11.35 r_z = .681 for p = .05. That's a large difference between .41 for women and .68 for men, before obtaining even the weak evidence of p = .05 of a population difference.

11.36 a. Diamond ratio = 2.0, which is large and matches how the diamond appears to change between the two models of meta-analysis; b. There is considerable heterogeneity, which is consistent with very large bouncing of the CIs in the forest plot, including a number of CIs that do not overlap at all.

11.37 The CIs closest to 1 are most asymmetric, as we expect.

11.38 a. The popouts explain that p refers to $\rho_0 = 0$. Nine of 16 studies have $p < .05$, and the overall result is $p < .001$, so we confidently reject the null hypothesis that $\rho_0 = 0$ and conclude there's a statistically significant positive correlation of amount of practice with achievement; b. Macnamara et al. estimated the correlation quite precisely and interpreted the value as indicating a correlation considerably smaller than the theory of Ericsson predicts; c. The estimation approach is more informative, and was used by Macnamara et al. to arrive at a conclusion that corresponds closely with the research question they were investigating.

12

Regression

In Chapter 11 we saw that correlation is a measure of the relationship between X and Y. Like correlation, *regression* is based on a data set of (X, Y) pairs, but it's different from correlation in that it gives an estimate of Y for a value of X that we choose. So correlation is a number that summarizes, overall, how X and Y relate, whereas regression takes a chosen single value of X and provides an estimate of Y for that X. Recall that Figure 11.1 was a scatterplot of Well-being (the Y variable) and Body Satisfaction (the X variable), for 106 college students. Figure 11.21 showed the separate scatterplots for women and men. Suppose Daniel scores X = 3.0 for Body Satisfaction: What Well-being score would we expect for him, assuming he comes from the same population of college students? We can use regression to estimate Y (Daniel's Well-being score) for X = 3.0. There are two steps:

1. Calculate from the data the *regression line for Y on X*.
2. Use that line to calculate an estimate of Y for X = 3.0.

Regression focuses on what X can tell us about Y. Almost always, X can tell us part of the story of Y, but not all of it. Informally, the full story of Y divides into two parts:

The story of Y = What X can tell us about Y + The remainder (12.1)

The informal story of Y.

First part, uses regression

Second part, what's left over

Regression is thus different from correlation, but the two are intimately linked. We'll see that X makes its contribution to the Y story (the first part) via the regression line, but it's r that determines how large this contribution is. If the correlation is large, X and the regression line give considerable information about Y; if small, they tell only a small proportion of the Y story.

I said that correlation is an effect size measure that has long been routinely reported and interpreted by researchers, which is excellent. For regression the news is even better, because researchers not only report and interpret regression effect sizes, but quite often report regression CIs as well—meaning they are already largely using the new statistics.

Here's the agenda for this chapter:

- The regression line for Y on X: minimizing the standard deviation of residuals
- Regression, correlation, and the slope of the regression line
- The proportion of variance accounted for: r^2
- Regression reversed: the regression of X on Y
- Assumptions underlying simple linear regression
- Confidence intervals and the uncertainty of estimation of Y
- A possibly strange natural phenomenon: regression to the mean

THE REGRESSION LINE FOR *Y* ON *X*

We use the ***regression line*** *of* Y *on* X to calculate an estimate for the *predicted* variable Y from a value of *predictor* variable X.

Figure 12.1 shows the scatterplot of Well-being (Y) and Body Satisfaction (X), for $N = 47$ men, as in Figure 11.21. First, some regression jargon. The line in Figure 12.1 is the ***regression line***. For correlation, as we discussed in Chapter 11, the two variables X and Y are interchangeable, meaning we can swap the labels X and Y and still calculate the same r. With regression, however, the two variables have different roles: X is the *predictor* variable and Y the *predicted* variable, also known as the ***criterion*** variable. Those terms reflect the frequent use of regression for prediction—we often regard the regression estimate of Y, for a particular value of X, as a prediction for Y. We speak of "the regression of Y on X" or "the regression of Y against X". (Later we'll consider the reverse, the regression of X on Y, for estimating X given a particular value of Y.)

For the regression of Y on X, X is the ***predictor variable*** and Y the ***predicted variable***, also known as the ***criterion variable***.

In Figure 12.1, the regression line of Y on X tells us that, if $X = 3.0$ for Daniel, then our best point estimate of Daniel's Well-being score is $\hat{Y} = 4.44$. We use $\hat{Y}$, which we say as "Y-hat", for the ***regression estimate***—the estimated value of Y calculated from the regression line. Later we'll find that there's large uncertainty in that estimate, which is hardly surprising given only a medium number of data points ($N = 47$) and the considerable scatter that gives a correlation of only medium size ($r = .53$).

The ***regression estimate*** is $\hat{Y}$, which is the value of Y calculated from the regression line for a particular value of X.

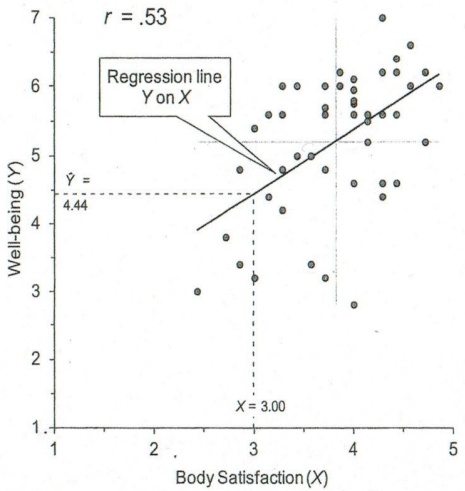


Figure 12.1. Same as Figure 11.21, left panel, for $N = 47$ men, with the addition of the regression line of Y on X. If Body Satisfaction is $X = 3.0$, the Well-being score calculated from the regression line is $\hat{Y} = 4.44$ as the dashed lines indicate. The cross of horizontal and vertical lines marks (M_X, M_Y), the

I'm talking about estimating Y for a particular value of X. That doesn't mean I'm making any assumption about X causing Y. Just as with correlation, there might be any pattern of causation between X, Y, and other variables, or no causation at all. With regression, as with correlation, we are working with the relationship between X and Y in some data set. There may well be interesting causation to investigate, but perhaps not.

