample. Samples can be small ($N = 23$) or large ($N = 10{,}823$). Smaller samples often lead to greater error in our results. So we prefer larger samples. Bad news: Large samples are labor intensive. And if that's not enough, there are other problems to consider. To keep you from demanding a refund at this point, we'll save those issues for later.

Sampling Error

All right, so we'll measure samples and not populations. But saving all of that work comes at a price: **sampling error**. Sampling error is the differ-

ence between the value of a statistic computed in a sample versus the population value of that same statistic. As an example, let's say that we wanted to investigate how well high school seniors know the capitals of the 50 states. Thus, the population consists of every high school senior (remember, the population isn't always everyone on the planet – it's everyone relevant to the study). It is clear that it's too much work to give our state capital test to every high school senior student. So, via a random process we select 100 students and test them. And let's say that their mean number correct is 34. Now that's a sample of people and their mean score represents our best estimate of the mean score for all the senior students. But this estimate is just that, an estimate, and it won't be perfect. Now, for the sake of argument, imagine that we collected data from every single high school senior (i.e., the population). And the mean population score turns out to be 22 correct. Whoa, there's a big difference between our sample value (34) and the population value (22). That difference is sampling error, and it's the price we pay for being lazy. Sometimes sampling error is big, or sometimes, by sheer luck, it works out to be zero for a given study. The rule to remember is this: Larger samples are likely to lead to smaller amounts of sampling error. So, we like large samples. The bigger, the better.

For the "larger samples lead to smaller amounts of sampling error" rule to work, everyone in the population must have an equal chance of being selected for the sample (such a sample is called a **probability sample**). There are a variety of techniques (e.g., simple random sampling, cluster sampling) available to collect probability samples. It's work, but it can be done. But what if the sample isn't a probability sample? The "larger samples lead to smaller amounts of sampling error" rule definitely does not apply if the sample is any type of **non probability sample** (i.e., samples of convenience; volunteer samples; collecting data

from friends, family, and pets). In a non probability sample, some members of the population have no chance of being selected. The classic example of a non probability sample is the use of college students in psychological research. Any sample taken from a college student subject pool will not be representative of any population broader in scope than college students for the simple reason that people who are not college students have zero chance of being selected. Data gathered from a non probability sample, regardless of size, should never be used to draw inferences regarding population characteristics; the validity of such generalizations is unknown and unknowable (Pedhazur & Schmelkin, 1991). No statistical magic exists which would fix the problems caused by a non probability sampling technique.

This next point should be obvious, but I'll state it anyway. The population from which the sample is taken must be the right kind of population. That is, it must be the population that is relevant to the study. Using our state capital example from earlier, if we wanted to know the average score of high school seniors, it wouldn't make sense to draw our sample from the membership of a plumber's union. If the sample is taken from Population A, we shouldn't generalize sample characteristics to Population B. It's not that any such generalization will be automatically wrong; it's that there is no way to know if it is correct.

To summarize, we can state rules regarding inferences from samples as follows. To use sample statistics to make inferences about the population with a minimum of error, a sample must be large, collected via a probability sampling technique, and drawn from the right kind of population.

Finally, these rules apply to all statistics. We used means in our example, but we could have used medians, standard deviations, correlations, or a whole pile of statistics whose names you and I have not yet even heard of. Sampling error af-

fects every statistic that we compute and the only sure way to completely avoid it is to measure the entire population. Since that's too much work, we can minimize the magnitude of sampling error by the use of large probability samples.

Central Tendency

You've probably already learned that there are three types of averages: mean, median, and mode. An average score describes the central tendency of a set of data. The mode is the most frequently occurring value. Consider the following table.

Person	Score
L. Sebastian	22
Kyle	18
S. Joe	29
Shauna	18
Ron	19

The modal score is 18 because it occurs more often (twice) than any other score (all just once each).

The median is the middle score. As an analogy, in a family with three children, who is the middle child? The second one, of course. If there are three scores, then the median is the value of the second score. So to compute a median, just find the middle score and note its value. In the above example, there are five scores so the middle score is the third highest one. The value of the third highest score is 19. Thus, the median score is 19 (not 3 or 29). It should be clear that to compute a median, (a) one must sort the data from highest to lowest (or lowest to highest), (b) find the middle score, and (c) obtain the value of the middle score. OK, new example. What if a family has four children, who is the middle child? It is a little tougher because two kids (the second and the third) tie for the middle spot. We could have the same issue with finding the median score. In

the above dataset, let's say we obtain data from a sixth person, who we will call Lucy. Lucy has a score of 20. That means we have six total scores. The middle scores are the third and fourth highest scores. Note that there are the same number of scores greater than and lesser than these two – that means that you've successfully found the middle value(s). The values of the two middle scores are Ron's 19 and Lucy's 20. To compute the median, split the difference. Thus, the median score is 19.5. To summarize, when we have an odd number of scores, just sort the data and find the value of the middle score. When we have an even number of scores, sort the data, find the values of the two middle scores, and split the difference.

You're probably most familiar with means. To compute a mean (symbolized as μ for populations, $\bar{X}$ for samples), add up the scores and divide by the number of scores. If you like equations, here's one:

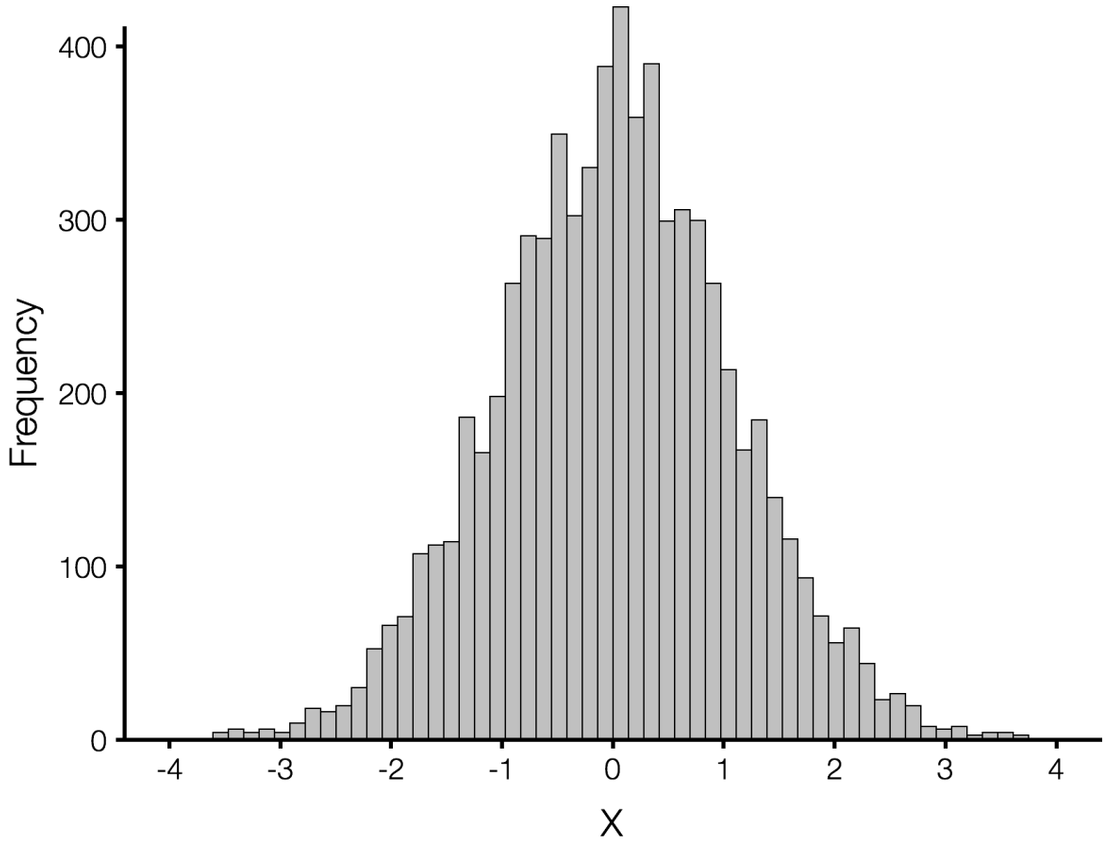
$$\bar{X} = \frac{\sum X}{N}$$

Now that you know three ways to compute average or central tendency, we should talk about the advantages and problems with each. There is no problem with mode, except that nobody uses it. And I mean nobody. Means can be overly influenced by a single extreme score, resulting in a value that is not representative of the dataset. Medians do not suffer from that problem. In fact, one might say that medians are not influenced enough by extreme scores.

Variability

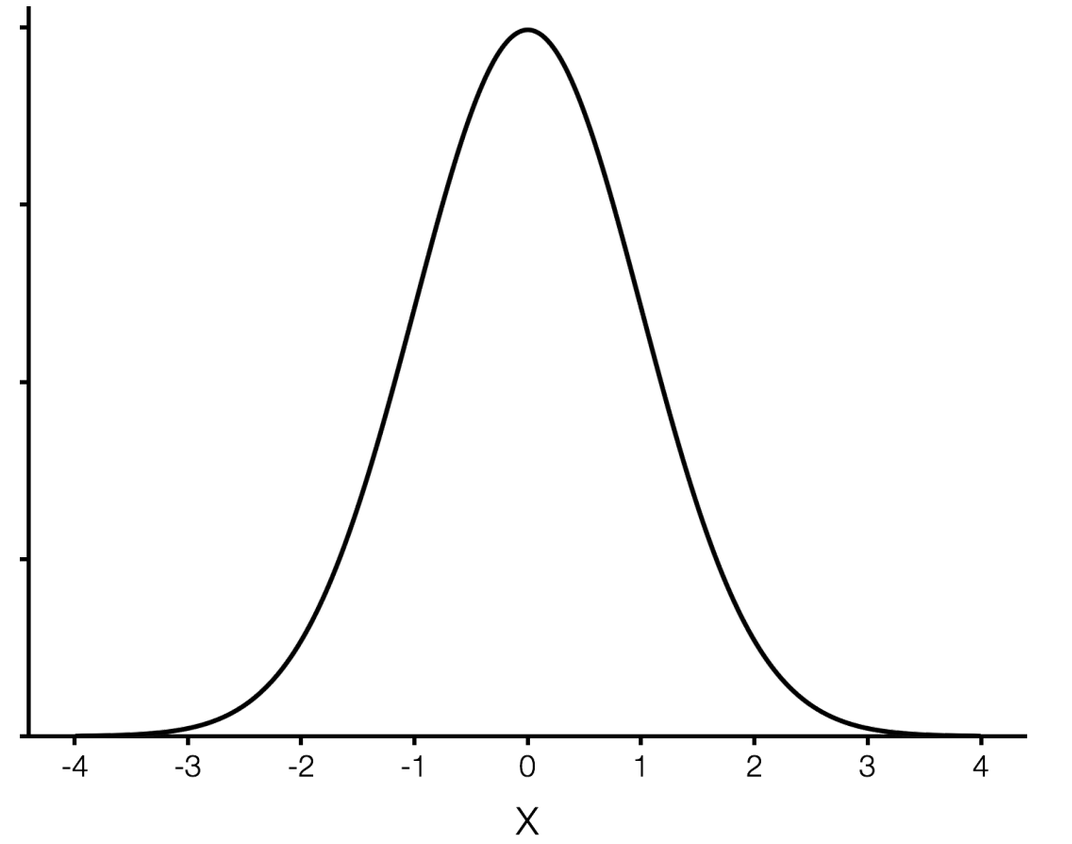
A frequency distribution (also called a histogram) is a graph of scores of a single variable (Figure 1). The *x*-axis indicates the various levels of the variable and the *y*-axis indicates the number of times each value is observed. It sounds fancy, but it's really just a bar graph, the sort of thing you

FIGURE 1 Frequency Distribution (Histogram)



made in third grade. The jaggedness of the bars is because the *X* variable is a variable which has discrete categories (like ACT scores, where the only possible score values are integers – there's no 21.7); in other words, the variable is not truly continuous. With datasets of infinite size (and continuous variables), frequency distributions smooth

FIGURE 2 Probability Density Function



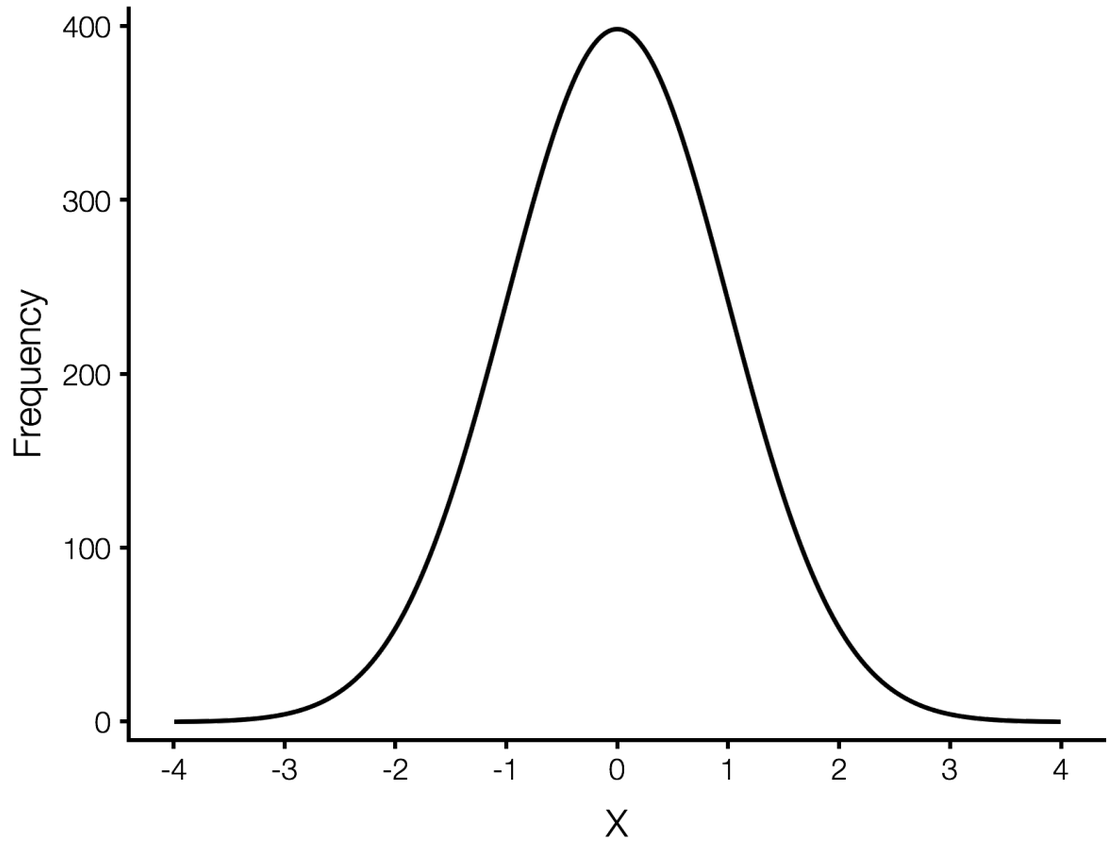
out to something called a probability density function, (see Figure 2). Much nicer, no?

Let us note a few of things in the two distributions. First, not many people have scores that are very low (-2 or -3) or very high (+2 or +3). Most people have scores in the middle ("The meaty part of the normal curve." Costanza, 1997). Second,

the distribution is symmetrical. If you draw a line down the middle, one side is a mirror image of the other. Go ahead, find a mirror and try it. Finally, when the distribution is symmetrical like this one, that line down the middle tells you where the mean is located. In this case, the mean is zero.

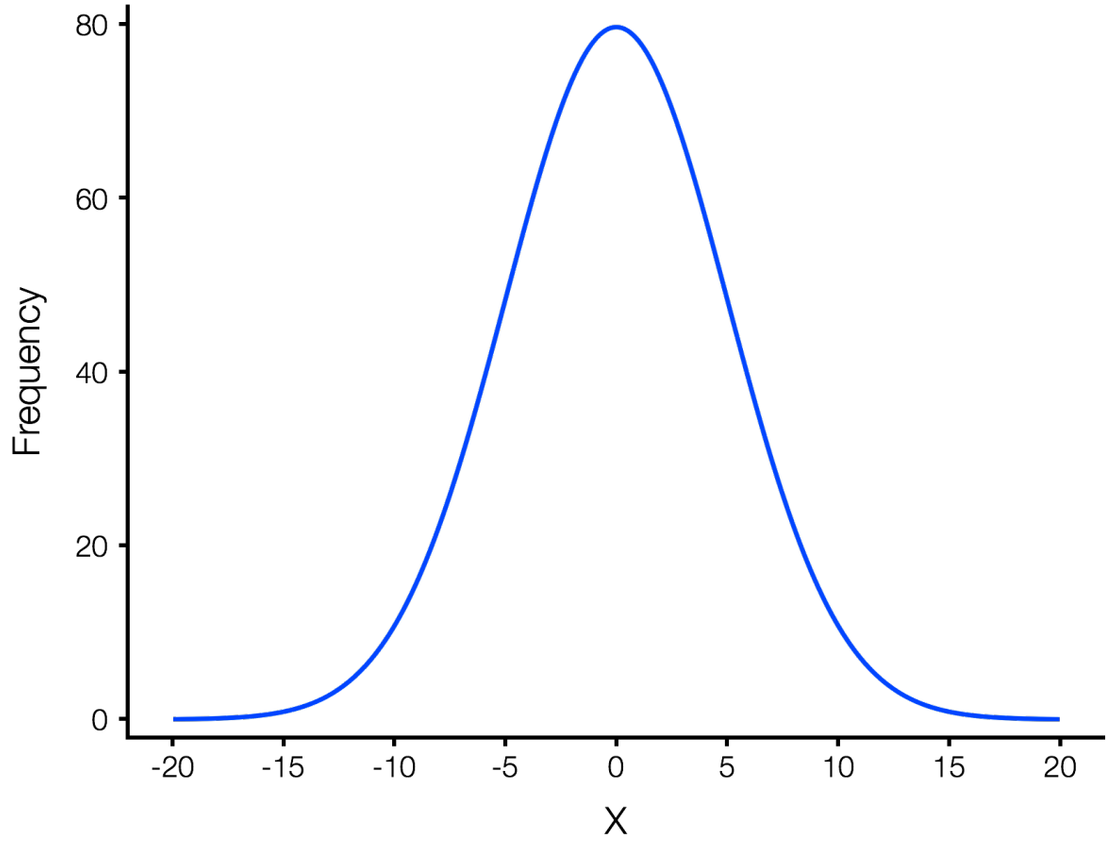
Moving on, distributions for two different datasets are displayed in Figure 3 and Figure 4. What's the difference between the first graph (black distribution) and the second graph (blue distribution)? When they are shown on separate graphs they appear to be the same. They have the same mean score. Notice how the midpoint of each is zero. They have the same sample size (trust me on this). If you've read the title of this section, then you've guessed that the difference is variability. In the first distribution (in black), most (approximately two-thirds) of the scores are within one point of the mean (the mean plus or minus one point). In the second distribution (in blue), very few of the scores are within one point of the

FIGURE 3 Variance Comparison: Low Variance



mean. You have to move out to five points away from the mean (the mean plus or minus five points) in order capture most of the scores. If we place both datasets on the same scale (Figure 5), it's clear that the scores are not spread out in the same way (if Figure 5 seems like a massive cheat, pay careful attention to the scale on the *x*- and *y*-axes on the three graphs).

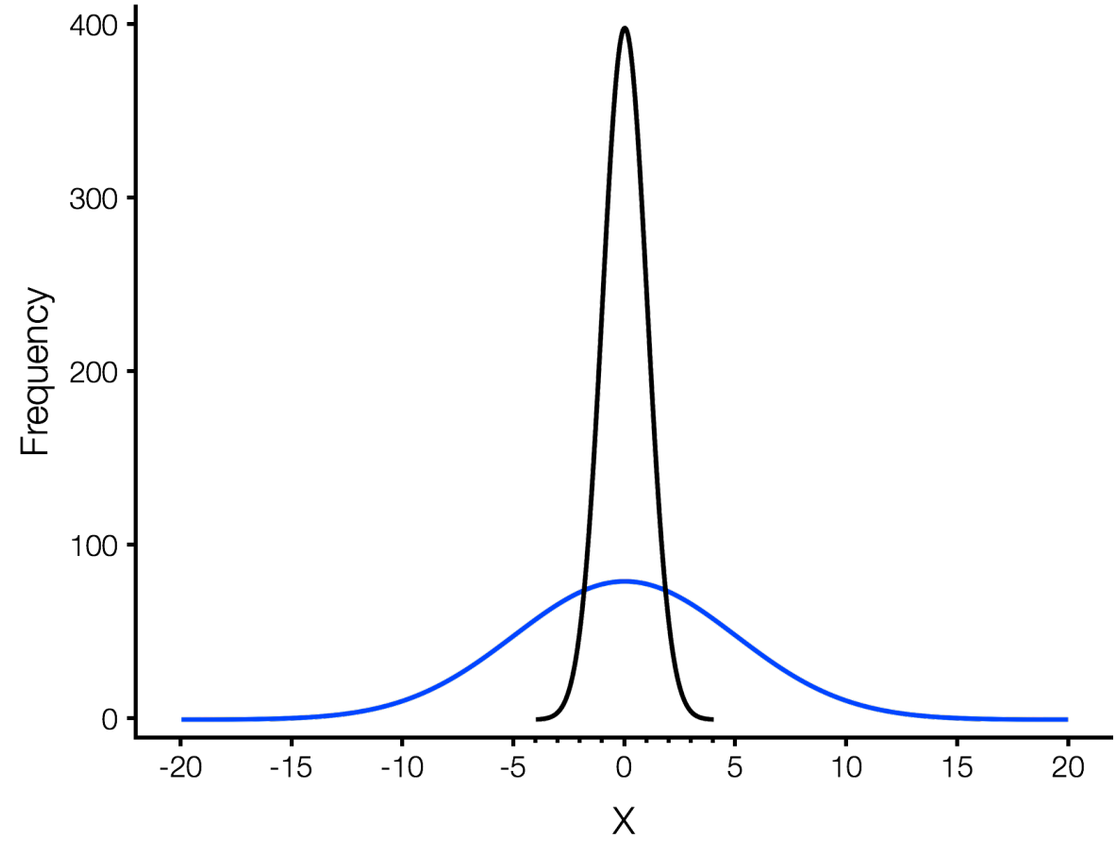
FIGURE 4 Variance Comparison: High Variance



Variability is greater for the blue distribution than for the black distribution. Variability is all about the differences between the scores. There are a number of ways to compute variability, but we'll end up using just two of them.

The simplest measure of variability is called range. The range is simply the difference between the highest and lowest scores. Easy to compute, but such a crude measure of variability. A single outlying score can result in a high range. Thus, it is not a sensitive measure of variability.

FIGURE 5 Variance Comparison: Both Distributions



The measure of variability that we like is called variance (symbolized for populations as σ^2). Yes, the name is a little confusing, so here's a hint.

Variability refers to all of these statistics (including range), whereas *variance* refers to a specific equation (given below for populations). (Just to be clear, the equation below computes variance for a population of data. But wait, you say, I thought people never measure an entire population. True. So why do we need this equation? Before I explain sample variance, I need to explain population variance. All things in due time.)

$$\sigma_X^2 = \frac{\sum (X - \mu)^2}{N}$$

This equation isn't that bad. In fact, it is really similar to the equation for a mean. To see that, take all the parenthetical stuff and call it Q (just to give it a name). The equation is now $\frac{\sum Q}{N}$. In essence, variance is the mean of this Q variable. So variance is the mean of something. Now let's look at the parenthetical component. It's $(X - \mu)^2$. Forget the squared part, focus on $(X - \mu)$. This is called a mean-deviation score and it is the simple difference between a score on X and the mean score. If X equals the mean score, then the mean-deviation score is zero. If X is greater than the mean score, then the mean-deviation score is positive. You get the idea. We'll be computing mean-deviation scores for all people in our dataset. An example is presented below. The mean of X is 6.

Person	X	(X - Mean)
Bennett	3	-3
Tommy	9	3
Todd	4	-2
Matt	8	2

Now we have to deal with the squared-ness. We'll be squaring the mean-deviation scores.

Person	X	(X - Mean)	(X - Mean)^2
Bennett	3	-3	9
Tommy	9	3	9
Todd	4	-2	4
Matt	8	2	4

Remember that Q thing we made up? That's the last column, the squared mean-deviation scores. As we said, variance is just the mean of this thing.

So variance is the mean of the squared mean-deviation scores. In this case, it's $(9+9+4+4)/4 = 6.5$. Another way to describe it is variance represents the average squared difference between each score and the mean. Here's another example.

Person	X	(X - Mean)	(X - Mean)^2
Julianna	9	0	0
Paul	9	0	0
Jennifer	9	0	0
Anthony	9	0	0
Brenden	9	0	0

Variance is, you guessed it, zero. Why? Every score is the same. Thus, the average distance between each score and the mean is nothing. Just for fun, diagram the frequency distribution of this dataset.

So that's the equation for population variance. What about the equation for computing variance when you have measured a sample (which, as we have discussed, is pretty much all of the time)? The equation to compute the variance of a sample of data (when you want an unbiased estimate of the population variance, and, trust me, this is what you want) is:

$$S_X^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

The only difference, aside from the symbol S_X^2 and the replacement of μ with $\bar{X}$, is that instead of dividing by N, we divide by $N - 1$. It is worth noting that the popular statistics programs (e.g., SPSS, SAS) use this $N - 1$ equation to compute variance. And, of course, the $N - 1$ version is the correct equation – unless you happened to have measured a population. And that won't happen on accident. So we'll stick with the sample variance equation from here on.

You might be wondering why we divide by $N - 1$ instead of N with the sample variance equation. Here's the short answer (and feel free to skip this paragraph if you don't care): the $N - 1$ denominator is necessary to obtain an unbiased estimate of the population variance. "Unbiased estimate?" you say. Well, think about it. We measure samples because it's inconvenient (well nigh impossible) to measure the entire population. But, and this is important, we want our sample statistics to represent the population statistic. All of the statistics we have discussed to this point (e.g., mean) were unbiased, meaning that the sample statistic would not consistently yield a value that was too high or too low (stated another way, there was about a 50% chance that the sample statistic would be too high compared to the population value and about a 50% chance that it would be too low). Variance computed in a sample using the N denominator is a biased statistic in that it will consistently yield a value that is less than the population value. And where does the $N - 1$ denominator come in? By dividing the squared mean-deviation scores by $N - 1$, the bias is eliminated and the sample variance equation produces an unbiased estimate of the population value. Aren't you glad you asked? If you want to know why the N denominator version of the equation produces a biased estimate in a sample, that's a much bigger question. There are

proofs for that. Take my word for it; they are not fun.

Our final variability statistic is called standard deviation (symbolized as S_X). If you know variance, then standard deviation is a snap because…

$$S_X = \sqrt{S_X^2}$$

That's right, standard deviation is just the square root of variance. If you know one, you can always compute the other. A clear sign of this is the symbol for each. The variance symbol (S_X^2) has a squared sign and the standard deviation symbol (S_X) doesn't.

You might be tempted to ask, given that variance and standard deviation are basically the same, why do we need both of them? Well that's a good question and I'm glad you asked it. Shows your intelligence. The answer relates to the metric of measurement. If scores on X are how much people weigh in pounds, and the variance comes out to be 85, then we say the variance is 85 pounds *squared* because variance is in squared units. Right away you can see the problem: *squared* pounds. Now imagine that we measured ACT scores. *Squared ACT points?* Variance just doesn't live in the land of regular units of measurement. But standard deviation does. With standard deviation, we're back to pounds, ACT points, and the like – the original metric of measurement. Operating in the original metric of measurement makes it a little easier to determine if a given value is big or small. In squared units, everything looks big.

Standard Scores: Linear z Scores

Standardizing a set of data changes the scores so that they have a useful mean and standard deviation. We call these rescaled scores standard scores. There are many forms of standard scores. We'll discuss a few. Before we get to that, why would anyone use standard scores? As we mentioned in our section on normative inference, test

score metrics (e.g., measuring race results in seconds versus hours, measuring job performance with a 5-point scale versus a 7-point scale) are arbitrary. Thus, it is difficult to interpret a score without knowing something about how other people score on the test. The mean and standard deviation are two pieces of information describing how well other people scored. Both statistics are used to transform raw scores into standard scores. Data expressed in standard scores allow us to interpret how high or low the score is as long as we know the characteristics of the standard scores. Think of standard scores as a neutral playing field for our test scores.

There are many types of standard scores, but the most popular is the linear *z* score (often referred to as just *z score*, but the *linear* word is important because there is a nonlinear version as well). The equation for computing a *z* score is simple.

$$z_X = \frac{(X - \bar{X})}{S_X}$$

X represents the person's score in question. $\bar{X}$ is the mean score and S_X is the standard deviation. So, all we need to know in order standardize a score is: the test taker's score, the mean score, and the standard deviation. That doesn't sound too difficult.

How about an example? Let's say that I took the SAT, and my verbal score (SAT-V) is a 400. The mean of the SAT-V section is 500, and the standard deviation is 100. Now we're ready to go. Plugging in these values into the *z* score equation, we find that my 400 on the SAT-Verbal becomes a *z* score of -1.0.

Let's take a closer look at my *z* score of -1.0. My *z* score is negative. The negative sign tells you something – I did worse than average. If my score was above the mean, my *z* score would have been positive. If my score had been exactly the same as

the mean, my *z* score would have been 0.0. The difference between my score of 400 and the mean is 100 points. The standard deviation is 100 points. Thus, my score of 400 is exactly one standard deviation below the mean. The *z* score is -1.0. Do you see where this is going? I'm not this redundant on accident. Here it comes: A z score is literally the number of standard deviations a score deviates from the mean. In case that's not clear, I'll restate the definition in the form of a question: How far (in terms of number of standard deviations) from the mean (above or below) is this score? If the *z* score is -2.0, then the person's score is two standard deviations below the mean. If the *z* score is 1.5, then the person's score is one and a half standard deviations above the mean. If the *z* score is 0.0, then the person's score is zero standard deviations above the mean – it is right on the mean. So when we talk about the number of standard deviations a score is from the mean, we're also using *z* scores. Very convenient.

One important point about the linear *z* score transformation (and all other linear transformations) is that the shape of the distribution does not change. If the data were normally distributed before the transformation, it will be normally distributed after. It the data were skewed before, they will be skewed after. The linear *z* score transformation changes the mean and standard deviation of the data, not the shape of the distribution.

There's another benefit to standard scores. Standard scores allow for easy comparisons of scores. Comparing two or more scores from the same test is child's play if the measurement is done at the ordinal level or better – the highest score represents the highest standing on the construct. Highest number wins. But what if we want to compare scores from one test to scores from a different test? This won't be as easy. Now you might ask, why would anyone want to do this? The answer is that we have many similar tests that do the same thing. The ACT and the SAT offer but

one example. Let's say that you took the ACT and scored a 30. We already know that I took the SAT and scored a 400 on the verbal section. Who did better, me or you? A layperson might look at the scores and say that I did better because 400 is bigger than 30. But we know better. We know that each test has a different metric of measurement – they use different numbers with different standards for good, average, and poor performance. What we need is a way to put both scores on the same metric of measurement. All we have to do is translate both scores to standard scores.

Back to our ACT-SAT example. We know my 400 on the SAT-Verbal translates to a *z* score of -1.0. What about your 30 on the ACT? What's its *z* score? Using the *z* score equation (we'll say that the ACT has a mean of 20 and a standard deviation of 5), your ACT score transforms to a *z* score of +2.0. Again, *z* scores are the number of standard deviations above or below the mean. So your score is two standard deviations above the mean.

REVIEW 1 Computing *z* Scores

Question 1 of 2

If a set of data has a mean of 50 and a standard deviation of 20, what is the *z* score for a person with a raw score of 40?

- ✓ **A.** -0.5
- **B.** +0.5
- **C.** +2.0
- **D.** -2.0
- **E.** +0.75
- **F.** -0.75

◀ Check Answer ▶

Now that both of our scores are in *z* score units, we can directly compare the numbers. It is clear that your *z* score of 2.0 is bigger than my *z* score of -1.0. You win. You did better on your test than I did on mine. Try to stay humble. It won't be easy.

One last bit on this comparison business. Some comparisons are not meaningful. Suppose you take a test of depression and I take the SAT-Verbal again. Your score is a 4 and mine is a 410 (I studied a bit harder this time). Who did better? The answer is: Who cares? The tests are completely different, measuring different constructs, existing for different purposes. It's a meaningless comparison.

Correlation

If there's a hall of fame for statistics, the correlation coefficient is in it.

Overview

You may have noticed that everything we have discussed so far has been related to scores on a single variable. That is, we've talked about a set of ACT scores, but we've never looked at the relationship between two variables (ACT and college GPA, just to throw out a wild idea) for a group of people who each have scores *on both variables*. Are the scores related? Unrelated? In what way are they related? Is it a strong relationship or a weak one? As you can see, life gets much more interesting when we measure multiple variables for each person. And we haven't even talked about *why* these two variables are related. That's a topic for another day (and another book). For now, let's focus on understanding how we quantify associations between two variables.

Bivariate Associations

When describing the association between two variables there are two issues to consider: the strength of the relationship and the direction of the relationship. One way to assess the association between two variables is to simply examine the raw data. Below is another one of our absurdly small datasets, which we'll use as an example.

Person	X (ACT)	Y (GPA)
John	12	1.1
Sal	23	2.8
Tim	24	2.9
Amy	31	3.4
Linda	22	-

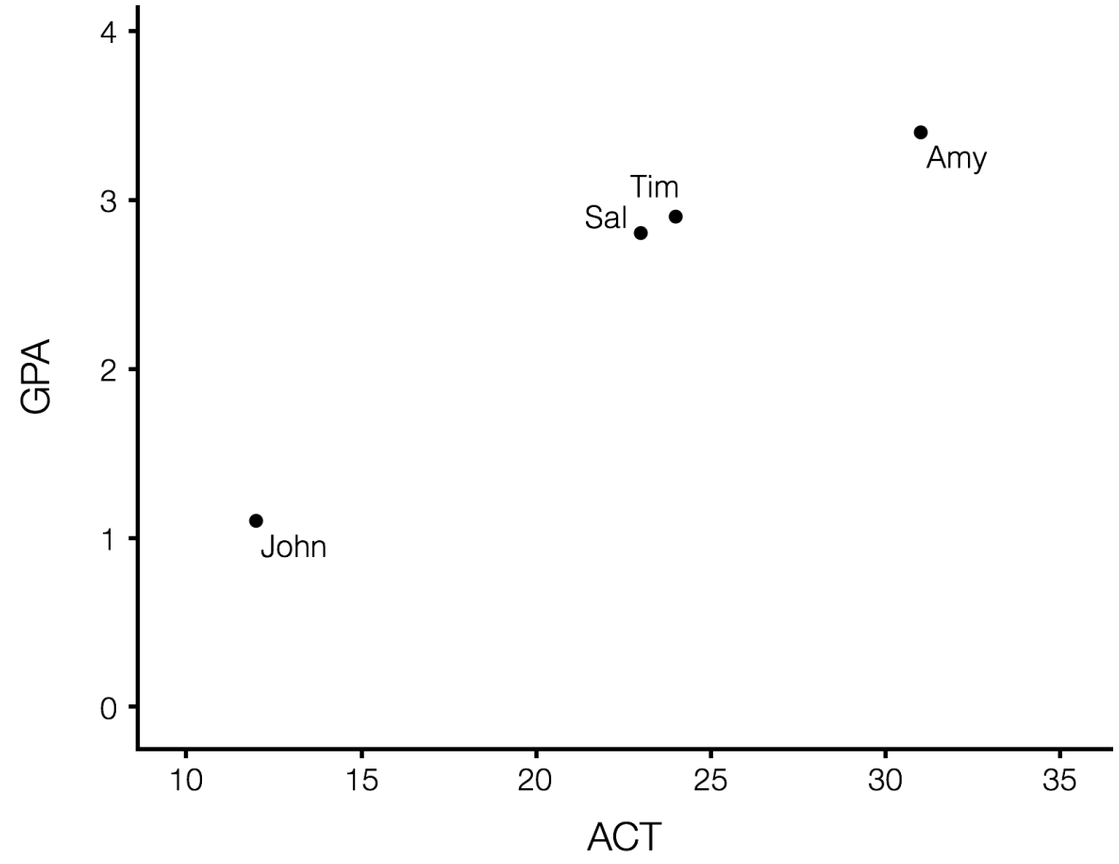
First off, we note that each person should have two scores: *X*, the ACT score, and *Y*, the GPA. If a person had only one score, we would be unable to include him or her in the analysis. Note that Linda

doesn't have a score on the *Y* variable, meaning that we are unable to include her in the analysis. A person must have scores on both variables to be included. We also note that I've sorted the remaining scores from lowest (John) to highest (Amy) on *X*. Now let's see if there's a trend in the data. And because I made up the data, there is. Lower scores on *X* are associated with lower scores on *Y*. Higher scores on *X* are associated with higher scores on *Y*. So it appears that there is a strong, positive relationship between *X* (ACT score) and *Y* (GPA). We say that it is a *strong* relationship because the rank order of scores is perfectly consistent. The person with the highest score on *X* (Amy) also has the highest score on *Y*. The person with the second highest score on *X* (Tim) also has the second highest score on *Y*. And so on. There are no exceptions to this perfect ordering of the scores. It is this consistency of rank order which is the primary determinant of the value of the correlation coefficient. Finally, we say the relationship is *positive* because higher scores on *X* are associated with higher scores on *Y*. If higher scores on *X* were associated with lower scores on *Y*, then the relationship would have been negative. Thus, we have addressed both aspects of bivariate associations: strength (the relationship between *X* and *Y* is strong) and direction (the relationship is positive). Now this is about all we can get from examining the raw data (don't try doing this with large datasets – it's borderline impossible); let's move on to a better way to examine the relationship, the scatterplot.

The Scatterplot

A scatterplot is a graph of the *X* and *Y* scores on two axes. It's the same old kind of *x*-*y* graph you've known since, oh, about third grade. The data from our example are graphed in Figure 1. On a scatterplot, each person receives a dot (or a square, or a plus sign, or a smiley face, or whatever you want). The dot indicates a person's score

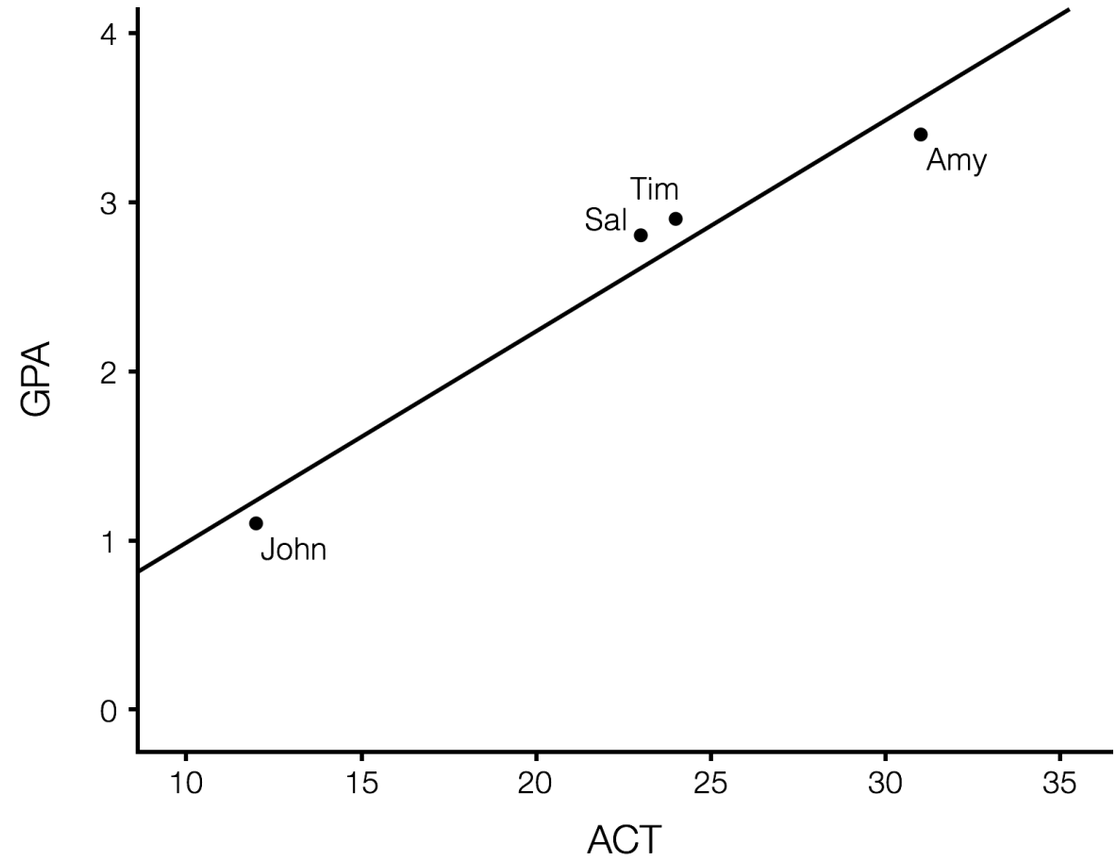
FIGURE 1 Scatterplot of Example Dataset



on *X* and *Y*. It should now be clear as to why we couldn't include Linda in the analysis. Where would we put her dot? It would be somewhere at 22 on the *x*-axis, but how high on the *y*-axis do we put the dot? We can't assume that she would have done poorly on *Y*. We're not in the business of assuming anything – we're in the business of using the available data to describe the relationship. Thus, she's gone. Looking at the scatterplot, we can see a trend, the same trend we saw when we looked at the raw data: higher scores on *X* are associated with higher scores on *Y*. And notice how the scores fall in the path of a straight line. The basic Pearson correlation tells us the strength of the *linear* relationship between two variables. What if the relationship is not linear? Another time, another chapter for that topic.

Getting back to how closely the scores match a straight line, let's draw the graph again, only this time with a straight line added as a reference (Figure 2).This line is called the *line of best fit*, or more commonly, the regression line. The regression line is the line that minimizes the vertical distance between the line and each point. You can imagine pulling out a ruler, measuring the vertical distance between each point and the line, averaging the distance, moving the line ever so slightly to try to improve things, and repeating until you

FIGURE 2 Scatterplot of Example Dataset with Regression Line

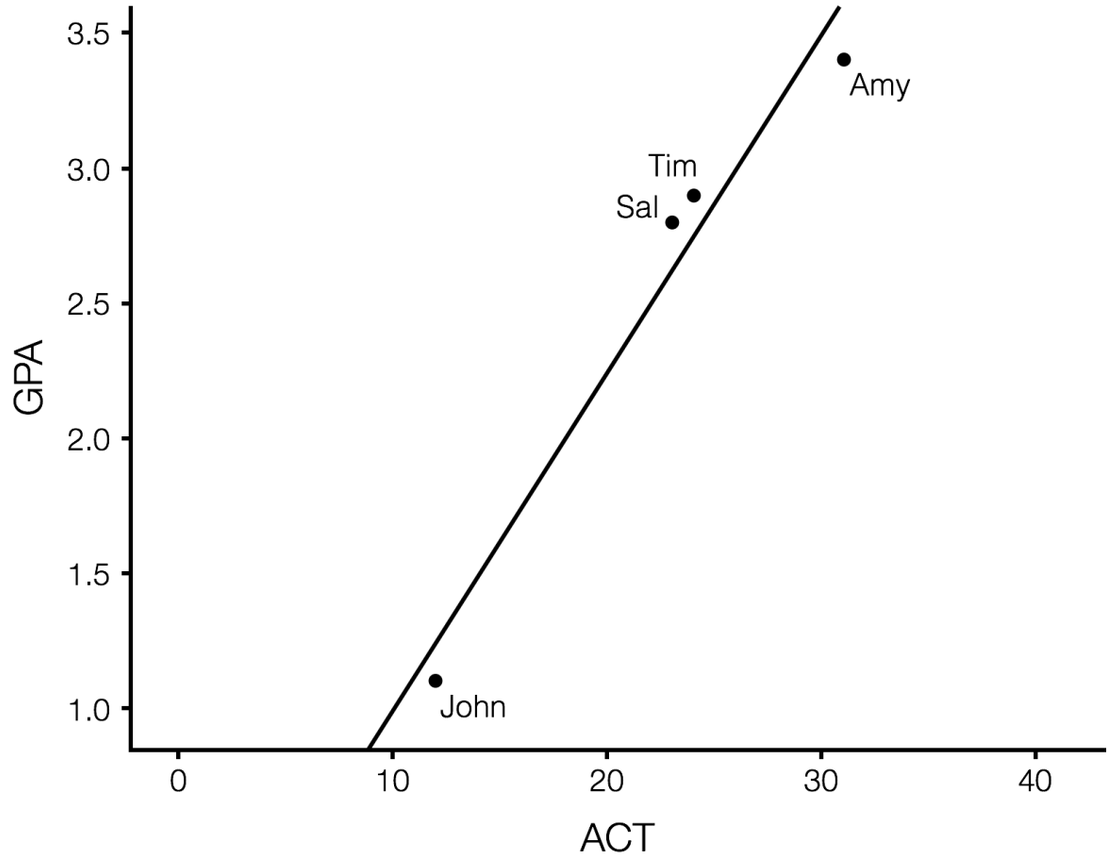


find the sweet spot. You can imagine doing this, but it sure doesn't sound like fun. Fortunately, we don't have to do this graphically (where we measure things with a ruler); we can do it mathematically with the raw data. Even more fortunately, we can let computers do all the work for us (more on this later). As mentioned, the strength of a relationship between two variables is indicated on the scatterplot by how close the points are to a straight line. As we will see, in weaker relationships, the points are far from the line. The direction of the line (pointing up or pointing down) tells you direction of the relationship (positive or negative). If the line is completely flat, there is no relationship. Very important point: The apparent slope of the regression line (aside from the case where it is completely flat) does NOT indicate the strength of the relationship. It seems like it should, but it doesn't (aside from one special exception with which we will not concern ourselves). As mentioned, the strength of a relationship between two variables is indicated on the scatterplot by the closeness of the points to a straight line, not the slope of this line. Why not the slope? The answer is that we can stretch or squash the *x*- and *y*-axes by a number of different methods to increase or decrease the slope of the line. The same

data are displayed again in Figure 3, this time with different ranges on both axes.

That new slope may appear impressive, but the strength of the association is unchanged. The correlation stays the same. So don't be fooled by the apparent slope of the regression line. Notice how I said that the *apparent slope* doesn't indicate the strength. If you were to compute slopes with the old slope = rise/run equation for the above graph and the previous one, you would find that the value is the same in both cases. By changing the range on the axes I've made the slope appear to be stronger. Always examine how close the points are to the line to assess the strength of the relationship. But this graphical stuff is just a visual representation of the data, something that we can eyeball to get a general idea of what is going on. To describe the strength of the association between two variables with any accuracy requires something more than a casual inspection of the raw scores or even a graph of these scores. We need a statistic to quantify the strength of the relationship. We have a few options. Before we discuss any of these statistics, let's discuss what properties a good measure of association statistic should have.