Using regression to estimate Y from X makes no assumption about possible causal links between X, Y, and other variables.

As you can see in Figure 12.1, the regression line goes through the point (M_X, M_Y), the means of X and Y, which is marked in the figure by the cross of horizontal and vertical lines. The regression line is designed to give the best estimate of Y for any particular value of X. How is its slope (sometimes called its gradient) determined? I'll take two approaches to answering that central question. The first considers estimation error and how we can minimize it.

Minimizing the Standard Deviation of the Residuals

We want the best regression estimates we can get, meaning we want to minimize estimation error. By ***estimation error***, I mean $(Y - \hat{Y})$, which is the vertical distance between a data point and the line, as marked by the red vertical lines in Figure 12.2. The Y refers to the data point—one of the dots in Figure 12.2—and the $\hat{Y}$ refers to the other end of its vertical line, which lies on the regression line. There are N data points and therefore N of the $(Y - \hat{Y})$ values, which are also called ***residuals***, the idea being that, while the $\hat{Y}$ estimates calculated from the line tell us the first part of the story of Y—the first part in informal Equation 12.1—the second part is told by the $(Y - \hat{Y})$ values and is the remaining, or residual, part of the story. We'll come back to this idea.

$(Y - \hat{Y})$ is an ***estimation error***, or *residual*, where Y is an observed value for some value of X, and $\hat{Y}$ is the estimated value given by the regression line for that value of X.

A ***residual*** is $(Y - \hat{Y})$, the difference between the value of Y for a data point (X, Y), and $\hat{Y}$, the regression estimate for that value of X.

The regression line is selected to minimize estimation error. More precisely, it's selected so that the standard deviation of the residuals is minimized. Imagine rotating

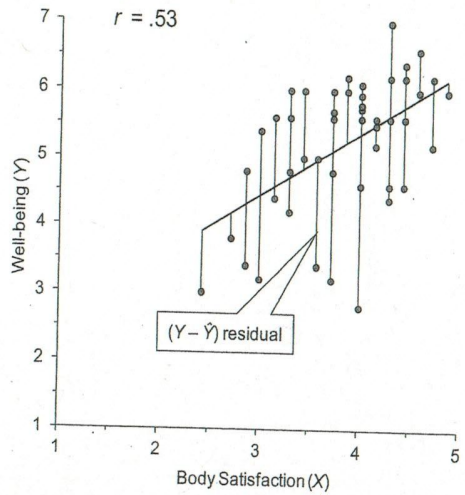


Figure 12.2. Same as Figure 12.1, but with

the regression line in Figure 12.2 about the point (M_X, M_Y) and noting the changes to all the residuals. Rotate it a little clockwise or counterclockwise and some of the red vertical lines become a little longer, others a little shorter. Almost certainly, the standard deviation of the residuals changes. In fact, the regression line in the figure is positioned so that the SD is as small as possible. Rotate the line either way and the SD increases—we'll see that happen in ESCI in a moment. Therefore, the regression line will, on average, give us better estimates than lines with larger or smaller slopes.

The SD of the residuals is written as $s_{Y.X}$, which is a measure of the variation of the data points from the line. The equation is

SD of residuals.

$$s_{Y.X} = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{(N-2)}} \qquad \begin{matrix} SS(\text{residuals}) \\ df \end{matrix} \qquad (12.2)$$

where the summation is over all N data points. The numerator under the square root is called the *sum of squares of residuals*, or SS(residuals). The denominator is the degrees of freedom, which is $(N-2)$. (Optional extra remark: The minus 2 reflects the fact that two degrees of freedom are used up by estimating both the intercept and the slope of the regression line, which are the determining features of the line, as we'll see in a moment.) I'll talk about minimizing the SD of residuals, on the left side in Equation 12.2. However, that's equivalent to minimizing the SS(residuals), the sum of squares of residuals, on the right side, which is what many textbooks discuss. In summary, the slope of the regression line is chosen to give us the best regression estimates based on our data set, and it does this by minimizing $s_{Y.X}$, the SD of residuals. Let's see how the minimization works.

I'll now shift to the small Thomason 1 data set, to keep things simple. Figure 12.3 shows the scatterplot for that data set, as in Figure 11.10, with the regression line and the red $(Y - \hat{Y})$ residuals displayed. At red 12 below the scatterplot is the equation of the regression line for Y on X:

Regression line, Y on X.

$$\hat{Y} = a + b \times X \qquad \begin{matrix} \text{Intercept} & \text{Slope} \end{matrix} \qquad (12.3)$$

esci **12.1** Fire up the **Scatterplots** page of **ESCI intro chapters 10–16**. If necessary, scroll right and click at red 16 to load the Thomason 1 data set. At the top, click at red 9 to reveal a panel with regression information. Click at red 10 to display the regression line, and at red 13 to mark the residuals with vertical red lines. Your screen should resemble Figure 12.3.

As you may know, the *intercept* is the value of $\hat{Y}$ when $X = 0$, and, therefore, is where the line, extended if necessary, intersects the Y axis. For the Thomason 1 data set, the estimates from the data—which are calculated by fitting the line so it goes through the means and minimizes $s_{Y.X}$, the SD of residuals—are $a = 4.22$ and $b = 0.78$. (As usual, I'm rounding the values displayed by ESCI.) That b value is the *slope* of the regression line. As you can see in Figure 12.3, the regression equation for predicting the posttest score (Y), given a pretest score (X), is:

Regression line for Thomason 1.

$$\hat{Y} = 4.22 + 0.78 \times X \qquad (12.4)$$

Now consider what we want to minimize—the SD of the residuals, which for the regression line is reported near red 13 to be $s_{Y.X} = 1.37$. Let's investigate