FIGURE 3 Scatterplot of Example Data with Adjusted Axes



Measures of Association

What properties should a measure of association have? First off, it must accurately convey the desired information: strength and direction of the relationship. If it does't do that, then there's really no need to proceed with it. Furthermore, it should be sensitive to small differences in strength. One of the problems with evaluating strength with an examination of a scatterplot or dataset is that we are able to determine only the biggest of big picture ideas about strength ("It's kinda strong. Maybe, medium strong."). There is simply no way to be precise that way. A measure of association must be precise, or why bother?

Building on this, a good measure of association should be easy to interpret. That is, upon computing the coefficient, we should be able to determine, without any other information, whether the relationship is strong or weak, positive or negative. We should be able to see the number and instantly know what it means.

Finally, a good measure of association should have a design that makes some sort of logical sense. It should be more than just a magic box in which the raw data is fed in the front, resulting in a coefficient falling out of the back end. I realize that this may not seem all that important, but it is. With these expectations set, let's examine our first measure of association, covariance.

Covariance

Covariance is not just our first measure of association, it's *the* first measure of association. Covariance is the start of it all. Aside from one incredibly annoying limitation, covariance is the simplest and most fundamental way to understand measures of association.

Covariance is like variance – for a pair of variables. To understand covariance we must take a step backwards and discuss variance again. Variance quantifies differences among scores for a single variable (remember that if everyone has the same score, variance is zero). If you recall, variance (in the population form) is defined as the mean of the squared mean-deviation scores:

$$\sigma_X^2 = \frac{\sum (X - \mu_X)^2}{N}$$

Where:
μ_X is the mean of X.

As long as we're looking at the variance equation, I'll do a little algebraic manipulation and expand the squared part.

$$\sigma_X^2 = \frac{\sum (X - \mu_X)(X - \mu_X)}{N}$$

There, same equation, just presented slightly differently. Just to refresh your memory a little more, a mean-deviation score is computed as the simple difference between a given score and the mean of that variable (i.e., $X - \mu_X$). A positive mean-deviation score indicates that the score is above the mean. A negative mean-deviation score indicates that the score below the mean. And a mean-deviation score of zero indicates that the score is, you guessed it, right at the mean.

To summarize how variance is computed, we transform every person's score into a mean-deviation score, square these mean-deviation scores, and compute the mean of these squared values. Covariance is computed like variance for a pair of scores for each person. That is, instead of multiplying a variable's mean-deviation score with itself (i.e., squaring), we multiply the first variable's mean-deviation score by the second variable's mean-deviation score. To make this happen, all we need to do is make a small modification to the variance equation listed above so that the mean-deviation scores of X are multiplied by the

mean-deviation scores of Y. Below is the equation for covariance.

$$\sigma_{XY} = \frac{\sum (X - \mu_X)(Y - \mu_Y)}{N}$$

Where:
σ_{XY} is the population covariance of X and Y
μ_Y is the mean of Y.

Below is a dataset demonstrating the calculations for covariance.

Person	X	Y	(X-Mean$_X$)	(Y-Mean$_Y$)	(X-Mean$_X$)* (Y-Mean$_Y$)
John	12	1.1	-10.5	-1.45	15.225
Sal	23	2.8	0.5	0.25	0.125
Tim	24	2.9	1.5	0.35	0.525
Amy	31	3.4	8.5	0.85	7.225

The mean of the last column (which works out to 5.775) is the covariance.

Covariance Logic

How does the covariance equation work as a measure of association? Consider what we learned about bivariate associations: A strong positive association is obtained when people with high scores on one variable (e.g., X) have high scores on another variable (e.g., Y) and when people with low scores on X also have low scores on Y. When we say *high scores* and *low scores*, doesn't that sound like mean-deviation scores? (High or low compared to what? The other scores – the mean being a great representation of the other scores.)

Now, all we need is a way to quantify the degree of consistency of these mean-deviation scores. Multiplying the two mean-deviation scores for each person results in a product which is maximized only when both scores are large numbers (either positive or negative). Take the mean of those products and you have a pretty sweet measure of association.

Let's talk a little more on how the mean of this product works. Consider that half of the the mean-deviation scores will be positive and half will be negative. If the people with high scores on *X* have high scores on *Y*, then you'll have two positive mean-deviation scores. Compute the product, and you have a nice, big, positive number (see Amy in previous dataset). Continuing with this example, if the people with the low scores on *X* have low scores on *Y*, then you'll have two negative mean-deviation scores. Take the product, and due to the old *negative times a negative equals a positive* property of numbers, and you'll have another nice, big, positive number (see John). The mean of all of the big, positive numbers is a big, positive number, indicating a strong, positive association. There's our index of strength and direction.

To continue to understand the logic of covariance, let's flip the previous scenario around. Now, the people with the high scores on *X* have the low scores on *Y* (and the converse). Thinking in terms of mean-deviation scores, that's a big, positive number multiplied by a big, negative number. Which results in a big, negative number. The mean of these is a big, negative number, indicating a strong, negative association.

And finally, what if there is no pattern? How does the covariance equation handle that? Some of the people with high scores on *X* have high scores on *Y*. Others have low scores on *Y*. Still thinking in terms of mean-deviation scores, that's some big, positive numbers multiplied by a big, positive numbers, resulting in big, positive products, *and* some other big, positive numbers multiplied by some big, negative numbers, resulting in big, negative products. Take the mean of these products, and you get a zero, indicating no association.

That's the logic of the covariance equation. That's how it quantifies the direction and strength of association. This statistic allows us to see how

these variables vary together, or co-vary (hence, the name covariance).

Totally irrelevant thought: What if the Y variable is just a copy of the X variable? Same scores and all. Doesn't that turn this part of the covariance equation: $(X - \mu_X)(Y - \mu_Y)$ into this: $(X - \mu_X)(X - \mu_X)$? A change which takes us back to the variance equation. As mentioned, covariance is just like variance, but it's for a pair of variables.

Since there was a population and a sample form of the variance equation, you just know that there had to be a population and a sample form of the covariance equation. So here it is, the sample form of the covariance equation.

$$c_{XY} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Where:
c_{XY} is the covariance of X and Y in a sample of data.

There's really nothing else we need to say about this one. Other than the $N - 1$ thing and the use of the sample mean instead of the population mean, everything else is the same.

Now that we've described the mathematical basis for covariance, let's talk about what it does. As mentioned a number of times, covariance indicates the strength and direction of the relationship between two variables. A covariance of zero indicates no relationship between the variables. A positive covariance indicates a positive relationship between the variables, and a negative covariance indicates a negative relationship between the variables. The only problem with covariance as a measure of association is that it is difficult to understand just how strong or weak these relationships are. For one set of data a covariance of 41.4 might be weak, but for a different set of data a covariance of .83 might be very strong. It's all very annoying. This lack of a consistent standard for strong and weak relations is the major limitation

of covariance, and it is the principal reason why covariance is seldom used as an index of the association between two variables. (It does have other value in terms of summarizing data, but don't worry about that.) Good news though, there is a statistic that does indicate the strength and direction of the relationship between two variables in a standardized, easy to interpret fashion. And that statistic is the correlation coefficient.

Correlation

Correlation is, like covariance, a measure of association between two variables. Unlike covariance, correlation describes the association in a way that allows us to easily interpret the strength of the association. Correlation is, in essence, standardized covariance. Correlation is defined as the covariance divided by the standard deviations of each variable.

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Where:
ρ_{XY} is the correlation of X and Y in a population.

We'll make this our last foray into the population-versus-sample version of something and just stick to sample versions of statistics. And here is that sample version of the correlation equation.

$$r_{XY} = \frac{c_{XY}}{S_X \cdot S_Y}$$

Where:
r_{XY} is the correlation of X and Y in a sample of data. To reiterate, the symbol for sample correlation is r, variable name subscripts optional.

For both equations, the principle is the same: Correlation is covariance divided by the standard deviations of each variable. What purpose does that serve? Dividing by the standard deviations res-

cales the statistic so that the maximum and minimum values are always 1.0 and -1.0, respectively (covariance maximums and minimums were a function of the product of the standard deviations – different standard deviations mean different maximums and minimums). Thus, a correlation of .6 always means the same thing in terms of strength, regardless of the standard deviations of the variables. That's the big advantage correlations have over covariance.

There are many types of correlation equations, but we'll focus on the most popular one, the Pearson Product Moment Correlation. If someone says that they correlated two variables, with no other information specified about the type of correlation, they are talking about the Pearson one. If you use one of the weird ones (e.g., phi, tetrachoric, Spearman), you mention them by name.

The Pearson correlation summarizes the strength and direction of the association between two variables in a single number. The correlation coefficient ranges from -1 to +1. A positive coefficient means that the relationship is, you guessed it, positive. A negative coefficient indicates a negative relationship. A -1.0 correlation indicates a perfect negative relationship and a +1.0 correlation indicates a perfect positive relationship. A 0.0 correlation indicates no relationship between the two variables. Thus, the strength of the relationship is indicated by how close the number is to +1 OR -1. A correlation of -.8 is just as strong as a +.8. I hope that it is clear that the sign of the correlation is irrelevant to the strength of association. The direction of the relationship is useful information worth knowing; it is just different information than the strength of the relationship.

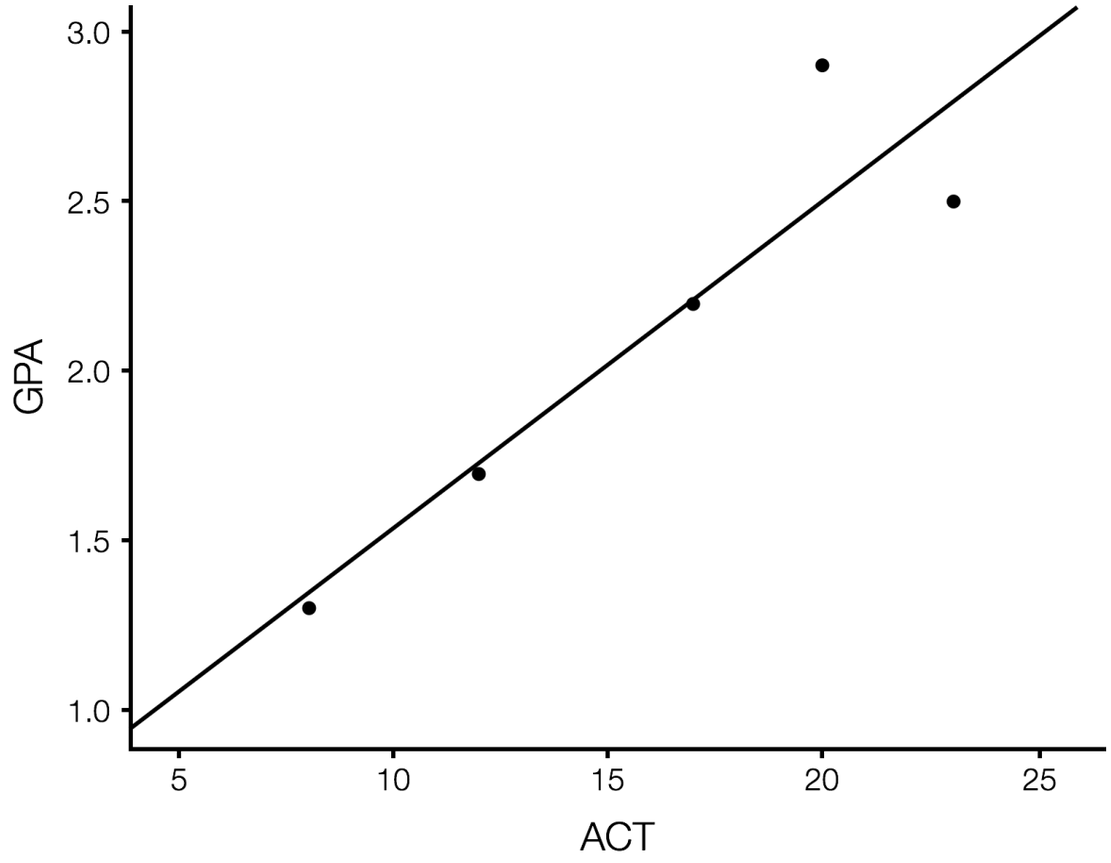
Consider the following dataset.

Person	X (ACT)	Y (GPA)
Frank	8	1.3
Kevin	12	1.7
Gianni	17	2.2
Warren	20	2.9
Judy	23	2.5

As you can see in Figure 4, the rank order, although good, is not perfectly consistent. The person with the highest score on *X*, Judy, has the second highest score on *Y*. The person with the second highest score on *X*, Warren, has the highest score on *Y*. They are out of order. Everyone else falls in line (third on *X* is third on *Y*, fourth is fourth, etc.).

Clearly the trend is positive, but compare this graph to any of the scatterplots of the previous dataset (Figure 2). Notice how the points in our new

FIGURE 4 Strong (But not Perfect) Association



scatterplot are not as close to a straight line. Weaker association. Computing the correlation confirms what we already know, $r = .92$. Still very strong, but weaker than the previous dataset.

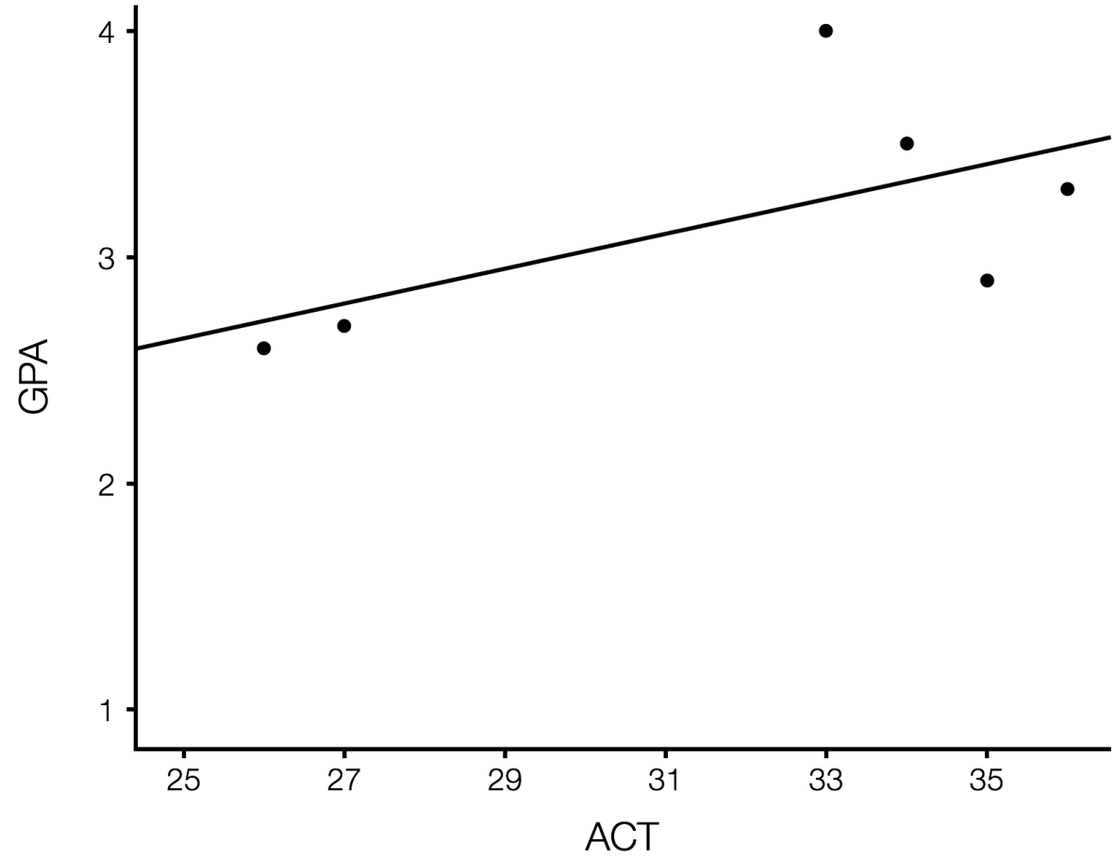
Time for a new example.

Person	X (ACT)	Y (GPA)
Rusty	26	2.6
Buck	27	2.7
Jeff	33	4.0
Dale	34	3.5
John	35	2.9
Margaret	36	3.3

Now we see even more exceptions to perfect rank ordering (Figure 5). The person with the highest score on *X*, Margaret, has the third highest score on *Y*. The person with the second highest score on *X*, John, has the fourth highest score on *Y*. More exceptions abound, but in spite of them, we can still see a general trend: Higher scores on *X* are associated with higher scores on *Y*.

The line points up, but the points are even further from the line than we have seen before. Thus,

FIGURE 5 Very Good (But Weaker) Association



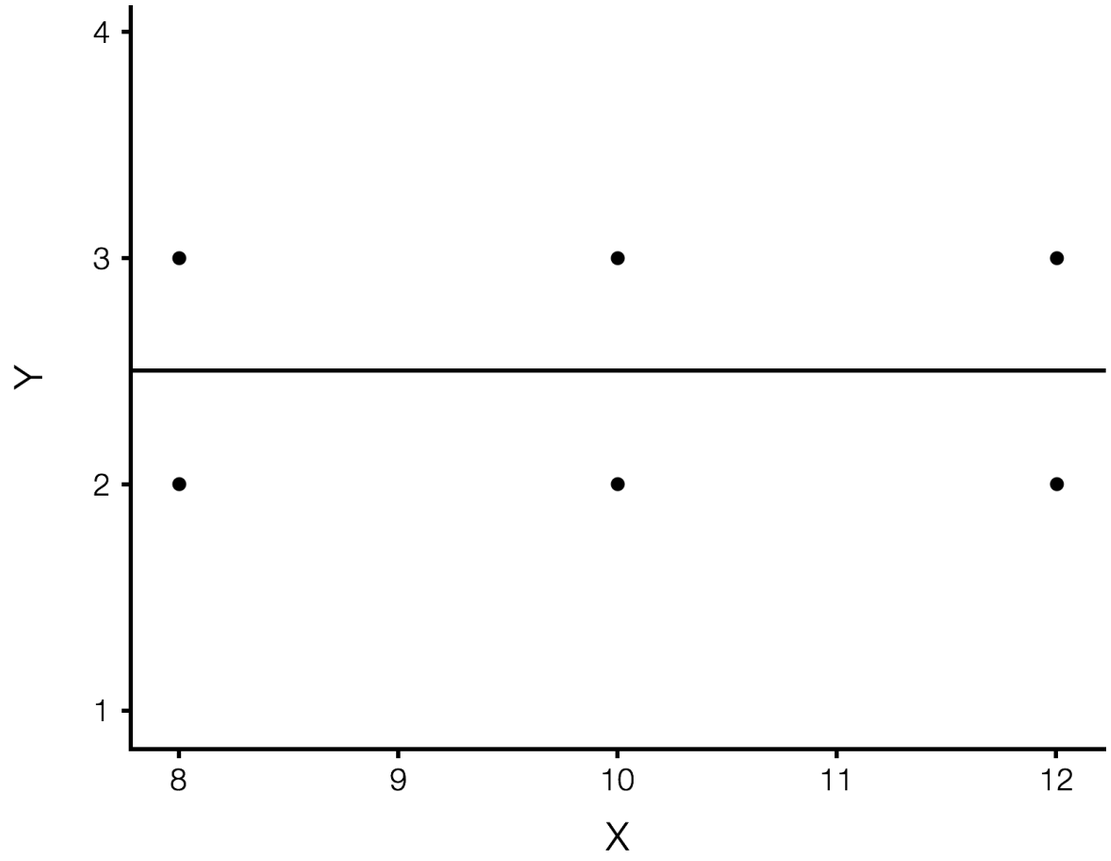
we have a positive association that's not perfect. How strong is it? $r = .61$. So it's positive and strong, but weaker still than the previous datasets.

You might be wondering what a zero correlation looks like. Well, wonder no more.

Person	X (ACT)	Y (SALES)
Hunter	8	3
Lonny	8	2
Charles	10	3
Craig	10	2
Danny	12	3
Kendall	12	2

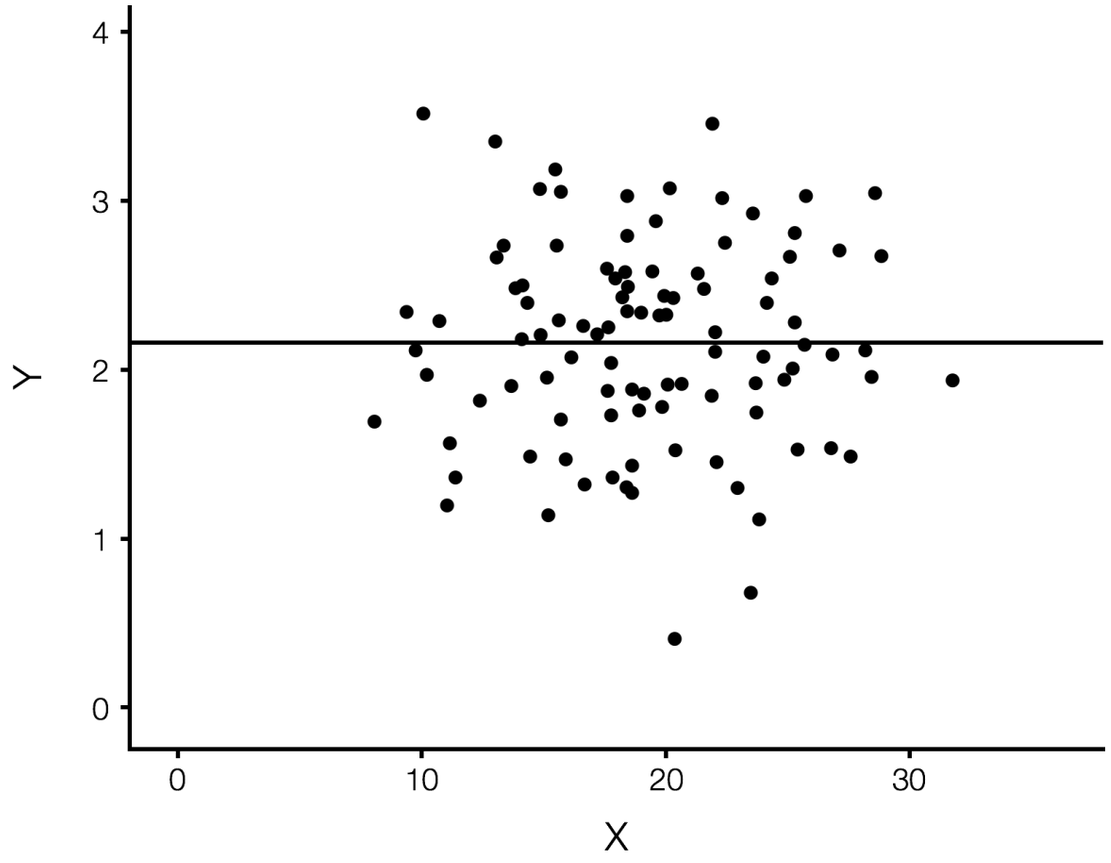
What do we see? No clear trend. High scores on *X* are associated with both high and low scores on *Y*. Low scores on *X* are associated with both high and low scores on *Y*. The scatterplot is shown in Figure 6 and looks like a rectangle of scores. I'll bet you didn't know what a rectangle of scores looked like before now. I'll also bet that you didn't care to know. And you still don't.

FIGURE 6 Zero Correlation Scatterplot



Of course, a six person dataset makes for a fairly uninteresting scatterplot when the correlation is zero, but, the point is made – there is no trend in the data.

FIGURE 7 Better Zero Correlation Scatterplot



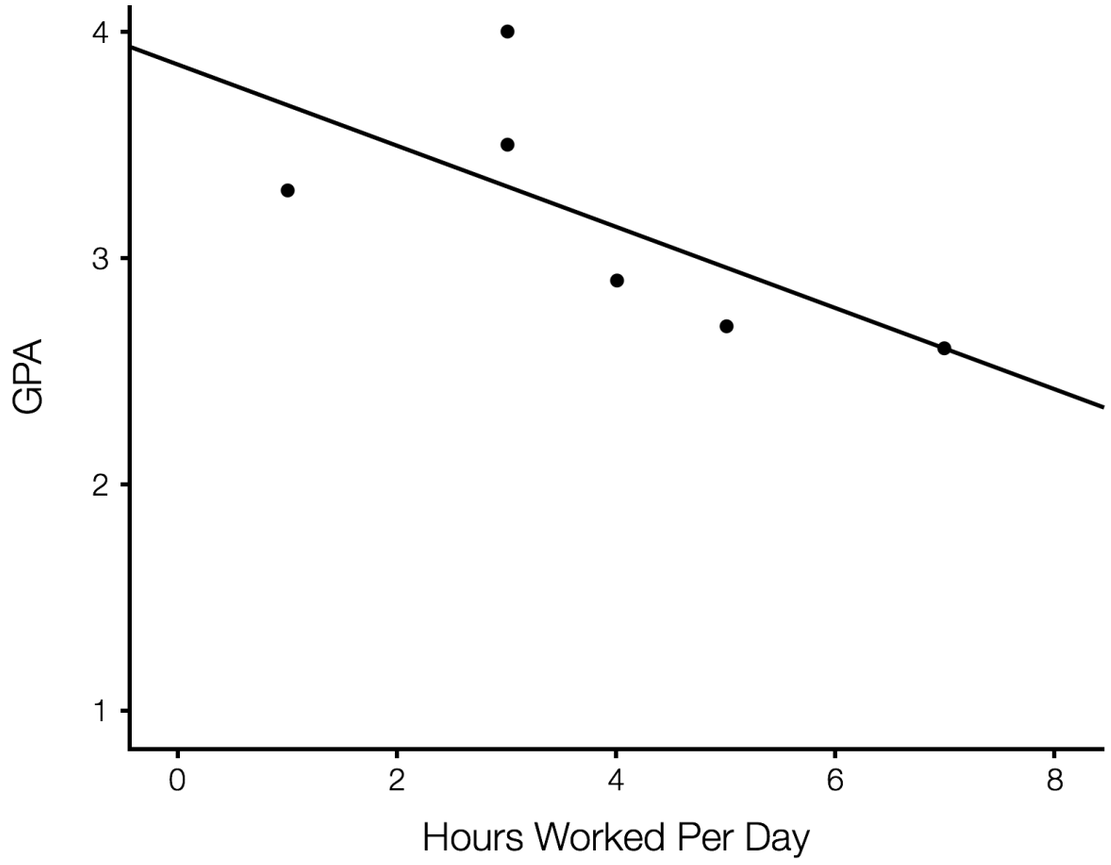
When you have a larger dataset, a zero correlation scatterplot resembles a circle (Figure 7). Notice how the regression line is perfectly flat. No slope at all. This is the land of $r = 0.0$. A bleak and desolate land. Unfit for both man and animal. No association of *X* and *Y* of any kind.

Finally, how about a negative correlation.

Person	X (Hours Worked)	Y (GPA)
Woodrow	7	2.6
Al	5	2.7
Evan	3	4.0
Eddie	3	3.5
Ernie	4	2.9
Kelly	1	3.3

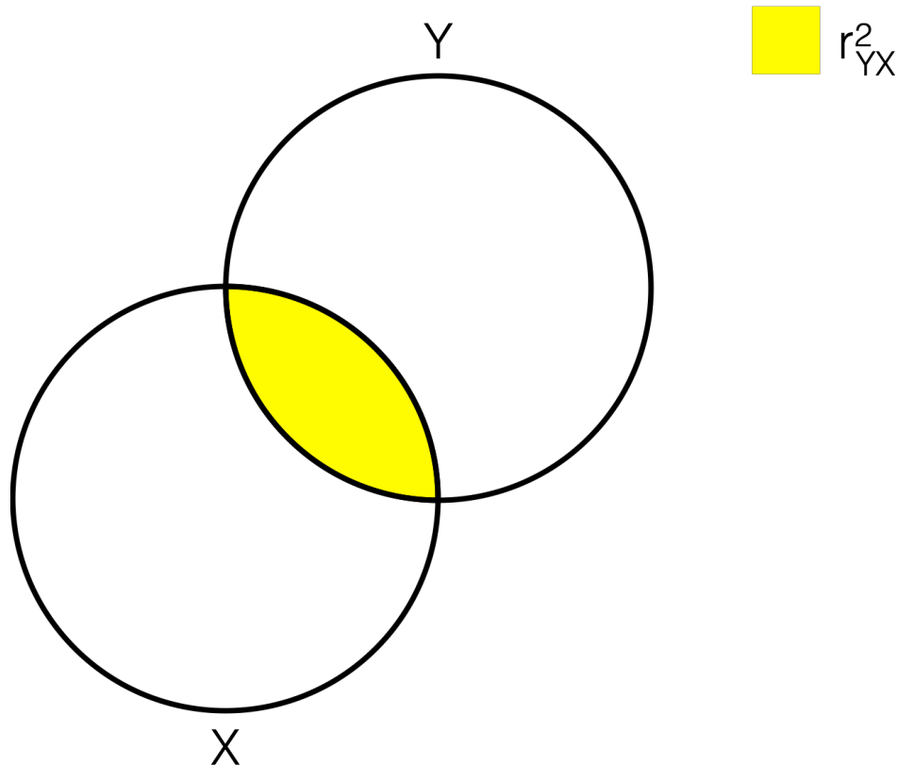
We can see a clear (although not perfect) trend: Higher scores on *X* are associated with lower scores on *Y*. The scatterplot is shown in Figure 8 and is different from our previous examples.

FIGURE 8 Negative Correlation Scatterplot



The line of best fit points down, indicating a negative association; the points are fairly close to the regression line, indicating a strong association. So we see a strong, but not perfect, negative association. The actual correlation is r = -.68.

FIGURE 9 Venn Diagram Illustrating Relationship between X and Y



At this point, it's time to introduce another way to illustrate the association between variables. If you like Venn diagrams, and who doesn't, one is shown in Figure 9. Venn diagrams illustrate the relationship between various concepts. When applied to correlations, the circles represent the variance of each variable. The overlap of the circles indicates the degree of association. Greater over-

lap indicates greater associations. To get technical, the percent of the area of Y overlapped by X represents the squared correlation between the two variables (i.e., r^2_{XY}). We'll talk more on squared correlations in the next chapter, but it's not like there's a lot of mystery there. You have a correlation. You square it. You get r^2.

Computing Correlation Coefficients

As mentioned, there are a variety of types of correlations, but we'll stick with the ever popular Pearson correlation. Since we live in a computer age, there is little to be gained by focusing on equations. Little, but not nothing. There are a few different, but equivalent, versions of the equation for the Pearson correlation. We've already seen one that starts with covariance. Let's examine the most intuitive form of the Pearson correlation equation, the average product of z scores. Listed below is the population version of it.

$$\rho_{XY} = \frac{\sum (z_X \cdot z_Y)}{N}$$

It's fairly simple; just compute the product of the z scores for each person, compute the mean of those products, and you're done.

To refresh our memory on z scores, the population form of the z score equation is listed below.

$$z_X = \frac{(X - \mu_X)}{\sigma_X}$$

And, of course, if we want to compute z scores for Y, the equation is be the similar, only with Y substituted for X at every opportunity.

Just for fun, let's take these z score equations for X and Y and substitute them into the correlation equation from above. Here's what we obtain with those substitutions:

$$\rho_{XY} = \frac{\sum (\frac{(X - \mu_X)}{\sigma_X} \cdot \frac{(Y - \mu_Y)}{\sigma_Y})}{N}$$

It's still the correlation equation, but it looks familiar. Where have we seen something similar? That's right, take away the standard deviations (σ_X,σ_Y), and it's the equation for population covariance.

$$\sigma_{XY} = \frac{\sum (X - \mu_X)(Y - \mu_Y)}{N}$$

Do you see it in the above two equations? The only differences between the correlation (ρ_{XY}) and covariance (σ_{XY}) equations are the standard deviations in the former.

Here's a thought exercise: What if X and Y both had standard deviations of 1.0 (which, of course, is the case with z scores)? Since anything divided by 1 equals itself, the standard deviation parts of the correlation equation disappear, leaving us with what we see in the covariance equation. (One quick lesson from this is that for standardized data, covariance equals correlation.)

As mentioned earlier, covariance is computed as the average of the products of mean-deviation (i.e., $X - \mu_X$) scores for each person. Correlation is computed as the average of the products of the z scores for each person. And what's the difference between a z score and a mean-deviation score? A division by a standard deviation. I hope that it's clear to you that correlation and covariance are very similar statistics. But correlation is better. There, I said it.

Those were population versions of the equations. For reasons that should be obvious by now, it will be much more useful if we discuss the correlation equation designed for samples. And here it is.

$$r_{XY} = \frac{\sum (z_X \cdot z_Y)}{N - 1}$$

What's the difference? Well, there's the symbol for correlation. It's now r instead of ρ. So, there's that. The only other difference is the denominator.

It's $N - 1$ instead of just N. Does this look familiar? It should. This is the variance story all over again. When computing the variance of a population of data, the denominator is N. When computing the variance of a sample of data, the denominator is $N - 1$. It's the same pattern with the Pearson correlation equation: N denominators for populations, $N - 1$ denominators for samples. So when computing z scores for the sample correlation equation, be sure to use the appropriate $N - 1$ variance equation. With samples, it's $N - 1$ denominators all the way down. Of course, we let computers do the dirty work for us, and they use sample versions of equations for everything. But just in case you have to do your computations by hand, you have enough information to do it right.

How Correlations Work

Let's put this all together so that we can really understand what makes correlations tick. Correlations are measures of association between two variables. When people who have high scores on one variable also have high scores on the other variable (and vice-versa), you get a strong, positive correlation. As discussed, the Pearson correlation equation quantifies the relationship by computing the mean cross-product of z scores. Interactive 1 demonstrates this process.

INTERACTIVE 1 How Correlations Work

Correlation and Causation

Correlation does not equal causation. It's important to remember that a correlation coefficient is just a statistic that describes the association between two variables. *Why* these variables are associated is another matter. *Why* is an issue of causality. In general, our statistics can't address causality. It is our research design that allows us to address causal issues. As but one example, consider the ACT/College GPA correlation. There is a positive correlation of about .5 between these two variables. Does that mean that your performance on the ACT causes your college performance (*X* causes *Y*)? Probably not. Does that mean that your college performance causes your ACT performance (*Y* causes *X*)? We can safely rule this out based on logic: ACT performance is measured months before college performance even begins. Statistically, the *Y causes X* inference is as valid as the *X causes Y* inference. It is our research design that allows us to rule out *Y* causing *X* in this case. So we've covered the causality issue in both directions. There is, however, a third possibility. It is possible that a third variable, which we'll call *Z* , is causing performance on both *X* and *Y* – making *Z* responsible for the correlation between *X* and *Y*. What is this third variable in our ACT-GPA example? Let's pick one: study habits. People with good study habits do well on the ACT and do well in college. People with poor study habits generally do poorly on both. So, it appears that *Z* is responsible for the correlation. No guarantees, but if I was betting person, which I am not, I'd bet on *Z*. (Just to be complete, there is also a fourth option in which *X* causes *Z* which causes *Y*, making *X* an indirect cause of *Y*. Don't worry about it, though. The previous explanation is far more relevant.)

New example: Ice cream sales (*X*) are correlated .7 with shark attacks (*Y*) at a certain seaside resort. Which seaside resort? That information is classified. Is *X* causing *Y*? Maybe, if people are eating a bucket of ice cream and then going swim-

ming right away. Is it possible that the sharks are attracted to the smell of ice cream? Can they even smell it? Does the flavor matter? All good questions, but let's switch gears. Is *Y* causing *X*? That is, are the shark attacks causing people to buy ice cream? Maybe the survivors of the shark attacks like to celebrate cheating death with some mint chocolate chip. Statistically, both are equally valid explanations – do you see why research design is so important? Also, do you see the dangers of a blind application of statistics (i.e., devoid of logic)? Now is it possible there is some third variable at work here? Yeah, probably.

As you know, a correlation of zero indicates that there is no relationship between *X* and *Y*. And you know that +1 and -1 indicate perfect relationships. But what are industry standards for strong, medium, and weak correlations? The classic resource on this issue is Cohen (1992). Cohen's standards for correlational strength are as follows: small is .10, medium is .30, and large is .50. Naturally, the same rules apply to negative correlations. As Cohen stated, .10 is small; it's far too weak to be useful under most circumstances. So consider .30 to be the minimum decent value for a correlation.

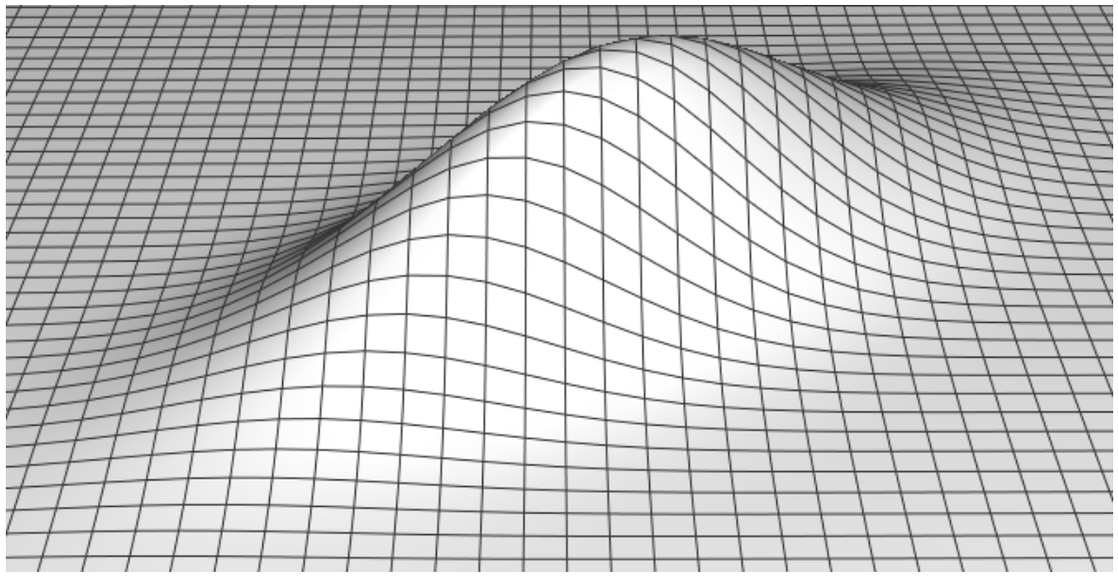
Correlation Assumptions

Correlation analysis has many assumptions. We'll merely list them here and save the explanations (with one exception) for the next chapter. The assumptions are as follows: *Y* (and optionally *X*) is a random variable (quick note: in this context the term *random variable* does not mean *a variable that is composed of random data*), the relationship between *X* and *Y* is linear, the variance of the residuals is constant across all levels of *X* (called homoscedasticity), and the joint distribution of *X* and *Y* is bivariate normal (i.e., bivariate normality).

The bivariate normality assumption is one that we need to discuss now because classically it has been applied to correlation in a slightly different manner than for regression. Now you may be thinking that bivariate normality means that *X* and *Y* are both normally distributed. Well, it's that and more. Bivariate normality concerns the joint distribution of *X* and *Y* with the assumption that this joint distribution is normally distributed. An examination of the distribution of a single variable is done with a simple *x*-*y* graph. A visual representation of a joint distribution requires a three-dimensional graph. Figure 10 displays an example of a bivariate normal distribution.

To understand bivariate normality imagine that what we see in Figure 10 is a cake. You heard me. Now imagine that we took a thin, vertical slice (running top to bottom) from the middle of this cake. If we viewed that slice from the side (you know, with the cakey part of the cake facing us), it would look like a regular, two-dimensional normal distribution. And we would see that no matter where we cut our slice from. That's bivariate normality.

FIGURE 10 Bivariate Normal Distribution



Significance Testing Overview

Unfortunately, we have to discuss significance testing. It's a nuisance, but there's no getting around it. Significance testing plays an important role in psychology. Before we address how to conduct a significance test for a correlation coefficient, we need to discuss the general concept of

significance testing. First, a review of three terms from Chapter 2: sample, population, and sampling error. I'm sure you recall that we measure samples because it's inconvenient or impossible to measure an entire population. We analyze data in our sample and make inferences from the sample to the population (e.g., 55% of the people in our sample watch football on TV; thus, we estimate that 55% of the people in the population watch football on TV). Unfortunately, measuring a sample instead of the entire population leads to problems. The statistics we compute in our samples will not be a perfect match to the statistics in the population. The difference between the two is called sampling error, and it is the price we pay for laziness.

Given our knowledge of sampling error, it should be clear that we should not infer too much from our samples. Quick example: let's say we collect a sample ($N = 127$) of ACT scores from high school athletes. We analyze the data and find that soccer players have a mean score of 20.5, and tennis players have a mean score of 20.3 (yes, it's a pointless study). It is obvious that soccer players outperformed tennis players in our sample of 127 students. But should we make an inference to the population and say that soccer players score higher than tennis players on the ACT? Probably not, you say, since the difference between the mean scores is so small. Good call. It is a mistake to think that a small difference in our sample statistics indicates that there is any sort of difference between the groups in the population. The small difference in our sample could be due to sampling error. Now what if there was a big difference in the sample means (let's say that the soccer players outscored the tennis players by 11.5 points)? Does this large difference in sample scores allow us to conclude that soccer players outscore the tennis players in the population? Yes. It is likely that they do. See how this works?

Where do significance tests fit in? Significance tests (also called *inferential* statistics) are used to

analyze the sample characteristics and indicate when it is wise (or unwise) to make inferences about the population based on the sample. They help us determine if we can conclude that a certain characteristic (e.g., a difference between test scores of girls and boys) actually exists in the population. Significance tests are probability analyses, and give an answer like, "There is only a three percent chance that a result like the one we found in our sample could have been found if there truly was no difference in the population. Therefore, we conclude that there is a difference in the population." (It helps if you read that sentence with a deep, authoritative voice.)

So what about significance tests for correlations? It's the same story except now that we examine the correlation in our sample (r_{XY}) and use it make inferences about the relevant population correlation (ρ_{XY}). Example: let's say we collect a sample of college students ($N = 93$) and find that time spent playing video games is positively correlated with GPA, $r = .07$. Sure, it's a weak correlation, but it's a positive correlation. It is indisputable that in our sample people who spent more time playing video games had a higher GPA. Go ahead, try and dispute it. Can we then conclude that more time spent playing video games is associated with higher GPAs in the entire population? I hope you're shouting, "No, our sample correlation is likely influenced by sampling error! The population correlation could be zero for all we know! The sample correlation is only slightly greater than zero!" That's enough shouting for now. Your instincts are correct. We need to conduct a significance test to determine whether our sample correlation is large enough to allow us to conclude that the population correlation is not zero.

Significance Test I: Inferences About a Population Correlation of Zero

There are a few ways to conduct correlation significance tests. The differences between these tests relate to the hypothesis you want to test. So figure out your hypothesis first, then choose the correct significance test. Good news, most of the time (and I mean almost all of the time), you'll be stating a hypothesis which reads as some version of this: The correlation between *X* and *Y* is greater than zero/is less than zero/is not zero. The first two options are directional hypotheses. You are picking a direction (let's say positive) and checking to see if your sample correlation supports that. For example, we might say: The relationship between study time and GPA is positive. (Notice that we didn't use the words *correlation* or *population*. It's implied that we are discussing the relationship, as indexed by a correlation, in the population. We don't make hypotheses about samples. We use sample data to test hypotheses about populations.) Or we could say that the relationship will be negative. Either way, we've picked a single direction in our hypothesis. Both of these tests will be one-tailed tests, a term you may remember from your statistics class. If you can't decide on a direction in your hypothesis (you expect a relationship between the two variables, but you don't know what direction the relationship will be), then you need to run a two-tailed test. The only problem with a two-tailed test is that it has reduced power as compared to the one-tailed test. Moral of the story: If at all possible, specify a direction of the relationship in the hypothesis. You have to do this in the hypothesis generation stage, well before you examine the data. No fair peeking at the data and then saying, "I predict a positive relationship!" Why isn't that allowed? Well, it's not much of a prediction when you already know the outcome. How could you ever be wrong?

Now that we have some general hypothesis testing issues out of the way, how does the signifi-

cance test work? It's a *t* test, and all you need to know to perform the test are the sample correlation and sample size.

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{N - 2}}}$$

Pretty simple. (By the way, *t* tests take the form of *sample statistic* divided by *standard error of sample statistic*. In other words, the denominator of this *t* test is the standard error of a correlation. I'm only mentioning this just in case you need it someday.) Once you have obtained your *t* value from the sample statistics (we often call this the *obtained t*), use a *t* table to find the critical *t* ($N - 2$ degrees of freedom) and compare the two. If the obtained *t* is greater than the critical *t*, then we say the sample correlation is significant (note: the previous term is statistician slang) and conclude that the hypothesis is supported. If our hypothesis specified a positive direction, then we would conclude that the population correlation is greater than zero.

One word of caution: Be sure that you use the correct alpha value and tailedness (more slang) when you look up the critical *t*.

Significance Test II: Inferences About a Non-Zero Population Correlation

As mentioned above, the previous significance test is used almost all of the time. But if you think about it, it's not exactly the most demanding standard. Saying that we have enough evidence to conclude that the population correlation is greater than zero (or less than zero or not zero, depending on the hypothesis) is not exactly a bold statement. For all we know, the population correlation could be +.02. Weak sauce, people. So why not attempt a slightly more demanding significance test, something that tells us a bit more about the population correlation? Our second correlation significance test concerns inferences about population correlations that are some value other than zero. For the sake of simplicity, let's consider .3. We

could conduct a test to see if our sample correlation is strong enough to allow us to conclude that the population correlation is greater than .3. (If you want to do the negative correlation thing, we would be testing whether the population correlation is less than -.3.) We could choose any value we want to test, but .3 is nice because it synchronizes nicely with Cohen's (1992) standards for a correlation of moderate strength.

There's one bit of bad news with this new test. It's more work. We have to do this irritating first step before we can get to the test proper. What is this first step? Something called the *r* to Fisher's *z* transformation. No, this is not a regular *z* score; it's a Fisher's *z* (I'll use the symbol *Fz* for it to help distinguish between the two). Two totally different things. (Although, it is scaled like a *z* score... Wait, forget that last sentence. You didn't see that.) Back to Fisher's *z*. You see, in order to do the sorts of operations we want to do to a correlation, the Pearson correlation that we all know and love must be transformed to a Fisher's *z*. Why? You don't want to know. Just accept that it has to be done and move on. (By the way, a Fisher's *z* transformation must be done even for the simple act of taking the mean of a set of correlations.) The equation to transform a Pearson correlation to a Fisher's *z* is given below.

$$Fz_r = \frac{\ln\left(\frac{1+r}{1-r}\right)}{2}$$

Assuming that you know that *ln* means natural log and that you know how to compute natural logs on your calculator, this transformation isn't too bad.

Now that we have the *r* to Fisher's *z* transformation down, let's get to the significance test. The equation is as follows.

$$z = \frac{Fz_r - Fz_\rho}{\sqrt{1/(N-3)}}$$

Where:
z is the obtained *z* score (which is compared to the critical *z* to determine significance).
Fz_r is the Fisher's *z* for the sample correlation.
Fz_ρ is the Fisher's *z* for the population correlation.

Not too bad, right? Guess where the Fisher's *z* transformation fits in? You must perform the *r* to Fisher's *z* transformation on both your sample correlation and the population correlation. Below is the *r* to Fisher's *z* transformation equation for the population correlation. It's the same equation as before, just set up for population values.

$$Fz_\rho = \frac{\ln\left(\frac{1+\rho}{1-\rho}\right)}{2}$$

Once you've done both *r* to *F*z transformations, the rest of the equation is simple. This is a *z* test, like the *z* tests for means that you had in stats class. If your alpha is the traditional .05, then the critical *z* values are 1.96 for a two-tailed test and 1.65 for a one-tailed test. And given that this test only makes sense with a directional hypothesis, the only critical *z* you need to use is 1.65. No need to consult a *z* table. Ever. So that's nice.

Back to when we discussed testing to see if the population correlation was greater than .3, if the test is significant, then the answer is yes. The population correlation may be .4 or it may be .5, but it's likely (only a five percent chance that we're wrong, assuming alpha is set to .05 and that the sample was collected via a probability sampling technique) that it is greater than .3. It could be as low as .31. That's a lot cooler than our first test where we could obtain significant result even when the population correlation is a puny .03.

Confidence Intervals for Correlations

Taking our significance testing concept to the next level, we can compute a confidence interval for our sample correlation. Confidence intervals indicate the likely (again, five percent chance we're wrong) value of the population statistic. Confidence intervals offer a lot of the same information that significance tests offer. If you set it up with the same tailedness, you can make confidence intervals perform the same function as the previous significance tests, only with more information.

How does this work? Here goes. Let's say that you compute a confidence interval (95% confidence) for a sample correlation of .34 ($N = 55$) and the confidence interval ranges from .08 to .55. Notice that this is a bidirectional confidence interval: The population correlation could be as low as .08 or as high as .55. This bidirectional confidence interval is capable of yielding information similar to the two-tailed version of our first significance test. How? Notice that the interval does not include zero. That means that the population correlation is greater than zero, a conclusion that is identical to what we would find if we performed the first significance test with an alpha of .05, two-tailed.

If the confidence interval can replicate the function of the first test, then why bother with it? (And a confidence interval is more work, as you will see.) The answer is that in addition to performing the job of the standard significance test, the confidence interval gives us more information than a mere significance test. After all, the confidence interval gives us the likely (only a 5% chance we're wrong) lowest possible value for the population correlation. Here's another way to think about the confidence interval. Consider the second significance test. It allows us to determine whether the population correlation is greater than some non-zero value (say .3). But it doesn't allow us to say exactly what the lower bound of the

population correlation could be. If all we had was the second test, and we wanted to know what the likely lowest possible value of the population correlation was, we would have to repeat the test with a variety of population values (Is it .35? Is it .40? Is it .50?) and stop when we get a non significant result. (Note: I do not recommend this procedure as it is annoying. And it has other problems as well.)

Let's illustrate that point with another example. We collect data from a sample of 103 people and find a correlation of .50. We run the second significance test (alpha = .05, one-tailed) against a population correlation of .30. Our obtained *z* is 2.40, which is greater than the critical *z* of 1.65 (you should do the calculations yourself to check these numbers). Thus, we conclude that the population correlation is greater than .30. But we don't know how much greater. Is it .35? Is it .40? What is the likely lower bound of this population correlation? The confidence interval gives us that information. In order to make this an apples to apples comparison, we need to compute a unidirectional confidence interval, meaning we'll compute just the lower bound of the interval. (Why? The test described above is one-tailed. Thus, we want only one half of a confidence interval. And, seriously, aren't we really just concerned with the lowest possible value for the population correlation? Who gets agitated about the maximum possible value? No one. That's who.) You don't know how to do this yet, so I'll just tell you that the lower bound of a unidirectional 95% confidence interval for a sample correlation of .5 based on a sample of 103 people is .366. There you have it. No need to run repeated tests using the equation from the previous section. We would have obtained a significant results if we tested to see if the population correlation is greater than .36, but not if it was tested against .37 (again, check this out). And it should be clear that there is no need to run the first significance test to see if the population corre-

lation is greater than zero. If it's greater than .36, then it is also greater than zero.

Now that I've convinced you of the value of computing confidence intervals, here's the bad news: It's a lot of work. The first step is an *r* to Fisher's *z* transformation for the sample correlation. The second step is compute the interval proper. I'll report the procedure for a bidirectional confidence interval (upper and lower bound). If you only want the lower bound, just do that part.

$$\text{Upper/Lower } Fz = Fz_r \pm z_{conf}\sqrt{\frac{1}{N-3}}$$

Where:
Fz_r is the sample correlation transformed to Fisher's *z*.
z_{conf} is the value from a *z* table corresponding to how confident we want our confidence interval to be. For a 95% confidence interval, we use the familiar 1.96 (bidirectional) and 1.65 (unidirectional) values, corresponding to our one- and two-tailed *z* tests from earlier.

At this point, we have the upper and lower bounds, so we're done, right? Or so it would appear. These upper and lower bounds are actually in Fisher's *z* terms. They need to be transformed back into regular correlations. Yes, that's right, you need to perform a Fisher's *z* to *r* transformation. Twice. Once for upper bound and once again for the lower bound. The equation to transform Fisher's *z* back into *r* is given below.

$$r = \frac{(e^{2Fz}) - 1}{(e^{2Fz}) + 1}$$

Where:
e^x is the inverse of the natural log.

Now that you have the equations, this would be a good opportunity to practice using them with the examples we have from earlier in this section ($r = .34$, $N = 55$, bidirectional confidence interval,

95% confidence; $r = .50$, $N = 103$, unidirectional confidence interval, 95% confidence).

Significance Test III: Correlation from Sample A Versus Correlation from Sample B

Our final significance test is really about one correlation compared to another correlation. These correlations are from independent (i.e., different) samples. For example, one researcher, Dr. Oldguard, correlated *X* with *Y* and found a correlation of .35 ($N = 84$). Another researcher, Dr. Newguy, also correlated a variable with *Y*, only this time it was a revised version of *X*. Dr. Newguy obtained a correlation of .45 ($N = 67$). Dr. Newguy wonders, "I wonder. Does my version of variable *X* predict *Y* better than Dr. Oldguard's version of *X*?" Now, obviously .45 is greater than .35, but these are sample correlations. The real issue is whether revised *X* predicts better than original *X* in the population. Stated formally: Is the observed difference in our sample correlations big enough to allow us to conclude that the relevant population correlations are actually different?

How do we do this test? Would you believe that the first step is an *r* to Fisher's *z* transformation? It's true. Transform both sample correlations to Fisher's *z* values. Then use those values in the following equation.

$$z = \frac{Fz_1 - Fz_2}{\sqrt{[1/(N_1 - 3)] + [1/(N_2 - 3)]}}$$

Where:

Fz_1 and Fz_2 are the Fisher's z values for the two correlations.

N_1 and N_2 are the sample sizes for the two correlations.

As is obvious, this is a *z* test, meaning our critical values are 1.96 (two-tailed) and 1.65 (one-tailed) if alpha is .05. In almost all cases, a one-tailed test is the appropriate test as most hypotheses are of the "This test will predict better" variety and not

"This test will predict better – or worse. I don't know, but it definitely will not be the same."

Concluding Remarks on Significance Tests

So there you have it. Two ways to do a significance test of a single correlation, a confidence interval for a single correlation, and a test to see if correlations from two different samples are different from each other. The one situation that we didn't cover is a version of the two correlations comparison where the correlations are from the same sample. We'll save that for another day.

There is one final point to discuss. This issue relates to the difference between probability samples and non probability samples, something covered in Chapter 2. All significance tests are predicated on the sample data coming from a probability sample. Apply a significance test to data collected from a non probability sampling technique and all bets are off. In the words of Pedhazur and Schmelkin (1991, p. 321), "The incontrovertible fact is that, in non probability sampling, it is not possible to estimate sampling errors. Therefore, the validity of inferences to a population cannot be ascertained." It is really hard to argue with an incontrovertible fact.

Range Restriction and What It Does to a Correlation

One last thing that you should know about correlations is that they are very sensitive to the variability in the scores on both *X* and *Y*: If the variance of scores on *X* in one sample is less than for another sample, all other things being equal, the strength of the correlation will be reduced.

Here is a practical example to clarify the issue. An unnamed college wants to determine the relationship between ACT scores and student GPA. To find the relationship, they need to have a sample of students with both ACT and GPA scores, which

means that they have to be actual college students. Now this college has been using ACT scores to make selection decisions for years, and this is a very selective school: The range of scores for applicants is 9 to 34, but the range of scores for admitted students is 26 to 34. Thus, the applicant sample has a much greater range of scores on the *X* variable than the actual student sample. In other words, any correlation performed on the student sample will have a restricted range as compared to the applicant sample. The bad news is that range restriction reduces the magnitude of (i.e., weakens) a correlation. Thus, a correlation computed on the selected student sample will be lower than if it were computed on the applicant sample. Interactive 2 presents an explanation of how range restriction weakens a correlation.

You may be thinking something like, "Well that's too bad for them. They deserve a reduced correlation. That's what happens when your sample is selected by using scores on the variable that you're correlating with the criterion, a concept known as direct range restriction." Well, to a certain extent, you're right. And how you knew all that amazes me. But the real question is not "What correlation do they deserve?" but rather "What is the real relationship between ACT scores and GPA at this school?" Here's the thing, in any study, the sample which we analyze needs to represent the relevant population. When this school excluded the lower scoring test takers (by not admitting them, and thus, preventing them from having

INTERACTIVE 2 How Range Restriction Weakens a Correlation

a GPA), they made it so that the student sample is not representative of the relevant population. All of those low scoring test takers, who would have had some low GPAs, are not included, weakening the resultant correlation. The real relationship between *X* and *Y* didn't change, but the relationship observed in the study changed because the study wasn't conducted on a sample representative of the relevant population – the applicants.

So what's the answer? It is clear that we should avoid range restriction to the extent possible. But how? Ideally, we would use a random process to make the selection decisions (remember that our college has been using the ACT for years to make selection decisions – that's what caused the problem for their correlation). If a random selection procedure is not possible, we have some alternatives that are beyond the scope of this book. Let's just say that they lead to a less harmful form of range restriction called indirect range restriction.

The moral of the story is that you should beware of possible traps, like range restriction, that can ruin your study. If you can't design your study so that range restriction doesn't occur (to be specific, it's direct range restriction that should be avoided at all costs), then you shouldn't do the study. A flawed study leads to the wrong conclusions (i.e., the test doesn't predict), which is worse than no study at all. If there is no study, at least you know better than to draw a conclusion can.

Simple Regression

4

If loving regression is wrong, I don't want to be right.

Introduction

First off, we need to address this word *regression*. Regression analysis is only vaguely related to the regular words *regression* or *regress*. The name is not important. They could have named it with a nonsense word like shnurffle. Or omegnacruz. Quavelcon analysis sounds pretty cool. Names aside, regression is a statistical procedure that describes the relationship between two variables (like correlation) and allows us to use this relationship to predict a person's score on *Y* given their score on *X* (unlike correlation). To reiterate: Correlation is a measure of association between two variables – regression is a measure of association *and* a method for predicting scores on one variable given scores on another. With regression, you get the bonus plan.

A few miscellaneous issues. First, regression and correlation are so closely related that it's hard to tell which one is derived from which (i.e., Which came first?). I like to conceptualize it as regression is an extension of correlation – it starts with correlation's function and takes it to a new level. That's how I like to think of it, but it doesn't matter a whit. Next, in this chapter (and the rest of the book), we won't address the sample-versus-population form of a statistic anymore. It was fun and all, but we're done with that. We'll just assume everything is a sample and present our equations accordingly. And you know enough by now that, if you had to generate a population version of any of these equations, you could figure it out without breaking a sweat. Finally, a note about terminology. This chapter is titled "Simple Regression." The full, proper name for what is described in this chapter is *bivariate linear regression*. Simple regression is a shorter name, and it fits well because bivariate linear regression is the simplest form of regression. Bivariate means two variables (*X* and *Y*). Linear means linear (like we discussed in the correlation chapter). And regression means,

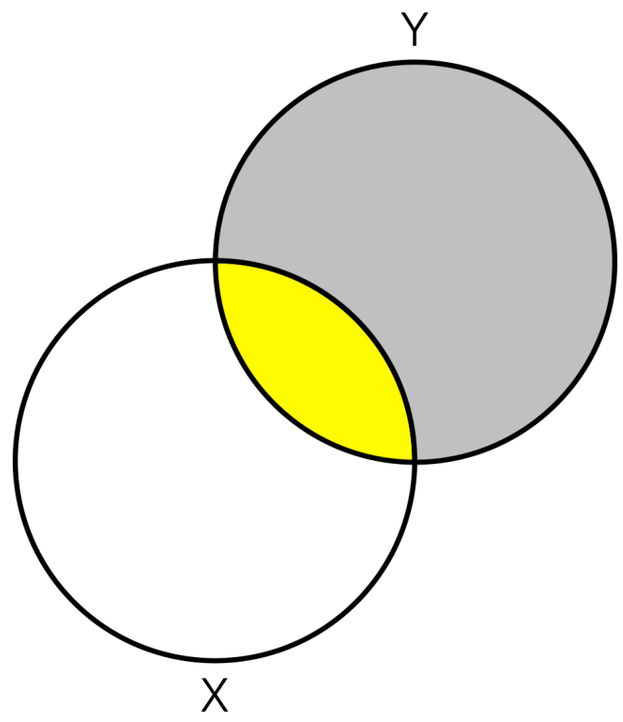
well, nothing helpful, as mentioned in the first paragraph.

Regression Philosophy

The basic principle of regression analysis is that the dependent variable can be conceptualized as being composed of two parts: a part that is related to the independent variable and a part that is not related to the independent variable. This philosophy is illustrated in Figure 1. It's the foundation upon which the regression house is built, and if we don't understand it, we'll be in serious trouble.

We can view this division of the dependent variable in terms of individual scores on *Y*, variance of the scores on *Y*, and even squared correlation coefficients. Let's start with individual scores on *Y*.

FIGURE 1 Regression Philosophy



Regression analysis breaks the dependent variable into a part related to the independent variable (yellow area) and a part unrelated to the independent variable (gray area).

A person's score on *Y* can be divided into part that is related to *X* and a part that is not related to *X*. In the form of an equation, this would look like:

$$Y = (\text{part related to } X) + (\text{part unrelated to } X)$$

Let's rename the *unrelated to X* component *residual* and give it the symbol *e*. Residual is a statistics term that means stuff not explained by the model. The *e* symbol is short for *error* (Who could have guess that one?), which itself is short for *error of prediction*. Furthermore, let's rename the *related to X* part *predicted Y* and give it the symbol Y'. Just new names and symbols for the same components. Using these new terms, our equation is:

$$Y = Y' + e$$

Same concept as before: A person's score on Y is composed of a part related to X (Y') and a part unrelated to X (e).

Now let's mathematically define e. Well, this turns out to be really easy because we can just use the previous equation, plus a little algebraic rearrangement, to make it happen.

$$e = Y - Y'$$

See, that is simple. e is defined as the difference between a person's actual score on Y and that person's predicted score on Y. We could substitute $Y - Y'$ for e in $Y = Y' + e$, giving us $Y = Y' + (Y - Y')$, and you would still see the same model as before: A person's score on Y is composed of a part related to X (Y') and a part unrelated to X ($Y - Y'$).

Now to the hard part, the mathematical basis for the part related to the independent variable (i.e., Y'). We can reason the solution from two just pieces of information. First, Y' must be derived from the relationship between X and Y. Second, we know that the correlation coefficient is an index of the strength and direction of the relationship between between X and Y. From these two points we can conclude that the correlation between X and Y will be involved in the prediction of scores on Y by scores on X.

Let's make our lives easier and assume that both X and Y are standardized variables. We could set up a prediction equation that looks like this:

$$z_{Y'} = r_{XY} z_X$$

Where:
$z_{Y'}$ is the standardized score on Y'.
z_X is the standardized score on X.

This equation states that standardized Y' scores, which reflect the relationship between X and Y, are the product of the standardized scores on X and the correlation between X and Y. Very simple. And guess what? This equation is the actual prediction equation for standardized data.

What if our data are not standardized? Because standardization removes any mean and standard deviation differences between variables, we'll just introduce means and standard deviation terms into our prediction equation. Our equation will take the form of:

$$Y' = (\text{something based on } r_{XY})X + \text{something}$$

So, it's similar to the standardized prediction equation. Scores on X are being multiplied by something involving the correlation between X and Y. To that we are adding a constant. Let's replace the "somethings" with the symbols b and a:

$$Y' = bX + a$$

The part that involves the correlation is called the regression coefficient (symbol: b) and is just the correlation times a ratio of the standard deviations.

$$b = r_{XY}\frac{S_Y}{S_X}$$

The other term in the equation is called the y-intercept (symbol: a) and contains the regression coefficient and the means of both variables.

$$a = \bar{Y} - b\bar{X}$$

There's really nothing all that complicated about these equations. They include the correlation, the standard deviations, and the means. Just a few basic statistics that we already know and love. (Fun exercise: Apply these equations to standardized data, which have means of zero and standard deviations of one. Check out what they reduce to when you do.)

Now that we know the symbols for each part, let us examine the regression equation again ($Y' = bX + a$). It states states that Y' scores, which reflect the relationship between X and Y, are the sum of the y-intercept (a) and product of the actual scores on X and the regression coefficient (b). And we know that the regression coefficient is really just the correlation between X and Y with some standard deviation stuff tacked on. And the y-intercept is just a boring scaling term (necessary because we are dealing with unstandardized data).

So now we have it. The equation above ($Y' = bX + a$) is the prediction equation for raw-score (i.e., not standardized), bivariate (meaning two variables, one independent and one dependent variable), linear (we'll get to linear later) regression. It's so important to what we'll be doing with regression, let's list it again. The prediction equation for raw-score, bivariate, linear regression is:

$$Y' = bX + a$$

Where:
b is the regression coefficient.
a is the y-intercept.

Now let's put this all together. In regression analysis, scores on Y are divided into a part that is related to X and a part that is not related to X. We expressed this philosophy with the following equation.

$$Y = Y' + e$$

Now that we know how to compute Y' scores ($Y' = bX + a$), let's substitute $bX + a$ for Y' in the above equation.

$$Y = bX + a + e$$

There it is. The philosophy of regression analysis in a fully defined equation. Scores on Y are composed of a part related to X ($bX + a$) and a part unrelated to X (e).

We could even go further and substitute $(Y - Y')$ for e. This substitution results in $Y = bX + a + (Y - Y')$. We don't need to do this, but we could.

Summary of Regression Analysis

Regression analysis views a dependent variable (Y) being composed of a part that is related to the independent variable ($bX + a$) and a part that is not related to the independent variable (e, defined as $Y - Y'$). This philosophy is defined with the following model:

$$Y = bX + a + e$$

The part of Y related to X is called predicted Y (and symbolized as Y') and is obtained with what can be called the prediction equation (from now on, we'll just call it the regression equation):

$$Y' = bX + a$$

It is this equation that is the basis for actual regression analysis, and it is this equation that we will use for computing predicted Y scores. The components of the regression equation (b and a) are rather simple are computed from the correlation between X and Y, the means of X and Y, and the standard deviations of X and Y.

A few final points and then we will move to some examples. First, I want to emphasize that Y' is not Y. Y' is the value of predicted Y (predicted

on the basis of X), whereas Y is a person's actual score on Y. Second, the regression coefficient (b) can be conceptualized as the weight applied to scores on X to get the best possible prediction of Y. Third, as with correlation, scatterplots will be used to illustrate the relationship between X and Y. Remember that line of best fit from the scatterplots in the correlation chapter? That's also called the regression line, and the regression equation is the equation for that line. b is the slope of the regression line and a is, you'll never guess this, the y-intercept for the line.

Regression Examples

We'll start with a sample dataset containing data from five people. No one in their right mind should collect a sample with only five people, but, like the other datasets in this book, this is just an example to illustrate principles. Using the correlation between X and Y, the means of X and Y, and the standard deviations of X and Y, I computed b ($b = 1.3$) and a ($a = 8.2$). Scores on X for this dataset are given below. Using these values for b and a our regression equation is $Y' = 1.3X + 8.2$ (or you could say $Y' = 8.2 + 1.3X$). We can compute Y' for each person by inserting each person's X score into the above equation.

Person	X	Y′
Hal	5	14.7
Fred	2	10.8
Eddie	6	16.0
Joe	8	18.6
Charles	2	10.8

That's it in all its glory. Predicting a person's Y score with a regression equation is a simple algebraic exercise (this is referred to *actuarial* or *statistical prediction*; in contrast there is *clinical prediction* which is a judgmental method). Note that the same X will always result in the same predicted Y

(see Fred and Charles). Our opinions don't count. Just the equation and the data.

Now let's say that we know each person's actual score on *Y*. We can compare these actual *Y* scores to our predicted *Y* scores.

Person	X	Y′	Y
Hal	5	14.7	16
Fred	2	10.8	9
Eddie	6	16.0	10
Joe	8	18.6	22
Charles	2	10.8	14

As we can see, our predictions were pretty close for some of the people (Hal and Fred) and were way off for the others (Eddie, who did a lot worse than we predicted, and Joe and Charles, who did a lot better than we predicted). The difference between the actual *Y* and the predicted *Y* (*e*, the residual) shows the amount of error in the prediction of *Y* for each person. Also note that Fred and Charles had the same predicted *Y* but had different scores on actual *Y*. This Fred/Charles situation illustrates how the same scores on *X* will always result in the same predicted *Y*; however, their actual scores on *Y* will likely turn out to be different.

Person	X	Y′	Y	(Y – Y′)
Hal	5	14.7	16	1.3
Fred	2	10.8	9	-1.8
Eddie	6	16.0	10	-6.0
Joe	8	18.6	22	3.4
Charles	2	10.8	14	3.2

I hope that it is obvious that we like it best when there are no errors of prediction. Such a situation would indicate that our predictions were perfectly accurate. But that doesn't happen in real life.

Let's talk about the accuracy of predictions made with regression analysis. Anyone can make

predictions. We want to make predictions when we have a good chance of being accurate. How can we know whether these predictions will be accurate? That's where correlation enters the picture. Stronger correlations between X and Y lead to more accurate predictions. (In the above dataset, $r_{XY} = .65$.) A perfect correlation (+1 or -1) would give us perfectly accurate predictions (0.0 residuals for all people). Of course, perfect correlations don't happen in the real world, but you get the idea.

A Regression Thought Experiment

Let's review the equations for b and a.

$$b = r_{XY}\frac{S_Y}{S_X}$$

$$a = \bar{Y} - b\bar{X}$$

An examination of the equation for b reveals something interesting: The most important part of b is the correlation between X and Y. The standard deviations are just scaling terms. (Consider that if the data are transformed to z scores, then S_X and S_Y both equal 1.0 and are irrelevant.)

Here's an interesting thought experiment: What happens if $r_{XY} = 0$? In this scenario, we won't even have to standardize our data. It can be raw data. Using the above equations, when $r_{XY} = 0$ then $b = 0$ (because anything multiplied by zero is zero). And if $b = 0$, $a = \bar{Y} - 0\bar{X}$, which simplifies to $a = \bar{Y}$. And, thus, the regression equation is $Y' = 0X + \bar{Y}$, which simplifies nicely to $Y' = \bar{Y}$. To restate: if $r_{XY} = 0$, then the regression equation is $Y' = \bar{Y}$. Thus, predicted Y is the mean of Y for all scores on X. It doesn't matter what your score on X is, we predict the mean of Y for you. Why? Because there is no association between X and Y. Thus, why should I care about your X score in my prediction of Y? And predicting people to be average is the safest prediction. That zero correlation tells us X is irrelevant. Earlier, we said that b tells

us how much to weight scores on X to get the best possible prediction of Y. If $r_{XY} = 0$, then X doesn't matter, and I should give it no weight in my prediction of Y (hence, $b = 0$ in this scenario).

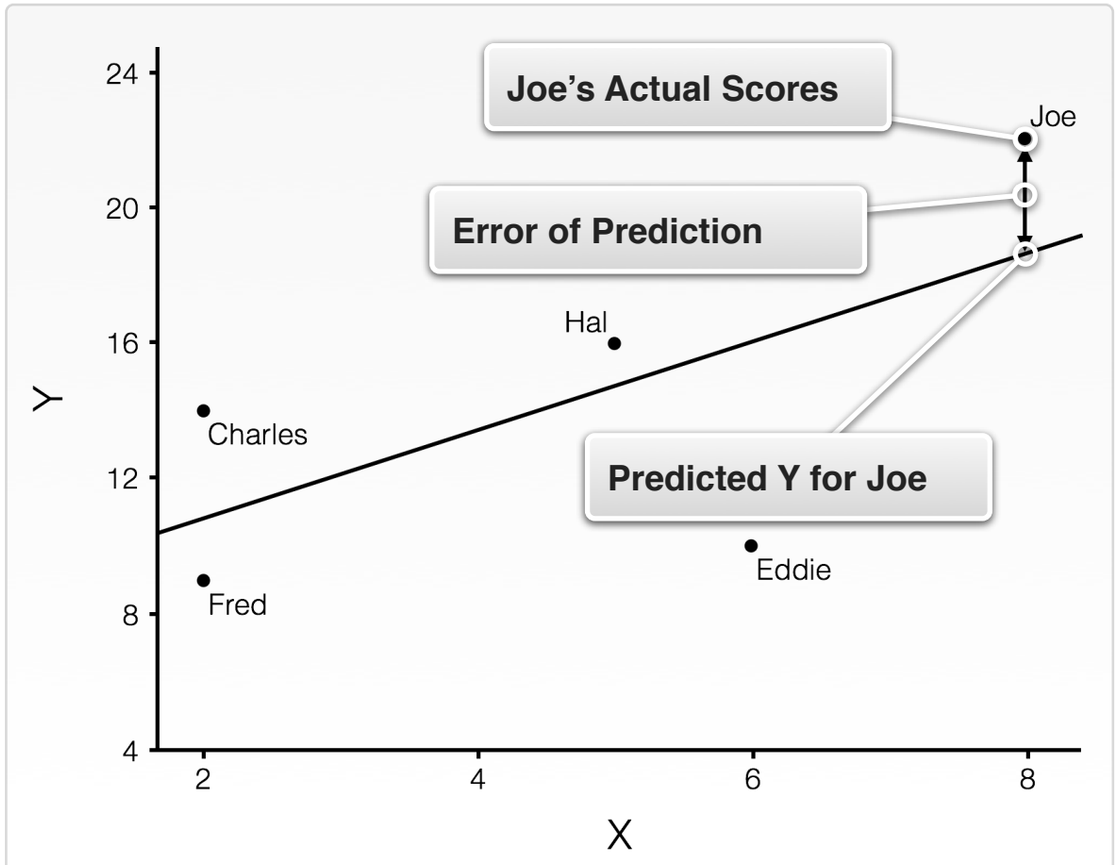
To the converse, what if $r_{XY} = 1.0$? To make matters easy on ourselves, let's also make X and Y standardized data so that the means are 0.0 and the standard deviations are 1.0. Inserting these numbers into our equations for b and a yields $b = 1.0$ and $a = 0$. (Try it – you'll see.) Thus, our regression equation is $Y' = 1X + 0$, which simplifies to $Y' = X$. With this equation, if X is 2.3, then Y' is 2.3. And if X is -1.9, then Y' is -1.9. Predicted Y is exactly as high or low as X.

Conclusions from our little thought exercises: The relationship between X and Y strongly determines the types of predicted Y scores generated from a regression equation. If $r_{XY} = 0$, then all predicted Y values are the same, regardless of the score on X. If r_{XY} is weak, then all predicted Y values are close to mean, even for people with extremely high or low scores on X. If r_{XY} is strong, then people with extremely high or low scores on X will have very high or low predicted Y values.

The Regression Line

Remember that line of best fit in the correlation graph? It was also called the regression line. It is a visual representation of Y' values for all scores on X. You can draw the line by plugging all possible values of X into the regression equation, obtaining Y' for each X and graphing each of the X, Y' points. Or you could just draw the line using the regression coefficient b as the slope and the a as the y-intercept. Our most recent dataset is graphed in Figure 2, and it includes the regression line. Note that the error of prediction (or residual) is indicated graphically by the vertical distance between each point and the regression line. Bigger distances mean worse prediction (more error). Let's examine Joe's data. Joe has a score of 8 on X.

FIGURE 2 Actual Y, Predicted Y, and Errors of Prediction (e)



His predicted Y is 18.6. His actual Y is 22. Thus, his error of prediction is +3.4. That is, he performed 3.4 points better than we predicted. By comparison, Hal has a much smaller error of prediction (1.3). Someone with scores located right on the regression line would have an error of prediction of zero. The old rule from correlation-land is relevant again: Stronger associations (which lead to more accurate predictions) are those with points closer to a straight line. Correlation and regression, two sides of the same coin. Or maybe they are the same side of the same coin. I never did understand that expression.

A quick note on interpreting the regression coefficient. The regression coefficient, b, tells us something useful, and unique, about the relationship between X and Y. For simple regression, b indicates the expected change in Y given a one point change in X. Here's an example. Let's say we conduct an experiment where we assign people to varying levels of study time and then measure their test performance. We regress test scores (Y) on study time scores (X, measured in hours) and obtain the regression equation $Y' = 21X + 11.5$. In this equation, b is 21. Using our interpretive rule, for every one hour increase in study time, we expect to see a 21 point increase in test scores. We

may find this information to be very useful in evaluating the relationship between study time and test performance. Also, the effects of *b* are cumulative; a three hour increase in study time will be expected to lead to a 63 point increase in test performance. Nice.

So the regression coefficient can be a useful indicator of the strength of the association between two variables. Just like correlations. Sometimes a regression coefficient can be every bit as useful (arguably more so) than a correlation. When is that? It all depends on whether the dependent variable is expressed in a meaningful metric. What kind of metrics of measurement are meaningful? Anything that has real world relevance, such as time (e.g., to complete a task), number of something (e.g., mistakes), dollar value (e.g., of items sold), and so on (e.g., and such and forth). Let's say that the dependent variable is expressed in dollars. If the regression coefficient is 50, then for every one point change in *X*, we expect *Y* to increase by $50. Based on the nature of the experiment, it will be easy to interpret whether $50 is a meaningful change (and thus, a strong relationship) or a trivial one.

As long as we're discussing strength of association, remember how we mentioned in our discussion of correlation that we shouldn't use the slope of the line on the scatterplot as an indicator of the strength of the association? Mostly because the *x*- and *y*-axes can be easily manipulated to produce the appearance of a strong slope. Remember that? Well, here's the thing. Even though it's unwise to use apparent slope of the regression line on the scatterplot as an indicator of the strength of association, it is fine to use the regression coefficient as an indicator of the strength of association. How is that OK, you're thinking? I once had a perceptive student ask me this very question. Well, we know why a visual inspection of the slope of the line on the scatterplot is a bad idea. But wait, didn't we just learn that the regression coefficient

is the slope of the regression line? How then is the regression coefficient useful? The answer is that it's not as easy to subtly manipulate the regression coefficient. You can't just change the scale on the axis of some graph and get the desired effect. You would have to change the scale of the data itself (e.g., multiply all of the scores on the dependent variable by 10). Such a change would be obvious. In short, if one wants to produce the appearance of a strong relationship, it's more difficult to manipulate the regression coefficient than it is to manipulate the apparent slope of the regression line on the scatterplot. The former is a number. The latter is a visual representation of that number.

To show how a regression has a real-world usefulness, let's pretend that we are in charge of admissions of a certain college. We'll call it Enormous State University (or ESU). At ESU we are considering using the ACT for freshman admissions. Somebody somewhere (maybe at arch-rival Enormous Tech) did a study and found a .5 correlation between ACT scores and college GPA. Based on that correlation, we decide to use the ACT at ESU. Naturally, we also need the means and standard deviations of X and Y to set up our regression equation. Once we get these data, we obtain a regression equation of $Y' = .1X + .5$. For every high school student that applies, we plug his or her ACT score into our equation, which generates a predicted Y for each person. If the predicted Y is high enough (say, greater than 2.0), then we admit the student. If not, then we send him the other letter.

Theoretical Basis for Regression Analysis

To really understand how regression analysis works, we need to further analyze the theoretical foundation. We start with a simple equation.

$$Y = Y$$

Equations don't get much simpler than that. A person's score on Y equals that person's score on Y. This is like saying $12 = 12$, a valid equation, but fairly useless. There's an algebra maneuver that allows you to do whatever you want to an equation as long as you do it to both sides of the equation. So let's add a few terms to each side of the equation.

$$Y + Y' + \bar{Y} = Y + Y' + \bar{Y}$$

Where:
Y is the actual score on Y for a given person.
Y' is the predicted value of Y for a given person.
$\bar{Y}$ is the mean of scores on Y.

And follow that with a little algebraic rearranging.

$$Y - \bar{Y} = Y' - \bar{Y} + Y - Y'$$

And put parentheses around some terms.

$$(Y - \bar{Y}) = (Y' - \bar{Y}) + (Y - Y')$$

In the final equation above, the difference between a person's score on Y and the mean of Y can be explained by what's on the right of the equal sign. Let's not get into that just yet, though.

Our next step will be to take this equation, which is in the form of a single person's scores, and write it as a sum of squared values across the entire sample of people. (There's a long, algebraic proof for the validity of this move somewhere, but I don't have it handy. Trust me – you wouldn't want to see it if I did.)

$$\Sigma(Y - \bar{Y})^2 = \Sigma(Y' - \bar{Y})^2 + \Sigma(Y - Y')^2$$

If these sum of squares things look familiar, it's because the numerator of the variance formula (see Chapter 2 for a refresher) is itself a sum of squares. Just think of a sum of squares as meaning *simplified variance*, because like all variability equations, it's an index of differences in scores. The

first term is the sum of squares of Y (which is identical to the numerator of the variance equation). The second term is called the sum of squares regression. The final term is called the sum of squares residual. (Remember the residual, or error of prediction, from earlier? There it is. $Y - Y'$ just like before.) Let's rewrite this equation using our new names.

$$SS_Y = SS_{reg} + SS_{res}$$

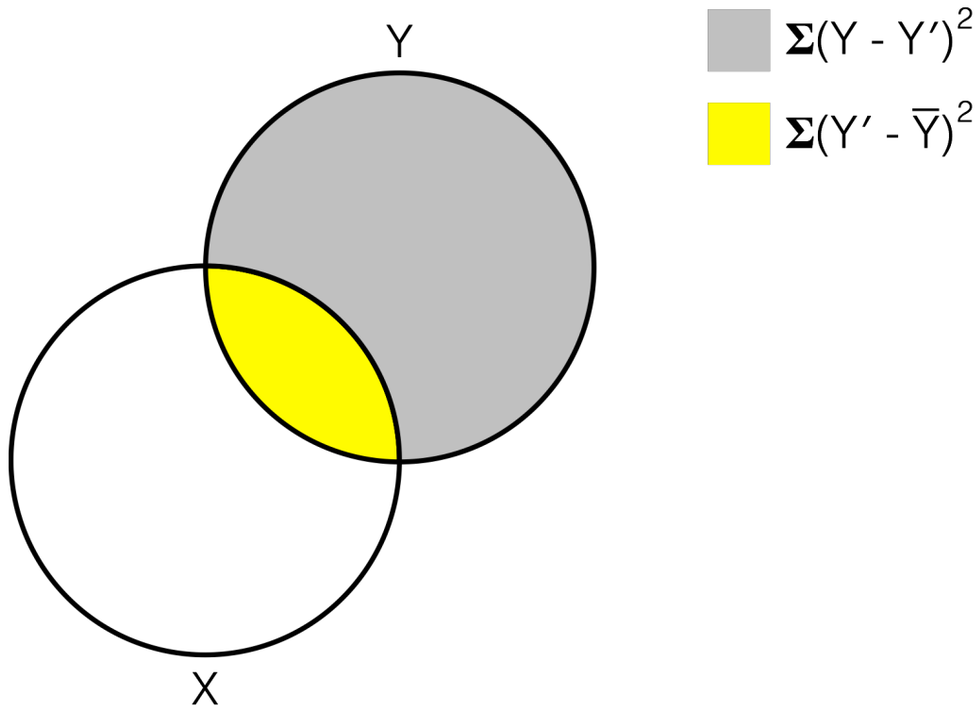
What this equation is saying is that variability (i.e., score differences) in Y (i.e., SS_Y) is a function of the regression and residual component of regression analysis. What's the regression component all about? We'll need to take a long look at it to find out. SS_{reg} is $\Sigma(Y' - \bar{Y})^2$, which is the sum of the squared differences between between predicted Y and the mean of Y. To understand the importance of this, we need to understand whence predicted Y emerged. Consulting our regression equations from earlier in the chapter tells us that $Y' = bX + a$, where $b = r_{XY}(S_Y/S_X)$ and $a = \bar{Y} - b\bar{X}$. As we discussed earlier, if $r_{XY} = 0$, then $b = 0$ and ultimately, $Y' = \bar{Y}$ for all values of X (i.e., everyone).

This is an important point, so I'll start a new paragraph and restate it: $Y' = \bar{Y}$ when there is no relationship between X and Y. Thus, we can conclude that Y' differs from $\bar{Y}$ only when there is a relationship between X and Y. With that principle in mind, we can understand what is going on with the SS_{reg} term (i.e., $\Sigma(Y' - \bar{Y})^2$). Stronger relationships between X and Y mean bigger SS_{reg} due to the greater differences between Y' and $\bar{Y}$. So regression refers to the relationship between X and Y, and the sum of squares regression is an index of the magnitude of the relationship between X and Y. When there is no relationship between X and Y, the sum of squares regression is zero. When there is a perfect relationship between X and Y, the sum of squares regression is huge – it equals the total sum of squares in Y (i.e., the total variability in Y).

What about the sum of squares residual? This part is easier to understand. Residual refers to error of prediction and is defined as $Y - Y'$ (i.e., the difference between actual Y and predicted Y). When there is a perfect relationship between X and Y, Y equals Y' for everyone, and all residuals are zero. Thus, the sum of squares residual is zero. Inference: The residual reflects the lack of a relationship between X and Y, and the sum of squares residual is an index of this lack of a relationship. When there is no relationship between X and Y, the sum of squares residual is, you guessed it, huge; it equals the sum of squares in Y. (You can take my word for it, but if you want to see it for yourself, remember that when $r_{XY} = 0$, Y' equals $\bar{Y}$ for every person. Substitute $\bar{Y}$ for Y' in the sum of squares residual and you can see why the residual equals the sum of squares in Y.)

To summarize our discussion of all things regression and residual (and their sums of squares), regression reflects the relationship between X and Y, and residual reflects the lack of a relationship between X and Y. These two sums of squares add to form of the total sum of squares in Y. Stated succinctly, the total variability in Y (SS_Y) can be divided into variability based on the relationship be-

FIGURE 3 Sums of Squares Regression and Residual Illustrated



Variability in Y can be divided into a sum of squares regression (yellow area) and a sum of squares residual (gray area). Dividing both of these sums of squares by $\Sigma(Y - \bar{Y})^2$ (the total sum of squares in Y) turns the sum of squares regression (yellow) into R^2 and the sum of squares residual into $1 - R^2$.

tween X and Y (SS_{reg}) and variability based on the lack of a relationship between X and Y (SS_{res}) Figure 3 is our familiar Venn diagram with the variability in Y divided into a sum of squares regression (yellow) and a sum of squares residual (gray).

One final algebra move: Divide all of the terms in our $SS_Y = SS_{reg} + SS_{res}$ equation by the sum of squares of Y.

$$SS_Y/SS_Y = SS_{reg}/SS_Y + SS_{res}/SS_Y$$

As far the left side of the equal sign goes, it should be obvious that SS_Y divided by itself reduces to 1.0. But what about the terms on the right side? What happens to them? Well, we already know that SS_{reg} is the variability due to the relationship between X and Y. Dividing that value by the total variability in Y gives us the percent of variance in Y that is due to the relationship between X and Y. Any ideas as to what that percent equals. Are you sitting down? $SS_{reg}/SS_Y = r^2_{XY}$. That's right, it's the same correlation that we started with, only squared. r^2_{XY} is has a cool name: coefficient of determination. And it has a cool definition: The percent of variance in Y explained by X. Very cool. What about the final term, SS_{res}/SS_Y? As we said before, the residual is the lack of a relationship between X and Y, and the sum of squares for the residual is the variability due to this lack of a relationship. Dividing SS_{res} by the total variability in Y gives us the percent of variance in Y not explained by X. In r^2_{XY} terms, that's $1 - r^2_{XY}$. With this information, we can rewrite the previous equation as:

$$1 = r^2_{XY} + (1 - r^2_{XY})$$

So, there we have it. The logic of regression analysis. It's actually kind of beautiful.

Concluding Thoughts on Regression Logic

Earlier in this chapter, without any explanation, I casually dropped the term *ordinary least squares regression* (OLS regression for short). We now know enough to explain what that means. The *least squares* part of OLS regression refers to the sum of squares residual (SS_{res}). As you recall, the residual is the error of prediction, the difference between the actual score on Y and the predicted Y. Bigger residuals mean worse prediction. This regression analysis that we have been discussing is based on minimizing the sum of squares residual. We want the regression coefficient (b) and y-intercept (a) that produces predicted Y scores with the minimum possible SS_{res}, hence the term *least squares*. (Presumably, you could figure out b and a by randomly trying various values for each, computing the sum of squares residual for each combination of b and a until you found the one combination that minimized the sum of squares residual. You could do it that way, but I wouldn't recommend it. It sounds like work.) I don't know where the *ordinary* part came from. Let's hope that's not important.

Back to the least squares part. I wouldn't blame you for thinking, "Wait, I thought you told me the b and a came from those equations earlier. They're simply a function of the correlation, standard deviations, and means." That's correct. Deriving the b and a in that fashion minimizes the SS_{res} for simple regression. There is no other combination of b and a that will result in a better SS_{res} in that sample.

I haven't exactly told you anything new here. The regression coefficient is a simple function of the correlation between X and Y and the standard deviations of X and Y. The y-intercept is a simple function of the regression coefficient and the means of X and Y. Using these statistics will allow us to predict Y with the minimum residual possible in that sample. No tough decisions. Since

there are no tough decisions to be made here, why all of the fuss over this "let's minimize the sum of squares residuals" stuff? Just wait until we get to multiple regression...

A Comment on the Use of the Word "Predict"

The regression equation is designed to predict scores on *Y*. The regression weights chosen are those that predict *Y* most accurately in that sample. More accurate predictions mean smaller residual scores. Smaller residual scores mean reduced residual variance, which means the independent variable is accounting for more of the variance in *Y* (which means a stronger R^2, the index used to determine the strength of the relationship between *X* and *Y*).

Please do not interpret the use of language such as "variable *X* contributes to the prediction of *Y*" to mean that we were only addressing predictive, as opposed to causal (or explanatory), research. This is simply the language of the regression equation. *Y′* is called predicted *Y*. The purpose of the regression equation is to weight the various independent variables to generate scores that match actual *Y* scores as closely as possible. That is, to *predict Y*. Whether you are using this regression equation for predictive or causal research is a separate issue.

Cool Regression Tricks

There are a few things we can do to illustrate that wonderful regression logic from the previous sections. Let's use our previous dataset for this.

Person	X	Y	Y′	(Y – Y′)
Hal	5	16	14.7	1.3
Fred	2	9	10.8	-1.8
Eddie	6	10	16.0	-6.0
Joe	8	22	18.6	3.4
Charles	2	14	10.8	3.2

For every person, we have scores on X and scores on Y. Our regression equation ($Y' = 8.2 + 1.3X$) gives us predicted Y scores and residual scores. Also, it is worth noting that $r_{XY} = .65$.

As we have stated more than a few times by now, scores on Y (and variance thereof) can be divided into a part related to X (predicted Y scores, and variance thereof) and a part unrelated to X (residual scores, and variance thereof). To illustrate this, let's explore the relationship between X, Y, Y', and residual scores with a few analyses.

First, a simple one to demonstrate how scores on Y are broken into the two parts mentioned above. In our example dataset, consider Hal's scores. Hal's actual score on Y is 16. That 16 on Y can be divided into a part related to X and a part unrelated to X. The part related to X is Y', and Hal's Y' score is 14.7. Interesting. The part unrelated to X is the residual score ($Y - Y'$), and Hal's residual score is 1.3. Add the two parts (predicted Y score and residual score) together and you get 16, which is Hal's score on Y. You are probably not impressed because all of this follows from the very definition of residual. But, it's one thing to say it; it's another thing to see that it really works that way with actual data.

Second, let us explore the correlation between X and predicted Y. Before I tell you the correlation, recall that predicted Y comes from the regression equation which, in this case, is $Y' = 8.2 + 1.3X$. Thus, every predicted Y score is somebody's score on X times 1.3 followed by the addition of 8.2. This sort of transformation is known as a linear transformation. We've seen linear transformations before when we discussed z scores. The interesting thing about a linear transformation is that it doesn't change the correlation. If that seems shocking to you, it shouldn't. Recall what we learned about indices of bivariate associations: A strong, positive association occurs when people with high scores on one variable (e.g., X) have

high scores on another variable (e.g., Y) and when people with low scores on X also have low scores on Y. In short, strong associations occur when we observe a consistent pattern of scores between the two variables. Adding 8.2 points to everyone's score isn't going to change that consistency. Multiplying every X score by 1.3 will not change that either. The only thing that a linear transformation could change is the direction of the correlation, and that would only occur if we multiplied by a negative number (i.e., when b is negative). That said, let's answer the question, what is the correlation between X and Y'? It should be obvious by now that $r_{XY'} = 1.0$ (or -1.0 when the b is negative, which only occurs when is r_{XY} negative) because, in bivariate regression, Y' is just a linear transformation of X.

Next, let's explore the correlation between Y' and Y. Given what we just learned in the previous paragraph, we can reason this out. If Y' is just a linear transformation of X (which it is), and if $r_{XY'} = 1.0$ (which it does), then it stands to reason that the correlation between Y and Y' will be the same as the correlation between X and Y. And that's what we observe: $r_{XY} = .65$ and $r_{YY'} = .65$. (Side note: $r_{YY'}$ is always positive, even if r_{XY} is negative. The b takes care of the negative relationship.)

For our final analysis, let's explore what happens when we correlate the residual scores with the X scores. As mentioned many times, the residual is the part of Y unrelated to X. Thus, any correlation between scores on X and the residual scores $(Y - Y')$ should be, well, pretty much zero. That's what *unrelated* means. No relationship. Therefore, it's no surprise when we compute the correlation and find that $r_{X(Y-Y')} = 0.0$. The thing that is so cool about this last analysis is that it's one thing to say that the residual is the part of Y that is unrelated to X, but it's far more impressive when you use a regression equation to first compute Y' scores, then the residual scores, and find that the correla-

tion between X scores and the residual scores is exactly zero. (By the way, feel free to enter the table data into a statistics program and compute these correlations. Go for it, it's fun.)