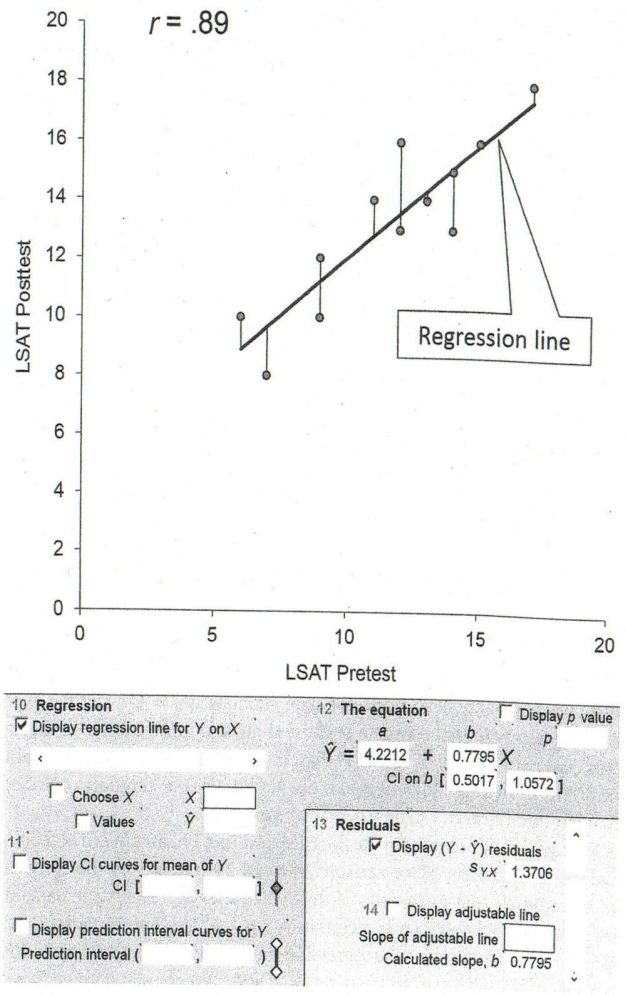


Figure 12.3. Scatterplot for the Thomason 1 data set, as in Figure 11.10, with the regression line and red $(Y - \hat{Y})$ residuals displayed. For the data set, $N = 12$ and $r = .89$. From **Scatterplots**.

esci **12.2** Click at red 14 to display a red adjustable line in place of the black regression line. Use the slider to change the slope of the red line, while still seeing the residuals marked by the fine vertical lines.

a. Over what range can you adjust the slope of the adjustable line?
b. Watch $s_{Y.X}$ near red 13. How does it change as you rotate the adjustable line?
c. What happens when the slope of the adjustable line is, as close as possible, equal to b, whose value in the equation is displayed also below red 14?

I hope you found that the minimum value of $s_{Y.X}$ was 1.37, and that this occurred when the slope of the adjustable line was equal (or very close to equal)

The regression line of Y on X minimizes

either way increased $s_{Y,X}$. Does the regression line look, roughly, as though it fits the pattern of points? Now for the second approach to thinking about the slope.

Regression Line Slope, and Correlation

I'll discuss some thought experiments about predicting the GPA of a student I'll call Robert. As in Chapter 4, I'll use the international GPA scale that ranges from 1 to 7. If I gave you no information about Robert, what would be your best guess of his GPA?

Thinking, thinking. Discussing, discussing.

It would be the mean GPA of all students, the mean usually being the best choice of a single value to represent a population. For your college, the mean may be, say, 4.5. Now suppose I told you that Robert is 176 cm tall. I also reported GPA and height data for a sample of 40 students, not including Robert, as pictured in Panel A of Figure 12.4. The correlation happens to be $r = 0$, so telling you Robert's height gives you no information about his GPA. Therefore, the college mean of 4.5 is still your best GPA prediction. If you didn't know that college mean of 4.5, you could use M_Y, the mean of Y for the sample, which is 4.4. That mean is marked by the heavy horizontal line, M_Y in the figure. In Panel A, because $r = 0$, whatever the value of X (a student's height) the horizontal line at a GPA of $M_Y = 4.4$ is the best point estimate based on the data. In other words, the regression line is that horizontal line, $\hat{Y} = M_Y = 4.4$, and its slope is zero. For any value of X, our regression estimate is $\hat{Y} = 4.4$, but we expect great uncertainty in that estimate because there's so much scatter in the data points, and $r = 0$. What I'm saying is that, if $r = 0$, height gives no information about GPA, so we can't do better than use $M_Y = 4.4$ as our estimate of Robert's GPA—but don't expect it to be a good estimate.

The slope of the $r = 1$ line is (s_Y/s_X), the ratio of the two SDs.

Next you decide to investigate the Nifty test (no, I hadn't heard of it before either) as a possible predictor variable. You test 20 students on the Nifty, and later record their GPA at the end of the year. Panel B presents the data, and shows that, remarkably, the correlation is $r = 1.0$. The line in Panel B goes through all the points and is the regression line that we'd use to estimate GPA. Its slope is (s_Y/s_X), where, as you may recall, s_Y and s_X are the standard deviations of Y and X respectively. That slope is the ratio of the two standard deviations, and so its units are (GPA units)/(Nifty scale units) or, in general, (units of Y)/(units of X). It's worth remembering that the slope of the $r = 1$ regression line is (s_Y/s_X). We'll be seeing it again.

Knowing a student's Nifty score gives excellent information about that student's GPA—sufficient information for us to make a fully accurate estimate of that student's GPA. Therefore, in this case with $r = 1$ all the points lie on the regression line. If Robert, not in our sample of 20, scores $X = 90$ on the Nifty, Panel B illustrates that the GPA estimate given by the line is $\hat{Y} = 5.9$. Assuming Robert is from the same population as our sample, we estimate his GPA will be 5.9.

Recall Equation 12.1, our informal expression for the story of Y. By "story of Y" I'm actually referring to the variability of Y about its mean, so, still informally, I can write:

The informal story of Y.

The story of Y =

(What X tells us about *the variability of* Y) + The residuals (12.5)

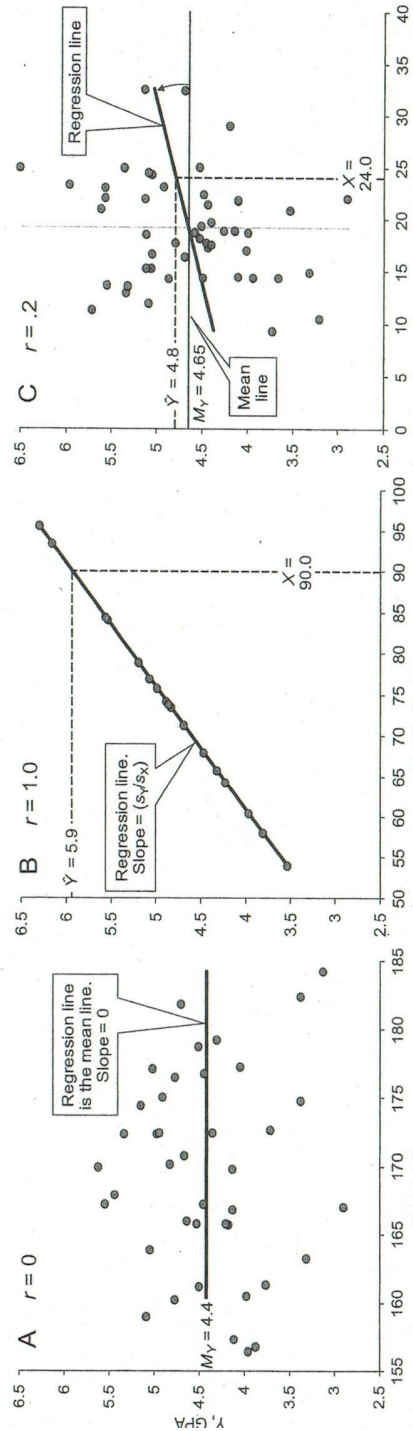


Figure 12.4. Three data sets for predicting a student's GPA, Y, which is plotted on the vertical axis in every panel. Panel A is the scatterplot of GPA and X, which is height, for a sample of $N = 40$ students, with $r = 0$. Panel B pictures GPA and X, which is Nifty test score, for a sample of $N = 20$ students, with $r = 1.0$. Panel C pictures GPA and X, which is initial impression rating, for a sample of $N = 50$ students, with $r = .2$.

Consider the two extreme cases, $r = 0$ and $r = 1$, that we discussed above.

- If $r = 0$, as in Panel A of Figure 12.4, we use M_Y as the estimate, for any X. The horizontal mean line is the regression line, with slope zero.
- Also when $r = 0$, knowing X gives no information about Y, and so the first term on the right in Equation 12.5 is zero and the second term, the residuals, must be giving us the full story of Y. Typically, as in Panel A, the residuals are large.
- If $r = 1$, as in Panel B, the line on which all points fall is the regression line, with slope (s_Y/s_X).
- Also when $r = 1$, knowing X gives complete information about Y, and so the first term on the right in Equation 12.5 tells the full story. The residuals are all zero, and they tell no part of the story of Y.

Panels A and B illustrate the two extreme cases for the slope of the regression line. What would you guess is the slope when $0 < r < 1$, and so X gives some but not full information about Y? That's worth pondering: Consider Panels A and B of Figure 12.4, and r close to 0, and r close to 1.

Take a break, discuss. Draw some sketches?

As r increases from 0 to 1, X gives progressively better information about Y, and the slope of the regression line increases from 0 toward (s_Y/s_X); the line rotates counterclockwise from horizontal.

If your intuition suggests compromise between the two extreme cases, give yourself a pat on the back. For $r = 0$, knowing X gives us no information about Y, then for progressively larger r, knowing X gives us progressively more information about Y, until $r = 1$ and we have complete information about Y. Correspondingly, the regression line rotates smoothly between the horizontal mean line and a line with slope (s_Y/s_X), as r changes from 0 to 1. (There's a similar smooth change for negative correlations.) The smooth change in slope as r changes is a key point, so I'll say it again: The regression line slope is somewhere between the zero slope of the horizontal mean line and the slope of the line through the points when $r = 1$, with correlation r determining where it lies between those two extremes.

Panel C of Figure 12.4 illustrates a case in which $r = .2$. I'm supposing that we rate our initial impression of a student, and assess how useful an estimate of GPA those ratings might provide. In a sample of $N = 50$ students, the correlation of those ratings with GPA at the end of the year was $r = .2$. The horizontal mean line at M_Y is shown and the regression line has rotated a small amount from horizontal.

To obtain the regression line of Y on X, rotate from the horizontal mean line, M_Y, until the slope is r times (s_Y/s_X).

The slope is actually $r \times (s_Y/s_X)$, so it's r that tells us how far to rotate from horizontal. When $r = 1$, the slope is $1 \times (s_Y/s_X)$, the slope in Panel B.

The dashed lines in Panel C show that, if the impression rating for Robert, who's not in the sample of 50, is $X = 24.0$, our GPA regression estimate for him would be $\hat{Y} = 4.8$. Because r is small, there's much scatter in the figure and so the residuals are large, and, therefore, we expect considerable uncertainty in that estimate.

We've now taken two approaches to specifying the regression line:

1. The line that minimizes $s_{Y.X}$, the SD of the estimation errors.
2. The line with a slope that's a compromise, based on r, between horizontal and the line for $r = 1$. Its slope is $r \times (s_Y/s_X)$.

The remarkable thing is that these two approaches give the *same* line. Our two approaches finish in the same place. We can now state the formula

for the regression line, and be happy that this is the line that minimizes estimation error.

The regression line passes through (M_X, M_Y), as illustrated in Figure 12.1, and has slope of $r \times (s_Y/s_X)$. That information plus some algebra gives us formulas for the intercept, a, and slope, b, in Equation 12.3, the equation of the regression line. The formulas are:

$$b = r \times \left(\frac{s_Y}{s_X}\right) \qquad (12.6)$$

Correlation; Slope of line for $r = 1$

Slope, regression of Y on X.

$$a = M_Y - b \times M_X \qquad (12.7)$$

Intercept, regression of Y on X.

Substituting for a and b in Equation 12.3 gives an alternative form of the equation of the regression line of Y on X:

$$\hat{Y} = \left[M_Y - r\left(\frac{s_Y}{s_X}\right)M_X\right] + \left[r\left(\frac{s_Y}{s_X}\right)\right]X \qquad (12.8)$$

Intercept, a; Slope, b

Regression line, Y on X.

That equation gives the regression line of Y on X, as displayed in Figures 12.1 to 12.4. To use the equation to calculate the regression line, we need just the following summary information from a data set of N pairs of (X, Y) values: M_X, M_Y, s_X, s_Y, and r.