To summarize everything up to this point, the regression model states that scores on Y (and its variance) can be divided into two parts: a part related to X (Y' and its variance) and a part unrelated to X ($Y - Y'$ and its variance). We see the first part with a correlation between X and Y', which equals 1.0. We see the second part with a correlation between X and Y', which equals zero. I think it's pretty cool. You may not. You would be wrong.

Regression Assumptions: Overview

Now is the time for a discussion about assumptions in regression analysis. And there are many. As with any statistic, you'll always get a result when you perform the analysis. But violating an assumption of that analysis means that the results will not necessarily match reality. So, yes, assumptions are a big deal.

The assumptions associated with regression analysis come in two flavors: those affecting the mathematical accuracy of the result and those affecting causal inferences made on the result. The first category includes linearity and homogeneity of variance. The second category includes model specification and the nature of the independent variable; these are relevant only for causal research. There is a final assumption, that the independent variable is measured without error, that sort of fits in both categories, but it works better with the latter group.

Regression Assumptions: Linearity

Simple linear regression has, as does correlation, an assumption of, you guessed it, linearity. Linearity means that the rate of increase (or de-

crease) for scores on *Y* remains the same across the range of scores on *X*. Another way of stating linearity is that the best fitting trend line is a straight line. Yet another way of stating the linearity assumption is that the most accurate summary of the observed relationship between *X* and *Y* is also the simplest: Higher scores on *X* are associated with higher scores on *Y* (or lower, if it's a negative relationship). Contrast that statement with the following: Higher scores on *X* are associated with higher scores on *Y* until a certain point at which scores on *Y* no longer increase. That statement is considerably more complicated, both mathematically and grammatically.

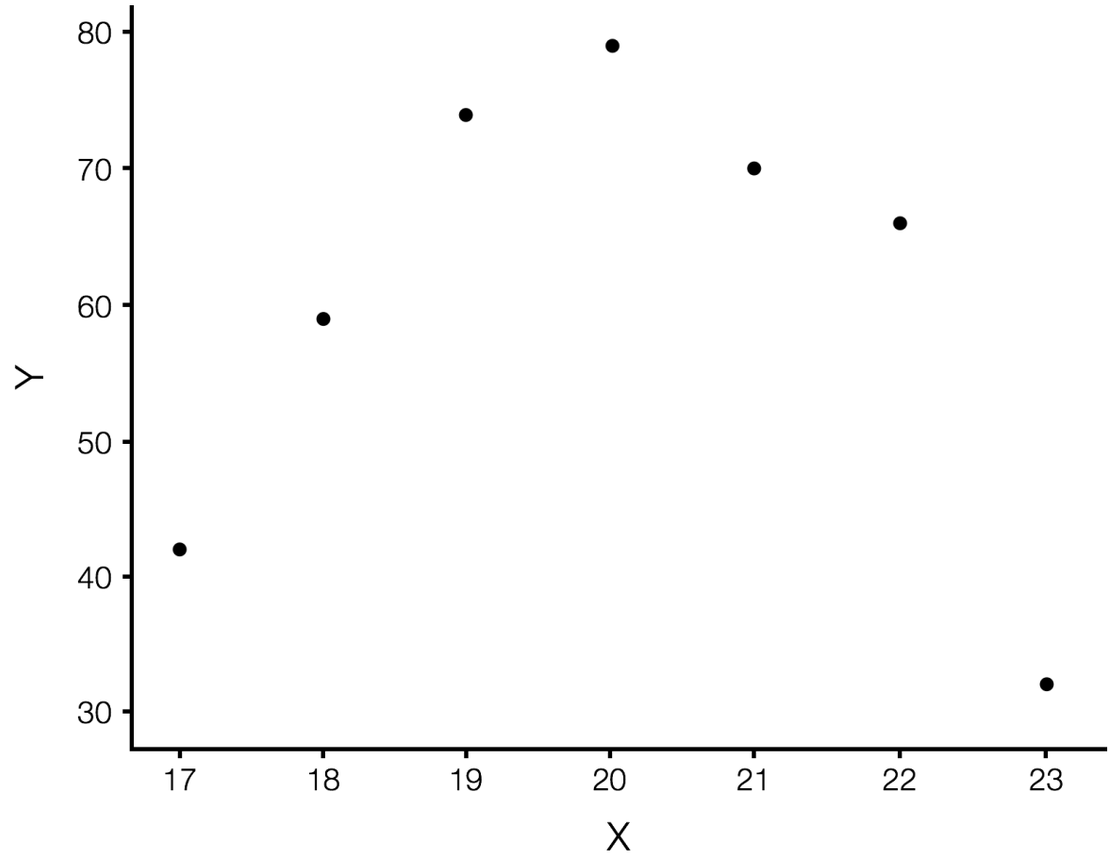
So linear regression has an assumption of linearity. What happens if this assumption is violated? If the linearity assumption is violated, a linear regression will underestimate r^2 (and *r* and *b*), and the regression equation will not accurately model the relationship between *X* and *Y*. Consider the following dataset and its scatterplot (Figure 4).

Person	X	Y
Gretchen	18	59
Steven	17	42
Jane	20	79
Mike	21	70
Brandon	23	32
Wendy	22	66
Pete	19	74

This relationship can be described as follows: Low scores on *X* are associated with low scores on *Y*, medium scores on *X* are associated high scores on *Y*, and high scores on *X* are associated with low scores on *Y* (note the complexity of this summary). Figure 4 shows a strong relationship between *X* and *Y*. That relationship just happens to be nonlinear.

As mentioned, a linear regression underestimates the strength of a nonlinear relationship (i.e., when the linearity assumption is violated).

FIGURE 4 Nonlinear Trend Scatterplot



Back at the beginning of the correlation chapter, we stated that for a scatterplot, the strength of the relationship is demonstrated by how close the points are to the regression line. Well, let's apply that principle to Figure 4. No matter where you draw a straight line on it, at least half of the points will have a large vertical distance between those points and the line. A linear regression of Y on X (Side note: We always say, "Regress the dependent variable on the independent variable." DV on IV, in that order.) results in an r^2 of .008 ($r = .09$). Thus, a linear regression of these data indicates an extremely weak relationship between X and Y when there is in fact a strong relationship between X and Y. The strong relationship just happens to be nonlinear. Thus, a linear model does not properly describe this relationship. That's the bad news. The good news is that there is a way to conduct a regression analysis that doesn't require a linearity assumption. We'll discuss this nonlinear regression analysis at a later date.

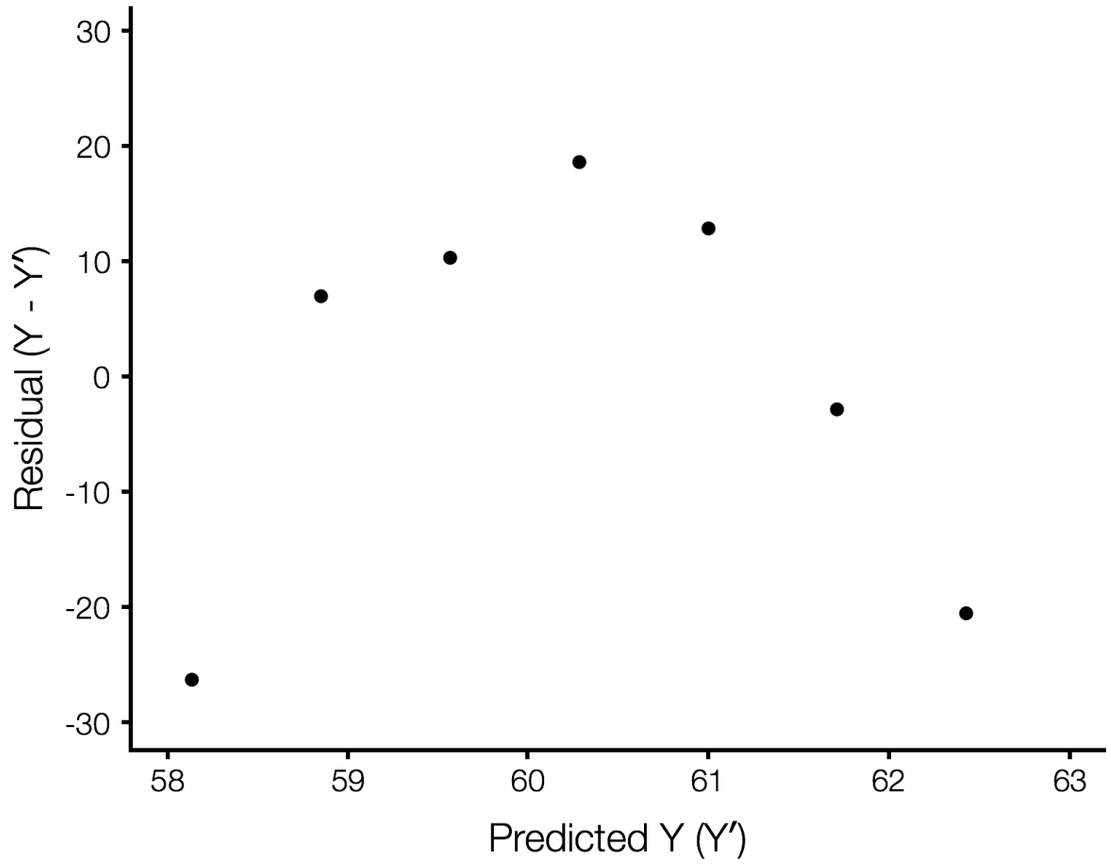
How do we check for violations of the linearity assumption? We just saw the nonlinear nature of this relationship in the standard scatterplot of scores on Y and X. Is that it? Just graph the scores on the standard scatterplot? Well that works here, in simple regression. But it doesn't work in multiple regression. So let's learn just one method that

works for both. Instead of graphing scores on Y and X, we graph residual scores (i.e., $Y - Y'$) on the y-axis and predicted Y scores on the x-axis. For the previous example, the predicted Y scores and residual scores are listed below (note that the linear regression equation is $Y' = -.71X + 74.6$).

Person	X	Y	Y′	(Y – Y′)
Gretchen	18	59	61.7	-2.7
Steven	17	42	62.4	-20.4
Jane	20	79	60.3	18.7
Mike	21	70	59.6	10.4
Brandon	23	32	58.1	-26.1
Wendy	22	66	58.9	7.1
Pete	19	74	61.0	13.0

The residual plot is shown in Figure 5. Seeing this nonlinear trend in the residual plot indicates that there is a violation of the linearity assumption, and thus, we need to proceed with a nonlinear regression analysis.

FIGURE 5 Nonlinear Trend Residual Plot



So that's what things look when the linearity assumption is violated. What do things look like when it's not? Figure 6 is a standard scatterplot (i.e., not the residual plot) for a new dataset (not shown) where $r_{XY} = .70$.

FIGURE 6 Strong Linear Association Scatterplot

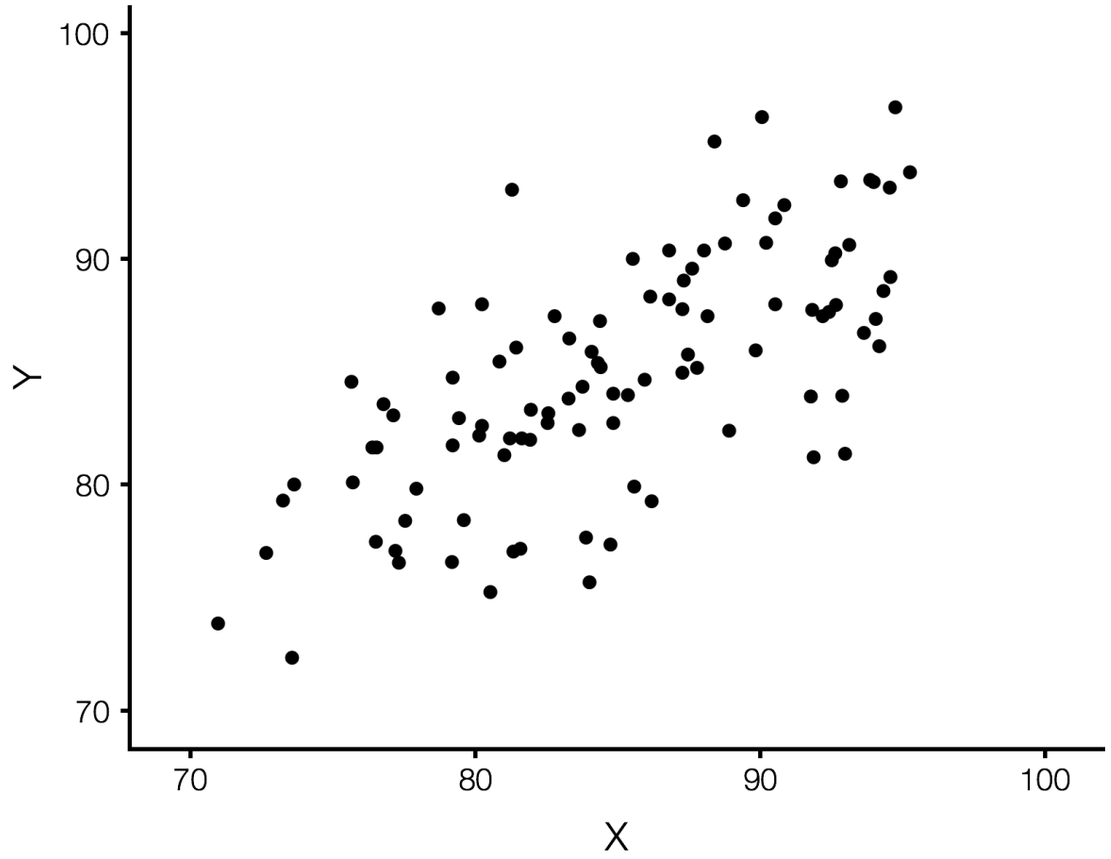
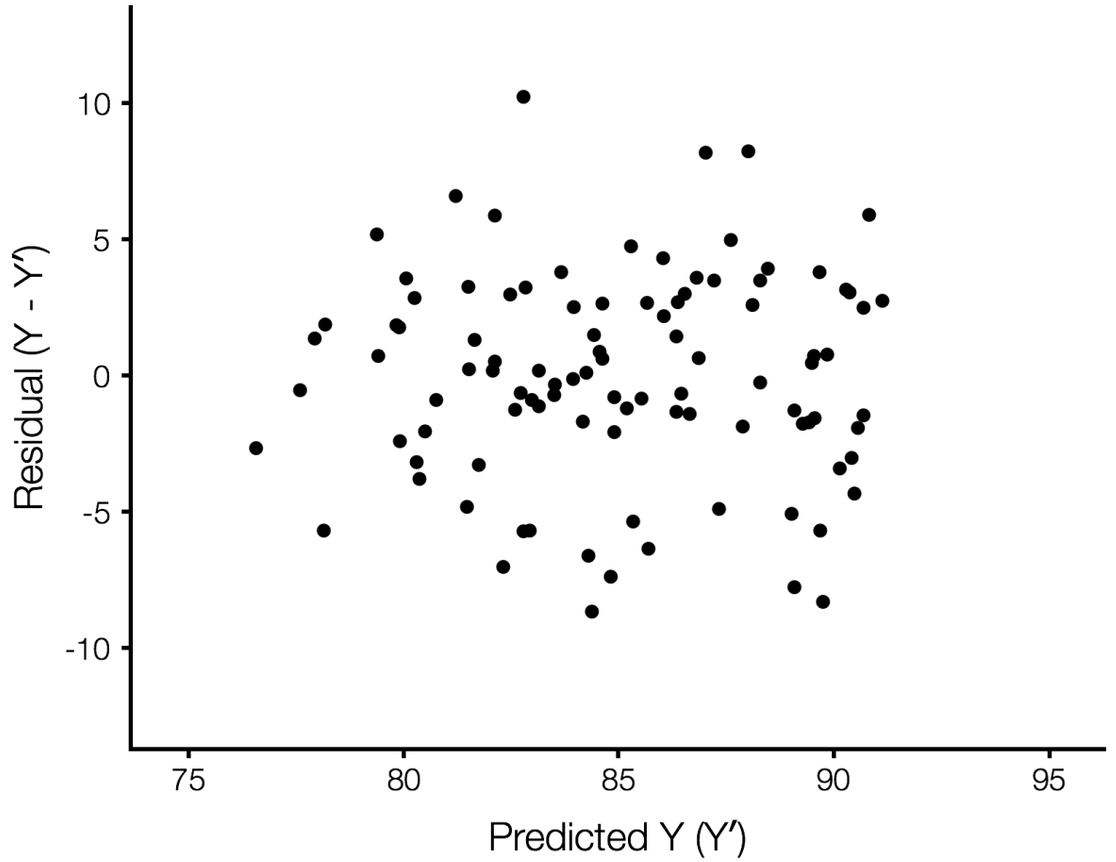


FIGURE 7 Strong Linear Association Residual Plot



And the residual plot for these scores in shown in Figure 7.

That's right, a residual plot for a dataset without a linearity assumption violation looks like scatterplot for a zero correlation dataset. This is what we want to see when we check the residual plot for violations of linearity assumption. No violations. (Unless we want a nonlinear association, which is arguably way cooler. But you won't know if it's there if you don't check for it.)

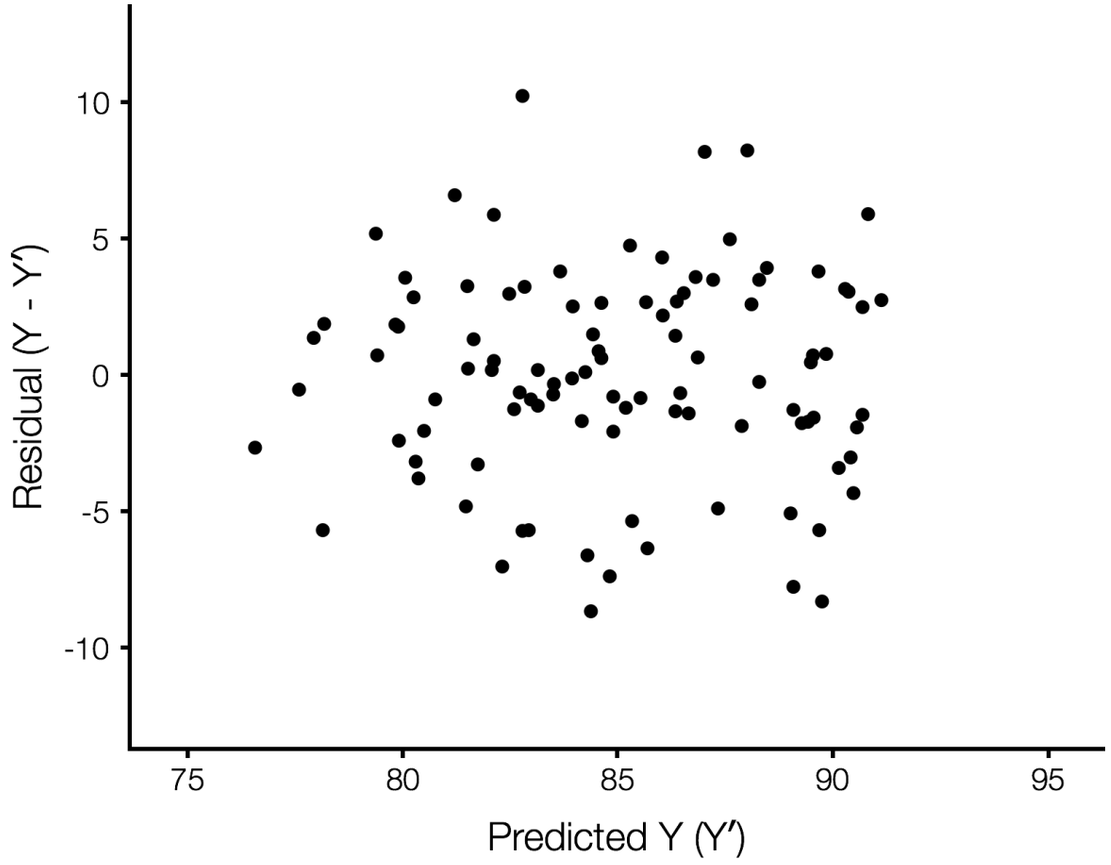
Before we move on to the second assumption of regression, what does this latter example tell us about residual plots in general? The scatterplot showed a nice linear trend, but the residual plot

showed a mass of points with no trend. The answer is that the residual plot shows the relationship between X and Y after the linear relationship between them has been removed. Remember that the residual is $Y - Y'$ and reflects the lack of a relationship between Y and X. It's what is unpredicted by scores on X. Thus, an examination of these unpredicted values can tell us what the linear prediction (i.e., Y') failed to capture. That's the y-axis. Why then are predicted Y scores plotted on the the x-axis? Why not just scores on X? The answer is that in simple regression, it makes no difference. Predicted Y is just a linear transformation of scores on X. But in multiple regression, it matters. Like I said, let's just learn one way to do this that will work for every condition.

Regression Assumptions: Homogeneity of Variance

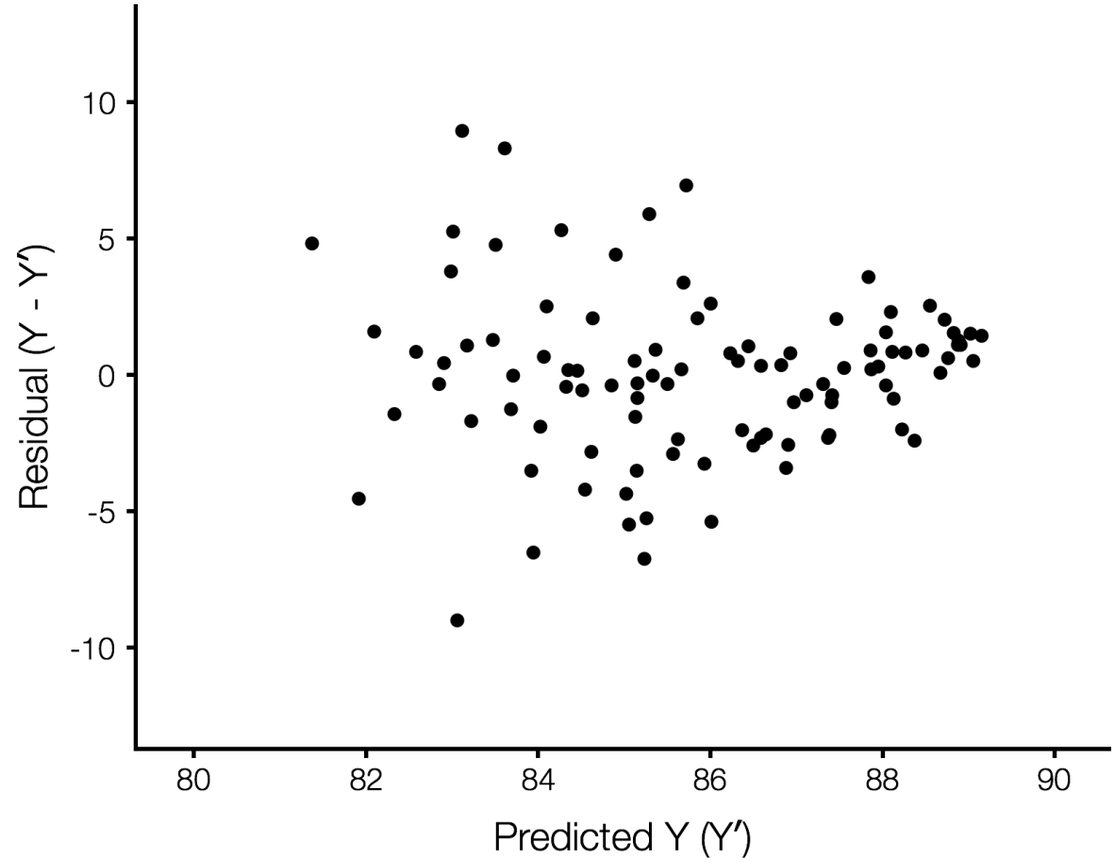
Our next assumption, homogeneity of variance (also referred to as homoscedasticity), is about the variance of the residual scores. The assumption is that residual variance is homogeneous, or constant, across all levels of X. As mentioned, residuals are the differences between Y and predicted Y. Bigger residuals indicate worse prediction. Regression assumes that the prediction equation works equally well across the range of scores on X. That is, the magnitude of the residuals is, on average, the same for high, medium, and low scores on X. What does homoscedasticity look like? Well, as with the linearity assumption, we check for violations of the homogeneity of variance assumption with the residual plot (hardly a surprise given that we're looking at the magnitude of the residuals). Figure 8 (which looks suspiciously like Figure 7) is an example of homoscedasticity.

FIGURE 8 Residual Plot Demonstrating Homoscedasticity



Here's what homogeneity of variance is not about. It's not assuming that the residuals are small for the entire range of scores on *X*. It is about having residuals that are of the same magnitude (be it small, medium, or large) for all scores on *X*. See the difference? Homoscedasticity is about a consistent size of the residuals.

FIGURE 9 Residual Plot Demonstrating Heteroscedasticity



Well, then what does a violation of this assumption look like? In Figure 9 we observe such a violation, a condition called heteroscedasticity. The magnitude of the residuals vary by scores on *X*. This *X* variable predicts better (i.e., smaller residuals) for high scores on *X* than it does for low scores on *X*. Big deal, you say. What does this

mean to us? It means that r^2 does not accurately describe the strength of the association between the two variables. In Figure 9, where this assumption is violated, r^2 overstates the strength of association for low scores on *X* and understates it for high scores on *X*. Not good.

Regression Assumptions: Model Specification

A third regression assumption is one that is relevant only for causal, or explanatory, research. The assumption is that the model is correctly specified. Aside from the linear/nonlinear issue from before, a correctly specified model relates to the choice of independent variables in the equation. There are a few ways for a model to be misspecified, but we'll focus on the most serious one: variable omission. As the name suggests, variable omission means that a necessary independent variable has not been included in the analysis. Three conditions must be met for model misspecification via variable omission to yield results that lead to incorrect conclusions. First, the omitted variable(s) must be an actual cause of the dependent variable. Second, the missing variable(s) must be correlated with *X*. Third, *X* must be correlated with *Y*. The consequences of this type of misspecified model are serious: Both r^2 and the regression coefficient for *X* will be will be too high (i.e., overestimated). Thus, *X* will appear to have a greater causal role with *Y* than it actually has. (Note: The effects of this problem are more complicated in multiple regression. As before, r^2 will be overestimated, but the regression coefficients may be too high or too low. Either way, they won't be correct.)

So those are the basic facts of model misspecification. Here's the odd part. It's only an issue for causal research. For predictive research, the model predicts as well as R^2 says it does (aside from one issue which we will discuss in a later chapter). If you left a relevant variable out of the model, that's your problem. You could have obtained better pre-

diction, but you didn't. If you achieved this r^2 by including an apparently irrelevant variable that just happened to work (because it's correlated with the actual cause), good news: It works. So r^2 is accurate for predictive research regardless of how badly you misspecified the model via the choice of independent variables (again, with one exception to be discussed later). With predictive research, one could hold to the attitude that "Prediction is all we care about. Who cares about causality?" and be just fine. (Side note: Understanding the causes of the dependent variable will always lead to better prediction.)

But what of causal research? The whole point of causal research is to understand, well, the causes of the dependent variable. Leaving out a relevant independent variable changes everything. A great example of this error is our ice cream-shark attack example from Chapter 3. In the example, we observed a .7 correlation between ice cream sales and shark attacks at some seaside resort (data collected in monthly intervals). The regression model could be set up with number of shark attacks (Y) as the dependent variable and ice cream sales (X, in tons) as the independent variable. Let's say the regression equation works out to be $Y' = 2X - 1$. For explanatory research the regression coefficient (recall that b indicates the expected change in Y given a one point change in X) is an index of the causal role of its independent variable. In this shark attack example, the regression coefficient is 2. That regression coefficient may look small to you, but all things considered, it's quite large: For every ton of ice cream sold, we expect to observe two shark attacks. If this is causal research, we would conclude that there is a very strong relationship between the two variables, and that shark attacks could be reduced, or even eliminated, if we stopped selling ice cream at this resort town (Amity, Massachusetts). That's the sort of a conclusion one draws from causal research: A recommendation for changes that

should be made with the expectation that changing the status on the independent variable will lead to the desired changes on the dependent variable.

To refresh, a misspecified model is one which is missing necessary variables. And our example is certainly missing a key independent variable. Sure, there is a strong association between shark attacks and ice cream sales, but that association is the by-product of the actual causal factors. Include the actual causal factors (number of people at the resort, in thousands, Z) and everything changes: $Y' = .0000000001X + .001Z + 0$. Thus, for every thousand, thousand people at the resort, we expect one shark attack. But for ice cream, I don't even want to figure out how many tons have to be sold to get one shark attack. The effect of ice cream is now seen to be so small that it's essentially irrelevant. Shark attacks are really related to the number of people in the resort: more people, more swimmers, more opportunities for sharks. Thus, a properly specified model leads to a completely different conclusion. If we had followed the first model, we would have restricted ice cream sales, expecting to see a reduction in shark attacks, and then been very dismayed when the expected results were not observed.

To close our discussion on model misspecification, let's remind ourselves that this doesn't matter a whit for predictive research. A regression model, specified correctly or incorrectly, predicts as well as r^2 says it does (aside from that one issue that I keep dodging). But, for explanatory, or causal, research, a misspecified model can lead to the wrong conclusions regarding which variables cause the dependent variable and the magnitude of their causal role. There are many other things that can be discussed about model misspecification, but let's leave it at this: It is a lot easier to correctly specify a model in a true experiment than with a quasi or non experimental design.

Regression Assumptions: Fixed Independent Variable

Another assumption is that *X*, the independent variable is a fixed variable. A fixed variable is one whose values (e.g., 1, 7, 8) would be observed each time the experiment is repeated. It should be clear that *X* will only be a fixed variable in an experiment where the values for *X* are assigned by the experimenter (e.g., Condition A receives two hours of study time, Condition B received four hours, etc.). As such, they are known before the experiment is conducted. The opposite of a fixed variable is a random variable. A random variable is one whose values are not determined or assigned by the experimenter (e.g., height, IQ). The values of a random variable are not known until the experiment is conducted. Just to be clear, in this context the term random variable does not mean that the variable is composed of random data. It simply means that the variable's values were not set by the experimenter. It's too bad this random variable thing for regression wasn't given a better name like *unfixed variable*. That would reduce the confusion.

In regression, *X* is assumed to be a fixed variable, but *Y* is assumed to be a random variable with a normal distribution. This brings to mind another assumption from correlation that applies to regression: bivariate normality. If *X* is a fixed variable, it is assumed that scores on *Y* will be normally distributed at every level of *X* and that each of these distributions will have equal variance (homoscedasticity).

Back to the *X* variable. The assumption is that *X* is a fixed variable. Based on the definition of a fixed variable, it appears that linear regression cannot be used when *X* is a random variable, which is the case for almost all non experimental research. Is that true? No regression for non experimental research? That's a heck of a restriction. I have good news. As long as the previous assumptions

are met, then OLS regression functions equally well for both random and fixed independent variables. But don't take my word for it. "It has been shown that when other regression assumptions, especially ones concerning model specification, are reasonable met, regression results hold equally for random variables" (Pedhazur & Shmelkin, 1991, p. 392). Conclusion: This fixed independent variable assumption is really just a causal research concern. It is not a mathematical thing, like linearity, but a "Do these results apply to other samples?" issue.

A Brief Discussion of Extrapolation Errors

This fixed variable issue brings to mind a regression analysis issue of which everyone should be aware. A regression equation should never be applied to values of *X* that weren't observed in the original dataset (e.g., the regression equation was developed on a dataset with *X* values ranging from 1 to 5, but now we want *Y′* scores for *X* values ranging from 7 to 12). Such an error is known as an extrapolation error and makes the assumption that the observed trend will hold for values which are beyond the range of those in the original dataset. There is a classic joke (or allegory or parable or something) about this kind of error: An old man had traveled all over America during his life. He noticed that temperatures were colder when he was in the Northern states. He also noticed that temperatures were warmer when he was in the Southern states. He concluded that the North Pole must be the coldest place on earth, and the South Pole must be the... Well, you can guess the rest. Back to our regression equation, if the equation was developed on a sample which had *X* scores that ranged from 1 to 5, it is unknown whether this same equation would also apply when scores on *X* range from 6 to 10. In short, the generalizability of the results is limited to the range of *X* values found in the original sample.

So how does this relate to the fixed variable/random variable issue? Well, not much. But just be careful when X is a random variable as the range of scores in the original sample may not match what is needed for the intended application of the equation. (Of course, this could also happen when X is a fixed variable, but that would be the result of a major error on the part of the experimenter.) A random variable is likely to have a slightly different range of values from sample to sample. Thus, when X is a random variable the results (regression weights, R^2, other stuff) may not translate well from one sample to another. It's not likely to be a problem, but it is something to consider.

Regression Assumptions: Independent Variable Is Measured Without Error

The final assumption is that the independent variable is measured without error. That's going to be a problem when X is not a fixed variable. Question: When is any variable measured without error? Answer: Never. And definitely not when it is a random variable. There will always be measurement errors. These measurement errors in X cause r^2 to be underestimated; they also have less predictable effects on the regression coefficients in multiple regression. Long story short, measurement error causes serious problems in regression.

Regression Assumptions: Summary

To summarize the assumptions of OLS regression, we can state that linearity and homogeneity of variance are always important and should be checked for every regression analysis. Violations of the linearity assumption can be addressed by using nonlinear regression analysis. Model specification and fixed independent variables assumptions apply to explanatory research only. Independent variables being measured without error is a problem for all types of research and violations of this assumption have the effect of lowering r^2.

So that's the giant collection of regression assumptions. To what end? What difference does it make? After all, you'll get results from your regression analysis regardless of whether the assumptions are supported or violated. The answer is that when a given assumption is violated, you'll get results (i.e., r^2, regression equation, significance tests thereof), but the results will not accurately represent reality (r^2 may be overestimated or underestimated; significance test results may be inaccurate; the regression coefficients will be over or underestimated; I could go on, but I don't think I need to). Long story short, assumptions matter. Violate them at your own risk.

Significance Testing in Regression

Significance testing just won't go away. Significance testing in regression is very similar to significance testing in correlation – with a twist. First off, the r_{XY} we obtain from a regression analysis is the same as the r_{XY} we obtain from a correlation analysis (except that it can't be negative). That much should be clear by now. Technically speaking, regression gives us r^2_{XY}, but r_{XY} is just a square root button on the calculator away. As these correlations are the same magnitude, it should not be a surprise that the outcomes of the significance tests are the same. Again, with a twist.

In regression, the significance test we conduct is actually a test of r^2. The equation is as follows.

$$F = \frac{r^2_{XY}/k}{(1 - r^2_{XY})/(N - k - 1)}$$

Where:
r^2_{XY} is the squared correlation between X and Y.

k is the number of independent variables.
N is the sample size.

Because this chapter is a treatment of simple regression, there is only one independent variable, meaning $k = 1$. This test is an F test with k, $N - k - 1$ degrees of freedom. Here's the cool part: The F test of r^2 is identical to the standard t test of r as long as the t test is a two-tailed test. Don't forget that last part: These two significance tests yield the same result if the t test is a two-tailed test. (Note: There is no tailedness to an F test. A clue to this can be found by noticing that, quite obviously, r^2 cannot be negative – a positive relationship between X and Y and a negative relationship between X and Y of the same magnitude will result in the same r^2.) A one-tailed t test will have greater statistical power than the F test (which, as stated, has no directional specificity and is equivalent to a two-tailed t test); thus, the F test of r^2 isn't the proper test to conduct if you have a directional hypothesis (i.e., "There will be positive relationship between X and Y.").

So that was pretty easy. Now the harder part. There are more significance tests in regression than just the test of r^2. There is also a significance test of b, the regression coefficient. This test is a t test, and you'd never guess this, but it yields results identical to the t test of r. Even better, with the test of b we can have positive bs or negative bs, meaning we can test directional hypotheses. Thus, we have the option of one-tailed or two-tailed tests. So here we get results identical to the t test of r. (Important note: Many things will change in the significance testing department when we discuss multiple regression. Some of the things I've just told you will no longer apply. Life will get more complicated. For now, let's enjoy the simplicity of simple regression.)

The test of the regression coefficient in simple regression is as follows.

$$t = \frac{b}{\frac{S_Y}{S_X}\sqrt{\frac{1 - r_{XY}^2}{(N - k - 1)}}}$$

Where:
b is the regression coefficient.
k is the number of independent variables in the model ($k = 1$ in simple regression).

That equation may look intimidating, but it's really just the t test for a correlation all dressed up for regression. (If you want to see how, recall that $b = r_{XY}(S_Y/S_X)$. Make that substitution in this equation, set $k = 1$, and simplify.) That's probably why they yield the same results. In simple regression, a significant t test of r means a significant t test of b.

If you enjoyed all of the discussion of the various significance tests in the previous chapter, you might be wondering exactly what kind of hypothesis we are testing. The answer is the standard "the population value (be it the squared correlation or the regression coefficient) is greater than zero/less than zero/not zero" test. In other words, it's the standard, common, easy significance test. No r to Fisher's z transformations involved.

One last significance test note. Statistical programs such as SPSS and SAS also report a t test for the y-intercept. Such a test is nonsensical, as in, it makes no sense. What do we conclude if the y-intercept is significant? Nothing. What do we conclude if the y-intercept is nonsignificant? Also nothing. Pay no attention to a significance test of a y-intercept.

Understanding Strength of Association in Regression

In this chapter, we've mentioned three ways to assess the strength of association in regression analysis: r, r^2, and b. We know that all three have their own significance tests. But strength of association isn't significance testing. In using these statistics to understand strength of association, all

three statistics have their merits. All have weaknesses as well.

We discussed the regular (i.e., un-squared) correlation in the previous chapter. Cohen's (1992) standards provide useful guidelines for assessing the strength of a regular correlation. We also have discussed the use of the regression coefficient as an index of the strength of association (earlier in this chapter). The regression coefficient is very useful, arguably more useful than a correlation, if the dependent variable is in a meaningful metric, such as time, money, number of accidents, and so on. As discussed earlier, the regression coefficient can be interpreted as follows: For every one point change in scores on X, we expect scores on Y to change by b points. For example, the equation $Y' = 35X + 130$ tells us that for every one point change on X, we expect scores on Y to increase by 35 points. If Y, the dependent variable, refers to days spent working, then this b of 35 has real meaning. We expect someone who scores a 10 on the test to work 35 days longer than someone with a score of 9.

Finally, let's discuss the use of r^2 as an index of the strength of association. As mentioned earlier in this chapter, r^2 has a cool name (coefficient of determination) and definition (percent of variance in Y explained or accounted for by X). So if X and Y are correlated .5, then r^2 is .25. This r^2 means that 25% of the variance in Y is explained by scores on X. Fans of simple math will note that this also means that 75% of the variance is not explained by X. Our .5 correlation (which Cohen describes as "strong") doesn't sound so strong anymore. Them's the breaks when you square numbers between 0 and 1. They get smaller.

It appears that we have a conundrum. A .5 correlation is strong, so says Cohen. But a .5 correlation fails to explain 75% of the variance in Y, so says r^2. A relationship can't be strong and weak at the same time. What's going on? The answer is

hiding in plain sight. The definition of r^2 states that r^2 indicates "the percent of variance in *Y* accounted for by *X*." The word *variance* is where the problem occurs. As we learned in some previous chapter, variance is in squared units (e.g., squared ACT points). Thus, if ACT scores are correlated .5 with GPA, the r^2 method of assessing relationship strength is saying that "differences among squared ACT points explain 25% of the differences in squared GPA points." This is not at all helpful. What we want is a way to understand how two variables are related to each other while retaining the regular metric of measurement (i.e., un-squared points). Brogden (1946) demonstrated that the regular, un-squared correlation is linearly related to how well one variable predicts another. Using our example, a variable correlated .5 with *Y* predicts *Y* half as well as a variable perfectly correlated with *Y*. A variable with a .4 correlation is 40% as efficient at predicting *Y* as is a variable with a 1.0 correlation. And so on.

Before I take any criticism from the gallery for my disdain for r^2 as an index of the strength of association between *X* and *Y*, let me say the following. I understand that predictive efficiency isn't relevant to every discussion regarding strength of association. However, when predictive efficiency is relevant, "percent of variance accounted for" is wholly inappropriate for understanding the strength of association. That said, even for regression analyses not focused on prediction (i.e., causal research), interpreting *r*, instead of r^2, is still the better way to understand strength of relationship. A correlation of .5 is 50% as strong as a perfect relationship. An interpretation of r^2 would lead you to believe that it is only 25% as strong because *X* only accounts for 25% of the variance *Y*.

Conclusion: r^2 may have a cool name and a definition that sounds useful, but it is not the best way to understand how well two variables are related to each other. Stick to statistics that remain in the original metric of measurement: *r* and *b*.

Regression Analysis Summary

Regression analysis extends the concept of correlation and applies it in new ways. A correlation coefficient simply describes the relationship between two variables. Like correlation, regression analysis describes the relationship between two variables. This description can be done with any of three different measures of association: r, r^2, and b. Unlike correlation, regression analysis can be used to predict scores on the dependent variable based on scores on the independent variable. These predictions are made based on the association between X and Y (and the means and standard deviations of both variables), and the accuracy of these predictions depends on the strength of the association between X and Y.

Multiple Regression

5

Multiple fun!

Introduction

The previous chapters were an extensive discourse of bivariate correlation and bivariate linear regression. It was always one *X* variable and one *Y* variable. Many interesting research questions can be explored with just one independent variable and one dependent variable. But why stop there? Why not use, oh I don't know, two independent variables? (Note: We'll always have just one dependent variable.) Maybe we would find a stronger relationship if we used two independent variables. Well, we can do that. And why stop at two? Why not use three? No problem. Or four? Can do. Or five? Slow down. Let's just discuss two independent variables for now.

Multiple Regression Basics

Just for fun, let's take a stroll down memory lane and examine the simple, bivariate linear regression equation.

$$Y' = a + bX$$

How do we turn this into a multiple regression equation, capable of using scores on two independent variables to predict *Y*? We'll just have to add a second *X* to the equation. And, of course, this new variable will need it's own regression weight.

$$Y' = a + b_1X_1 + b_2X_2$$

Just like simple regression, there is just a single *y*-intercept (*a*). So no change there. What's different is that each independent variable gets its own regression coefficient, which we will call a partial regression coefficient. These partial regression coefficients weight each independent variable. Bigger

partial regression coefficients mean greater weights.

Let's explore a multiple regression equation with a data example. A regression of Y on X_1 and X_2 results in the following equation: $Y' = -6 + 6.1X_1 + 2.5X_2$. As you can see from this equation, the coefficients are -6 for a, 6.1 for b_1, and 2.5 for b_2. Listed below are the scores on X_1 and X_2 for this sample. (I have scores on Y too, but we'll keep those hidden for now.) When we apply each person's scores to the regression equation, we compute predicted Y for each person.

Person	X_1	X_2	Y'
John	7	10	61.7
Molly	9	10	73.9
Neil	9	20	98.9
Chris	5	16	64.5
Jordan	6	11	58.1

Thus, computing Y' for each person in multiple regression is just a simple algebraic exercise. It's not much more complicated for equations with more than two independent variables. Just a little more algebra. Five independent variables? No problem. Just a regression equation with an a and five partial regression coefficients. Plug in scores on the five variables and solve. No surprises.

Let's talk about Y' in multiple regression. As with simple regression, Y' represents our prediction of Y based on the scores on the various independent variables. But this is multiple regression, so there's a little more to it. Y' represents a weighted average of the scores on X_1 and X_2 (along with some scaling stuff that's not interesting). The idea of taking an average of the scores on the independent variables makes some sense as averaging allows us to reduce the multiple independent variable scores down to a single score for each person. Many variables (scores on all of the independent variables) become one variable (Y'). Every per-

son in our example dataset has two scores on the independent variables. Reducing these scores down to a single score for each person simplifies matters. But this isn't a simple average score – it's a weighted average. How are they weighted? The regression weights are chosen to obtain the best possible prediction of *Y*. There are an infinite number of possible weights that could be used but only one set that yields the best possible prediction. We refer to these weights as optimal weights. The partial regression coefficients are the weights, and they reflect each variable's unique relationship with the dependent variable (and some other stuff we'll discuss later). Other factors held constant, independent variables with stronger relationships with *Y* get greater partial regression coefficients.

Remember the residual scores from the previous chapters? Residuals were the errors of prediction; they were literally $Y - Y'$ for each person. We can compute residual scores in multiple regression too. Naturally, we'll need to know scores on *Y* to do this. Good thing I have those lying around.

Person	X_1	X_2	Y'	Y	$(Y - Y')$
John	7	10	61.7	42	-19.7
Molly	9	10	73.9	80	6.1
Neil	9	20	98.9	99	0.1
Chris	5	16	64.5	62	-2.5
Jordan	6	11	58.1	75	16.9

Just as with simple regression, bigger residuals mean worse prediction. And there are a couple of big residuals here (John and Jordan).