Quiz 12.1

1. In regression, the variable being predicted is $X/Y/\hat{Y}$, the variable being used to make the prediction is $X/Y/\hat{Y}$ and the prediction we make is $X/Y/\hat{Y}$.
2. In the regression equation, b is the _____ and a is the _____.
3. What is a residual? How is a residual calculated?
4. Which of the following r values would be the *most* useful for using with regression? Which would be the *least* useful?
   a. $r = 0$
   b. $r = .3$
   c. $r = -.6$
   d. $r = .05$
5. The _____ of the regression line of Y on X is $r \times \left(\frac{s_Y}{s_X}\right)$.
6. The regression line is chosen to minimize
   a. the slope.
   b. the intercept.
   c. the correlation.
   d. the sum of squares of residuals.

The Linear Component of a Relationship

In Chapter 11 we saw that r measures just the linear component of the relationship between two variables, X and Y, and that we need to see the scatterplot to be sure we have the full story. We can calculate a value of r whatever the scatterplot, but Figure 11.4 illustrated two cases in which r could easily mislead. The regression we're discussing is called *linear regression* because it

too, expresses the linear component of the relationship, by means of a straight line. Once again this may not be the full story and thoughtful inspection of the scatterplot is always necessary. Further, it's called *simple linear regression* because there's just one predictor, X.

Simple regression uses a single predictor (X), and ***multiple regression*** more than one predictor (X_1, X_2, ...), to predict Y. Use and interpret multiple regression with great caution.

Beyond that is *multiple linear regression*, which estimates Y from two or more predictors, X_1, X_2, ... That's beyond the scope of this book, but sufficiently important to be worth a brief mention. If, for example, you use number of years of education (X) to predict annual income (Y), you are using simple regression as we discuss in this chapter. If, however, you use not only years of education (X_1), but also, say, age (X_2) and a measure of socioeconomic status (X_3) to predict income, you would be using multiple regression. Multiple regression often, although not always, gives a better prediction of Y. It also provides estimates of the relative contributions of the various predictors (X_1, X_2, ...) to the prediction of Y. It can be highly valuable to have, for example, an indication that years of education makes a larger (or maybe smaller?) contribution than socioeconomic status to expected income. I need to warn you, however, that multiple regression is often misused and can be very tricky to interpret. One major issue is that the estimated contribution of, say, X_1 depends on the full set of predictors (X_1, X_2, ...). Drop one of those predictors—or add another—and the relative contributions of *all* the predictors to the prediction of Y may change drastically. I won't try to explain in detail, but the take-home message—all you need to keep at the back of your mind—is that multiple regression can be highly useful, but using it properly requires a fair bit of knowledge. Be very cautious about any conclusion you read that's based on multiple regression.

Now back to simple regression. Recall that one or two outliers can be very influential on the mean and standard deviation, and one or two points that don't fit the general pattern can make a big difference to r, as Figure 11.12 illustrates. One or two such points can also make a big difference to the regression line. Figure 12.5 illustrates this by displaying the Thomason 1 data set, and the same with the addition of a point at (18, 7) for a student who did very well on the pretest, but terribly on the posttest. That single aberrant point drops r from .89 to .40 and reduces the slope of the regression line from 0.78 to 0.36. The scatterplot on the right might lead us to doubt the usefulness of linear regression for these data, or we might consider calculating regression twice, once for the full data set, and once without the aberrant point. As always, examine the data carefully and make a reality check for anything statistical that you calculate.

You may be thinking that journal articles rarely include scatterplots, so how can we be sure that the correlation and regression analyses we read are appropriate? That's a good question. Unless the authors tell us that the scatterplots show no signs of departure from linearity, they are asking us to take their analyses and interpretation on trust. We should always, however, keep in mind possible limitations. For correlation and regression, always remember that it's the linear component being assessed. Keep a sharp lookout for any indication of outliers.

12.3 Consider a data set with $N = 12$, $M_X = 11.58$, $M_Y = 13.25$, $s_X = 3.32$, $s_Y = 2.90$, and $r = .89$.

a. Use Equation 12.8 to calculate the regression equation of Y on X.
b. What Y would you estimate for $X = 9$? Would you expect small or large uncertainty in that estimate? Why?
c. What Y would you estimate for $X = 0$? Might there be some problem

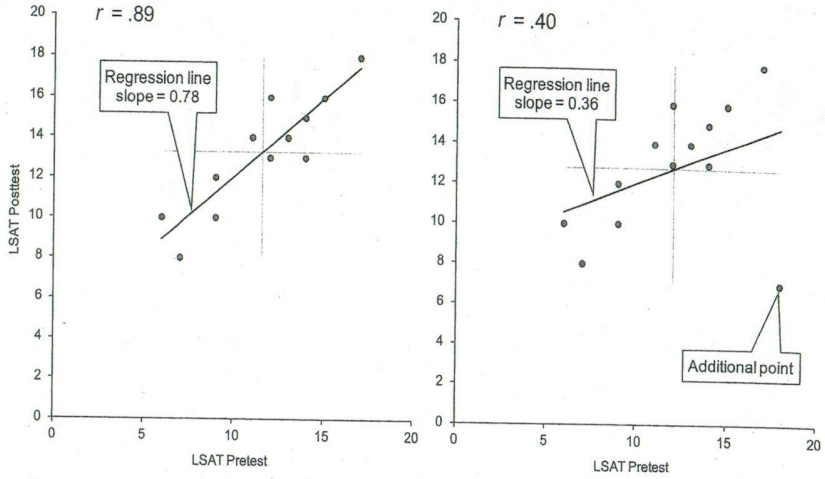


Figure 12.5. Scatterplot at left is for the Thomason 1 data set, as in Figure 12.3. Scatterplot on the right is the same, but with the addition of an outlier data point at (18, 7).

esci **12.4** Open **Scatterplots** and, if necessary, scroll right and click at red 16 to load the Thomason 1 data set.

a. Reveal the regression panel and click the three checkboxes at red 10 to display the regression line and mark an X value and the corresponding $\hat{Y}$.
b. Use the slider to set $X = 9$, as closely as possible. What is $\hat{Y}$? Explain what that $\hat{Y}$ value tells us.

esci **12.5** Examine the descriptive statistics reported at red 3 and compare with the values I stated for Exercise 12.3.

a. Compare your calculated regression equation with that shown at red 12.
b. Compare your calculated $\hat{Y}$ for $X = 9$ with your answer to Exercise 12.4.

esci **12.6** Below red 2 find the two students who scored $X = 9$ on the pretest.

a. What posttest scores did they obtain?
b. Compare with our $\hat{Y}$ estimated value for $X = 9$. Is there a problem? Explain.

esci **12.7** You discover that the results for one student were omitted from the data set. That student scored $X = 12$ on the pretest and $Y = 7$ on the posttest. Enter those additional data values below red 2 and watch what changes. Use **Undo** and **Redo** to note the changes as the additional point is removed and added.

a. What happens to r? Explain.
b. Does the slope of the regression line change much? Does the regression line seem to represent the points reasonably? Explain.
c. Compare the changes given by addition of (12, 7) with the changes illustrated in Figure 12.5 for the addition of the point (18, 7). Which point is more influential? Explain and draw a general conclusion.