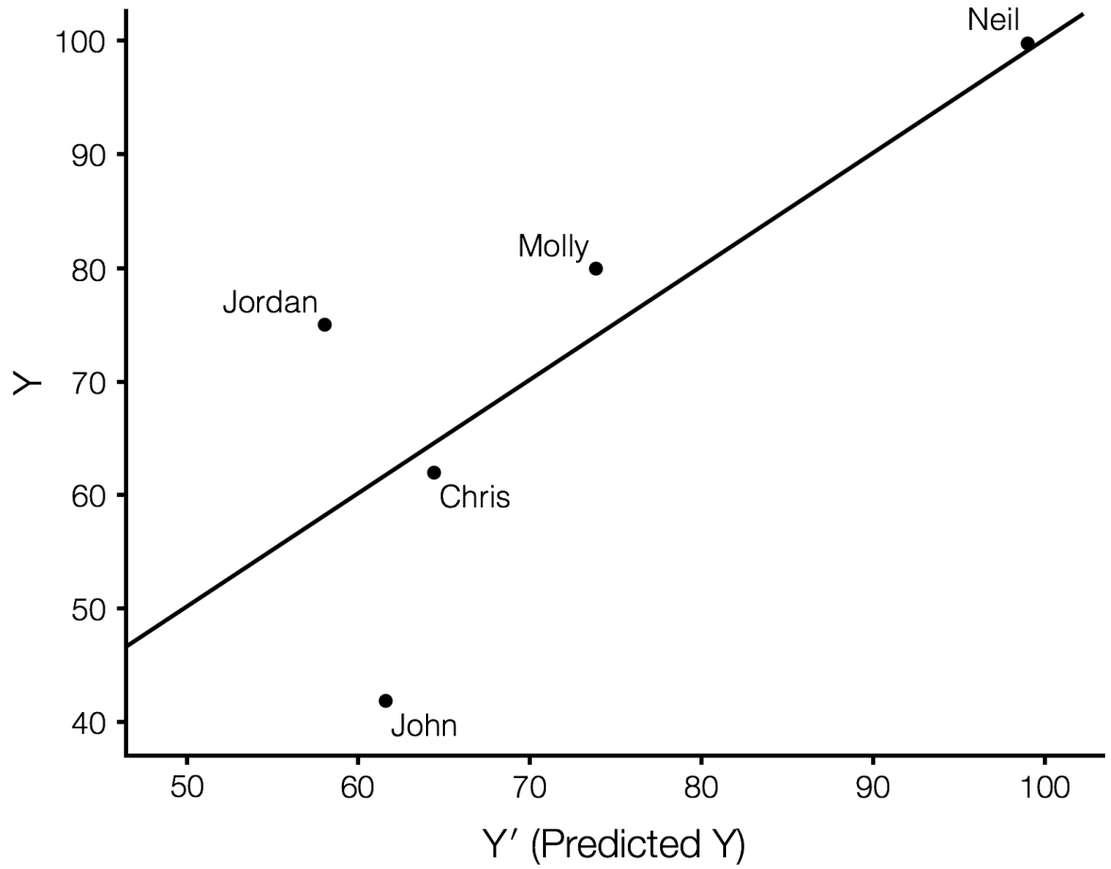
How can we compute a correlation to describe the strength of the association between X_1, X_2, and *Y*? You can't just correlate *Y* with X_1 because that would only tell you how well X_1 predicted *Y*. Same with correlating X_2 with *Y*. We need to know how well X_1 and X_2, when combined, correlate with *Y*. So how do we do this? You already know how. You

may think you don't, but you do. Here it is: We correlate Y with Y'. Why does this work? Well, in the old days of simple regression Y' was a linear transformation of X (recall that $r_{XY} = r_{YY'}$ for simple regression). In multiple regression, we have multiple X variables, meaning that scores on Y' are a weighted average of the scores on X_1 and X_2 for each person. Thus, it is Y' that is actually being used to predict Y, and our correlation between all of the independent variables and Y is actually just a correlation between Y' and Y.

What shall we call this correlation for multiple regression? How about a multiple correlation? Sounds good. The symbol is R (capital R instead of lower case r from bivariate correlation days). R is just like r except that it ranges from 0 to 1. No negative values. For our above dataset, R is .78. It's important to understand subscripts in the multiple correlation coefficient. For our example, the multiple correlation symbol is $R_{YX_1X_2}$. (By the way, is it now obvious that $R_{YX_1X_2} = R_{YY'}$?) For multiple correlations, always list the dependent variable (i.e., Y) first, followed by the independent variables. Sometimes people put a dot between the two, but there's no reason to do that. Finally, double subscripting can get a little tedious, so it's not uncouth to write the previous multiple correlation as R_{Y12}.

So how are we going to graph this? In the days where we had one X and one Y, we graphed those two variables on a two dimensional graph. Here, we have three variables. Do we graph the scores on a three dimensional graph? Well, we could, but what about multiple regression with six X variables and a Y variable? A seven dimensional graph? You've already figured it out. Because we can compute the multiple correlation by correlating Y with Y', where Y' was a weighted average of the scores on the various X variables, we'll graph the results by graphing Y against Y'. Figure 1 is a graph of the scores from the previous dataset.

FIGURE 1 Multiple Regression Scatterplot



As you can see, a scatterplot in multiple regression is nothing more than a graph of Y against Y' for each person. Check for yourself. I've also drawn the regression line so that you can see the accuracy of our predicted Y. As usual, $Y - Y'$ is the error of prediction. Notice that Neil has the smallest error of prediction (0.1). And he's closest to the line. Thus, evaluating strength of association is no different here than it was with simple regression.

Back to our multiple regression equation. Thought question: How should the independent variables be weighted? What factors should we consider in setting the regression weights? The simple answer is to weight them according to how strongly they are associated with Y. Stronger association leads to greater assigned weights. The slightly more complicated answer states that we also have to consider the standard deviations of the variables. So that's two factors: association strength and standard deviation. That's not so bad. I hate to say this at this point, but even the prediction quality part is more complicated than it appears. To explain why, we'll take a slight detour.

Multiple Regression with Two Uncorrelated Independent Variables

Consider the following scenario. Two variables are used to predict Y. The bivariate correlations of X_1 and X_2 with Y are both .5. Because we have three variables (X_1, X_2, and Y), we must also consider the correlation between X_1 and X_2, the two independent variables. In this scenario, we'll set that correlation to 0.0. To summarize, X_1 and X_2 are both strongly related to Y and are unrelated to each other. We'll explore two things: the multiple correlation and the regression equation. The regression equation describes how these two variables are combined to predict Y, and the multiple correlation tells us how strongly they are related to Y. First, the multiple correlation. It's easy to see how well each variable predicts Y on its own – the bivariate correlations tell us that ($r = .5$ for each in this case). We need a way to assess how well X_1 and X_2 predict Y when the two variables are used in combination.

When the independent variables are uncorrelated with each other (i.e., $r_{X_1X_2} = 0$), the squared multiple correlation can be computed with the following equation.

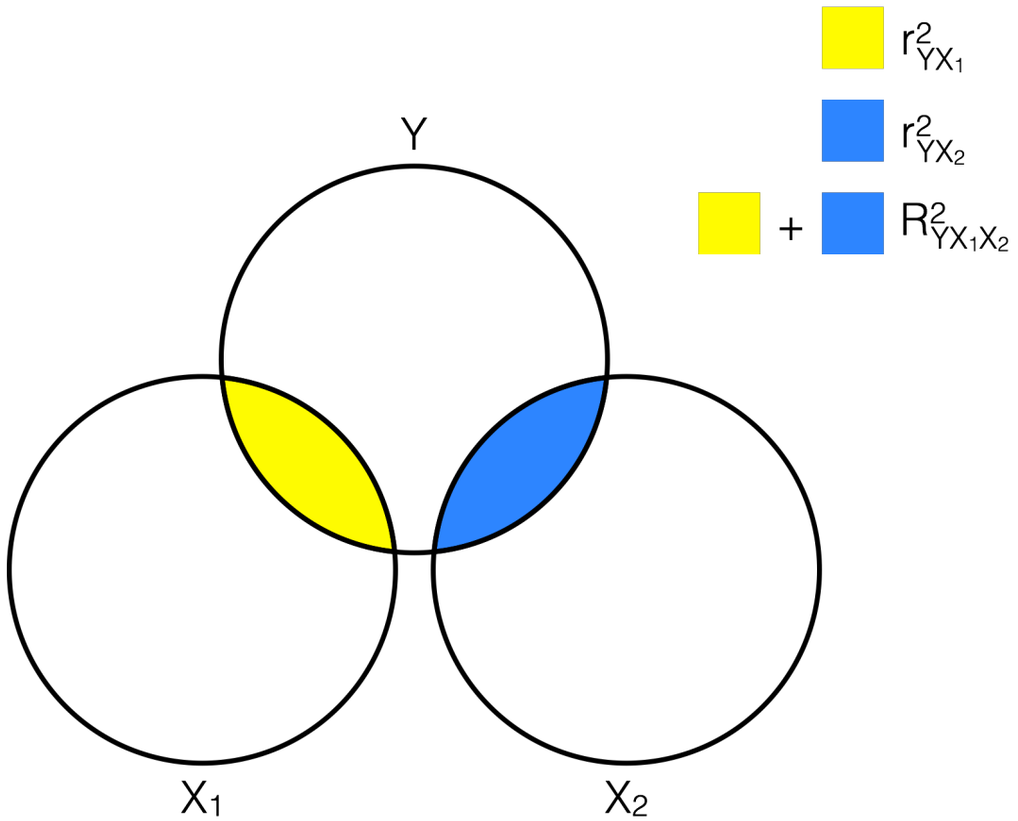
$$R^2_{YX_1X_2} = r^2_{YX_1} + r^2_{YX_2}$$

Thus, the squared multiple correlation is the simple sum of the squared bivariate correlations. Figure 2 is a illustration of this principle.

Applying this equation to our example shows that $R_{YX_1X_2} = .71$ ($R^2_{YX_1X_2} = .5^2 + .5^2 = .50$; take the square root to obtain $R_{YX_1X_2}$). Now let's think about these results. We can use either independent variable by itself and obtain a correlation of .5, or we can use both of them in combination and obtain a multiple correlation of .71. That second variable sounds pretty useful.

As for the regression equation, we'll keep this simple and use standardized variables so that all of our variables have means of zero and variances

FIGURE 2 Multiple Correlation with Uncorrelated Independent Variables



When the independent variables are uncorrelated the squared multiple correlation is the sum of the squared bivariate correlations.

of one. For our present example, the regression equation is $z_{Y'} = .5z_{X1} + .5z_{X2}$. (If you're wondering where the *a* went, there is no *y*-intercept in standardized regression.)

Multiple Regression with Two Correlated Independent Variables

Everything gets worse when the independent variables are correlated with each other (i.e., $r_{X_1X_2} \neq 0$). So let's all agree to make our lives easier and never use correlated independent variables. Deal? Well, maybe we promise a bit too much. Unless we're conducting research with random assignment to conditions (i.e., a true experiment), independent variables have a nasty habit of being correlated. So we're going to have to deal with it.

To compute the multiple correlation, we'll need a new equation, something that takes into account the relationship between the two independent variables. And here it is:

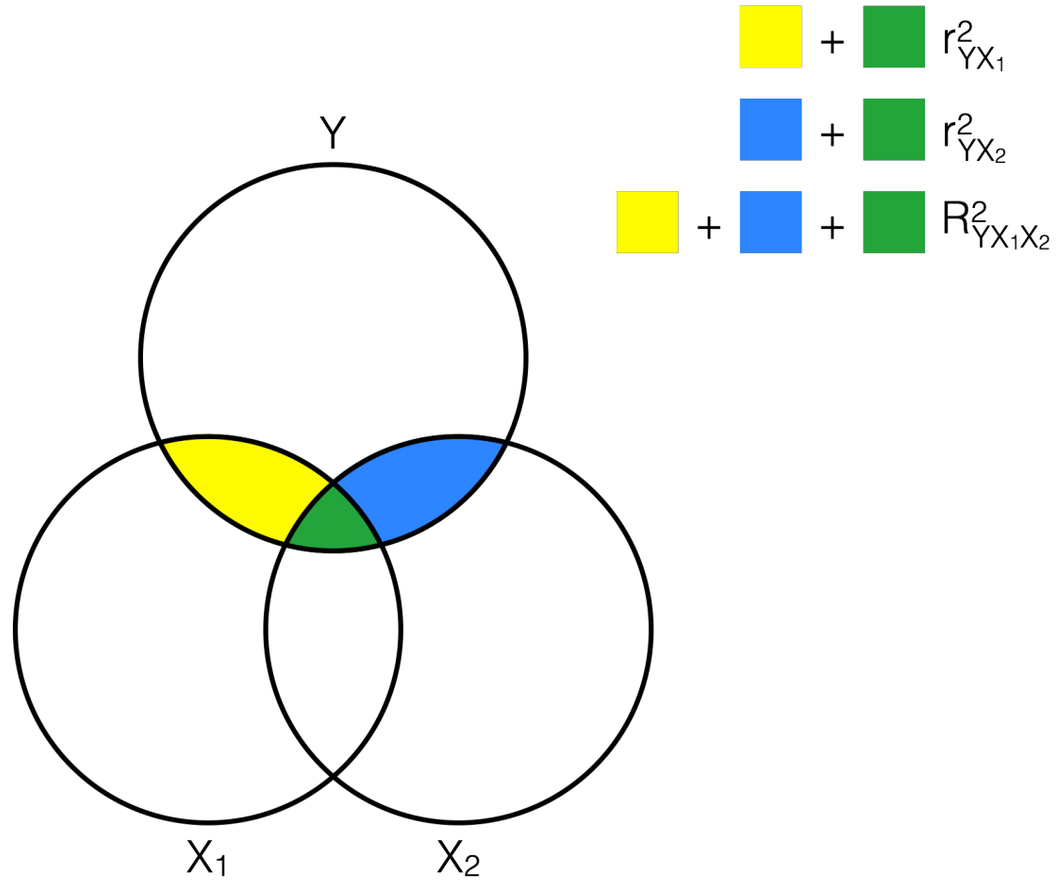
$$R^2_{YX_1X_2} = \frac{r^2_{YX_1} + r^2_{YX_2} - 2r_{YX_1} \cdot r_{YX_2} \cdot r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

Greater correlations between the independent variables diminish the usefulness of a second inde-

pendent variable. A quick example will demonstrate this principle. As before, each variable has a .5 correlation with Y (i.e., $r_{YX_1} = .5$, $r_{YX_2} = .5$). But this time the two independent variables are correlated with each other: $r_{X_1X_2} = .4$. Inserting these values into the above equation (and taking the square root to obtain $R_{YX_1X_2}$) yields an $R_{YX_1X_2}$ of .60. Figure 3 illustrates the concept of multiple correlation with correlated independent variables. Notice how the portion of Y predicted by both X_1 and X_2 (green area) can be counted only once in the multiple correlation.

Looking at these results logically, we can use either X_1 or X_2 to predict Y with a correlation of .5, or we can use X_1 and X_2 together and obtain a multiple correlation of .60. There isn't as much added value in using that second variable when the independent variables are related to each other. Compare this R^2 of .60 to the previous example with uncorrelated independent variables which yielded an R^2 of .71. At some point, the second variable

FIGURE 3 Multiple Correlation with Correlated Independent Variables



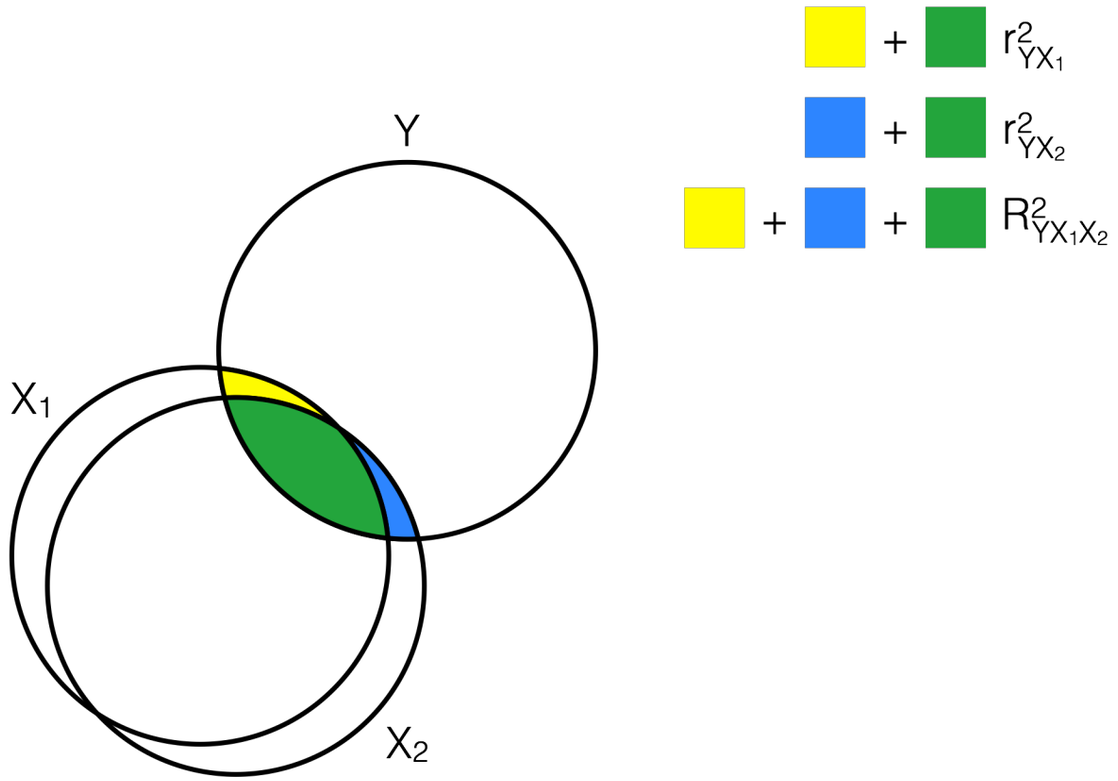
The squared multiple correlation is the sum of the squared bivariate correlations with the qualification that variance in Y related to both X_1 and X_2 (green area) is counted only once.

may not be worth using at all. If you're bored on a rainy day and want to see the effects of intercorrelated independent variables, play around with the previous equation, varying the values of $R_{YX_1X_2}$. It's

actually kind of fun. (When you do this, be sure to consult **McNemar's triangular inequality** to evaluate the permissibility of the values you use. I've caused myself some trouble by forgetting to do just that. Also, check out this other **cool correlation equation**.)

Let's pause here and rehash the principal principle: other things being equal, lower correlations among independent variables are better; they allow for a better multiple correlation with Y. A rather extreme example of this principle is given in Figure 4. Note that in Figure 4, both X_1 and X_2 have the same bivariate correlation with Y as in Figure 2. The obvious difference between the two is that in Figure 2, $r_{X_1X_2} = 0$; whereas, in Figure 4, $r_{X_1X_2}$ is strong (let's just make up a number and say .8). Notice that far more of the area of Y is covered by the two independent variables in the first graph than in the second. Here's another way to look at it: Given the strength of R^2 obtained by use of X_1 alone, what does X_2 offer in each case? In Figure 2, X_2 contributes as much to R^2 as X_1, whereas in Figure 4, X_2 offers next to nothing that we didn't already have from X_1. Yet another thought question: What would the Venn diagram look like if $r_{X_1X_2} = 1.0$?

FIGURE 4 The Effect of Independent Variable Intercorrelation on Multiple Correlation



Highly correlated independent variables: Each variable correlates well with Y on its own. Because the independent variable are highly correlated with each other, they do not combine to make a strong multiple correlation.

What about the regression equation? Sticking with our practice of making all things easier by using standardized data (all means equal zero, all variances equal one), the equation for a standardized partial regression coefficient looks like this:

$$B_1 = \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

Where:
B_1 is the standardized partial regression coefficient for X_1.

Thus, the partial regression coefficients are affected by the same factors as the multiple correlation. Increased correlations between the independent variables results in decreased partial regression coefficients. Using our example from this section ($r_{YX_1} = .5$, $r_{YX_2} = .5$, $r_{X_1X_2} = .4$), the standardized regression equation is now $z_{Y'} = .36z_{X1} + .36z_{X2}$. Note the reduced regression coefficients as compared to the uncorrelated independent variables example where the standardized partial regression coefficients were equal to the bivariate correlations. In the present case each variable is weighted less because each variable's unique contribution to the prediction of Y is now less. Thus, greater correlations between independent variables lead to reduced partial regression coefficients.

You may want to know how to compute the partial regression coefficients and y-intercept when dealing with unstandardized data (i.e., raw scores). It's surprisingly easy. Of course, it's even easier if you just let a computer do it for you. Just combine the previous equation with what we know from the bivariate regression chapter. The following formula computes the unstandardized partial regression coefficient for a two-independent variable equation.

$$b_1 = \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r^2_{X_1X_2}} \cdot \frac{S_Y}{S_{X_1}}$$

Where:
b_1 is the unstandardized partial regression coefficient for
X_1.

(Note the symbols for regression coefficients: B_k for the standardized regression coefficient and b_k the for unstandardized regression coefficient.)

As for the *y*-intercept in multiple regression, it's equation bears a remarkable resemblance to the bivariate regression formula:

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

Multiple Regression with Any Number of Independent Variables

When multiple regression is used with three or more independent variables the principles are all the same, just mathematically more complicated. Computers make great tools for those analyses. Actually, they make great tools for simple regression too. The regression equation look like this:

$$Y' = a + b_1X_1 + b_2X_2 + \ldots b_kX_k$$

Where:
b_k is the partial regression coefficient for X_k.

Any number of variables you want. Same general form of the regression equation. What about the multiple correlation and the equations for the regression coefficients and *y*-intercept? Remember those computers we talked about? They sure come in handy.

There is one question you may be asking yourself. And that is, "Does the order of the independent variables in a regression equation matter?" That is, does changing the order of the variables in the regression equation change the regression coefficients for those variables? Let's answer that question by reminding ourselves that the regression weights are assigned to the independent variables

so as to maximize the prediction of Y. Other things being equal, the best predictors are assigned the greatest weights. Knowing this, why would the order in which the variables are listed affect the weights? It wouldn't. So the answer to the original question is no. Y regressed on X_1, X_2, and X_3 (in that order) will result in the same regression weights assigned to those variables as will Y regressed on X_3, X_1, and X_2 (in that order).

Significance Tests in Multiple Regression

Significance testing in multiple regression is similar to significance testing in simple regression: We can test R^2 with an F test, and we can test the regression coefficients with a t test. We have multiple independent variables, each with their own regression coefficient, so we'll have t tests for each of these regression coefficients. These multiple t tests allow us to see which independent variables have a significant association with the dependent variable. We'll explore this issue further in future chapters. As for actually computing the the t test, below is the equation for the two-independent variable situation (set up for X_1); any more variables and it's too complicated to deal with by hand. And that's why we have computers.

$$t = \frac{b_1}{\frac{S_Y}{S_X}\sqrt{\frac{1 - R^2_{YX_1X_2}}{(1 - r^2_{X_1X_2}) \cdot (N - k - 1)}}}$$

As for the F test of R^2, almost everything is the same as with the days of simple regression. It's still a test of how well the independent variables, combined as per the regression equation, predict Y. The equation for the F test is the same. I'll list it again, customized with multiple regression symbolology, just for the memories.

$$F = \frac{R^2_{YX_1 \ldots X_k}/k}{(1 - R^2_{YX_1 \ldots X_k})/(N - k - 1)}$$

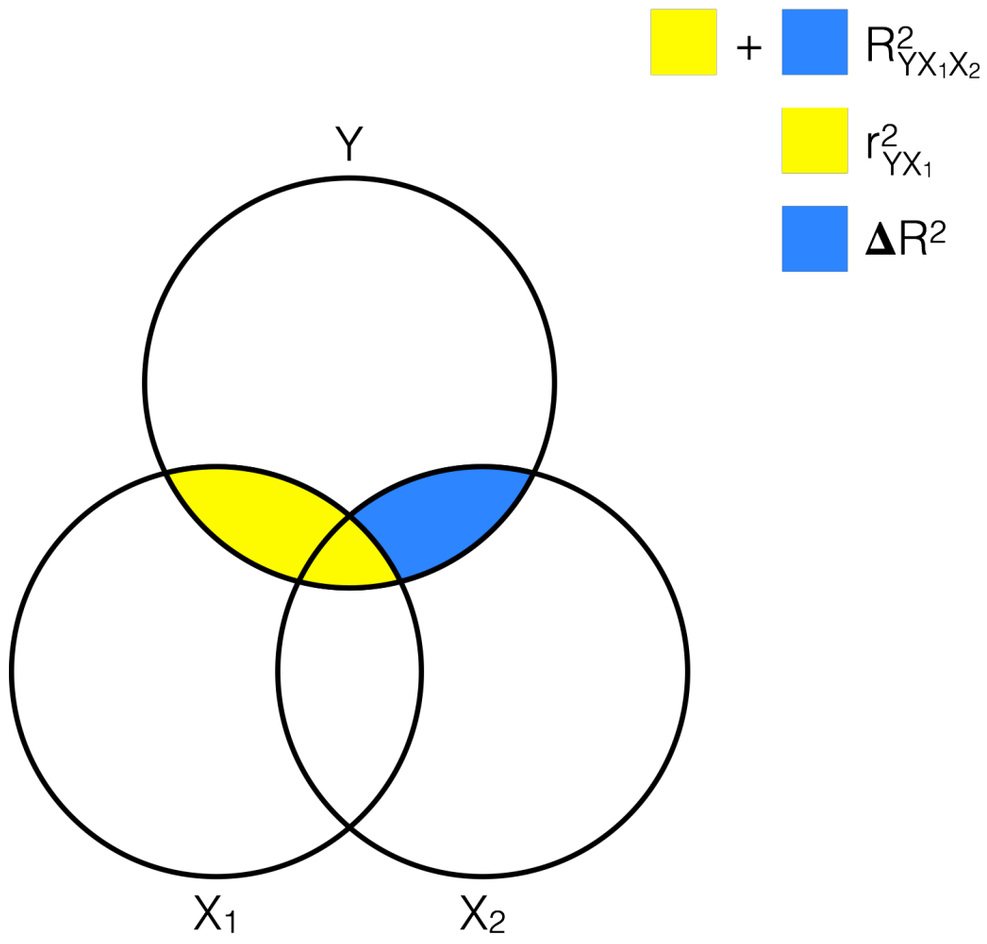
As with the F test of R^2 in simple regression, degrees of freedom are $k, N - k - 1$.

Significance Tests for Changes in R^2

There is another significance test worth discussing in multiple regression. It's a derivative of the F test of R^2. Rather than a test of whether R^2 is significant, this test determines whether the change in R^2 due to the addition of an independent variable to the model is significant. This test is for a specific situation and won't occur on accident. The situation is as follows: We have a regression equation with a certain number of independent variables (we'll use just one for this example). Having conducted our regression analysis on this model (Y regressed on X_1), we know R^2, the F test of R^2, the regression equation, and the t tests of the regression coefficients. In short, we know everything about this model. But then we get the idea of adding a new variable (or two or twenty, but let's just say one for now) to the model. Now we're regressing Y on X_1 and X_2. We have two questions. How much did R^2 change, and is this change significant? Answering the first question is easy; we simply compute the difference between the R^2 from the bigger model and the R^2 from the smaller model (in this case, $\Delta R^2 = R^2_{YX_1X_2} - R^2_{YX_1}$). If the addition of X_2 to the model did not raise R^2 at all, then $\Delta R^2 = 0$. If the addition of X_2 to the model did raise R^2, then ΔR^2 will be some number greater than zero. This change in R^2 can be defined as (a) the percent of variance in Y explained by the addition of X_2 to the model, (b) the percent of variance in Y that is explained by X_2 beyond what was already explained by X_1, and (c) the unique contribution of X_2 to the prediction of Y. All of these definitions are pretty much the same, but sometimes one is more helpful than the other. Figure 5 is an illustration of ΔR^2.

Now that we know how much R^2 changed, we need to know if this change is significant. There are bunches of variables we could add to the equation that would raise R^2 by a trivial amount (i.e., where the change in R^2 is so small that it could be expected to be purely a product of sampling er-

FIGURE 5 ΔR^2 Illustrated



Computing the change in R^2 involves conducting two regression analyses. The first is the regression of Y on X_1 (yellow area). The second is the regression of Y on X_1 and X_2 (yellow plus blue area). The difference between the two R^2 values is ΔR^2 (blue area).

ror). So we'll test whether ΔR^2 is big enough in our sample to allow us to conclude that the change in R^2 in the population is greater than zero. The equation to test for this is a sort of an expanded version of the F test of R^2.

$$F = \frac{(R^2_{big} - R^2_{small})/(k_{big} - k_{small})}{(1 - R^2_{big})/(N - k_{big} - 1)}$$

Where:

R^2_{big} is the R^2 from the larger model (i.e., the regression equation with more independent variables).

R^2_{small} is the R^2 from the smaller model (i.e., the regression equation with fewer independent variables).

k_{big} is the number of independent variables in the larger model.

k_{small} is the number of independent variables in the larger model.

$k_{big} - k_{small}, N - k_{big} - 1$ are the degrees of freedom.

Aside from the formulae, how is this ΔR^2 F test different from the the standard F test of R^2?

The answer is found in what they are testing. The standard F test of R^2 is used to test how well the collection of variables in the regression equation, with their assigned regression weights, are associated with Y in the population (i.e., how well they predict Y). To use an example with four independent variables, it's a test of whether $R^2_{YX_1X_2X_3X_4}$ is greater than zero in the population. In contrast, the ΔR^2 F test is a test of whether the increase in R^2 associated with adding variables to an equation is greater than zero in the population. If our four-variable example is compared to a smaller model with only X_1 and X_2, it's a test of whether $R^2_{YX_1X_2X_3X_4} - R^2_{YX_1X_2}$ is greater than zero in the population. It's easy to confuse these two tests, so be careful. Just remember what you want to test and choose accordingly.

One last note on the ΔR^2 F test. If only one independent variable is added to the model, the ΔR^2 F test yields the same result at the t test for the partial regression coefficient associated with that variable (assuming a two-tailed t test). Not only is this nugget of information a potential labor saving shortcut, it also reveals what is going on with the t test and the partial regression coefficient: They are both based on that variable's unique contribution to the prediction of Y.

Closing Thoughts on Multiple Regression

We have discussed a few concepts in this chapters. First, we can use any number of variables to predict Y. Scores on these variables are combined to form a weighted average. We call scores formed by this weighted average predicted Y (i.e., Y'). The regression coefficients, now called partial regression coefficients, are the weights applied to the various independent variables to get the best possible prediction of Y. The multiple correlation describes how well this combination of independent variables predicts Y. The magnitude of the multiple correlation depends on how well the individual variables predict Y as well as the relationship be-

tween the independent variables. The multiple correlation is identical to the correlation between Y' and Y. Finally, the partial regression coefficient for a given variable is a function of that variable's correlation with Y as well as that variable's correlation with the other independent variables (and some relatively uninteresting standard deviation stuff).

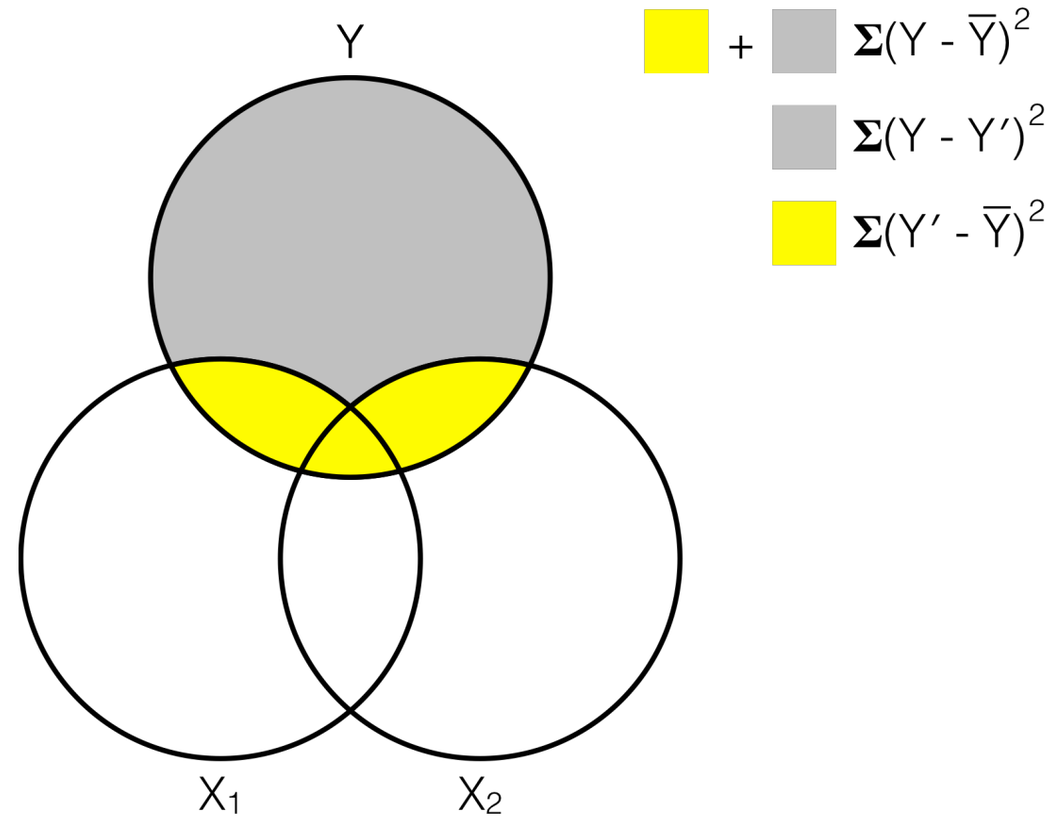
Let's see, I think that about wraps it up for multiple regression. Wait, what about assumptions? I have good news. The assumptions for multiple regression are the same as they were for simple regression. And I already showed you how to check for violations of the linearity and homogeneity of variance assumptions in a way that works for simple and multiple regression (i.e., the residual plot is a graph of residual scores on the y-axis and Y' scores on the x-axis). The other assumptions apply too. No graphs to check for them, though.

One last multiple regression issue concerns the partial regression coefficient. How do we interpret the meaning of a partial regression coefficient? Determining the meaning of a partial regression coefficient is similar to interpreting a regression coefficient in simple regression. In the case of simple regression, b indicated the expected change in Y given a one point change in X. In multiple regression, b_1 indicates the expected change in Y given a one point change in X_1, *holding the other X variables constant*. It's that last part that's new. It's also that last part that can cause problems. If X_1 and X_2 are correlated variables (e.g., height and weight), then it may not be possible to manipulate X_1 without also having scores on X_2 change as well. Thus, the net change in Y may be more than what b_1 indicates. Not so bad, you say? Well, what if b_2 is negative? The change in Y you expect to obtain by increasing X_1 is offset by a decrease in Y of b_2 for every point X_2 changes. Your actual increase

in Y may turn out to be a whole lot less than what an examination of b_1 led you to believe.

What's going on in Figure 6, you ask? Oh, just another way to think about multiple correlation using Venn diagrams. Nothing new here. Just pulling a few concepts from our first discussion of regression analysis (regression and residual sums of squares) to illustrate the multiple correlation.

FIGURE 6 Gratuitous Use of Venn Diagrams to Illustrate Regression Concepts



As with simple regression, the variability of Y (represented in the diagram as the total area of Y; statistically it is $\Sigma(Y - \bar{Y})^2$, the sum of squares of Y) can be divided into a residual component (gray area) and a regression component (yellow area). Dividing each these terms by $\Sigma(Y - \bar{Y})^2$ yields $1 - R^2$ for the residual component (gray) and R^2 for the regression component (yellow).

Partial & Semipartial Correlation

Real magic.

6

Overview

Determining the causes of a given variable (e.g., success in school) is tricky business. The reason for this difficulty is that there are an infinite number of potential causes. We need a way to eliminate the irrelevant ones so that we can identify and assess the impact of the actual causes. This chapter addresses two statistical methods for controlling for irrelevant variables, something that helps both causal and predictive research.

There are many ways to control for the influence of a variable. The best method is to randomly assign people to groups. With random assignment, the groups are likely close to equal on every conceivable variable, measured or unmeasured. (Random assignment to groups is one of the key features of the true experiment. The other is the exercise of experimental control, where there experimenter treats the groups the same except for the variables the experimenter want to manipulate. True experiments are great because if done well, they solve so many of the problems we face with causal research.)

If random assignment isn't possible, and it frequently isn't (Randomly assign people to various heights?), there are other ways to control for variables. One method is matching, which involves matching subjects on one or more variables. A strict matching method would work as follows: for every person with a low score in one group, there is a person with the same low score on that variable in the other group (cases without matching scores are removed). Thus, whatever effect that variable (i.e., the matched variable) has on the dependent variable, the effects are the same for both groups (e.g., matching smokers and non smokers on family history for cancer). The problem with matching as a means of control is that only the matched variables are controlled. It is always possible (highly probable, actually) that an unmatched

variable is the actual cause of the dependent variable.

Another method of controlling for the effects of a variable is with statistical control. Statistical control is a bit like matching, but with a lot less hassle. With statistical control all of the data is used, but the controlling is done with some statistical magic. Statistical control can be done with a variety of methods: partial correlation, semipartial correlation, analysis of covariance (ANCOVA), and more. We'll discuss partial and semipartial correlation in this chapter. ANCOVA will be discussed in Chapter 11.

Partial and semipartial correlation may just be the coolest thing ever invented in the entire world. With partial and semipartial correlation, we can examine the relationship between two variables while controlling for the relationship they have with a third variable. *Controlling for the relationship with* means *removing any association with* this third variable. Stated yet another way, we can assess the relationship between two variables independent from the effects of a third variable.

Statistical Control Example

An example may help illustrate the nature of statistical control. Let's say we want to investigate whether grade school student achievement test scores are associated with the time a student spends one-on-one with a teacher or teaching assistant. We think that classes with more teaching assistants are able to give more one-on-one time with students, leading to improvements in student performance. We sample students from classes in a couple of school districts, measure their scores on both variables, compute a bivariate correlation, and find a strong correlation between the two variables. Sure enough, achievement test scores are positively associated with solo instruction time with a teacher or teaching assistant ($r_{XY} = .55$). (Important note: All numbers in this example are

fictitious. No real research was consulted in the course of creating this example. It's more fun this way.)

Before we start making recommendations to people, school boards, state agencies, people on the internet named DrWhoSuperfan9000, etc., it occurs to us that there may be other variables at work. We consider the variable of student intelligence. Of course, we know that there is a relationship between student intelligence and student achievement ($r_{ZY} = .70$). Could intelligence also be associated with the number of teaching assistants per class, which then makes possible more solo instruction time? There is no reason to think it is, but students were not randomly assigned to condition, so we can't rule it out. It turns out that one of the school districts is located next to a major university and many of their students are children of the university professors. Sure enough, this same school district also has a policy of encouraging parents to volunteer as teaching assistants. We compute the correlation between student intelligence (measured at the start of the year) and solo instruction time and find a correlation of .60.

To summarize the situation at this point, more solo instruction time is associated with greater student achievement test scores ($r_{XY} = .55$). But greater intelligence is also associated with greater achievement test scores ($r_{ZY} = .70$) and with more one-on-one instruction time ($r_{XZ} = .60$). We can't just ignore these intelligence test correlations. If only we had randomly assigned the students to classrooms with varying numbers of teaching assistants, then the intelligence-solo instruction time correlation would be zero (or close to it) and would not be a problem. We need to remove the relationship that intelligence has with our variables. We could try matching, but that's a lot of trouble and will cost us data (unmatchable cases will have to be removed). What we'll do is statistically remove any association intelligence has with the other two variables. Once that is done, we'll re-

assess the relationship between achievement and number of assistants per class.

So that's the setup. How do we statistically control for a variable? That is, how do we remove any association a variable has with another variable? You actually already know the answer. Remember the definition of residual scores? It's the part of *Y* unrelated to *X*. Are you seeing it? To remove any association between *Y* and *X*, regress *Y* on *X* and compute residual scores for everyone in the dataset. These residual scores are the part of *Y* unrelated to *X*. Scores on *Y* have been divided into two parts: a part related to *X* (predicted *Y*) and a part unrelated to *X* (residual). There is, very literally, less of *Y* in the residual scores. If you compute the variance of scores on *Y* and compute the variance of residual scores you will find that the residual scores have less variance. Any association *Y* had with *X* has been removed from *Y*. That's the genius of statistical control. There are actually a few ways to do it. We'll discuss two of them in this chapter. First up, partial correlation.

Partial Correlation

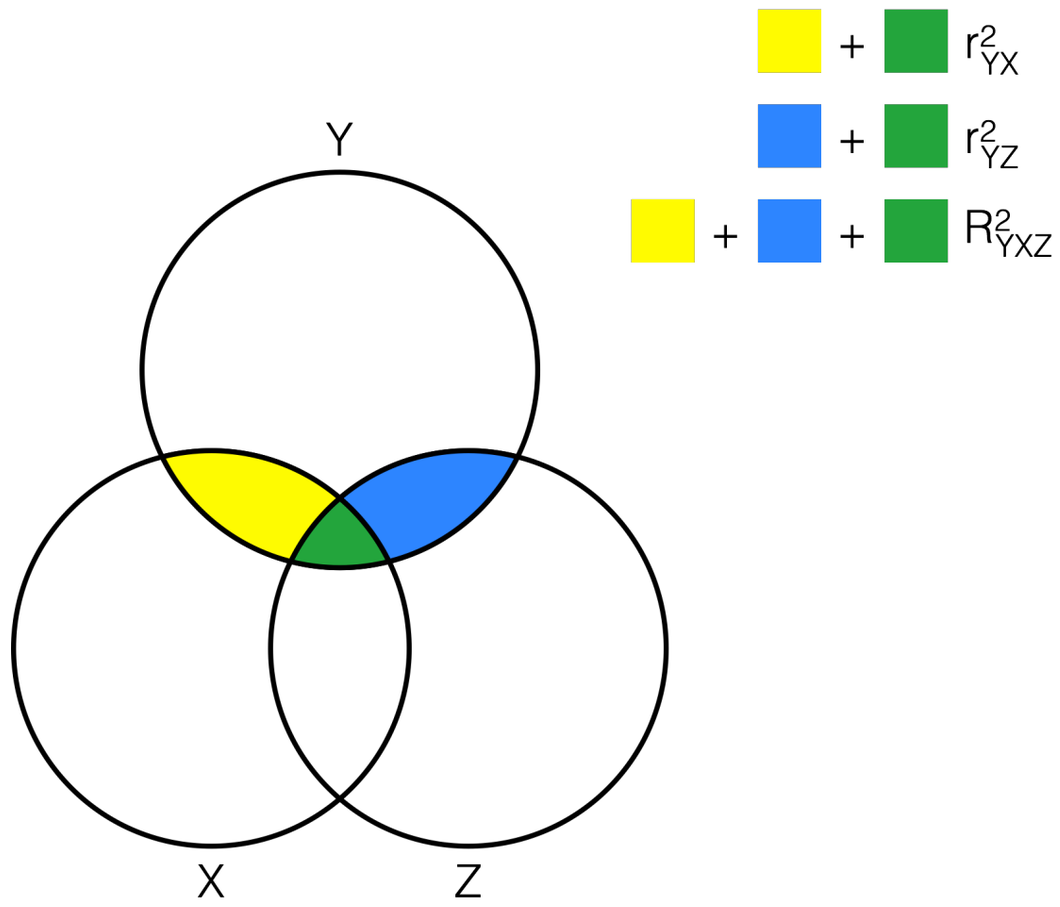
We'll explain partial correlation using the simplest data scenario possible, three variables. *Y* is the dependent variable. *X* is the independent variable. *Z* is the control variable. A control variable is the variable, well, that you want to control. We want to remove any association the other variables have with the control variable. To reflect back on our example, intelligence is the control variable. The goal of the study is to see if achievement test scores (*Y*) are related the solo instruction time (*X*). We want to know this correlation independent of the relationship that each variable has with intelligence. A partial correlation is just that – a correlation between *X* and *Y* independent of with any association that either variable has with *Z*.

Computing a partial correlation is just a few easy steps. Because we want to remove the association that both *X* and *Y* have with *Z*, we need to residualize both variables on *Z*. Yes, I made up a new word.

> *Residualize* (verb): To regress a variable on another variable and compute the residual scores for all people in the dataset.

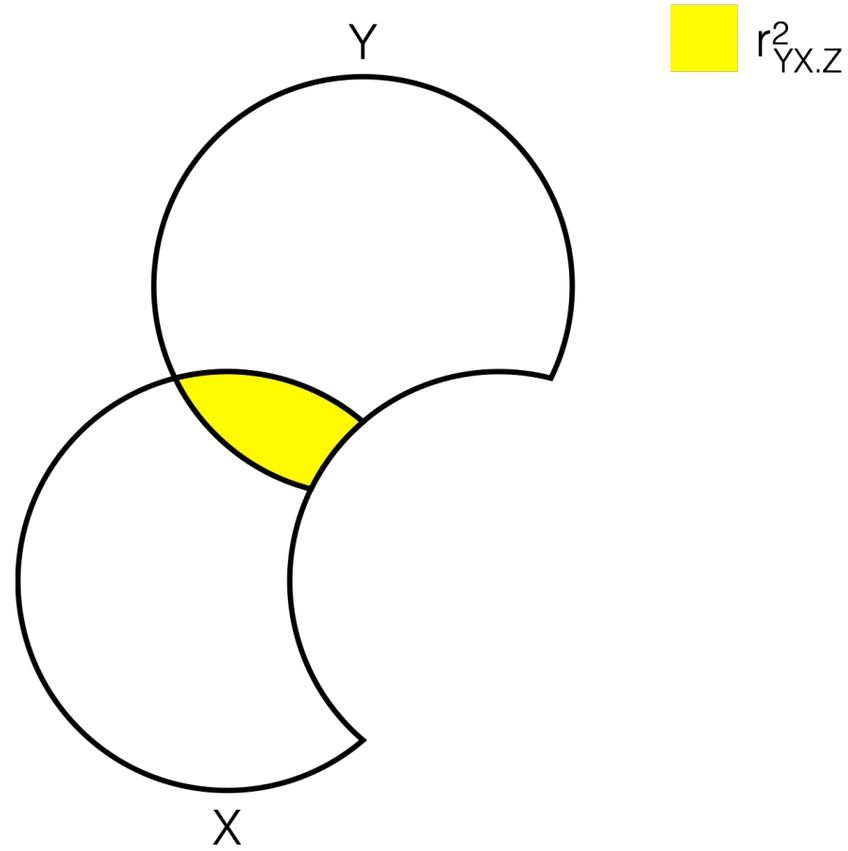
So the steps are as follows: (a) regress *Y* on *Z* and compute the residual scores, (b) regress *X* on *Z* and compute the residual scores, and (c) correlate the residual scores. When we correlate the residual scores, we are correlating the part of *Y* unrelated to *Z* with the part of *X* unrelated to *Z*. Thus, the resultant correlation, called a partial correlation, reflects the relationship between *X* and *Y* independent of *Z*. The symbol for partial correlation is $r_{YX.Z}$. Note that this looks like a multiple correlation symbol with the addition of a dot. Everything listed after the dot is a control variable.

FIGURE 1 Relationship Between *Y*, *X*, and *Z*



An illustration of the squared partial correlation is shown in Figure 1 and Figure 2. As always, the percent of *Y* covered by the independent variable(s) represents the squared correlation. Note how in Figure 2 the removal of any association with *Z* leads to very literally less of *X* and *Y*. Both variables have reduced variance. All that is left for

FIGURE 2 Partial Correlation Between Y and X Controlling for Z



each is the part that is unrelated to Z (i.e., the residual from the regressions upon Z). The overlap between the remaining parts of X and Y indicates the relationship between the two.

Back to our example. The partial correlation between Y and X, controlling for Z, is .23. This is considerably less than the zero-order bivariate correlation between X and Y ($r_{XY} = .55$). To summarize, the correlation between achievement test scores and solo instruction time appears to be strong, but after controlling for student intelligence, the association is rather weak.

A comparison of Figure 1 and Figure 2 illustrates the difference between a multiple correlation between Y, X, and Z (i.e., R_{YXZ}) and a partial correlation between Y and X, controlling for Z (i.e., $r_{YX.Z}$). The former describes the association Y has with Z and X. The latter describes the association between X and Y independent of any relationship with Z.

A quick word on partial correlation terminology. A correlation where we do not control for any variables is called a zero-order correlation. A correlation where we control for one variable is called a first-order correlation. You can probably figure out the rest. This zero-order, first-order terminology

also applies to our next item, the semipartial correlation.