Regression Using *z* Scores

A scatterplot of (X, Y) data pairs is usually displayed using the original units

units. Alternatively, as we saw in Chapter 11, a scatterplot may display the corresponding standardized scores, Z_X and Z_Y. Figure 12.6 shows both scatterplots for the Thomason 1 data set: original units on the left, as in Figure 12.3, and z scores on the right. The z scores are calculated so that the mean is zero and the SD is 1, for both Z_X and Z_Y. Therefore, in the scatterplot of z scores the regression line passes through (0, 0), the means point. The slope of the line for $r = 1$ is (s_Y/s_X), which here is $(1/1) = 1$, so that line would slope at 45° upward to the right. The slope of the regression line is r times that, or $r \times 1 = r$. The equation of the *standardized regression line* of Z_Y on Z_X is therefore:

Regression line, Z_Y on Z_X.

$$Z_{\hat{Y}} = r \times Z_X \qquad (12.9)$$

Slope

The *standardized regression line* of Z_Y on Z_X has intercept zero, and slope r.

This equation (12.9) is the conversion of Equation 12.3 to standardized scores. It has no term for the intercept, because the line passes through (0, 0) and so the intercept is zero.

I need to mention a customary symbol that's very poorly chosen, as if designed to make life confusing for all of us. In the regression world, the slope of the standardized regression line expressed by Equation 12.9 is almost always given the symbol β. It's a highly valuable convention to use Greek letters for population parameters, but here β refers to a slope estimate calculated from a data set, using standardized scores, so β is a sample statistic, not a population parameter. We could use it as our point estimate of the standardized regression slope in the population, but we would need to find some other symbol, perhaps $\beta_{population}$, for that population parameter.

It's crazy to use a Greek letter for a sample statistic, but in this case, unfortunately, it's the norm. If you see reference to β in the context of simple linear regression, think of Equation 12.9 and remember that $\beta = r$. A regression slope is sometimes referred to as a *regression weight*, so β is often referred to as the *standardized regression weight*. Despite it being a Greek letter, we have to remember that it's a sample statistic, not a population parameter. Sorry about that.

The slope of the standardized regression line of Z_Y on Z_X is referred to as β, called the *standardized regression weight*.

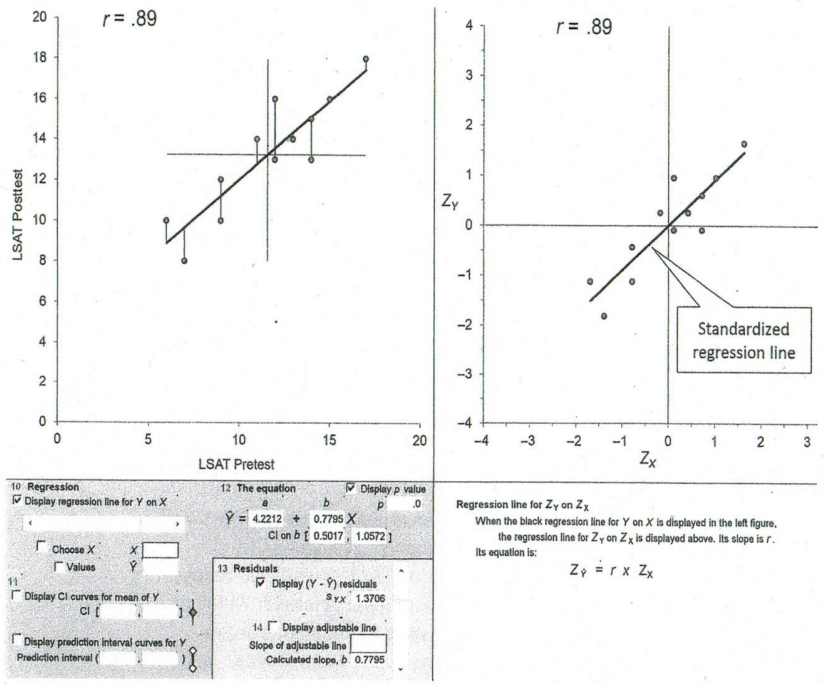


Figure 12.6. Scatterplots for the Thomason 1 data set. On the left is the scatterplot in original units, as in Figure 12.3, and on the right the same data transformed to z scores. In each figure, the regression line and cross through the means are displayed. From Scatterplots

esci **12.8** Observe the Thomason 1 data set in **Scatterplots** and click at red 15 to reveal the second scatterplot. Display the regression line and cross through the means. Compare with Figure 12.6.

a. Type in the additional data point (18, 7). Again use **Undo** and **Redo**, and note changes in both scatterplots.
b. What is β for the original 12 points? For 13 points? Explain.

12.9 Suppose that for this month in your city the daily maximum temperature is approximately normally distributed with mean 20 °C and standard deviation 4 °C. The correlation of maximum temperature from one day to the next is $r = .6$.

a. Suppose the maximum today is 14 °C, and we wish to estimate tomorrow's maximum. What would you choose as X and Y?
b. Find Z_X, and use Equation 12.9 to calculate $Z_{\hat{Y}}$. What is your prediction of tomorrow's maximum?
c. If $r = 0$, what's your prediction for tomorrow's maximum?
d. For r larger than .6, would tomorrow's prediction be closer to today's maximum, or further away? Explain.