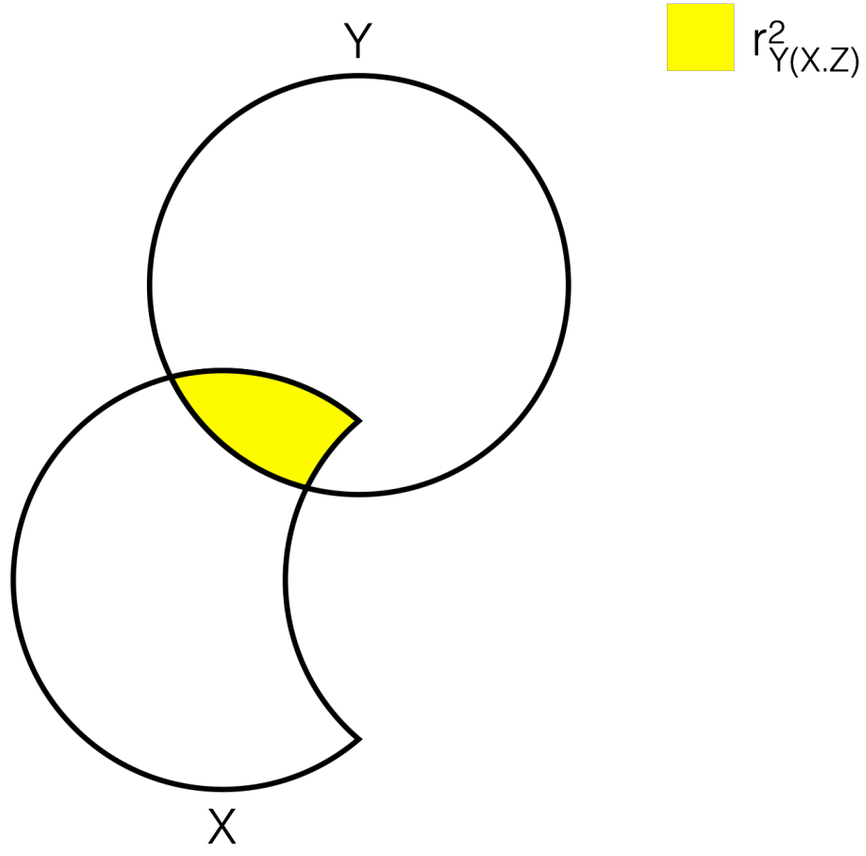
Semipartial Correlation

The other kind of correlation involved in statistical control is the semipartial correlation (also called a *part* correlation, but that's a terrible name). The semipartial correlation is similar to the partial correlation. The only difference is that with the semipartial correlation, the association with the control variable is removed from the independent variable only. In a sense, the semipartial correlation is sort of halfway to a partial correlation. Maybe that's how it got its name.

The steps to compute a semipartial correlation are as follows: (a) regress *X* on *Z* and compute the residual scores and (b) correlate with *Y* with the residualized *X*. That's it. Notice that only *X* is residualized; *Y* is untouched. Thus, the part of *X* unrelated to *Z* is correlated with *Y*. An illustration of the semipartial correlation is given in Figure 3.

How does the semipartial correlation compare to the partial correlation? Semipartial correlations are always less than or equal to partial correlations (assuming correlations greater than zero). In our example, where the partial correlation is .23, the semipartial correlation between *X* and *Y*, controlling for *Z*, is .16. Thus, the semipartial is less than the partial correlation. This reduction in magnitude is demonstrated in the diagrams of each. As a reminder, the magnitude of a squared correlation in a Venn diagram is indicated by the percent of area of *Y* (a visual representation of the variance of *Y*) that is covered by the independent variable(s). Comparing the same data for partial and semipartial correlations (e.g., Figure 2 and Figure 3), there is the same amount of *X* area covering *Y*. But with the partial correlation, the total area of *Y* has been reduced; there is less of *Y* that can be explained, but we're explaining the same absolute

FIGURE 3 Semipartial Correlation Between Y and X Controlling for Z



amount. Thus, the ratio of explained variance to total variance (i.e., r^2) is greater for the partial correlation.

And finally, the symbol for the semipartial correlation is $r_{Y(X.Z)}$, and is similar to the partial correlation symbol; the only difference are the parentheses around the independent and control variables.

Computing Partial and Semipartial Correlations with Bivariate Correlations

As much fun as residualizing is, and it's pretty great, there are other ways to compute partial and semipartial correlations. Both the partial and semipartial correlations can be computed with bivariate correlations. The equations below are designed for the three variable case. If you have more than three variables, you better learn to love residualizing.

The partial correlation between X and Y, controlling for Z, can be computed with the following equation.

$$r_{YX.Z} = \frac{r_{YX} - r_{YZ} \cdot r_{XZ}}{\sqrt{1 - r^2_{YZ}}\sqrt{1 - r^2_{XZ}}}$$

The semipartial correlation between X and Y, controlling for Z, can be computed with the following equation.

$$r_{Y(X.Z)} = \frac{r_{YX} - r_{YZ} \cdot r_{XZ}}{\sqrt{1 - r_{XZ}^2}}$$

That last equation, the one for the semipartial correlation looks awfully familiar. I know we've seen it before. Well, not exactly it, but something very close. What was it? Oh yes, the equation for the standardized partial regression coefficient. In the event you forgot, I'll list it below (customized for our variables names of Y, X, and Z).

$$B_X = \frac{r_{YX} - r_{YZ} \cdot r_{XZ}}{1 - r_{XZ}^2}$$

So the difference between the standardized partial regression coefficient and the semipartial correlation is just a square root in the denominator. This similarity offers some insight into the nature of a partial regression coefficient. It's like we're staring into its soul. Both the semipartial correlation and the partial regression coefficient reflect the unique relationship (the relationship apart from the other independent variables) that a given independent variable has with Y.

Think we're done computing partial and semipartial correlations? Not a chance.

Computing Squared Partial and Semipartial Correlations with Squared Correlations

Yet another way to compute partial and semipartial correlations (squared partial and semipartial correlations in this case) involves using squared bivariate and multiple correlations. We'll call this method the ΔR^2 method for reasons that will be obvious in about three lines.

The ΔR^2 method for computing a squared partial correlation is listed below.

$$r^2_{YX.Z} = \frac{R^2_{YXZ} - R^2_{YZ}}{1 - R^2_{YZ}}$$

And the ΔR^2 method for computing a squared semipartial correlation is listed below.

$$r^2_{Y(X.Z)} = R^2_{YXZ} - R^2_{YZ}$$

As a reminder of the obvious, these equations yield the *squared* partial and semipartial correlations. One must take the square root to obtain the un-squared correlations. A word of caution: Computed any other way, partial and semipartial correlations can be positive or negative. With the ΔR^2 method, you'll never know if the partial correlation is positive or negative. To err on the side of caution, avoid using this method if any of the zero-order bivariate correlations are negative.

Did you notice something interesting about computing the semipartial correlation as a difference between two R^2 values? Every time we compute a change in R^2 associated with adding independent variables to a regression equation, we're computing a squared semipartial correlation. The semipartial correlation reflects a variable's unique relationship with the dependent variable after taking all of the other independent variables into account. So all of the ΔR^2 stuff in the previous chapter was also a squared semipartial correlation. The independent variables in the first regression equation are the control variables (i.e., *Z*) in the semipartial correlation, and the added variables in the second regression equation are the independent variables (i.e., *X*) in the semipartial correlation.

Multiple Partial and Semipartial Correlations

All of the examples to this point have involved three variables: the dependent variable, the independent variable, and the control variable. What if we want to control for multiple variables (Z_1 to Z_k)? Believe it or not, the process is the same as before. We can use the residualizing method or the ΔR^2 method. Because it's more interesting, let's

go with the residualizing method. First, we residualize Y on Z_1 to Z_k. Because the control variables are independent variables in this regression equation, we can have as many as we want – it's just a multiple regression equation. The residual scores here are the part of Y unrelated to all of the control variables. The next step is to residualize X on the control variables. Same concept. Now we have the part of X unrelated to the control variables. Finally, we correlate residualized Y with residualized X. All very easy and, this process can accommodate any number of control variables. Same procedure for semipartial correlations except we don't residualize Y.

OK, you say, what about multiple independent variables? Well, you can residualize, but that gets a little complicated. It may be better to just use the ΔR^2 methods to compute the squared partial and semipartial correlations. That method, in case you forgot, is simple enough. Just run two regression equations. The first is Y on all of the control variables. The second is Y on the control variables and all of the independent variables. Then use the appropriate ΔR^2 formula from before to compute the squared partial correlation or squared semipartial correlation. As an example, consider the following partial correlation scenario: We want to correlate X_1, X_2, and X_3 with Y controlling for Z_1 and Z_2. We can compute $r^2_{YX_1X_2X_3.Z_1Z_2}$ with just two regression analyses. The first is a regression of Y on Z_1 and Z_2. The second is a regression of Y on Z_1, Z_2, X_1, X_2, and X_3. Take the R^2 values from these regressions and plug into the partial or semipartial ΔR^2 equation.

Significance Testing of Partial and Semipartial Correlations

Since the ΔR^2 method showed us that every change in R^2 is a squared semipartial correlation, and we already have a significance test for ΔR^2 (which we called the ΔR^2 F test), care to guess

what significance test we'll use for the semipartial correlation? If you guessed the ΔR^2 F test, you are correct. We'll list it again for old time's sake, only with terminology customized for the semipartial correlation.

$$F = \frac{(R^2_{av} - R^2_{cv})/(k_{av} - k_{cv})}{(1 - R^2_{av})/(N - k_{av} - 1)}$$

Where:
R^2_{av} and k_{av} are from the regression of Y on all of the variables.
R^2_{cv} and k_{cv} are from the regression of Y on only the control variables.

Problems with Statistical Control

Statistical control is a powerful and amazing concept. Some might even say a magical concept. It sounds like the cure for all of our methodological problems. Can't control for a variable with a true experiment though randomization? No problem, just compute a partial or semipartial correlation and remove its effects. However, before we start thinking of statistical control as the solution for all of the limitations we face, we must note a few serious issues with it. The major issue, called the omitted variable problem, is that we didn't control for all of the relevant variables. It may not have even occurred to us that we should have controlled for a given variable; thus, we didn't measure it. And we can't control for a variable if we don't measure the variable. So if you've finished your study and you didn't measure some variable that you now want to control, you're out of luck. Either repeat the study or give up on the study.

To make matters worse, we also have the regression assumption that variables are measured without error. Thus, even if you measure this variable so that you can control for it but have a poor measure of the variable, you didn't really control for it. You may think you did, but you didn't. You controlled for some variable that's only weakly related to the real variable. That doesn't cut it.

Thus, the assumption is that you have controlled for all relevant variables – relevant variables that you measured poorly were not really controlled.

The net result of this omitted control variable problem is that if we didn't measure the control variable, or if we use a poor measure of it, then the partial or semipartial correlation doesn't really reflect the association between independent and dependent variables free from the control variables. We didn't remove as much variance from the independent and/or dependent variables via residualizing as we should have, yielding in a result that indicates a stronger effect for the independent variable than is true.

The opposite problem with statistical control is that it is possible to control too much, making the independent variable appear to have a weaker relationship with the dependent variable than is true. This problem occurs when the control variable is correlated with the independent variable, and, although not an actual cause of the dependent variable, it is correlated with the dependent variable as well. In essence, this is the opposite of the omitted variable problem (but doesn't have a cool name like, say, the included irrelevant variable problem). Researchers often cause this problem in an attempt to avoid the omitted variable problem by controlling for every variable that might possibly be relevant. (Researcher: "Hmmm, I'm not sure if this variable should be controlled, but better safe than sorry. I'm including it.") The moral of the story is that one shouldn't include irrelevant control variables in an attempt to make a pre-emptive defense against charges that a relevant variable was left uncontrolled. Such a move will only yield inaccurate conclusions.

In summary of the last few paragraphs, statistical control will not yield the correct answers if (a) a relevant control variable is omitted, (b) a relevant control variable isn't measured properly, or (c) an irrelevant variable (that is correlated with

the other variables) is included as a control variable. The first two problems results in an underadjustment of the zero-order correlation, making the partial or semipartial correlation greater than it should be. The latter problem results in an overadjustment of the correlation, making the partial or semipartial correlation weaker than it should be.

(Another irritation with the entire concept of statistical control is that it can be difficult to convince people that you did it right. Let's say you did your homework and controlled for all of the variables that you should have. You still have problems if a reviewer says, "Sure you controlled for variables A, B, and C, but you forgot about D." Now you have to argue with this person as to whether variable D really is relevant and needs to be controlled. Either win the argument or repeat the study with D controlled as well – even if controlling for variable D results in the included irrelevant variable problem.)

Final Thoughts on Statistical Control

In conclusion, statistical control, although cool beyond words on a technical level, is not without problems. The long and short of it is that if you want to attempt to establish causality with a minimum of problems, conduct a true experiment. Use statistical control only when a true experiment is not possible. The unfortunate reality is that true experiments are not possible in many situations (e.g., smoking research). What does all of this mean? It means there are no easy answers. Maybe that's the real lesson here. And finally, regardless of the experimental design (i.e., true experiment, quasi experiment, non experiment), you should start with a well-reasoned theory or be prepared to face a tidal wave of questions about the integrity of your conclusions. Actually, you'll face a tidal wave of questions either way, but you'll have better answers if you started with a well-reasoned theory.

One more note. Statistical control isn't just for causal research. It's also used to determine a variable's unique predictive power (i.e., how well a given variable predicts after controlling for other predictors). This sort of research question is common in research regarding the prediction of school and job performance. A typical research question concerns whether a given predictor has any predictive power beyond what is obtained from some other predictor (e.g., Do interviews predict job performance beyond what we can already predict with an intelligence test?). It's one thing to say that my test predicts job performance. It's far more impressive to say that my test predicts job performance even after controlling for other commonly used predictors.

Prediction

7

In the world of prediction, regression is a powerful ally.

Overview

Let us return to the concepts of predictive and explanatory research, a topic originally covered in the first chapter. Explanatory research is about determining causality whereas predictive research is about predicting something. The great thing about predictive research is that the results are right there in front of you. Unlike explanatory research, there are no hidden third variables whose existence, once revealed, force you to completely reinterpret your results and throw your conclusions out of the window. The beauty of predictive research is that a variable or set of variables predicts the criterion as well as R^2 says it predicts. Leave out important variables? Not a problem as long as the R^2 you obtained from the predictors you actually used is satisfactory. Measure the wrong variable? Again, not a problem if the R^2 is satisfactory.

Nothing in the preceding paragraph should be taken as an excuse to measure the wrong variables. Prediction will be better if you measure the right variables. However, in spite of any errors made in the choice of predictor variables, if R^2 is strong enough for the set of variables actually measured then you have successfully predicted the criterion.

A set of variables predicts as well as the R^2 says it does. No room for opinion or second-guessing. All very simple. Well, I lied a little bit. It's not quite that simple. A set of predictors, weighted as they are in a regression equation, predicts as well as the multiple correlation says it does *in that sample*, but they won't predict that well when applied to *future* samples. More on this issue later.

Prediction Efficiency

In our initial discussion of multiple regression, we noted that we can improve the prediction of the dependent variable by using more predic-

tors (assuming that the predictors have a non-zero semipartial correlation with the dependent variable). Now here's an additional principle: Some of these predictors won't work very well, which is to say that they won't contribute much to the prediction of Y. Thinking in terms of ΔR^2, the addition or removal of a weak predictor to the regression equation results in a small change in R^2. To settle the "How small is too small?" issue, we'll just use a significance test of ΔR^2. To summarize: A weak predictor will contribute a small, nonsignificant amount (i.e., ΔR^2) to the prediction of Y. To refresh your memory, the significance test of ΔR^2 is repeated below.

$$F = \frac{(R^2_{big} - R^2_{small})/(k_{big} - k_{small})}{(1 - R^2_{big})/(N - k_{big} - 1)}$$

Where:
big and *small* refer to the model with greater and fewer predictors, respectively.

Should we keep a predictor that does not significantly contribute to the prediction of Y? A person might answer that any increase in R^2, even if small and nonsignificant, is worth having. That person would be wrong. A small, nonsignificant contribution to the prediction of Y in a given sample is extremely unlikely to be found in subsequent samples. Long story short, a nonsignificant increase in R^2 is a phantom increase – one whose alleged benefit will never be realized.

The previous paragraph addressed a topic we discussed many chapters ago: regression equations are frequently used to make predictions in future samples. A regression equation that was developed on one sample can be used to make predictions in other samples. Although it's good to know how well a regression equation predicts in the sample in which it was developed, what really matters is how well it predicts when used in future samples. And predictors that don't predict well in the initial sample have a bad habit of pre-

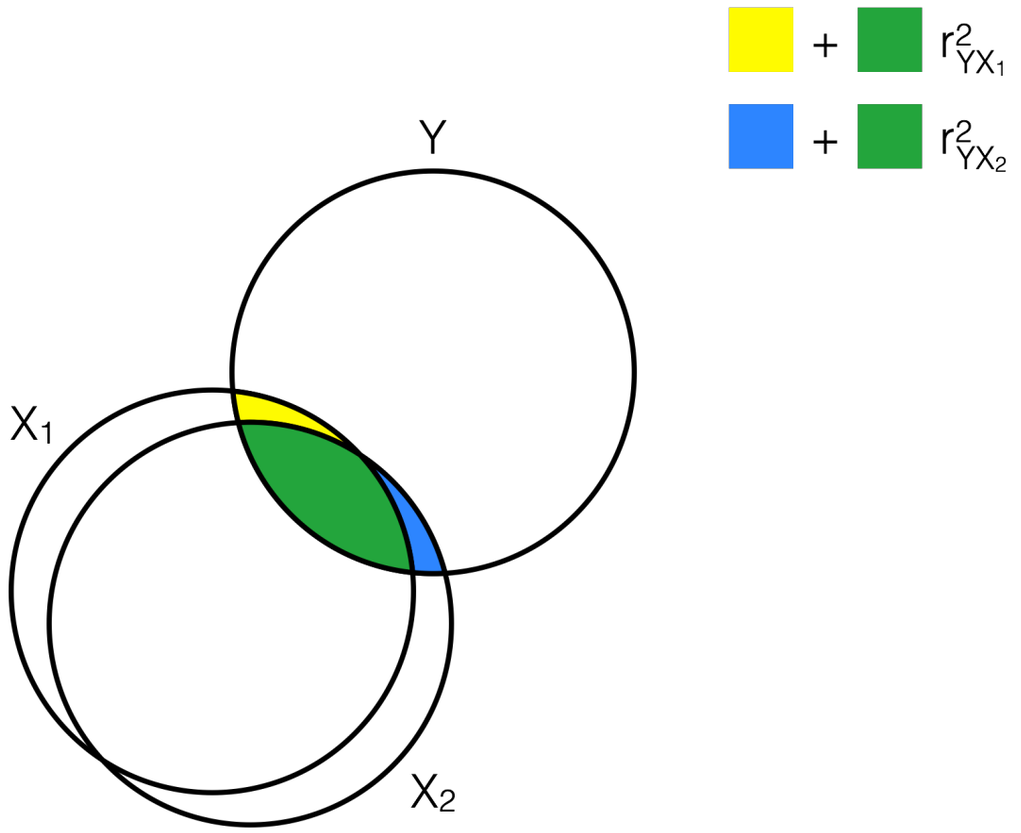
dicting even worse when that regression equation is applied to future samples.

Predictor Selection Techniques

Now that we've established that there is little benefit to keeping predictors that aren't doing much predicting, let's discuss how we go about developing a regression equation that includes only the useful predictors. And "useful" will be defined as contributing a significant amount to the prediction of Y (i.e., a significant ΔR^2).

Let's get the bad ideas out of the way. We could choose our predictors by examining the bivariate correlations between each predictor and the criterion variable. The problem with this approach is that it looks at each predictor in isolation and doesn't take into account the relationships among the predictors when they are used together (i.e., in a multiple regression equation). We can illustrate this problem two ways. First, consider Figure 1. Either predictor predicts Y well enough on its own, but once one predictor (e.g., X_1) is in the regression equation, the second one (e.g., X_2) will contribute very little to the prediction of Y. A second way to illustrate the problem of using bivariate correlations to determine

FIGURE 1 The Problem of Highly Correlated Predictors



whether a given predictor is useful in a multiple regression equation is to consider the equation for a partial regression coefficient. To simplify matters, we will once again use the formula for a standardized partial regression coefficient (listed below).

$$B_1 = \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{1 - r^2_{X_1X_2}}$$

Note that as the correlation between X_1 and X_2 increases, the magnitude of the partial regression coefficient for X_1 decreases. At a certain point, the strength of the correlation between X_1 and X_2 results in a partial regression coefficient (for one or both variables) so low as to be useless. (To tie this into ΔR^2, recall that the significance test for the change in R^2 associated with adding a certain predictor to a regression equation is the same as the significance test of the partial regression coefficient.) Once again, this information would not be obtained by a simple inspection of the bivariate correlations.

So what's a better way to do this? The lesson from before is that you can't use a bivariate correlation to address issues in a multiple regression equation. It sounds like the smart way to do this would be to throw all of the predictors into a regression equation and see what works and what doesn't. This is actually a pretty good method, although it will help to conduct it in a systematic fashion.

Here is a systematic way to do what is described above: First, regress Y on all of the predictors. Second, examine the significance tests of the partial regression coefficients. As mentioned above, and in a previous chapter, and probably again some time in the future, the t test of a partial regression coefficient is the same as the F test of the change in R^2 associated with adding that predictor to an equation already containing the other predictors. This little nugget is an enormous labor saving device. It means that all we have to do is examine the t tests of the partial regression

coefficients. If all of the predictors are significant, then we keep them all. If one is nonsignificant, then we drop that predictor. But what if two or more are nonsignificant? We drop the least significant one. Which one is that? It's the one with the lowest obtained t (or highest p value, whichever you prefer). In terms of ΔR^2, it's the one whose removal will result in the smallest drop in R^2. Important note: Even if two predictors are nonsignificant, we drop only one at a time. Why? Because dropping one may allow the other to become significant (think of Figure 1). How do we drop that nonsignificant predictor? Just re-run the regression analysis without it. Those are the steps. Regress, check, and drop. Repeat this process until all predictors remaining in the regression equation are significant.

That wasn't so bad, but wouldn't it be cool if this process was automated? Well, good news, it is on the major computer statistical analysis programs. On SPSS and SAS, this option is called backward selection, and it works in the same manner described in the previous paragraph. One word of caution: Be sure to check the p value these programs use to drop predictors. For SPSS and SAS the default p value for dropping a predictor is .10; it should be .05.

That was backward selection. What's the opposite of that? Forward selection. And as the name implies, forward selection is an procedure (automated on the major statistics programs) that proceeds in the exact opposite fashion as backward. Start with no predictors in the model. Add the most significant predictor (i.e., the predictor that will raise R^2 by the most, assuming that ΔR^2 is significant). Continue to add predictors that will increase R^2 by a significant amount. Stop when there no significant predictors left to add. The p value for adding a predictor to the model should be .05 (which is the default value in SPSS and SAS).

You may have one question on your mind: Shouldn't forward and backward selection yield the same results? Well, they should, but sometimes they don't – even when the *p* values for adding or dropping a variable are set to the same level for both procedures. There are some subtle issues (borderline cases, high correlations between predictors, and the like) that can cause the results to be different. That said, even though the final choice of predictors may be different, the resultant R^2 values will be about the same for either procedure for the simple reason that if there was another predictor that should be in the model that would increase R^2 by more than just borderline significance, it would in the model. Neither forward or backward will miss a strong predictor. The differences, when they occur, always relate to using one predictor instead of the other when they both work about the same and keeping the second predictor wouldn't help either model (a Figure 1 situation – both X_1 and X_2 work equally well but whichever one you use relegates the other to useless status).

Finally, wouldn't it be cool if we could combine forward and backward selection into some sort of hybrid model? Well, sort of. This combination procedure is called stepwise selection and proceeds like forward selection with a backward check after each step (e.g., after every predictor is added during the forward procedure, we check to make sure that none of the predictors added to the model have dropped to a nonsignificant level). Does this happen? Do predictors, once added to the model, sometimes become nonsignificant when other predictors are added to the model? The answer is yes, sometimes. Is this likely to be a problem? No.

A note on terminology. All of these variable selection techniques are a type of something called hierarchical regression. Hierarchical regression refers to a analysis in which more than one regres-

sion analysis is conducted with variables added to or removed from the equation at each step. Furthermore, the results from the various analyses are compared to each other. Thus, every ΔR^2 is the result of a hierarchical regression analysis. (Hierarchical regression is sometimes called stepwise regression; however, that name can cause problems. Stepwise is the name of a specific variable selection procedure.) The opposite of hierarchical regression is simultaneous regression. With simultaneous regression, only one regression analysis is conducted – no variables are added to or subtracted from the model.

One final note on all variable selection procedures. Let us remind ourselves of the goal of these variables selection procedures: We want to obtain a regression equation with the best possible (i.e., maximum) R^2 with a minimum of predictors. This goal is consistent with the goal of having a parsimonious model. Parsimony is a concept that means when there are two models of similar effectiveness (e.g., equal predictive power), the simpler model is to be preferred; the more complicated model is seen as being unnecessarily complicated. Back to our variable selection procedures, we use these variable selection procedures to purge our model of predictors that fail to significantly contribute to the prediction of the dependent variable. Thus, the choice to retain a predictor is purely a statistical one. The significance test of ΔR^2 determines whether a predictor is retained in or is dropped from the model. As such, variable selection procedures are appropriate only for predictive research and should never be used for explanatory research. With explanatory research, theory guides the choice of variables in the model.

Regression Overfitting

Earlier in the chapter we discussed how a regression equation developed for predictive purposes is ultimately intended to be used on future samples. We also said that the key issue is not

how well this equation predicts in the sample in which it was derived but how well it predicts when applied to new samples. The bad news is that regression equations have a bad habit of predicting worse in future samples as compared to the original samples on which they were derived. There are three issues we should address on this point. First, why does this happen? Second, how do we obtain an accurate estimate of how well this equation will predict when applied to new samples? And third, how can we prevent this problem from being such a problem?

This problem where a regression equation does not predict as well in future samples is known as regression overfitting. The reduction in R^2 when the equation is applied to future samples is called shrinkage. Regression overfitting refers to the fact that a regression equation developed in a sample is customized to that sample. If you recall our discussion in earlier chapters, we stated that the regression weights are the best possible weights for that sample. No other weights will work better *in that sample*. These regression weights are optimized for the sample in which they were derived. (I know I'm being redundant here. Sorry. I just want to make sure it's clear.) Recall that linear regression analysis is also referred to as *ordinary least squares* regression, where *least squares* refers to minimizing the sum of squares residual. This is just another way of saying maximize R^2, and R^2 is maximized by choosing the best possible set of regression weights. None of this would be a problem if our samples represented the population perfectly. The unrelenting thorn in our collective sides is sampling error. Sampling error cannot be avoided when we measure anything less than the entire population, and it affects every statistic we compute, including correlations and regression coefficients. So a regression equation is optimized (or customized, if you like that word better) for the sample on which it was derived. That sample has characteristics that are not pre-

sent in other samples from the same population. As such, a regression equation developed on one sample simply will not work as well when applied to other samples.

You might ask how exactly a regression equation is optimized for a given sample. We find the answer by analyzing the partial regression coefficients. One thing partial regression coefficients do is weigh predictor variables relative to each other in terms of predictive power. Other factors being equal, variables that predict the dependent variable better are assigned greater weights than other predictors. As we learned in an earlier chapter, partial regression coefficients are a function of a given predictor's correlation with the dependent variable, that predictor's correlation with the other predictor variables, and the standard deviations of these variables. If the correlations, both among predictors and with the dependent variable, in a given sample deviate from the population values (and they will), then our regression equation has improperly weighted the various predictors. (Note that this is only a problem with multiple regression. For simple regression, there can be no regression overfitting as there is only one variable and it cannot be over or under weighted as compared to other predictor variables.)

Here's an example of regression overfitting in action. We start with a dataset representing a population consisting of a million cases with scores on three predictors and a criterion variable, all standardized. A multiple regression analysis on this population dataset yields an R^2 of .31 and a regression equation: $Y' = 0 + .32X_1 + .25X_2 + .18X_3$. Those are the population values. We would like for our sample regression statistics to match the population values perfectly. But they won't because of sampling error. It's always sampling error.

Let's see what happens when we randomly draw a sample of 100 cases from the population and execute a regression analysis. First, the R^2 is

.39. Note that it is greater than the true population R^2 of .31; that's the result of overfitting. Next, the regression equation is
$Y' = .15 + .48X_1 + .27X_2 + .01X_3$. Notice how X_1 is overweighted as compared to the population equation (.48 in the sample equation versus .32 in the population equation). Also notice how X_3 is underweighted as compared to the population equation (.01 versus .18). The sample equation is the best fitting equation in that sample, but it is not the best set of regression weights when the equation is applied to new samples or the entire population. The equation is overly customized to the characteristics of the sample on which it was derived and will not yield the same level of predictive power for future samples.

Cross-Validation

So a regression equation will not work as well when applied to future samples as it does in the sample in which it was derived. How then can we know how well it will work in future samples? The answer is a process called cross-validation which allows us to estimate an R^2 free from the biasing effects of regression overfitting (this new estimate of R^2 is also called the *shrunken* R^2). There are two ways to estimate the cross-validated R^2. One way is with an empirical (i.e., loads of real data) cross-validation process. For the sake of clarity, let's call this procedure *two-sample cross-validation*. The other method is to use an equation-based estimate. We will discuss both methods.

A two-sample cross-validation is based on collecting and analyzing two samples of data (hence the useful name) and proceeds as follows.

I. A regression analysis is executed on one sample of data.
II. A new sample of data is collected. The same predictors used in the first sample are used in the second sample (i.e., we obtain

scores from everyone in the new sample on the same set of variables).

III. The regression equation from the first sample is applied to the second sample. That is, predictor scores from the second sample are plugged into the regression equation obtained from the first sample. This process results in a set of predicted *Y* scores for everyone in the second sample.

IV. Within this second sample, the actual scores on *Y* and the predicted *Y* scores are correlated.

This correlation, once squared, is the cross-validated R^2 and indicates how well the regression equation actually predicts when applied to new samples of data. We hope that the cross-validated R^2 is not much lower than the original R^2.

Cross validation can be a little confusing. Here's what is not happening with it. We are not conducting a second regression analysis in the second sample. Doing so would give us a new regression equation optimized for the second sample. Such an analysis would tell us nothing about how well the first equation predicts when applied to future samples. Consider that if there was some sort of predictor selection (i.e., forward), a different set of predictors might be chosen in the second sample. How would that evaluate the effectiveness of the regression equation from the first sample? So put that thought out of your head. A proper two-sample cross-validation takes a regression equation developed in one sample and uses that equation to make predictions in a second sample. The correlation between the predicted *Y* and actual *Y* in the second sample tells us how well it predicted.

The great thing about a two-sample cross-validation study is that it answers the question we had by doing exactly what we should do to answer that question. You want to know how well a regression equation predicts when applied to future

samples? Get a second sample of people, give them the same tests, use the equation obtained in the original sample to compute predicted *Y* in that new sample, and correlate predicted *Y* with actual *Y* in the new sample to see how well it predicts.

The only negative aspect of this two-sample cross-validation procedure is that we must collect two samples of data. It's difficult to collect one sample of data of sufficient size. Collecting two samples of data of sufficient size is twice the work (and it may feel like ten times the work). One solution to this problem is to take a single sample of data and randomly split it into two subsamples, one for running the initial regression analysis and a second for estimating the cross-validated R^2. The problem with this approach is that the sample size has effectively been cut in half, increasing sampling error within each subsample. The regression equation developed in the first sample is now based on half of the available sample.

So that's two-sample cross-validation. What of this other way to estimate the cross-validated R^2? This second method consists of a family of equations which use the results of a single regression analysis conducted on one sample to estimate the shrunken R^2. These equation-based methods have an obvious advantage. There is no need to collect a second sample of data, nor is there a need to divide one sample into two subsamples (with the resultant increase in sampling error). If these equation-based estimates of the cross-validated R^2 are as accurate as a two-sample cross-validation study, then they should be preferred.

There are many equations available to estimate the shrunken R^2. We will discuss two. The first is an equation developed by Ezekiel (1930; commonly credited to Wherry, 1931).

$$R_p^2 = 1 - (1 - R^2)\frac{(N - 1)}{(N - k - 1)}$$

As you can see, there are only three components to this equation: R^2, N, and k (the number of predictors). This equation is quite popular and is the default method for estimating the shrunken R^2 in SAS and SPSS (where it is called Adjusted R^2 or R^2_{adj}). Unfortunately, the Ezekiel/Wherry equation doesn't truly estimate the cross-validated R^2, although it comes close. The Ezekiel/Wherry equation is actually an estimate of the population multiple correlation and not the cross-validated multiple correlation (both of which can be legitimately called a shrunken R^2). The conceptual difference is subtle and the mathematical difference is often rather trivial in magnitude. But as long as we're going to do this, let's do it right. Many other equations exist to estimate the cross-validated R^2. Raju, Bilgic, Edwards, and Fleer (1999) compared the accuracy of the various estimator equations and found that an obscure equation by Burket (1964) was superior to or equal to the others in most conditions. Even better, the Burket equation is rather computationally simple. The Burket equation for estimating the cross-validated multiple correlation is given below.

$$R_{cv} = \frac{NR^2 - k}{R(N - k)}$$

All terms are defined as before. Note that this equation yields R and not R^2. To obtain R^2, you have to square it yourself.

Raju et al. (1999) found excellent accuracy for these formula-based estimates of the cross-validated R^2. Based on what we have presented, it appears that an equation-based estimate of the cross-validated R^2 should always be used instead of a two-sample cross-validation study. Before we get too excited, a word of caution. Equation-based estimates of the cross-validated R^2 may not be accurate when any sort of predictor selection (i.e., forward, backward, stepwise, or using only predictors with significant bivariate correlations) has occurred. At this point, it is premature to draw any

conclusions. However, it is possible that predictor selection techniques may increase sample-specific regression overfitting due to the fact that the choice of predictors retained in the equation is itself affected by sampling error. Additional research should be conducted on this issue. At this point, the safe conclusion is that if predictor selection has occurred, then two-sample cross-validation is to be preferred to equation-based estimates of the cross-validated R^2.

Minimizing Shrinkage

So far we've explained why regression overfitting occurs, the effects of regression overfitting, and how to estimate the predictive power of a regression equation free from the biasing effects of regression overfitting. In a sense, we know the nature of the sickness, we know the cause of the sickness, we know the effects of the sickness, but can we prevent the illness? The answer is... sort of. An examination of the equations designed to estimate the cross-validated R^2 indicate that shrinkage is a function of the number of subjects and predictors in the initial regression analysis. As with any discussion of sampling error, larger sample sizes reduce the magnitude of sampling errors. Thus, larger sample sizes in our initial regression analysis yield an equation that predicts almost as well when applied to future samples. The second factor is the number of predictors. More predictors in the regression equation mean more opportunities for sampling error and, thus, greater shrinkage when that equation is applied to future samples. Long story short, a high ratio of subjects to predictors means less overfitting and, thus, less shrinkage when that equation is applied to future samples. In short, when subjects to predictor ratios are high, the R^2 obtained in the initial regression analysis is a better estimate of how well that equation will predict when it is applied to future samples than when subjects to predictor ratios are low. What's a high subject to predictor ratio? An-

swers to this question range from 30:1 to 10:1. The conclusion we can draw from this is that any ratio less than 10:1 is too low; we are likely to be disappointed when an equation developed in such a situation is applied to future samples.

The Part Where I Sneak in a Major Discussion About Science

Back to how predictor selection techniques function, recall that when predictors are added to or dropped from a model, the basis for these selections is the ΔR^2 F test. It may not have occurred to you, but with all predictor selection techniques, a series of ΔR^2 F tests is being conducted. Is that bad? Maybe. Here's why. Consider how this thing we call science is supposed to proceed.

> I. Based on theory, previous research, or a wild guess, a hypothesis is proposed.
> II. Following a specified procedure, data are collected on the relevant variables.
> III. Those data are analyzed.
> IV. If the results are consistent with the hypothesis, then the results are interpreted as supportive of the hypothesis.
> V. If the results are inconsistent with the hypothesis, the results are interpreted as indicating that the hypothesis is incorrect.
> VI. If the hypothesis wasn't supported, the researcher has option to: identify fatal flaws in the design of the study (making the results of the study irrelevant), abandon the hypothesis (i.e., give up and admit that it's worthless), or modify the hypothesis and test this modified version with a new set of data (i.e., conduct a new study to test this revised hypothesis).

Let's focus on that last part where we modify and re-test the hypothesis. One option that is definitely not available to us is to test the modified hypothesis on the original set of data (the data that demonstrated that the original hypothesis was incorrect). Why not? Given that the modifications

are based on the what the dataset revealed, how could a retest of the modified hypothesis *on that same dataset* ever fail?

Here's an example of the problem. Based on theory, an unnamed researcher hypothesizes that Treatment A will result in greater success on the dependent variable than Treatment B. The researcher conducts the study, analyses the data, and finds the opposite case, B has outperformed A. Our researcher then revises his theory, reformulates his hypothesis to state that Treatment B will result in a greater success rate than Treatment A, tests this newly modified hypothesis on the original dataset, and finds, much to his delight, that his hypothesis is supported.

If that sounds too nefarious for you, here's a slightly more palatable version of events. Our researcher is not so sure about how things are going to turn out, so starts without a hypothesis. He conducts the study and analyzes his data. He sees that Treatment B outperformed Treatment A. At this point, he officially hypothesizes that Treatment B will be more successful than Treatment A. The data analysis that has already been done is used as support for his hypothesis.

You see what I mean? There's no way for this to go wrong. Our researcher will always end up with experimental support for his hypothesis. No matter what. This process of examining the data, and then forming a hypothesis is known as *post hoc theorizing* (and the hypothesis is a *post hoc hypothesis*). Just to clarify, if these newly modified hypotheses were tested on a new sample of data, all would be well. But that's not what was described in the two preceding paragraphs. It was: data analysis, hypothesis modification/formation, test of that hypothesis on the same sample of data.

I once heard a speaker defend this practice with the statement that, "The data don't know when the hypothesis was formed." Yes, those were

his actual words, anthropomorphism and all. Of course he meant that forming a hypothesis before or after examination of the data doesn't change the data. And he's right on that count. But it sure does change your hypothesis. And it's not hard to find a significant effect somewhere in a dataset if you keep looking long enough. The consequence of this is that you end up building a hypothesis around something that is often just sampling error. It won't replicate because it's not a real effect.

Do you see how this post hoc theorizing issue ties into predictor selection in regression analysis? If we start with ten predictors in the model, then our initial hypothesis is that these ten predictors are related, as an optimally weighted composite, to the dependent variable. Once we start kicking predictors out of the model because they are not significantly related, then we have changed our hypothesis. I'll present this situation as a one person play.

INTERIOR RESEARCH LAB

A dimly lit room. A researcher of indeterminate age has his face buried in a computer screen and is waiting for SPSS to finish loading. His leg shakes with a nervous energy. The blinds are down but not completely closed. A late afternoon light seeps through the gaps between the slats. In the background is the faint sound of a grant application being rejected somewhere off in the distance.

RESEARCHER
I hypothesize that these nine variables will predict the DV.

SPSS finally finishes loading. The researcher starts a regression analysis with a backward variable selection procedure.

Before he can blink, three independent variables are deleted from the regression equation.

RESEARCHER
Did I say nine? I meant six. These six variables will predict.

His face begins to show signs of anxiety.

RESEARCHER
I never liked those variables anyway.

As he speaks another independent variable is deleted. Sweat trickles down his face and into the collar of his ill-fitting golf shirt.

RESEARCHER
Yes, sir. These six, I mean, five variables will predict the DV in a significant fashion. No doubt about it.

One more variable is deleted. The analysis stops, retaining only four of the original nine independent variables. He makes a note of the remaining variables.

RESEARCHER
These four IVs will predict the DV. That's my hypothesis. Now, it's time to test it.

On the same dataset, he executes a regular regression analysis on the four variables retained from the backward selection procedure.

RESEARCHER
Would you look at that? I was right. Those four predict Y. And all four were significant Just like I hypothesized.

E N D

Now, most of the time it's not as bad as all that, but there are some tough scenarios. What if the overall R^2 in the full model is not significant; however, after dropping weak predictors, it becomes significant upon retesting? Isn't that the very situation we described earlier? In such a situation a new sample of data would need to be collected to determine if this revised equation actually predicts. Because the original equation sure didn't.

Curvilinear Regression

Clever solutions to
irritating problems.

8

Introduction

As we have discussed many times to this point, linearity is a key assumption of correlation and regression. Linearity means that the best fitting model of the relationship between the independent and dependent variables is in the form of a straight line. But what happens when this assumption is violated? Sometimes a dataset exhibits a relationship that is not adequately summarized by a straight line. What to do then? The answer is that we'll have to use a different kind of regression analysis.

Two Options

This new kind of regression analysis will be some sort of nonlinear regression analysis. If linear regression analysis was designed to fit a straight line to the observed data, this nonlinear regression analysis will have to fit a curve to the data. Thus, the parameters (i.e., the various *b*s, the *a*) in the regression equation will be nonlinear. The days of multiplying scores on *X* by a constant and adding a constant will be over. Scores on *X* will be multiplied by something newer and cooler. This new form of regression analysis will be fundamentally different from anything we have discussed before.

But there is another option. What if, instead of using a new type of regression analysis, we use the same old type of regression that we know and love (i.e., OLS regression), but we use it in a different way? Different how? Instead of altering the regression equation parameters, we'll alter the variables. This second option has some nice benefits. For one thing, we can use a form of regression analysis with which we are already familiar. Second, it's an easy procedure to implement. So we'll go with this option. And we'll call it polynomial regression analysis.

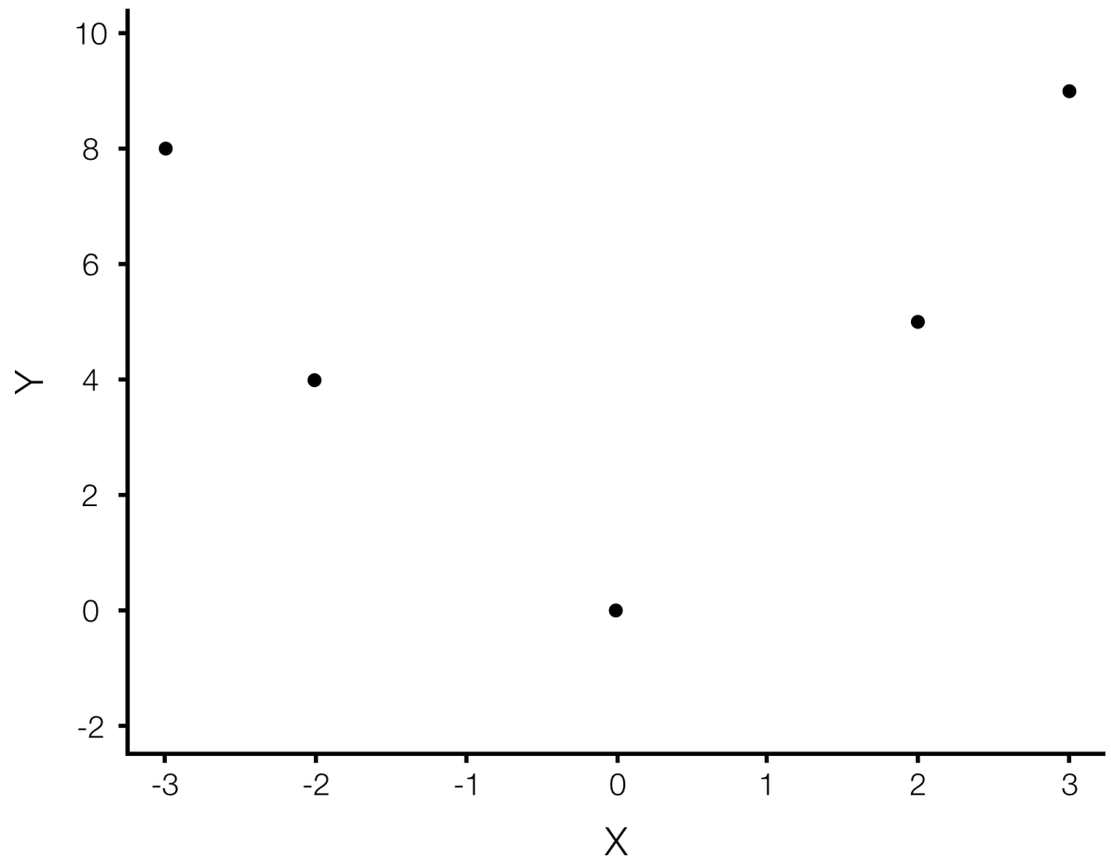
A Brief Demonstration

Here's a quick demo of how a simple data transformation solves our linearity (or lack thereof) problems. Consider the following dataset.

Name	Y	X
Albert	8	-3
Whitney	4	-2
Jewell	0	0
Martha	5	2
Odell	9	3

An examination of the scatterplot (Figure 1) shows that the relationship is decidedly non linear. The correlation between *X* and *Y* is .14.

FIGURE 1 Data Transformation Demo: Raw Data

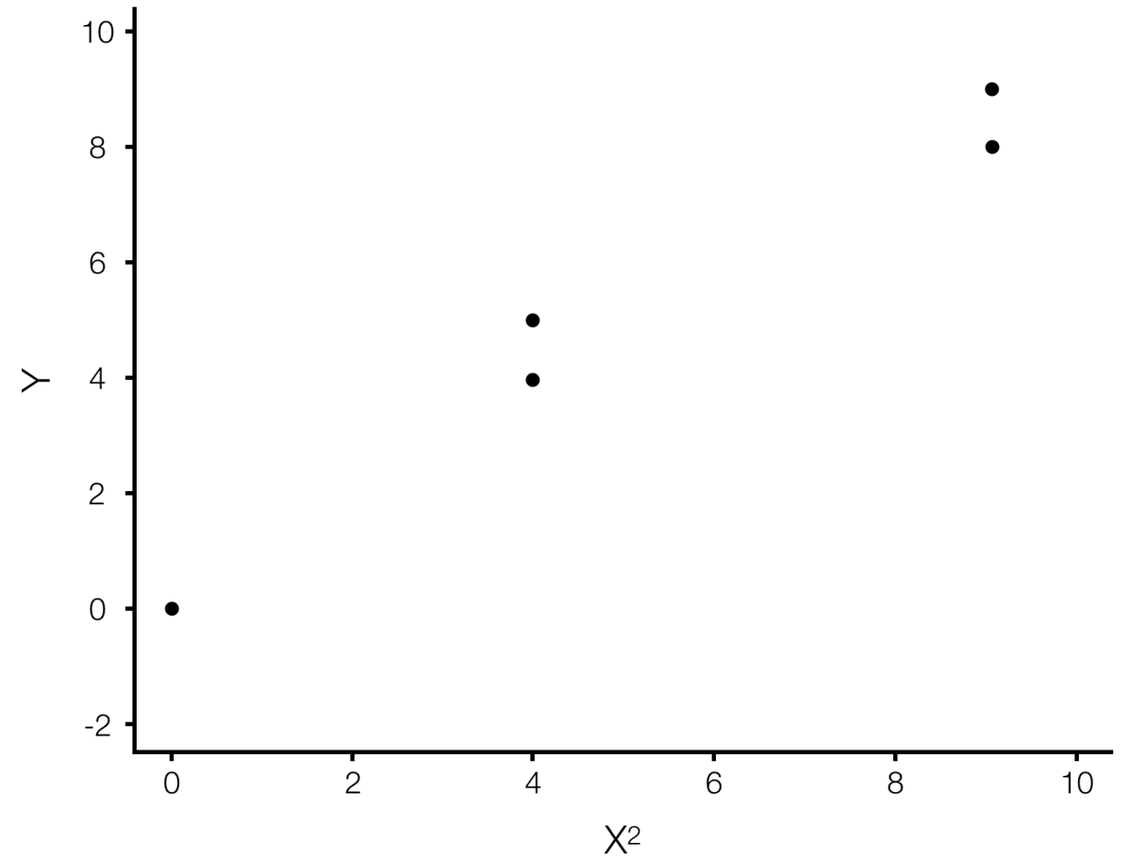


Watch what happens when we simply square the scores on *X* and re-run the analysis.

Name	Y	X^2
Albert	8	9
Whitney	4	4
Jewell	0	0
Martha	5	4
Odell	9	9

The correlation between X^2 and Y is now .98. That's quite an improvement. Those low, negative scores on X are now high scores on X^2. Even more striking than the correlation is the scatterplot (Figure 2). Note how what was once a non linear trend is now a linear one.

FIGURE 2 Data Transformation Demo: Squared Scores on X



There are a few more steps (and important details) to an actual polynomial regression analysis than what we did for this demonstration. But this exercise shows how a simple data transformation can solve a major problem for us by turning non linear relations into linear ones.

Polynomial Regression Analysis

Polynomial regression analysis involves transforming our independent variables and using them in a standard linear regression equation. Transformed how? Assuming one independent variable, scores on this variable will be squared, cubed, taken to the fourth power (quaded?), and so on. These new forms of this variable will be called powered vectors. These powered vectors are not new variables, rather they are multiple representations of a single variable. These powered vectors will be entered into a hierarchical regression analysis, one at a time, to see if the strength of the relationship between X and Y improves with their addition. Thus, the philosophy of polynomial regression analysis is clear: A variable exhibiting a nonlinear association with Y will undergo a nonlinear transformation; this variable and it's higher-powered representations will be entered into a linear regression analysis to find out just how many of these powered vectors are needed to adequately model the relationship between X and Y.

To summarize, a polynomial regression analysis proceeds as follows. First, various powered vectors are created from the independent variable. Second, the dependent variable is regressed onto the independent variable (i.e., Y is regressed on X). Third, powered vectors are added to this regression equation, one at a time. ΔR^2 is computed and evaluated for significance at every step (using the standard F test for the change in R^2 test that we know and love). This procedure stops after two or three nonsignificant additions to the model. Finally, the last model featuring a significant ΔR^2 over the previous model is retained as the final regression equation.

To illustrate this procedure, consider a sample dataset.

Name	Y	X	X^2	X^3	X^4
James	20	4	16	64	256
Randall	25	5	25	125	625
Celia	29	6	36	216	1296
George	31	7	49	343	2401
Henry	32	8	64	512	4096
Mary	32	9	81	729	6561

Notice that we only have two variables: X and Y. We have created various powered vectors for X: X^2, X^3, and X^4. Our next concern is determining which powered vectors are needed. For this, we first regress Y on X. At this stage this is just regular regression – nothing fancy yet. A linear regression of Y on X results in an R^2 of .857. As high as .857 is, this linear relationship may not be the best model of the relationship of Y on X. So we add X^2 to the equation and regress Y on X and X^2. This regression results in an R^2 of .999. The ΔR^2 for the addition of X^2 is .142, a sizable and significant increment ($\Delta R^2 = .142$, $F = 402.6$, $p < .05$). Thus, the best fitting model of the relationship between Y and X requires, at the least, the scores on X to be squared, which is a nonlinear model.

Continuing with our example, R^2 is already .999, so it probably won't increase any more. But let's see what happens when X^3 is added to the equation. A regression of Y on X, X^2, and X^3 results in an R^2 of .999. So there's no change ($\Delta R^2 = 0$, and of course, it's not a significant increase). Same with the addition of X^4 to the model. Let's return to the last model that had a significant increase in R^2 as compared to the previous model, the model with Y regressed on X and X^2. It is this equation that models the best fitting (and simplest as the addition of higher-powered vectors did not increase the strength of association between X and Y) relationship between X and Y. The regression equation for this model is

$Y' = -13.2 + 11.0X - .66X^2$. An inspection of the t tests of each regression coefficient shows that the regression coefficients associated with both X and X^2 are significant.

But what if the regression coefficient for X wasn't significant? Would we drop X from the model? The answer is no – all lower powered components of a powered vector must stay in the model. If X^7 is in the model, then X, X^2, X^3, X^4, X^5, and X^6 must remain in model regardless of the significance of each of these lower powered components. (By the way, these lower powered components are called constituent variables, probably the cutest name in all of statisticsdom.)

Scatterplots and Residual Plots in Polynomial Regression Analysis

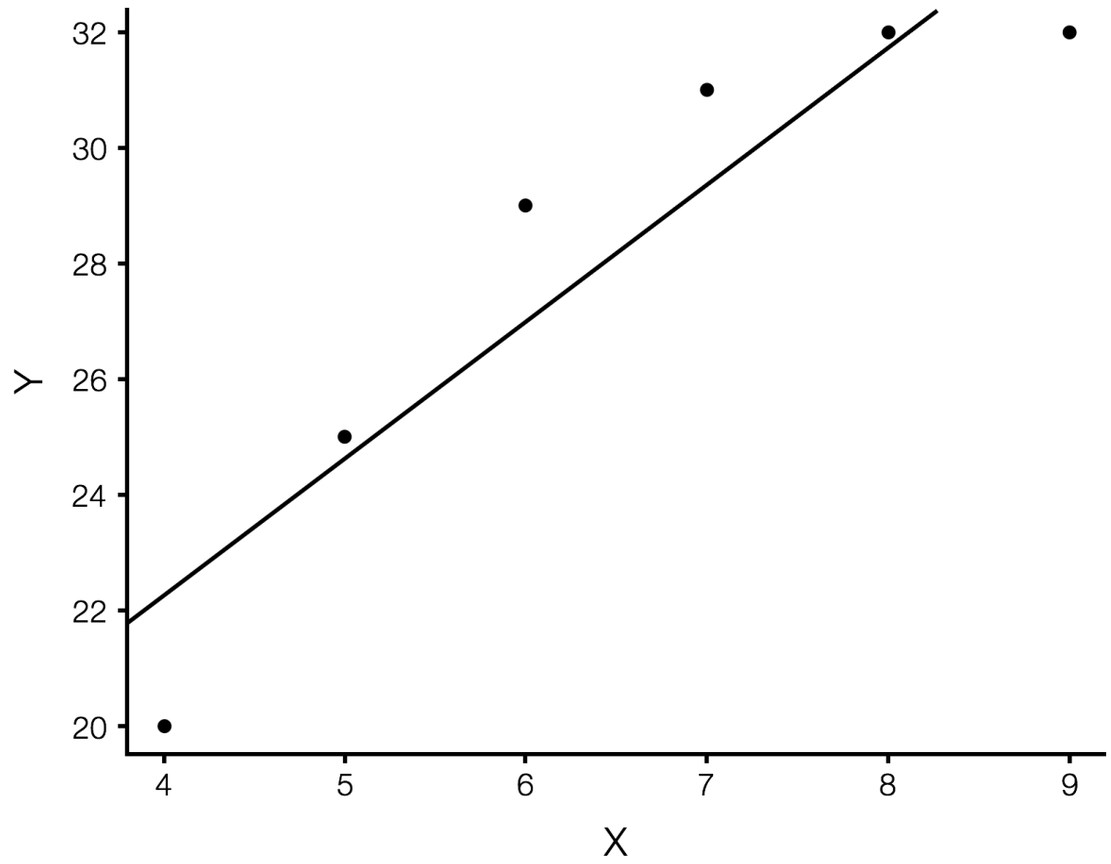
To summarize what we have found in our example, there is a relationship between X and Y. Although a linear regression revealed a rather strong relationship ($R^2 = .857$), the best fitting model of this relationship was nonlinear and yielded an R^2 of .999. The resultant regression equation was $Y' = -13.2 + 11.0X - .66X^2$, and it functions just like any other regression equation we have encountered to date (i.e., insert scores on X to obtain predicted Y). Long story short, we ran a curvilinear regression analysis and found some interesting results. But how would a researcher know whether she or he should run a curvilinear regression analysis in the first instance? If your answer involves examining the R^2 obtained from a simple linear regression of Y on X (e.g., R^2 for the regression of Y on X is weak so maybe there's a nonlinear relationship), you are wrong. In our example, that R^2 was .857, a number so high that you would be forgiven for thinking that the linear regression equation is the correct model for this dataset. But we know it's not the correct model for the data.

Back to the question. How do we know when we should conduct a polynomial regression analy-

sis? Is this something we should do every time we run a regression analysis? Well no. We already know the answer to this question. When we first discussed regression analysis, we also discussed the assumptions of regression analysis. Two key assumptions were linearity and homoscedasticity. And, as discussed, the way to check for violations of these assumptions is by examining the residual plot.

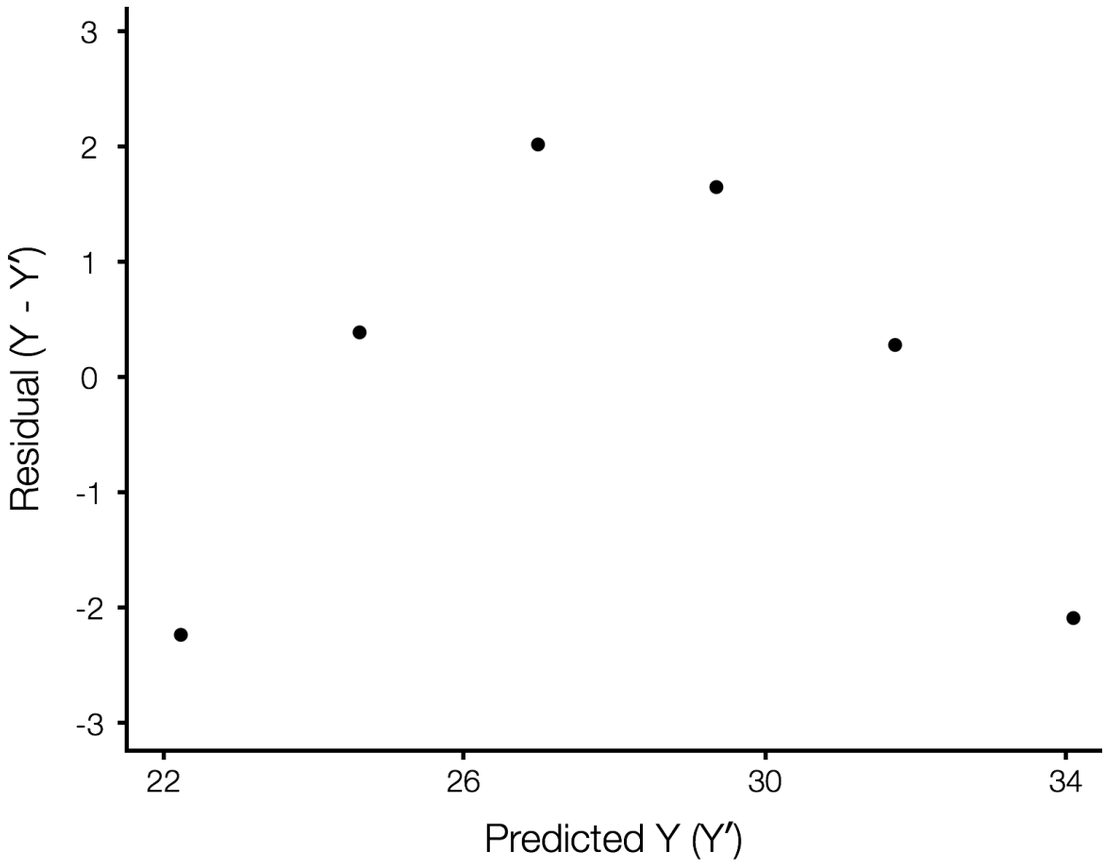
Back to our example dataset. Figure 3 shows the simple scatterplot between *X* and *Y*. Just for good measure, I threw in a regression line. You can tell that this particular regression is from a linear regression because, well, it's a straight line. As this scatterplot makes clear, a linear model does not adequately describe the relationship between *X* and *Y*. It's not a total failure because there is a general trend of higher scores on *Y* being associated with higher scores on *X*. But that pattern tails off at the high end of scores on *X* where increased scores on *X* no longer leads to increased scores on

FIGURE 3 Scatterplot for Regression of *Y* on *X*



Y. What about the residual plot? The residual plot is shown in Figure 4. A residual plot, when taken from a linear regression analysis, shows you what is left over after the linear association has been extracted from the relationship between *X* and *Y*. In this case, there is a clear nonlinear association remaining. Now we know that a curvilinear regres-

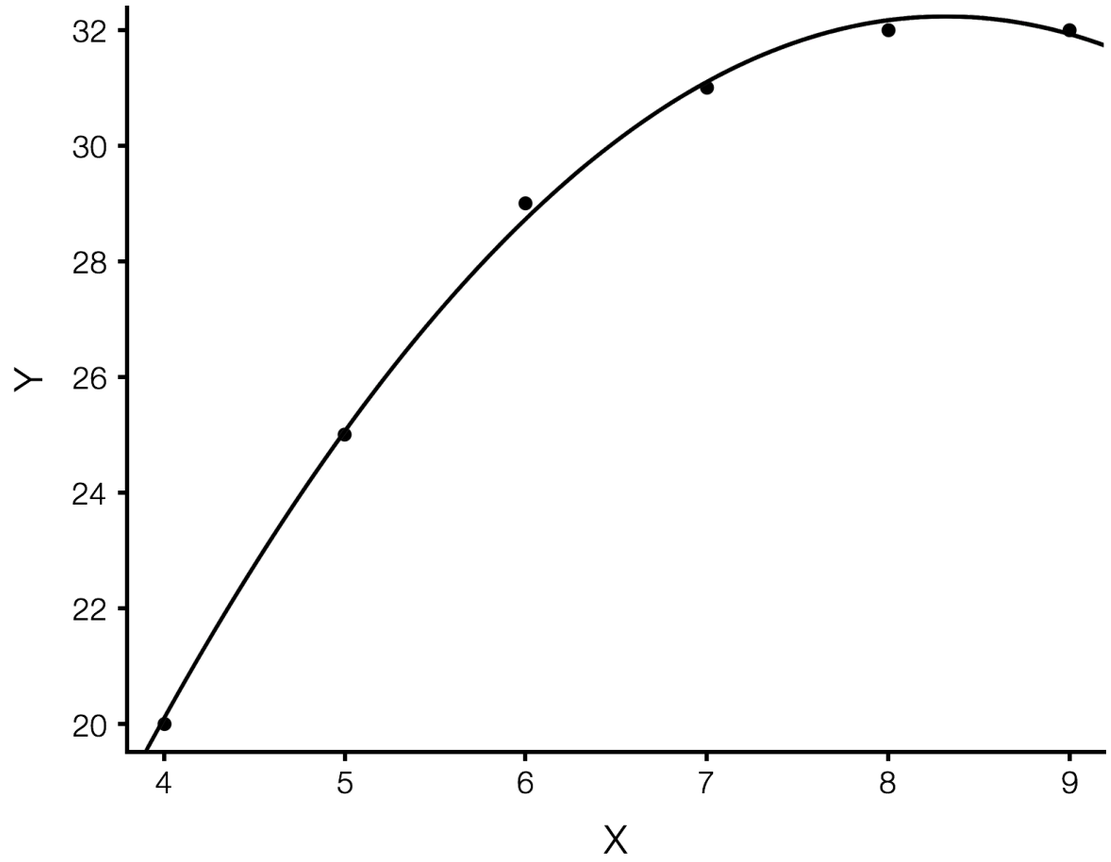
FIGURE 4 Residual Plot for Regression of Y on X



sion analysis is called for. The moral of the story is that you should always generate a residual plot. Failing to do so may cause you to miss one of the coolest things you could ever find, a curvilinear relationship.

As long as we're talking about graphs, you might ask what the regression line looks like

FIGURE 5 Scatterplot with Regression Line for the Regression of Y on X and X^2

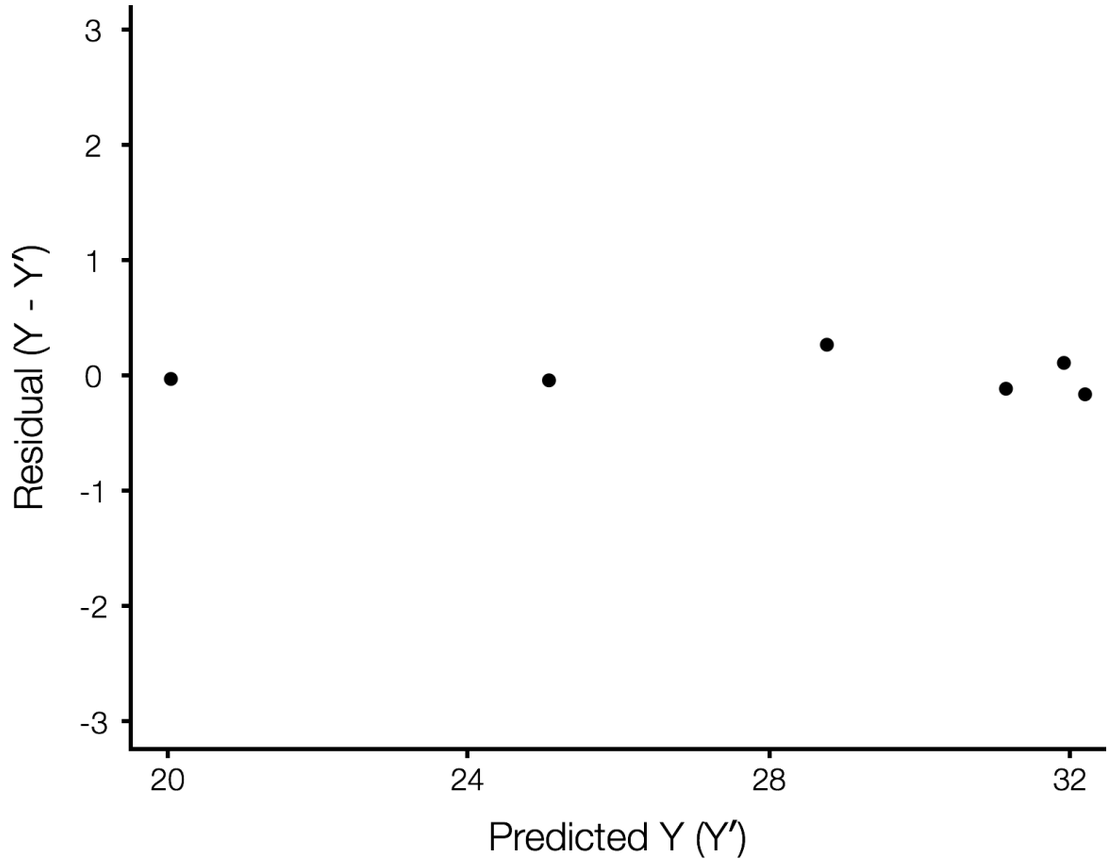


when we execute a proper curvilinear regression analysis. Do we get a cool curve that fits the scatterplot we observed in Figure 3? Why, yes we do. Figure 5 shows a simple scatterplot of our data with the new regression line drawn (which is now not really a line, but a curve). How do we generate this line? We graph it using the regression equa-

tion, just like we would graph any regression line. If you recall from an earlier chapter (I can't remember which one, but it was earlier), one way to draw the regression line is by plugging a range of values for X into the regression equation, solving for Y', and plotting the series of (X, Y') scores on the regular scatterplot. If this procedure sounds cumbersome, modern statistics programs automate it. But it's important to know how it really happens.

One last issue about our example. We found that the best fitting model of the relationship between X and Y is described with equation $Y' = -13.2 + 11.0X - .66X^2$. We saw what the residual plot looked like before we added X^2 to the equation. We also saw how well this curvilinear model fits the scatterplot. But one thing we haven't examined is what the residual plot looks like when Y is regressed on X and X^2. Will we still see a nonlinear trend in the residual? If so, then our regression equation will need additional

FIGURE 6 Residual Plot for the Regression of Y on X and X^2



higher-powered powered vectors (i.e., X^3, X^4). Of course, we already know from our significance tests of ΔR^2 that these higher powered terms are not needed. The residual plot for the regression of Y on X and X^2 is shown in Figure 6, and it looks

about as good as a residual plot could hope to look.

Other Issues with Polynomial Regression Analysis

One problem that might have occurred to you is that the various powered vectors might be highly correlated with X (and each other). In our example X correlates .99 with X^2. Point nine-nine! A high correlation among independent variables in a multiple regression equation (I know the powered vectors are not separate independent variables, but they appear as such to the mathematical number crunching of regression analysis) is called colinearity and can be a bad thing. Excessively high colinearity can cause a regression analysis to have serious problems being executed. Indeed, a 1.0 correlation among any combination of independent variables (as might be found when various subtest scores and a composite score of these subtests are entered as independent variables) is called a linear dependency. The regression of a set of linearly dependent variables cannot be executed. Due to rounding issues, a correlation close to 1.0 can have the same problem. Various statistics programs will refuse to execute a regression analysis when the colinearity among independent variables is so high as to approach a linear dependency, presumably to prevent the occurrence of some sort of cosmic catastrophe.

Thus, we are in a bind. The mere creation of various powered vectors leads to high colinearity. If colinearity is intolerably high, we can't run the analysis we need. What are we to do? The answer comes from a simple mathematical principle: a positive number when squared stays positive, but a negative number when squared becomes positive. If half of the numbers in a set are negative, squaring them results in the sort of transformation that kills a correlation. Thus, our solution will be to transform scores on X before computing powered vectors so that half of the scores are posi-

tive and half are negative. Standardizing to z scores does this. But we don't have to go that far. Simply computing mean-deviation scores $(X - \bar{X})$ for all of the scores does this as well. This latter procedure is known as centering. Both work fine. Choose whichever you prefer. Back to our example, the centered scores on X correlate 0.0 with the squared centered scores on X. Much better in the colinearity department. Problem solved. (Before moving on, I want to emphasize the steps: If centering is needed, first center, then compute the powered vectors from the centered scores on X.)

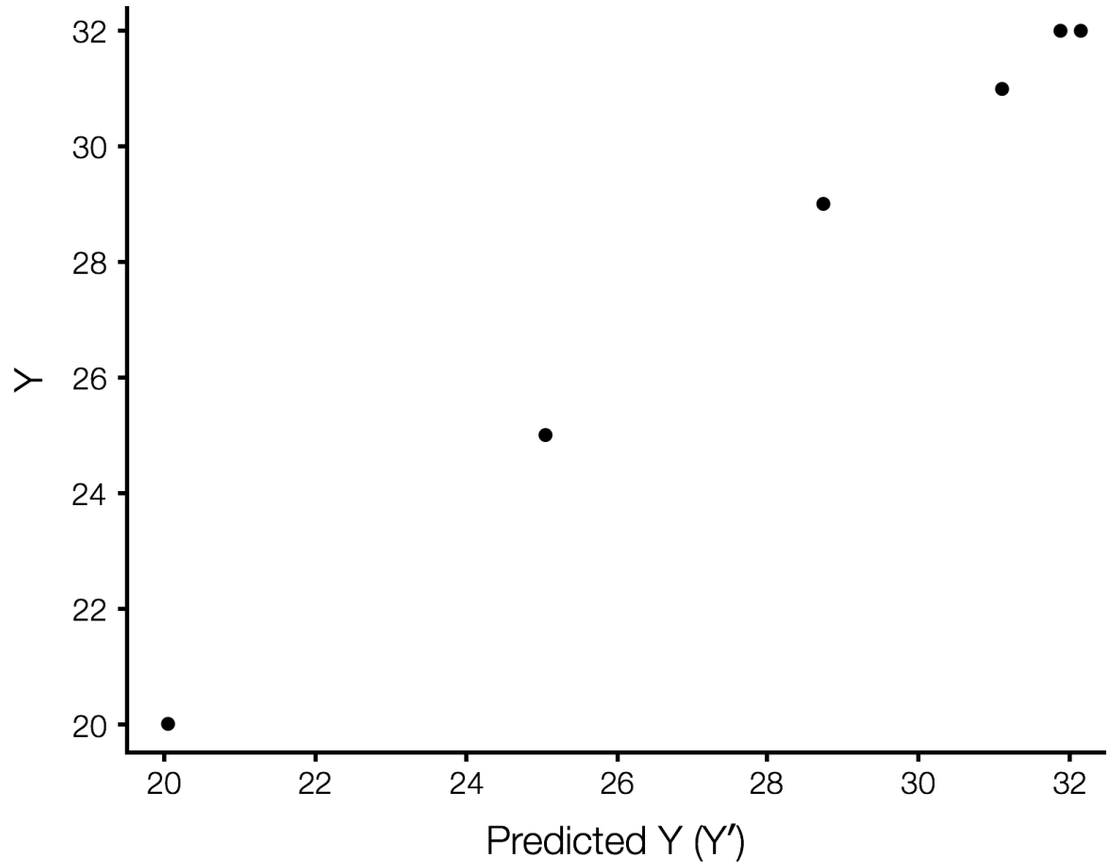
This colinearity issue is important to address because you may find that when you execute a polynomial regression analysis, your statistics program of choice will not do what you request of it. You must know where to look to find out if the analysis you wanted was executed (the regression equation) and know how to handle it when it was not (via centering). Indeed, when I analyzed the data from our example, the statistics program I used refused to run an analysis with X, X^2, X^3, and X^4 in the equation. I had to center X, recompute the powered vectors, and rerun the analysis to find out that X^4 was not needed for the model (i.e., nonsignificant ΔR^2).

Finally, how do we interpret the partial regression coefficients in a polynomial regression analysis? The old rule of "b_k indicates the expected change in Y given a one point change in X_k, assuming the other independent variables are held constant" won't work because there is simply no way to manipulate X without also manipulating X^2. And what if the regression coefficients for X and X^2 have different signs? A gain on one is offset by a loss on the other. This is a more extreme version of the problem we observed when we first discussed interpreting a partial regression coefficient back in the multiple regression chapter. Long story short, there is no easy way to interpret regression coefficients with polynomial regression analysis.

Closing Thoughts

At the beginning of this chapter we discussed how polynomial regression analysis is really just linear regression at heart. Sure, we're using variables transformed in a nonlinear fashion, but it's still OLS regression. By making Y' a nonlinear representation of X, we allow the relationship between Y' and Y to be linear. If you want proof of that, just look at the graph of Y against Y' (recall that one of the cool things about regression is that $r_{YY'} = R_{YX_1 \ldots X_k}$). Figure 7 is a graph of Y against Y' for our regression equation of $Y' = -13.2 + 11.0X - .66X^2$. The regression equation, with it's squared scores on X, has straightened out the nonlinear relationship between X and Y into something thoroughly linear. A linear correlation of Y' with Y results in a correlation of .999, which is the same value we obtained when we regressed Y on X and X^2. One wouldn't think that a simple transformation of the scores on X allows OLS to do something that appears antithetical to its nature, but there it is.

FIGURE 7 Scatterplot of Y Against Y' Where Y' is from the Regression of Y on X and X^2



Moderated Multiple Regression

9

Finding a moderator variable is like finding gold.

Statistical gold.

Overview

Moderated multiple regression is a technique designed to test for the effect of a moderator variable. A moderator variable is a variable for which the relations between two other variables are different at various levels of a third variable. Perhaps a story could better convey the nature of a moderator variable than this awkward definition. Before I get to that story, it may help to know that a moderated relationship is similar to an interaction. The only real difference between the two is causality. Interactions imply causal relationships. Moderated relationships do not.

A Moderator Variable Story

Back in the day, there was a researcher named Edwin Ghiselli. Ed, as he was known by his friends (I'm guessing), was developing an ability test to predict job performance. The test was the independent variable. Job performance was the dependent variable. Ed gathered a sample of people, gave them the ability test, asked them some other questions, hired almost a hundred of them, and measured their job performance (Ghiselli, 1956). When Edwin correlated test scores with job performance he obtained a correlation of .22. Not good. Rather than give up and not get paid, Ed got creative. In addition to the ability test, Ed also asked applicants how interested they were in the job. He split the sample into two subgroups on the basis on this interest variable. There were about thirty people who gave the high-interest answer and about sixty people who gave the low-interest answer. For the thirty people in the high-interest subgroup, the correlation between ability test scores and job performance was .66. Ed didn't report the correlation for the sixty people in the low-interest subgroup but it would have been considerably less than .22. Let's just call it zero.

Let's recap what Ghiselli found. The correlation between ability test scores and job perform-

ance across all 93 people was weak ($r_{XY} = .22$). But when this sample was split into two subgroups on the basis of a third variable (interest in the job), he found that the relationship between ability test scores and job performance was stronger for some people ($r_{XY} = .66$) than for others (estimated $r_{XY} = 0.0$). Notice how the relationship between X and Y is different at different levels of the third variable (which we could call Z). This third variable is a moderator variable. Interest in the job moderates the relationship between ability and job performance. As a final point to this story, the subgroup correlations need to be significantly different to support our claims of a moderated relationship (Ed didn't report a significance test). Small, non significant differences between subgroup correlations (e.g., .25 versus .30) are likely the product of sampling error and do not indicate a moderated relationship. As such, they are not the slightest bit interesting.

Summarizing What We Know To This Point

Let's review what we can call the Ghiselli procedure for testing for a moderated relationship. First, divide the sample into subgroups on the basis of the presumed moderator variable. Second, compute correlations within each subsample. Third and last, test the difference between the correlations for significance. Which significance test do we use? The difference between subgroup correlations can be tested using the test for differences between correlations from independent samples (see Significance Test III in Chapter 3). A non significant result would indicate that the correlations are the same for both high and low scoring people on the third variable (and, thus, no moderated relationship).

So it's a fairly easy procedure. But wait, Ghiselli's moderator variable was dichotomous, making it obvious that we would split the sample into two subgroups. What if the moderator variable is

continuous? Do we dichotomize it? I'll answer that question with a question: Should we ever dichotomize a continuous variable? Of course not. So how are we going to test for a moderated relationship? We'll have to use a whole new procedure. And it will involve regression analysis. Anyone see that coming?

Moderated Multiple Regression Analysis

In the present scenario we have three continuous variables: a dependent variable (Y), an independent variable (X), and a moderator variable (Z). The regression-based test for a moderated relationship is as follows. First, create a product vector that is the simple product of scores on the independent variable and the moderator variable (i.e., $X \times Z$). Second, regress the dependent variable on the independent and moderator variables. Third, regress the dependent variable on the independent variable, moderator variable, and the product vector. Fourth and last, compute the change in R^2 and test with the familiar (at least, it should be familiar by now) ΔR^2 F test (shown below).

$$F = \frac{(R^2_{big} - R^2_{small})/(k_{big} - k_{small})}{(1 - R^2_{big})/(N - k_{big} - 1)}$$

If the change in R^2 is significant, then we conclude that there is a moderated relationship.

Note what happened when we computed ΔR^2. The first regression analysis indicates how well X and Z predict Y. This regression tells us how well X and Z, used in the normal way, predict Y. The second R^2 indicates how well X, Z, and the product of X and Z predict Y. Any change in R^2 is due to the addition of the product of X and Z to an equation that already had X and Z in it. This ΔR^2 reflects the unique value of the (non casual) interaction of the two variables in the prediction of Y.

Consider the following dataset.

Name	Y	X	Z	X × Z
Napoleon	4	2	9	18
Jones	5	10	5	50
Frederick	6	4	6	24
Muriel	7	5	6	30
Benjamin	9	6	7	42
Pilkington	9	6	8	48

Note how the product vector is the simple, well, product of scores on X and Z. Nothing fancy there. A quick analysis of this dataset shows that neither X nor Z correlates well with Y (both r_{XY} and r_{ZY} are less than .2). The regression of Y on X and Z results in an R^2 of .09. The regression of Y on X, Z, and the product vector yields an R^2 of .95. The change in R^2 is .86, a huge increase. The F test for ΔR^2 results in an F statistic of 34.5, $p < .05$. The regression equation, including the product vector, is $Y' = 14.63 + -2.5X + -1.56Z + 0.465(X \times Z)$.

A quick summary of the example. A regression of Y on X and Z, in the usual multiple regression fashion, resulted in a nice R^2 of .09. But the addition of the product of X and Z to the regression equation increased R^2 to .95. Notice that we didn't add a new variable. A product vector isn't a new a variable – it's just a different representation of the variables already in the equation.

Based on the results of this analysis, we conclude that we have a moderated relationship between Y, X, and Z. We could call Z the moderator variable if we want, but both X and Z qualify for the title. In a moderated multiple regression analysis, the distinction between independent and moderator variables is arbitrary.

What about predicted Y? How is that computed in a moderated relationship? It's nothing unexpected; just substitute the scores on X and Z into their respective places in the equation. We'll do an example for the first case in our example da-

taset. Napoleon has scores of 2 on X and 9 on Z. Inserting those scores into the regression equation yields:
$Y' = 14.63 + -2.5(2) + -1.56(9) + 0.465(2 \times 9)$. A little basic math tells us that predicted Y for Napoleon is 3.96.

Issues With Moderated Multiple Regression

There are a few issues with moderated multiple regression. They should sound familiar to readers of the curvilinear regression chapter. In fact, all three are exactly the same. The first issue relates to interpreting the partial regression coefficients in a moderated multiple regression equation. The usual multiple regression interpretation of partial regression coefficients states that, "b_k indicates the expected change in Y given a one point change in X_k, assuming the other independent variables are held constant." As with curvilinear regression, the problem with this is that there is no way to manipulate X without also manipulating the product vector (and vice versa). So there's no easy way to interpret the partial regression coefficients. That's unfortunate.

The second problem relates to colinearity. As with powered vectors in curvilinear regression, there can be high correlations between X and the product of X and Z. If this correlation is too high, statistics programs may not react well and will refuse to run the analysis we wanted. The solution to this problem is once again, centering. Centering is the transformation of raw scores on a variable into mean-deviation scores. If you're not comfortable with that, then just standardize (i.e., z scores) the scores on a variable. Standardization is centering plus a little more. As a reminder, when centering you have to center (or standardize) X and Z before computing product vector.

There is one other issue to discuss and that concerns non significant constituent variables. If you forgot, constituent variables are the compo-

nent variables of the product vectors (and power vectors). In the case of the product vector $X \times Z$, the constituent variables are X and Z. Well, what of these constituent variables? If the addition of the product vector is a significant addition to the equation, then the product vector stays in the equation. But what if you take a peek at the t test for any of the constituent variables and see that X, for example, is non significant? Do we drop X from the equation? The answer is no. If a product vector is in the equation, then its constituent variables must also stay in the equation, regardless of their significance.

Closing Thoughts

Why are moderator variables cool? Let's go back to Ghiselli. Forget the moderator variable for a second. Did his test predict job performance? Not very well. After he identified a moderator variable, did his test predict job performance? Yes, for some of the people. It predicted job performance well for those people with high scores on the moderator variable (i.e., high interest in the job). That's certainly a lot better than if we didn't have the moderator variable. If that doesn't make it sound cool for you, consider the second example. X and Z, used in the usual multiple regression fashion, can be used to predict Y with some effectiveness. Adding the product vector to the equation increases prediction strength from good to great. This thing that is so great about both of these examples is that with a moderator variable, we obtain a stronger relationship with the dependent variable without adding any new variables. No new tests. No new independent variables. Just using, in new ways, the same ones we already had lying around.

One last issue to consider. This chapter covered two ways to test for a moderated relationship. There was the Ed Ghiselli two-correlation method and the moderated multiple regression method. If you have two continuous independent

variables, then you should always use the moderated multiple regression method. If one of the independent variables is a dichotomous variable, then you can use the two-correlation method. But can you use the moderated multiple regression method with a dichotomous independent variable? You can. If you know how to dummy code the dichotomous variable…

Dummy Coding

10

Ever wanted to conduct a regression analysis with a categorical independent variable?

Overview

Here's the scenario: Your dataset consists of a continuous dependent variable and a dichotomous independent variable. Your hypothesis is that there are differences between the two groups. How do you analyze the data? If you answered *t* test or ANOVA, you are correct. But did you know that you can also use regression analysis? It's true. What if the independent variable has more than two categories? ANOVA still works, but *t* tests are out. What of regression? Yes, regression still works. In this chapter we'll explore how we can use regression analysis to analyze data with a categorical independent variable.

Continuous and Categorical Variables

Before we continue, it is important that we define the difference between a continuous and a categorical variable. We'll start with categorical variables. A categorical variable has various values, maybe just two, maybe more. The values assigned to these categories lack any order. There is no *more* or *less* with a categorical variable, just *different*. People with scores of 3 are different on the variable than people with scores of 4. In fact, the values of the scores are completely arbitrary. We could recode the data so that everyone who had a score of 1 now has a score of -11. And the people with scores of 2 now have π for their score. It would be weird, but we could do it that way. Even a dichotomous variable is categorical – it's just a simpler set of categories. Back in the day, gender was coded dichotomously as 0 or 1. It was up to the researcher to decide which is 0 and which is 1. Males could be 0 and females 1. Or the other way around. Or something completely different from 0 and 1. It matters not. There is no more or less with a categorical variable, only different.

As for continuous variables, there is an ordering to the scores; there is a *more* and *less*. In addition, there are no longer discrete categories (e.g.,

1, 2, 3, etc.). In theory, there is an infinite number of possible scores between 1 and 2. In practice, our measurement techniques do not allow for a true continuum of response options and are frequently constrained to the use of limited options (e.g., the five options on a five-point response scale). Even though the lack of a true continuum might be considered a fatal flaw, no one seems to mind, and we proceed with these *n*-point type scales with discrete categories as if they really are continuous. Finally, for you Stevens's scales fans, a true continuous variable must be measured at either the interval or ratio level of measurement.

Regression Analysis with a Categorical Independent Variable: The Wrong Way

As mentioned, regression analysis can be used to analyze data with a categorical independent variable. Let's consider an example dataset, analyze it with an ANOVA, and analyze it with regression analysis without changing the dataset.

First, our example dataset.

Person	Y	X
George	21	1
Tyrus	24	1
Mordecai	22	1
Joshua	31	2
Walter	30	2
Joseph	27	2
Cornelius	28	3
Theodore	26	3
Jack	29	3

Notice that there are three groups in our independent variable. We recognize that this is just one independent variable with three possible categories. It's not three independent variables. These three categories could represent three different treatments in our study. We could have coded these categories any way we wanted, but here they are

coded as 1, 2, 3. There is also a dependent variable. This dependent variable is a continuous variable. That's enough of an overview. Let's get to the analysis.

We'll start with an ANOVA since that is the traditional way to analyze such data. The ANOVA um… analysis (think about it) shows that there are significant differences among the three groups, $F = 13.37$, $p < .01$. Standard rules for interpreting ANOVA results lead us to conclude that at least one group's mean is different from the other group's means. Also, eta-square (i.e., η^2), an index of the strength of the relationship between the two variables (reflecting the differences between groups relative to total variability), is .82, indicating a strong relationship. If we had any planned comparisons to do among the groups, we'd do them now. Also, if we wanted to do any post hoc tests, we'd do those now as well.

Moving on to our regression analysis, we regress the dependent variable on the independent variable, coded as shown in the table as 1, 2, 3, and find that there is a significant relationship between X and Y, $F = 5.38$, $p > .05$. The strength of the relationship, as indexed by R^2, is .66. All of this is interesting. Even more interesting is the fact that none of this matches the results of our ANOVA in any way.

So that was fun. All nice and simple. Two ways to analyze this dataset but with different results. Wait a second, I just realized that this dataset was coded wrong. Groups 1 and 2 were mixed up. I have fixed the coding and listed the revised dataset below.

Person	Y	X
George	21	2
Tyrus	24	2
Mordecai	22	2
Joshua	31	1
Walter	30	1
Joseph	27	1
Cornelius	28	3
Theodore	26	3
Jack	29	3

A re-execution of the ANOVA yields the same results ($F = 13.37$, $p < .01$). Let's repeat the regression analysis and see if anything changed there. Hmm, that's odd. Now the F test is different, yet again, $F = .31$, $p > .05$. And the R^2 is different too, $R^2 = .21$. Why is all of this different? It's the same dataset. The only thing I've changed was how the categorical variable was coded. Which coding scheme was the right one? I'd like it to be the first one since I obtained better results that way. And why didn't the ANOVA results change when I changed the codes?

As you may have surmised by now, neither coding was correct as far as regression analysis goes. And neither was incorrect as far as the ANOVA goes. The reason for this is that the independent variable is a categorial variable. ANOVA treats a categorical variable as a categorical variable regardless of how it's coded. With ANOVA there is no more or less, just different categories. Group 2 doesn't have more of something than Group 1. However, with regression the independent and dependent variables are treated as continuous variables. Greater numbers mean more. When I changed the coding scheme, I changed who had more or less of whatever the independent variable measures as far as regression is concerned.

To take this to absurd lengths, consider the following coding scheme for our dataset.

Person	Y	X
George	21	-11
Tyrus	24	-11
Mordecai	22	-11
Joshua	31	3.14
Walter	30	3.14
Joseph	27	3.14
Cornelius	28	9000
Theodore	26	9000
Jack	29	9000

The ANOVA results were unchanged, but a regression analysis revealed the following: $F = .52$, $p > .05$, $R^2 = .26$. You know, if we can get creative enough, I'll bet we find a coding scheme that gets R^2 above .90. I hope it's obvious that such an exercise would be silly, and the results would have no connection to reality.

So what are we to do about regression analysis? It wants every variable to be a continuous variable. Do we just give up if we have a categorical variable? Of course not. (Otherwise, this would be a very short chapter.) We'll just have to find a new way to code it so that the categorical variable is treated as a categorical variable. And this new way involves vectors (again with the vectors) and is commonly called dummy coding. Actually, it is a family of techniques including dummy coding, effect coding, and orthogonal coding. All of these techniques involve the use of coded vectors.

Regression Analysis with a Categorical Independent Variable: The Right Way

Coded vectors, like powered vectors and product vectors, are not new variables; rather, they are new ways to represent existing variables. The coding scheme that we will use is dummy coding. As mentioned, there are other coding schemes, but dummy coding is the easiest to implement and understand.

Dummy coding works by representing the categorical with various coded vectors. To be specific, we will need $k - 1$ coded vectors, where k is the number of categories or groups in the categorical independent variable. For our example dataset, we have three categories, so we'll need two coded vectors. Each coded vector will consist of zeroes and ones. In essence, each coded vector is a dichotomous variable. Values are assigned to the coded vectors as follows: Everyone in the first group gets ones for the first vector and zeroes for the others, everyone in the second group gets ones in the second vector and zeroes for the rest, and so on, until the last group, which gets zeroes for all vectors. Dummy codes for our example data are given below. Note how people in Group 1 are no longer represented by a single score of 1 on X, but rather by scores of 1 on D_1 and 0 on D_2.

Person	Y	X	D_1	D_2
George	21	1	1	0
Tyrus	24	1	1	0
Mordecai	22	1	1	0
Joshua	31	2	0	1
Walter	30	2	0	1
Joseph	27	2	0	1
Cornelius	28	3	0	0
Theodore	26	3	0	0
Jack	29	3	0	0

Dummy coding is well suited for an experiment where one of the groups is a control group. Making the control group the last group allows for easy comparisons between the various groups and the control group. More on this later. For now, let's focus on how to analyze this with regression. The independent variable, X, has now been recoded into two coded vectors. To find the relationship between X and Y, we regress Y on the two coded vectors (i.e., Y on D_1 and D_2). We do not include X, the original categorical variable, in this analysis – the whole point of dummy coding is to use these coded vectors in place of X.

Enough with the preamble, how did it work? A regression of Y on the two coded vectors resulted in an R^2 of .82, which is significant, $F = 13.37$, $p < .01$. Where have I seen those numbers before? Oh yes, that's the same F statistic we obtained from the ANOVA (same p value too). And the R^2 is a match to the eta-square. Our two effect size indicators (R^2 and eta-square) are the same, as are our two significance tests. So we've done it. With dummy coding, we were able to take a categorical variable and represent it in a way that forces regression to treat it as a categorical variable instead of a continuous variable. Pretty cool.

Now that we've passed that hurdle, what about the regression equation? Is that useful? Well, yes it is. Here's how. The y-intercept is the mean of the last group. Big deal, you say. There are twenty other ways, all much easier, to compute the mean score on Y for a given group. Well, there's more. The regression coefficients for each coded vector indicate the difference between that group's (whichever group is coded as 1.0 for that vector) mean and the last group's mean. For example, b_1, which is the regression coefficient for the first coded vector (i.e., D_1), indicates the difference between the means of Group 1 and the last group. b_2 indicates the difference between the means of Group 2 and the last group. Big deal, you say, anyone can compute group means and the differences

between them. You can even do the difference part with a pencil on the back of an envelope, you say. Let me finish. Here's the kicker: The significance test of the regression coefficient tells us if this difference is significant. Checkmate, Mr. Complainer.

Let's look at our example data to understand these features. For our example, the regression equation is $Y' = 27.67 - 5.33D_1 + 1.67D_2$. Thus, the mean of the third group (the last group) is 27.67. As for the other groups, Group 1's mean is 5.33 points less than Group 3's mean (because the regression coefficient for D_1, which corresponds to Group 1, is -5.33), and Group 2's mean is 1.67 points greater than Group 3's mean ($b_2 = 1.67$). What about the significance tests of the regression coefficients? Here are the results: b_1 is significant ($p < .01$ for t test of b_1), b_2 is not ($p > .05$ for t test of b_2). Thus, we can conclude two things. First, because the F test of R^2 is significant, there are differences among the groups, a standard ANOVA conclusion. Second, because b_1 is significant, Group 1 is different from Group 3, a standard planned comparison/post hoc test type conclusion.

Other Dummy Coding Details

You may have a few questions at this point. Like, how do we obtain predicted Y? And does the squared correlation between predicted Y and actual Y equal what we obtained from our dummy coded regression analysis? Let me assure you that everything we learned from the old days of regression analysis still applies to the dummy coded days of regression analysis.

Let's start with predicted Y. How do we obtain this? The answer is like before: Plug in scores on X (this time the dummy coded vectors) and solve for Y. For the first person in the dataset, George (a member of Group 1 with scores of 1 for D_1 and 0 for D_2), this works out as follows:

$$Y' = 27.67 - 5.33(1) + 1.67(0)$$

Solving for Y' results in a predicted Y of 22.33. All very easy. Wait a second, I just realized that 22.33 is also the mean for Group 1. Why? Remember how we discussed that a is the mean of the last group and b_1 is the difference between Group 1's mean and the last group's mean? Once you put those two coefficients together, you have the mean of Group 1.