12.10 The height of adult women is approximately normally distributed with mean 162 cm and SD 6 cm.

a. Susan is $X = 174$ cm tall. Find Z_X.
b. Suppose $r = .5$ for the heights of a woman (X) and her adult daughter (Y).Use Equation 12.9 to find $Z_{\hat{Y}}$ for Susan's daughter. Find her estimated height, $\hat{Y}$.
c. Find the estimated height of Susan's granddaughter, when adult. And her great-granddaughter.

12.11 Now let Y be the height of Susan's mother.

a. Find $Z_{\hat{Y}}$ and $\hat{Y}$.
b. Explain what that $\hat{Y}$ is, and how it relates to the starting information that Susan is $X = 174$ cm tall.

The pattern of height of a woman, and estimated heights of her mother, daughter, granddaughter, and so on, is worth pondering and discussing. It was discussed extensively around a century ago when these statistical ideas were being developed, although then it was mainly about fathers and sons. Some people thought the regression analysis seems to imply that, after a few generations, all the descendants of Susan would have about the same height—the population mean. That's tricky to think about, and was often described as a paradox. How does it strike you? We'll discuss it further a little later.

THE PROPORTION OF VARIANCE ACCOUNTED FOR: r^2

I now want to discuss a useful interpretation of r^2, the square of the correlation. Consider Figure 12.7, in which I'm imagining we wish to estimate Y, Maria's height. Panel A reports data for $N = 30$ pairs of women who are identical twins, with $r = .80$. If Maria (not in that sample) has an identical twin, then the high correlation means that knowing the twin's height, X, gives us good information about Maria's height. The regression line is rotated most of the way toward (s_Y/s_X), the slope of the $r = 1$ line, and residuals are generally not large.

By contrast, Panel C reports heights for $N = 30$ pairs of women who are best friends, with $r = .20$. If we are told X, the height of Maria's (female) best friend, the low correlation means that the regression line gives an estimate,

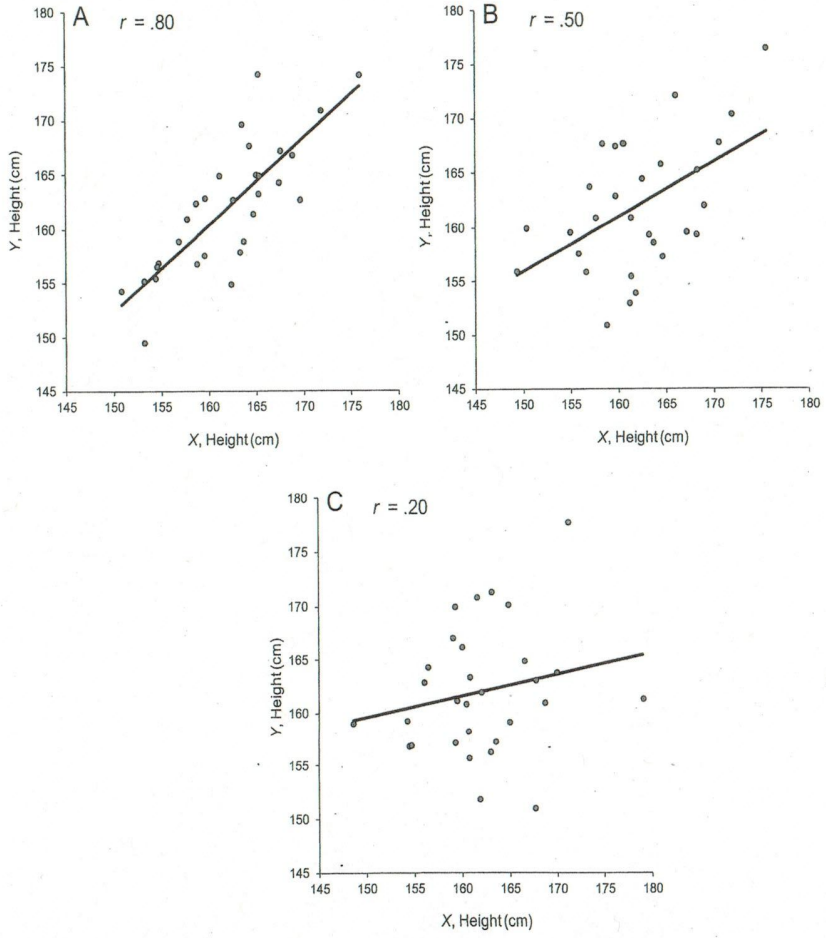


Figure 12.7. Scatterplots for fictitious data for three samples, each comprising $N = 30$ pairs of women, and each displaying the regression line for Y on X. Panel A displays women's heights for pairs of identical twins, with $r = .80$. Panel B displays

$\hat{Y}$, of Maria's height that's likely to be poor. The regression line is rotated only a small amount from horizontal and residuals are generally large. Panel B displays data for $N = 30$ pairs of sisters, with $r = .50$, so the regression slope is intermediate between those of Panels A and C.

Our informal idea, in Equations 12.1 and 12.5, is that the regression line provides $\hat{Y}$ estimates calculated from X that tell part of the story about Y, the remainder lying with the residuals. In Panel A, correlation is high, the regression line tells much of the story of Y, and residuals are small. In Panel C, correlation is low, the line tells only a small portion of the Y story, and residuals are large.

Slightly more formally, our story becomes:

The variance story of Y.

$$\text{Variance of } Y = \text{Variance of } \hat{Y} \text{ estimates} + \text{Variance of } (Y - \hat{Y}) \text{ residuals} \qquad (12.10)$$

Why use variance? You don't need to know, but if you're curious: As I mentioned in Chapter 4, variance has the valuable statistical property that variances add—as in Equation 12.10—when the different sources of variability are independent. Now to formulas. If these are becoming eye-glazing, think back to the informal account above.