What about predicted Y for someone in Group 2? Just insert Group 2's coded vector scores:

$$Y' = 27.67 - 5.33(0) + 1.67(1)$$

Solving for Y' leads to 29.33. And finally, what about Group 3? Group 3's coded vector scores are 0 and 0. Inserting these values into the equation gives you:

$$Y' = 27.67 - 5.33(0) + 1.67(0)$$

Which of course becomes 27.67. The full dataset with predicted Y scores is listed below.

Person	Y	X	D_1	D_2	Y′
George	21	1	1	0	22.33
Tyrus	24	1	1	0	22.33
Mordecai	22	1	1	0	22.33
Joshua	31	2	0	1	29.33
Walter	30	2	0	1	29.33
Joseph	27	2	0	1	29.33
Cornelius	28	3	0	0	27.67
Theodore	26	3	0	0	27.67
Jack	29	3	0	0	27.67

What about the correlation between predicted Y and actual Y? Just like with every other use of regression, that correlation is the same as what we obtained from a regression analysis. In this case, when squared, it's .82, the same value that we obtained earlier.

What's the message here? Predicted Y works just like before. And predicted Y is the same as the

mean *Y* score for a given group. (This, by the way, unlocks one of the secrets of regression analysis: The hidden meaning of predicted *Y*. Predicted *Y* is not just the score we predict for a person with a given score on *X*. It's also the mean of *Y* scores for all people with a given score on *X*.) We won't bother computing residual scores, but that's the same as before too.

The Special Case of the Dichotomous Independent Variable

Do we have to dummy code when the independent variable is dichotomous? If this dichotomous independent variable is coded as 0 and 1, then it is already dummy coded. Think about it: What's your code if you're in Group 1? One. And what's your code if you're in Group 2? Zero. The *y*-intercept still gives you the mean of the last group (i.e., the group coded as 0). The regression coefficient still tells you the difference between the last group and the group coded as 1. And the *t* test for this regression coefficient still tells you if this difference is significant. So, every dichotomous variable coded as 0 and 1 has already been dummy coded and yields all of the associated benefits.

Other Coding Schemes

Earlier I mentioned that dummy coding is just one of a family of coding techniques that we can use for our categorical independent variables. Also mentioned were effect coding and orthogonal coding. How are they different? Not much. Are we going to discuss them? Not really. Effect coding is pretty much pointless – it's almost identical to dummy coding but with different subgroup comparisons. Orthogonal coding allows the researcher to test any conceivable comparison among specific groups (i.e., Is the average of Groups 1 and 3 greater than than Group 2?), something that can be done with a standard ANOVA via planned comparisons. It's a nice bonus, but the main value of

any coding scheme is the enabling of regression analysis to be used with a categorical independent variable. The extra stuff (specific comparisons among the various groups) is superfluous.

Closing Thoughts

We've seen how to use dummy coding to enable regression analysis with a categorical independent variable, something that can be analyzed quite well with ANOVA. If you think about it, this is actually more work than ANOVA. Seems like a waste of time. What's the point? The point is this: We're just setting the stage here. Sure, there's nothing special about using regression analysis to replicate the functions of ANOVA. But can an ANOVA handle a continuous variable (including evaluating whether a linear relationship exists)? Under limited circumstances*, yes. But even for those conditions analysis with ANOVA is rather cumbersome. (To make matters worse, ANOVA is sometimes employed when the independent variable is wholly unsuited for it, resulting in a categorization of an unmapiulated continuous variable – something no one should ever do).

What is the point of this discussion of continuous variables in a chapter on categorical variables? What if you have a dataset with a continuous independent variable *and* a categorical independent variable? ANOVA is definitely not the best tool for the job. Regression is the way to go.

*What circumstances? Where the independent variable has discrete values (e.g., 0 minutes, 30 minutes, 60 minutes – no values in between – with n people at each level) assigned by the researcher.

ATI and ANCOVA

??

Categorical and continuous independent variables in the same equation? Madness.

Or is it?

Overview

In this chapter, we will combine procedures from previous chapters to build something new: the analysis of data with both categorical and continuous independent variables. Depending on the research design (e.g., true experiment, quasi experiment, non experiment), there are a number of names and goals for these procedures. The two most prominent are Attribute-Treatment Interaction (ATI) and Analysis of Covariance (ANCOVA). Before we get to those, we must first lay the foundation.

Two Types of Continuous Variables

There are two types of continuous variables in the data analysis world. One type is where the continuous variable is measured as a continuous variables, with more or less all possible values. For an example of this type consider height. We may not measure height as a true continuum (where no two people have the exact same height), but our measure of height approximates a continuum to some degree (e.g., rounding to the nearest millimeter). Contrast this to the other type of continuous variable, an independent variable in a true experiment where selected values along the continuum are chosen by the researcher. For example, people in a learning experiment could be assigned to one of four levels of study time: 0, 1, 2, or 3 hours. The continuous independent variable has four clear categories and these categories are the result of a decision made by the experimenter when he or she assigned people to various levels of study time (using terminology from the old days, we would call this a fixed variable), four distinct levels with no values in between. We don't need to do any artificial categorizing; the data already exist in four clean categories. To summarize, there are two types of continuous variables: those that are not manipulated and are (more or less) measured as a true continuum and those that are assigned

and only include selected values from the continuum.

Continuous Variables and ANOVA

As mentioned in earlier chapters, ANOVA is designed to analyze the relationship between a continuous dependent variable and a independent variable (or variables) with discrete levels (either a categorical variable or a continuous variable with values assigned by the researcher). How well does ANOVA handle a continuous independent variable? Well, if it's a variable manipulated by the researcher with selected levels chosen by the researcher, then ANOVA can handle this just fine. A plain vanilla analysis of variance will tell you whether there are differences on the dependent variable for different levels of the continuous independent variable. All well and good. But wait, if you analyzed that variable with regression analysis, you would be able to determine whether there were not just differences on the dependent variable but whether there was a linear trend to the data (a fairly important issue given that this is a continuous independent variable). In fact, regression pretty much exists to do this sort of thing. (That said, people develop habits and like to use their favorite tool for every purpose. Can ANOVA be modified to determine whether there is a linear trend to the data? Yes, but it is definitely more trouble than a regular regression analysis.)

To summarize, at this point we can say that for a certain type of continuous independent variable (one where people are assigned to various pre-determined levels of the continuous variable), you can properly analyze the data with ANOVA. But I don't know why anyone would want to when we live in a world where regression analysis exists.

But what of the other type of independent variable, the one that is measured as a truly continuous variable? Well, there is a way to do that with

ANOVA, but doing so requires one to force an artificial categorization upon it, a grievous sin. Why? Because categorizing a continuous variable results in the removal of important information. Just as an example, consider height. Height is a continuous variable. Now, let's dichotomize it. All scores greater than six feet will be coded as tall (numerically, we could use a 1 for the code), and all scores less than six feet will be coded as short (we could use 0 for these). Thus, someone with a height of 5 feet 11 inches is treated the same as someone at 5 feet 2 inches; both are scored as 0. These very different scores are being treated as if they were the same. Of course, the same problem applies to the 6 foot 1 inch person and the 6 foot 5 inch person. When data are dichotomized (or sliced into as many categories as you want), we throw out real information. This information removal reduces the accuracy and sensitivity of our measurement and analyses. Plus, the data no longer represent reality very well. Dichotomizing or categorizing continuous variables just so that we can analyze them with ANOVA results in a far worse analysis than keeping them continuous and analyzing them the right way.

So if the continuous variable is not one with discrete, selected values (e.g., 0, 1, 2, or 3 hours) assigned by the researcher, then forget about ANOVA – if you want to analyze data where a continuous variable is measured in all of its continuous glory, then regression is the only game in town. And if we have a continuous independent variable *and* a categorical independent variable in the same analysis, regression analysis is the best tool for the job. We'll keep the continuous variable as it is and dummy code the categorical variable.

Continuous and Categorical Independent Variables: Testing for Interactions

The previous two chapters introduced us to testing for moderation (Chapter 9) and dummy

coding (Chapter 10). In our discussion of moderated multiple regression, we mentioned that moderation means almost exactly the same thing as interaction. The only difference is one of causality: interaction implies causation, whereas moderation does not. Thus, we will test for interactions in the same way that we tested for moderation: create a product vector, regress the dependent variable on the two independent variables, regress the dependent variable on the two independent variables and the product vector, and test the change in R^2 for significance.

To this procedure we're going to add dummy coding (see Chapter 10 for a refresher on the process of dummy coding). The combination of these two procedures, interaction testing and dummy coding, will allow us to test for interactions and main effects for research designs that include both continuous and categorical independent variables. In essence, this will allow us to conduct the classic multi-factor ANOVA-type analysis (check for interactions first, then check for main effects for each independent variable) for data situations that can not be addressed properly via ANOVA.

The steps for testing for an interaction in multiple regression are quite familiar by now. What follows are those steps, customized for data with a continuous and a categorical independent variable.

I. Dummy code the categorical independent variable.
II. Create product vector(s). This will be the simple product of each coded vector with the continuous independent variable. And yes, if there are multiple coded vectors, then there will be multiple product vectors.
III. Regress the dependent variable on the continuous independent variable and the dummy coded categorical independent variable.
IV. Regress the dependent variable on the continuous independent variable, the dummy

coded categorical independent variable, and the product vector(s).
V. Compute the change in R^2 and test it for significance using the familiar F test for ΔR^2.

$$F = \frac{(R^2_{big} - R^2_{small})/(k_{big} - k_{small})}{(1 - R^2_{big})/(N - k_{big} - 1)}$$

Continuous and Categorical Independent Variables: Testing for Main Effects

The test for main effects is even simpler. There are two independent variables, so there are two main effects to test. However, for reasons that will be clear later, the main effect with which we are more concerned is likely the one for the categorical independent variable. The steps for testing for the main effect for the categorical independent variable are:

I. Make sure that there is not an interaction. Yes, this means that the interaction test is step one for a main effect test.
II. Regress the dependent variable on the continuous independent variable.
III. Regress the dependent variable on the continuous independent variable and the dummy coded categorical independent variable.
IV. Compute the change in R^2 and test it for significance using the familiar F test for ΔR^2.

So those are the procedures for testing for interactions and main effects on data with both categorical and continuous independent variables. We will reference these two procedures multiple times in the discussion of how they are applied to various research designs.

Treatments-By-Levels Design

Just to be clear, this *treatments-by-levels design* name is not a standard name. Not many people

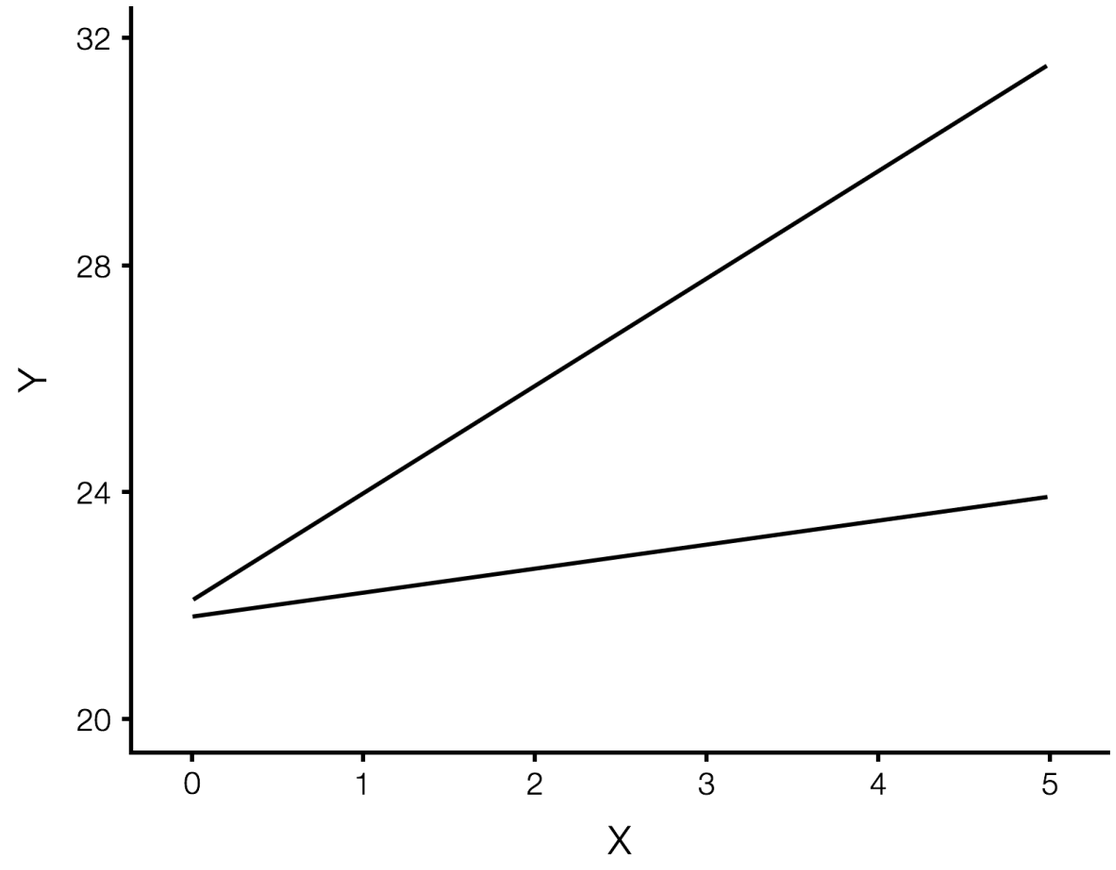
use it, but I think it describes the design of the previous example fairly well. To refresh our memories, the situation is a true experiment where people are randomly assigned to various classes (or groups) of the categorical independent variable (i.e., the treatment) and various pre-determined levels of the continuous independent variable. Thus, the values on the continuous variable do not have a continuous range from some value to another; rather, they have clearly separated levels.

So that's the setup for a treatments-by-levels design. How do we analyze it? Well, let's see, how do we handle a categorical independent variable in regression analysis? We dummy code it (or use effect or orthogonal codes, if that's your thing). And because there might be an interaction between our two independent variables, we'll need to test for that first.

The steps for testing for an interaction when there are both continuous and categorical independent variables were given at the start of the chapter. Long story short, if the addition of the product vector(s) to the model results in a significant R^2, then we have a significant interaction.

If the interaction is not significant, we may continue with analyses of main effects (e.g., Is there a significant relationship between the dependent variable and either of the continuous independent variables?). The test for main effects was also listed at the beginning of the chapter. Here's the short version: Regress the dependent variable on the continuous independent variable, regress the dependent variable on the continuous independent variable and the dummy coded categorical independent variable, test the change in R^2 for significance. If the change is significant, then there is a main effect for the categorical independent variable. To test for a main effect for the continuous independent variable, there are a couple of ways, but the shortcut version is to just examine the t test for the continuous independent variable in the pre-

FIGURE 1 Separate Regression Lines by Group: Slope Differences



Graph of the relationship between a categorical independent variable (with separate lines for each class of the categorical variable; top line is Group 1, bottom line is Group 0), a continuous independent variable (X) and a continuous dependent variable (Y). Note how the lines have different slopes.

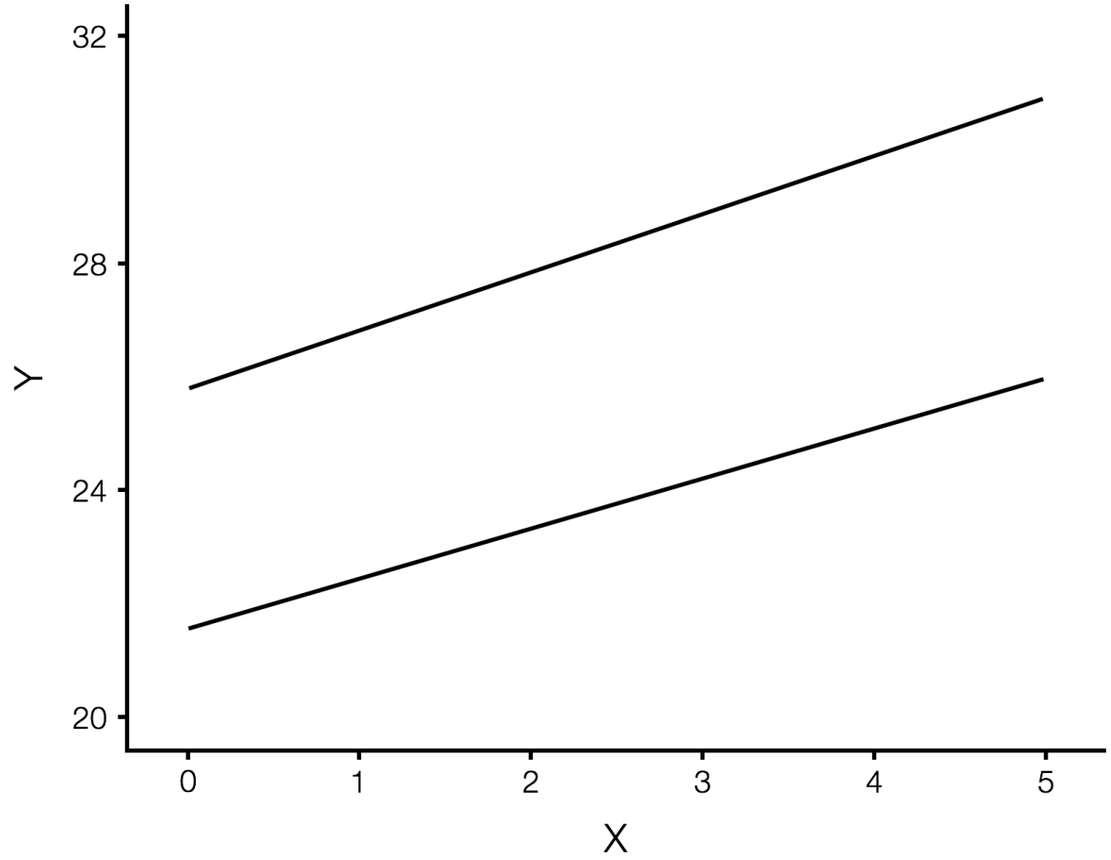
vious regression analysis (the one with both independent variables). (If you don't like shortcuts, here's the long version: Regress the dependent variable on the dummy coded categorical independent variable, regress the dependent variable on the categorical and continuous independent variables, and test the change in R^2.)

Let's get back to the interaction. What does an interaction mean in an analysis where one of the independent variables is continuous? It means that if we graph separate regression lines for each group, we would observe that the lines are not parallel (i.e., there are differences in slope). An example of a graph displaying an interaction is shown in Figure 1. Note how each class (i.e., group) of the categorical independent variable has its own regression line. As a point of contrast, Figure 2 shows a graph of data where there is not an interaction. Another way to think about interactions is this: An interaction means that the relationship between the dependent variable and continuous independent variable is different (different regression coefficients) for different categories of the categorical independent variable.

And finally, what if the categorical independent variable has more than two groups? In our learning theory example we had just two training conditions. Let's say there are three conditions. How do we analyze this? The steps are still the same: dummy code, create product vectors, regress, regress again, and analyze ΔR^2. The only difference is that we'll have more coded vectors (two coded vectors in the three-condition example) and more product vectors (two of these for this example). There is a product vector for every combination of coded vectors and continuous independent variables (e.g., $D_1 \times X$ and $D_2 \times X$; where X is the continuous independent variable).

An example dataset showing coded vectors and product vectors is given below. Let's say that the group to which each person is assigned is one of three different reward conditions and that the continuous independent variable is study time in minutes. The dependent variable is the score on a recall test. Group has been dummy coded into two

FIGURE 2 Separate Regression Lines by Group: Same Slopes, Different Intercepts



Graph of the relationship between a categorical independent variable (with separate lines for each class of the categorical variable; top line is Group 1, bottom line is Group 0), a continuous independent variable (X) and a continuous dependent variable (Y). Note how the lines have the same slopes but different *y*-intercepts.

coded vectors (D_1 and D_2). Product vectors have been formed.

Person	Y	Group	X	D_1	D_2	$D_1 \times X$	$D_2 \times X$
Cutler	22	1	25	1	0	25	0
James	21	1	50	1	0	50	0
Hector	13	2	25	0	1	0	25
Edward	31	2	50	0	1	0	50
Bill	13	3	25	0	0	0	0
Mercer	12	3	50	0	0	0	0

All that remains are the two regression analyses. The first analysis is the regression of Y on X, D_1, and D_2. The second is the regression of Y on X, D_1, D_2, $D_1 \times X$, and $D_2 \times X$. The difference between the two R^2 values indicates the magnitude of the interaction effect. The ΔR^2 F test determines the significance of it.

Attribute-Treatment Interaction

So that takes care of how we analyze data in a treatment-by-levels design. To refresh, the treatment-by-levels design occurs in a true experiment where people are randomly assigned to various, pre-determined classes or levels on both the continuous and categorical independent variables. The analytic procedure for regression analysis was not too complicated. Now it's time to move to a new area, one where people are not randomly assigned to various levels on the continuous independent variable. No longer a fixed variable, this variable is a random variable. Analysis of this design is called an attribute-treatment interaction, or ATI. It's also called an aptitude-treatment interaction. I'd like to call it a few other names, but let's just stick with ATI. The good news is that as far as our analysis goes, nothing changes from the days of the treatment-by-levels design. Exact same analyses.

An ATI analysis can be done with data from two types of experimental designs. Before I describe these, let me reiterate that in these designs people are not assigned (randomly or otherwise) to various levels of the continuous independent variable. A person's score on the continuous independent variable is a property of the person – the person brings that score to the table. We don't assign it. If that's clear, let's get back to the experiment types. First, there is the true experiment, where people are randomly assigned to the various classes (i.e., groups) of the categorical independent variable by the experimenter. Second, there are the quasi experiment and the non experiment (both designs lumped together for this discussion) where people are not randomly assigned to the various classes (i.e., groups) of the categorical variable by the experimenter; whatever manipulation being done here (and there is no manipulation with the non experimental design) is done with pre-existing groups.

There is some good news and bad news about these different research designs. The good news is that the ATI analytical procedure is the same as it was for treatments-by-level regardless of the experimental design (true experiment, quasi experiment, and non experiment). The bad news is that drawing conclusions from these analyses is much more difficult with the quasi and non experimental designs. Which is pretty much always the case with those two designs. They've always been troublemakers.

So why the weird names for ATI? It relates to the continuous independent variable, a variable which is not manipulated by the experimenter. Scores on this variable are a property of the person, like height or intelligence. No one assigned people to a height of six feet or an intelligence of 110. These scores are an attribute of the person. A common study using this design involves a treatment of some sort being performed on people with varying levels of the continuous variable. The

question in an ATI analysis is whether there is an interaction between the characteristic of the person (the continuous variable) and the treatment (the categorical variable) as they relate to the dependent variable.

Two ATI Examples

Consider the following basketball-themed example. We could perform a study on how two techniques for shooting free throws (the treatment) for people of various heights (the attribute) relate to free-throw percentage (the dependent variable). We might think that there will be difference in the effectiveness between these two techniques. We might also think that there will be an effect for height such that taller people will be more successful. Neither of these ideas represent an interaction; they are both main effects. The interaction would be something like: Taller people perform better with Technique A, but shorter people perform better on Technique B. The ATI analysis tests for this interaction.

And for that second example, let's say we want to explore whether there is a difference in achievement test scores between two school types (i.e., private versus public schools). School type is our categorical variable; however, in this study no one is assigned, randomly or otherwise, to one school type or another. Our continuous variable could be something like student intelligence, an attribute of the person. This research design clearly falls into the quasi/non experiment category. Once again, we are interested in whether there is an interaction between the categorical and continuous independent variables. Perhaps the relationship between student intelligence and scholastic achievement is different for the two school types. Maybe the intelligence-achievement relationship is stronger for students at public schools than for private schools.

ATI Details

We already know how this analysis proceeds from multiple previous parts of the chapter, so there is no need to list it again. Suffice it to say that it ends with a change in R^2 significance test. If this change in R^2 is significant, then we have a significant interaction – the regression lines, when computed separately for each of the groups, do not have the same slope (see Figure 1 for an example of slope differences). It's just the same old procedure, only applied to different research contexts.

You may be wondering how exactly we obtain separate regression equations for each class (i.e., group) of the categorical variable. There are a few ways. I'll share two. First, we could do exactly what the description suggests: Perform a regression analysis on members of each class separately. That is, we regress the dependent variable on the continuous independent variable for members of Group 1 only. From this we obtain a regression equation. We repeat this procedure for the rest of the groups. Each of these regression equations describe how the continuous independent variable relates to the dependent variable for members of a given class of the categorical independent variable. As mentioned, a significant interaction means that the slopes are different among the different groups.

As for the second method for obtaining separate regression equations, we can compute them from the overall regression equation. By overall regression equation, I mean the one containing the interaction term. Consider the following equation:

$$Y' = 27.1 + 5.2X + 1.6D_1 + 13.4(D_1 \times X)$$

Where:
X is the continuous independent variable.
D_1 is the dummy coded vector for the categorical independent variable (because there are only two groups in this example, only one coded vector was necessary).

$D_1 \times X$ is the product vector of the categorical and continuous independent variable.

So how do we get those separate regression equations? By substituting the codes for the categorical variables and simplifying the equations. We'll start with the easy one first. Using terminology from our basketball example, let's say that the Technique A group is coded as 0 for D_1. Substituting 0 every instance of D_1 in the equation yields:

$$Y' = 27.1 + 5.2X + 1.6(0) + 13.4(0 \times X)$$

Because zero times anything is zero (the greatest property in mathematics – I'll fight anyone who says otherwise), this simplifies to:

$$Y' = 27.1 + 5.2X$$

Members of the Technique B group are coded at 1 for D_1. Thus, every time we encounter a D_1, we'll insert a 1.

$$Y' = 27.1 + 5.2X + 1.6(1) + 13.4(1 \times X)$$

Which simplifies to:

$$Y' = 28.7 + 5.2X + 13.4X$$

But wait, we can simplify a bit more by combining X terms:

$$Y' = 28.7 + 18.6X$$

Thus, the regression equation describing the relationship between the continuous independent variable (height) and free-throw percentage for people taught how to shoot with Technique A is $Y' = 27.1 + 5.2X$ and the equation for people taught with Technique B is $Y' = 28.7 + 18.6X$. Note how the y-intercepts for these regression lines are about the same, but the slopes are very different. The relationship between between the continuous independent variable and the dependent variable is different for members of different classes of the categorical independent variable.

ATI Assumptions

The most important assumption of the ATI model is that there is not a causal relationship between the treatment (categorical independent variable) and control attribute variable (continuous independent variable). To state this differently, a person's standing on one independent variable does not affect, or play any causal role, in his or her standing on the other variable. In a treatment-by-levels true experiment (random assignment to groups), a causal relationship between the independent variables isn't possible as a person's standing on both variables is under control of the researcher (assuming a fully-crossed design). In the version of ATI where people are randomly assigned to classes of the categorical (but not the continuous) independent variable, it is possible that the treatment could affect people's scores on the attribute variable (although the converse isn't possible in this design). Just think how easily that could happen if the attribute variable was a personality measure. Such an effect would be a serious problem for our analyses. Fortunately, there is an easy solution: Measure the attribute variable (the continuous independent variable) before exposing people to the treatment. Long story short, do not measure the continuous independent variable after administering the treatment to people in the experiment.

That's the good news version of this. What about the bad news? If this is a quasi or non experiment, where people are not assigned to either variable, then there is no way to be certain that one variable didn't affect the other. Like we've said many times before, life is tougher with those designs. The analysis may be the same, but drawing a conclusion is a more difficult proposition.

Analysis of Covariance

Alright, so enough about interactions. If this were a two-factor design in an ANOVA, and your interaction test was non significant, what would you do next? You would examine main effects. Well that's what we'll do with regression. The only difference is that when one of the variables is a continuous variable (and is considered to be a control variable), as is the the case with the attribute variable in an ATI design, the analysis of main effects takes on a special significance. This procedure is called Analysis of Covariance, or ANCOVA. To understand how this works, let's go back to our two examples from our ATI discussion, the basketball example and the scholastic achievement example.

Two ANCOVA Examples

First, the basketball example. The dependent variable is free throw percentage. The independent variables are shooting technique and height. People were randomly assigned to one of two shooting techniques. No one was assigned a height (they brought that score with them). So we tested for an interaction (the ATI analysis) and found none. Time to look for main effects. The big issue concerns whether there is a difference in free throw percentage rate between the two shooting techniques *after controlling for height*. It's that last part that's important. We measured height for a reason – we thought it might be related to success and we wanted to control for it. But wait, you say, weren't people randomly assigned to groups? And as such, shouldn't the heights of the members of these groups be about the same? And wouldn't that make height pretty much an irrelevant variable? The answers are yes, yes, and yes. And maybe another yes. I lost track of how many ques-

tions you asked. So why go to trouble of measuring and controlling for height? The answer is that there are still likely to be differences between the two groups (small though they may be) and controlling for these differences increases the precision of the analysis. Random assignment to groups makes the groups equal on every variable *in theory*. In reality, the groups will not be exactly equal. Flip a coin a hundred times. The mathematical expectation is that you will obtain 50 heads. In theory. In practice, you are not very likely to observe exactly 50 heads. But you are very likely to see something close to 50 heads. Long story short, random assignment isn't perfect. ANCOVA can be used to pick up the slack.

You see, we could analyze the data from our basketball study a different way. We could just forget the height variable and analyze the data with an ANOVA. Think about it. Without the continuous independent variable, all we have left is a dichotomous independent variable and a continuous dependent variable, the very task for which ANOVA was designed (because there are only two groups in this example, we could even use a *t* test). But by controlling for this variable (height), we remove some irrelevant variance and end up with a more precise analysis. We are actually more likely to find an effect for training technique by controlling for this variable (with ANCOVA) than by ignoring it (ANOVA).

As for our other example, the scholastic achievement example, the concept is the same as it was for the basketball example. Drawing a conclusion is a little tougher. The control variable is student intelligence. The categorical independent variable is school type: public or private. The research question is whether one type of school leads to better student achievement after controlling for student intelligence. Because students are not randomly assigned to school type, they are likely to be very different in terms of many individual difference variables, like intelligence. In this

analysis, ANCOVA is used to control for these pre-existing group differences. This means that the groups are different in terms of the continuous independent variable and then are statistically adjusted so that those differences are removed. Thus, when we investigate scholastic achievement by school type, we might find a large effect for school type, with private school students outperforming public school students. But we also find that the private school students have higher intelligence test scores than the public school students have. When we control for this intelligence test score difference, we may find that the achievement test score difference between the schools has disappeared. We would then conclude that school type, independent of student intelligence, is not related to student achievement. If, on the other hand, we still find a difference between the schools after controlling for intelligence, we would conclude that private schools, independent of student intelligence, do lead to higher student achievement than public schools. As cool as this process sounds, there are too many assumptions that must be met (and aren't met) for our conclusions to be sound. More on these assumptions later.

Considering our two examples, we can see the two reasons for conducting an ANCOVA. The basketball example, a true experiment, used ANCOVA to increase the precision of an analysis. The scholastic achievement example, featuring a non experimental design, used ANCOVA to control for pre-existing differences between groups as a way to make comparable what is not, in its native condition, comparable. The first purpose works well, and the second, although sounding cool beyond words, does not.

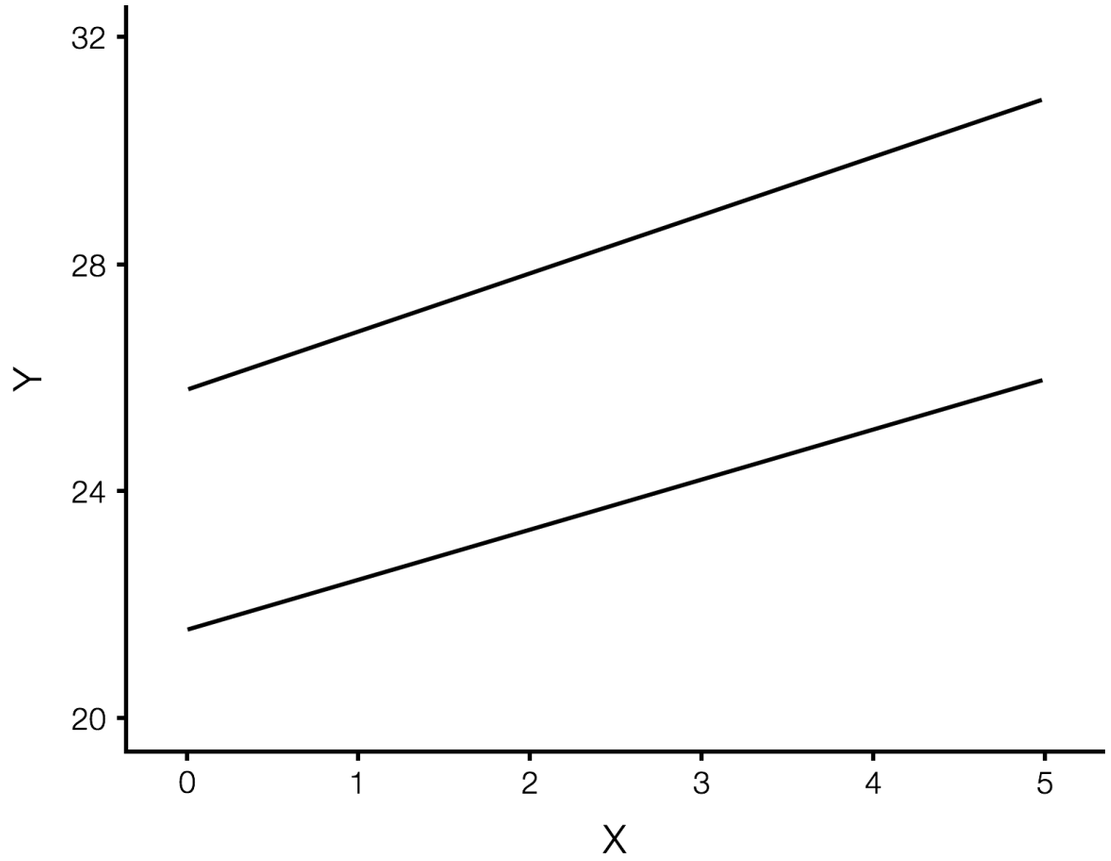
ANCOVA Details

So how do we execute the ANCOVA? It's nothing more than the test for main effects that we discussed at the beginning of the chapter. I'll briefly list the steps. First, and this is most important, we check the to see if there is an ATI. If we find an interaction (i.e., slope differences), we stop. We can't control for a variable if it doesn't have a consistent relationship with the other variables (the very nature of an interaction is that the relationship between two variables changes as a function of a third variable). This is the same logic that is found with any multi-factor ANOVA: Test for interactions first. With the ATI procedure; if an interaction is found, interpreting the main effects is greatly complicated (and pretty much impossible). If no interaction is found, then we can proceed with the ANCOVA proper: Regress the dependent variable on the continuous independent variable, regress the dependent variable on the continuous independent variable and the dummy coded categorical independent variable, and test the change in R^2 for significance. If ΔR^2 is significant, then we conclude that there are score differences on the dependent variable among our groups (i.e., the categorical independent variable) after controlling for scores on the continuous independent variable. In terms of regression lines, these differences would be observed as differences in the *y*-intercepts.

Figure 3 is displays regression lines from data where the ANCOVA was significant (this is the same graph from Figure 2 – I'm repeating it here for convenience). Note how the regression lines are parallel (i.e., have the same slope). The similar slopes are found when the ATI is non significant, a first step to conducting an ANCOVA. Also note how the *y*-intercepts are not the same; that's why the ANCOVA was significant.

As a point of contrast, Figure 4 displays regression lines taken from a dataset with a non significant ATI and a non significant ANCOVA. Note

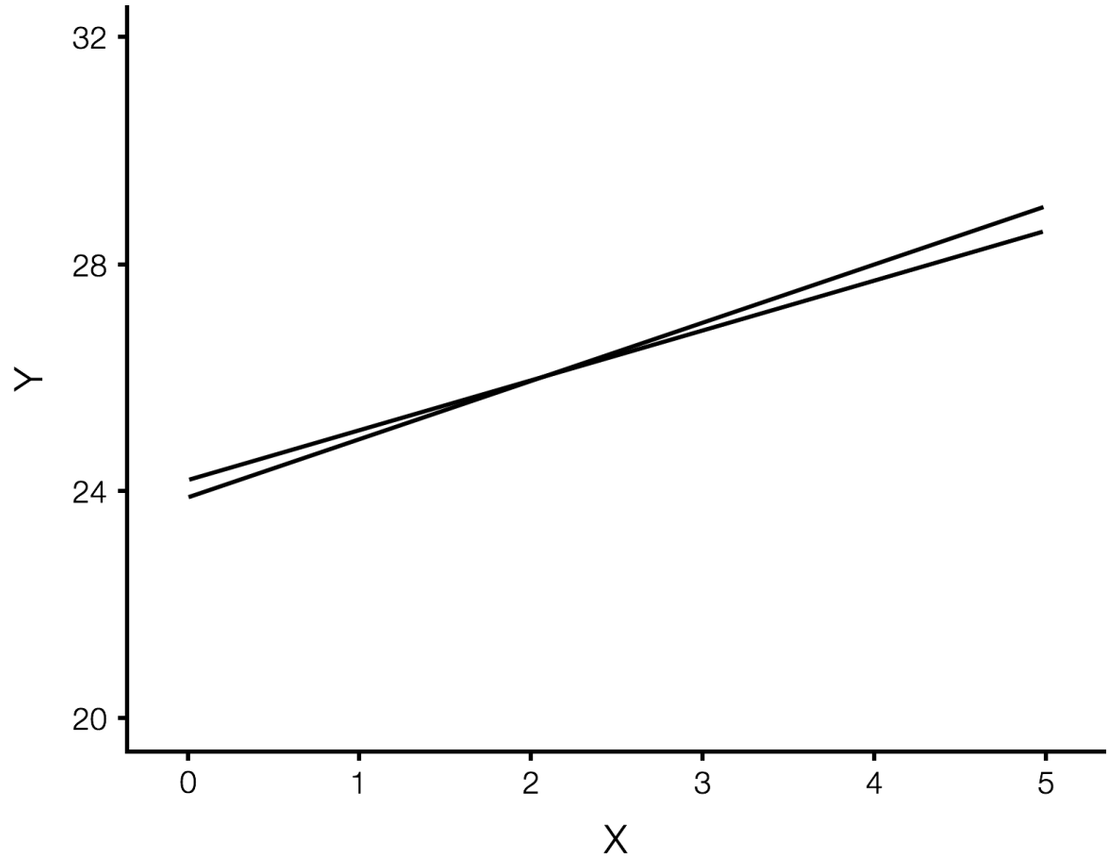
FIGURE 3 Separate Regression Lines by Group: Same Slope, Different Intercepts



Graph of the relationship between a categorical independent variable (with separate lines for each class of the categorical variable; top line is Group 1, bottom line is Group 0), a continuous independent variable (X) and a continuous dependent variable (Y). Note how the lines have the same slopes, but different *y*-intercepts.

how both lines have more or less the same slope and *y*-intercept. They're not exactly the same, but

FIGURE 4 Separate Regression Lines by Group: No Differences in Slope or Intercept



Graph of the relationship between a categorical independent variable (with separate lines for each class of the categorical variable), a continuous independent variable (X) and a continuous dependent variable (Y). Note how the lines have the same slopes and the same *y*-intercepts.

the differences are trivial. Hence, the non significant ΔR^2 F tests.

Let's get back to the big picture of the ANCOVA procedure. There's an R^2 based on one independent variable. Then there's an R^2 based on both independent variables. We're interested in the difference between these R^2 values to see if the added independent variable has a significant relationship with the dependent variable after controlling for the other variable. Where have we heard this before? This is the same concept as semipartial correlation. In fact, one of the ways we discussed for to computing a semipartial (a squared semipartial, to be exact) was with the following equation.

$$r^2_{Y(X.Z)} = R^2_{YXZ} - R^2_{YZ}$$

In this equation, Z is the control variable, and X is the independent variable. If Z is a continuous variable and X is a dummy coded categorical variable, then you have the ANCOVA procedure we just described. Y is regressed on Z. Y is then regressed on Z and X. The difference between the R^2 values indicates the relationship X has with Y after controlling for Z. Thus, the ΔR^2 from ANCOVA and the squared semipartial correlation are the same. Well, the procedures are the same. The goals of the study, the nature of the variables, and the interpretations of the results may be very different.

ANCOVA Assumptions

As mentioned above, there are two reasons why researchers conduct ANCOVAs. The first is to increase the precision of an analysis of data within a true experiment. The second is to control for the sort of pre-existing group differences that would be found within a quasi or non experimental design. And, as mentioned, there are so many assumptions that must be met that any conclusions drawn from an analysis done for the second purpose are not likely to be correct. What are these assumptions, you wonder. Careful what you ask for…

We'll start with the most important assumptions. One of these is something we have already discussed: no causal relations between the two independent variables. As mentioned, this is easy to establish if people are randomly assigned to various treatment classes and the continuous variable is measured before the treatment is administered. This is difficult to establish if the groups are preexisting (as is the case with the quasi and non experimental designs). So that's strike one against those designs.

A second assumption is that we have not omitted a relevant variable from the analysis. That is, did we control for all of the variables for which we should have controlled? If we left something relevant out, then we will conclude that there is too much of an effect for the categorical independent variable. This error is a type of specification error, if you like fancy names. The only way to know if you controlled for everything relevant is to research and identify the correct theoretical framework for the problem at hand. If, in the course of your research, you missed something important and didn't include that variable as a control variable, you're in trouble. Consider our public/private school example from earlier: The only control variable was intelligence. Now, intelligence is an excellent control variable in this situation, but don't you think there might be a few other variables that are also important? On what other variables might public and private school students differ? Maybe parental emphasis on scholastic achievement. Maybe access to tutors outside of school. Maybe the number of years spent in pre-school. Maybe about fifty other things. If our groups differ on one or more of these variables (i.e., correlate with them), and these variables have an effect on the dependent variable, then failing to control for them results in an overstated effect for the categorical independent variable. It is primarily for this reason that explanatory research within a quasi or non experimental design is so difficult.

It gets worse. Another assumption is that the independent variables are measured without error. This isn't much of a problem with categorical independent variables. But it is a big problem with continuous independent variables (unless it's a fixed independent variable in a true experiment, like study time). If a hundred years of measurement theory and research has taught us anything, it's that any time we measure anything, there's error. So that assumption is violated.

What are the effects measurement error in the independent variable? First, R^2 will be reduced, but that's not the bad part. The bad part is that when we use a poorly measured (i.e., too much random error) variable as a control variable, we do not make enough of an adjustment, and we conclude that the categorical independent variable has a greater effect on the dependent variable than it really does. Thus, the net result of this error is just like a model specification error (leaving our a relevant variable); it causes us to conclude that the categorical independent variable has a greater effect than it really does.

The remaining assumptions are child's play by comparison. They include nonlinearity, extrapolation errors, and the use of proxy variables. Let's focus on the last one. Proxy variables are variables that we measure instead of the actual variables that we should be measuring. Proxy variables are merely correlated with the variables we should be measuring. As an example, consider the study mentioned in the first chapter (Armor, 1972). The results shows that owning a refrigerator was correlated with student verbal achievement. It should be obvious that refrigerator ownership is a proxy for parental wealth. Although manipulating the real variable (wealth) may result in a change in verbal achievement (and I'm not saying that it would, just may), manipulating the proxy variable (by, say, buying people refrigerators) is extremely unlikely to result in a change in verbal achievement. Using a variable correlated with the actual rele-

vant variable is, in essence, another form of measurement error and will have the same result: an insufficient degree of adjustment and an erroneous conclusion regarding the role of the categorical independent variable.

After considering all of these assumptions and the profound unlikeliness that some of them will ever be supported, I think the big question is: Why does anyone use ANCOVA for adjustment? And when they do, why would they think that anyone else would accept the conclusions they draw from their analysis? I guess that's two big questions, but you get the idea.

Concluding Thoughts

In this chapter we've discussed how to use regression analysis to analyze data gathered within a variety of experimental designs. The common threads through these analyses are as follows. First, we have at least one continuous independent variable and at least one categorical independent variable. If all of the independent variables were categorical, then this would be an ANOVA chapter (although you could still do it with dummy coding and regression…). Second, regardless of experimental design and hypotheses, the first step is to check for an interaction between the independent variables and the dependent variable (i.e., the ATI analysis). Third, if no interaction is found, then we can examine main effects, including a type of analysis called ANCOVA where we examine whether there is a effect for one variable after controlling for another. Fourth, the procedure for these analysis (ATI and ANCOVA) involves dummy coding the categorical independent variable, creating product vectors (for ATI), conducting a regression analysis, repeating the analysis with a new term added to the regression equation (product vectors for ATI, coded vectors for the categorical independent variable for ANCOVA), and testing the change in R^2 for signifi-

cance. And finally, what we can conclude from these analyses depends greatly on our research design (true experiment, quasi experiment, or non experimental design). It's a simple matter to draw conclusions from the analysis of data collected within a true experimental design. It's much more difficult to draw sound conclusions when the design is quasi or non experimental.

In closing, have you ever considered that sometimes we earnestly desire to find an interaction, and at other times an interaction is the last thing we want to see? If we hypothesized an interaction, conduct our ATI analysis, and fail to find an interaction, we're disappointed. To the converse, if we hypothesized that there will be group differences after controlling for some variable, we test this hypothesis with an ANCOVA, which can only be done if there is not an interaction. If we find an interaction, we're disappointed because the interaction precludes us from conducting the analysis we desired and from testing the hypotheses we wanted to test. Life's funny that way.

References

12

If someone else hadn't already thought of it, I'd never figure it out.

References

Brogden, H. E. (1946) On the interpretation of the correlation coefficient as a measure of predictive efficiency. *Journal of Educational Psychology, 37,* 65-76.

Burket, G. R. (1964). A study of reduced rank models for multiple prediction. *Psychometrika Monograph Supplement,* No. 12.

Cohen. J. (1992). A power primer. *Psychological Bulletin, 112,* 155-159.

Ezekiel. M. (1930). *Methods of correlational analysis*. New York: Wiley.

Ghiselli, E. E. (1956). Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology, 40,* 374-377.

McNemar, Q. (1969). *Psychological statistics* (4th ed.). New York: Wiley.

Pearson, K. (1907). *On further methods of determining correlation*. London: Cambridge.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Pedhazur, E. & Schmelkin, 1991 *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Earlbaum.

Raju, N., Bilgic, R., Edwards, J., & Fleer, P. (1999). Accuracy of population validity and cross-validity estimation: An empirical comparison of formula-based, traditional empirical, and equal weights procedures. *Applied Psychological Measurement, 23,* 99-125.

Wherry, R. J. (1931). A new formula for predicting the shrinkage of multiple correlation. *Annals of Mathematical Statistics, 2,* 440-457.