Equation 12.2 gave the basic formula for $s_{Y.X}$, the SD of the residuals:

SD of residuals.

$$s_{Y.X} = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{(N-2)}} \qquad (12.2)$$

Here's a second equation for $s_{Y.X}$ that gives the same results as Equation 12.2, although I won't try to explain why it's true:

SD of residuals.

$$s_{Y.X} = s_Y \times \sqrt{1 - r^2} \qquad (12.11)$$

which is an equation involving two standard deviations. Square it to give an equation about variances:

Variance of residuals.

$$s^2_{Y.X} = s^2_Y \times (1 - r^2) \qquad (12.12)$$

which can be rearranged to give:

$$s^2_Y = \left[r^2 \times s^2_Y\right] + s^2_{Y.X} \qquad (12.13)$$

Total variance of Y | Variance of $\hat{Y}$ estimated from X | Variance of $(Y-\hat{Y})$ residuals

The variance of Y is the sum of variance of $\hat{Y}$ estimated from X, and variance of the residuals.

Equation 12.13 is just a formalization of informal Equation 12.10. It tells us that the total variance of Y in our data set is the sum of two components: the variance of $\hat{Y}$ values estimated from X, and the variance of the $(Y - \hat{Y})$ residuals. The first component is referred to as the variance of Y that's *attributable to X*, or *accounted for by X*. Equation 12.13 also tells us that, in terms of variance, r^2 is the proportion of the Y story that's told by the regression line; the remaining $(1 - r^2)$ is with the residuals.

As usual, we're not making any assumptions about causation—we're not, for example, claiming that X *causes* the regression component of the

If |r| is zero or small, where |r| is the absolute value of r, does X account for much of the variance of Y? If |r| is large, does X account for much of the variance of Y?

Enter small or large values of r into Equation 12.13... Reflect, discuss...

If $r = 0$, the $\left[r^2 \times s_Y^2\right]$ term in the equation is zero and the variance of the residuals equals the total variance of Y, as in Figure 12.4, Panel A. The regression line is horizontal and doesn't help, and residuals are generally large. At the other extreme, if $r = 1$, the $\left[r^2 \times s_Y^2\right]$ term is s_Y^2, and so $s_{Y.X}^2 = 0$, meaning there's no variance in the residuals. As in Panel B, the estimates are perfectly accurate, and all residuals are zero.

r^2 is the proportion of s_Y^2, the variance of Y, that is accounted for by X.

As r increases from 0 to 1, the regression line rotates from horizontal toward (s_y/s_x), the slope of the $r = 1$ line. Also, the $\left[r^2 \times s_Y^2\right]$ term increases, meaning that the regression line can use X to account for more and more of the variance in Y, and the variance of the residuals decreases. In a nutshell, r^2 is the proportion of s_Y^2, the total variance of Y, that can be attributed to X, or accounted for by X.

In Panel B of Figure 12.7, $r = .5$ for the heights of sisters. Therefore, a woman's height (X) accounts for $r^2 = .5^2 = .25$, or 25% of the variance in her sister's height (Y).

12.12 a. What percentage of the variance of a woman's height is attributable to the height of her identical twin, based on the r value from Figure 12.7? What percentage is variance of the residuals? Considering Panel A of Figure 12.7, does that seem reasonable?

b. Answer the same questions for a woman's best friend and Panel C.

I conclude that, when X and Y are strongly correlated, a regression line allows us to make better estimates of Y based on X, and the residuals are smaller. Large (in absolute magnitude) correlations are indeed useful.

REGRESSION REVERSED: THE REGRESSION OF *X* ON *Y*

It's customary to arrange things so X is the predictor and Y the predicted variable, and to consider the regression of Y on X. I did say, however, that I'd mention regression the other way round: the regression of X on Y. This refers, of course, to estimating a value of X for a particular value of Y, so the roles of predictor and predicted variable are swapped. Let's consider the Thomason 3 data set.

For that data set, Figure 12.8 displays the scatterplot for $N = 39$ students who gave HCTA scores before and after training in critical thinking. The correlation is $r = .60$, and in each panel a regression line and the cross through the means are displayed. I'm not going to give formulas, but think of a regression line as a compromise between a mean line and the $r = 1$ line, with r determining the degree of compromise. The familiar regression line of Y on X is displayed on the left. It's the heavy line that's rotated counterclockwise from the horizontal mean line through M_Y, with the amount of rotation determined by $r = .60$.

Now consider the right panel and estimation of X for a given value of Y. If $r = 0$, knowing Y gives us no information about X, so our best prediction is M_X, which is the vertical mean line at about 67. For $r > 0$, the regression line will be

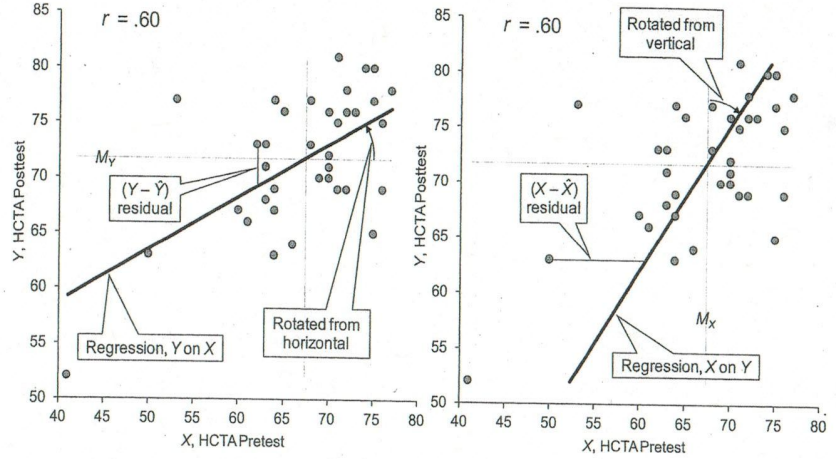


Figure 12.8. Scatterplot for the Thomason 3 data set for $N = 39$ students who gave HCTA scores at pretest (X) and posttest (Y). In each panel, a regression line and the cross through the means are displayed. The left panel illustrates the regression line of Y on X, which is rotated counterclockwise from the horizontal mean line at M_Y. An example $(Y - \hat{Y})$ residual is displayed red. The panel at right illustrates how the regression line of X on Y is rotated clockwise from the vertical mean line at M_X. An example $(X - \hat{X})$ residual is displayed red.

rotated clockwise, and the larger the r, the more the rotation from the vertical. The result is the heavy line labeled as "Regression, X on Y".

It may seem weird to have two different lines for a single data set, as in Figure 12.8. Both lines give estimates with minimum SD of the residuals, but residuals are defined differently in the two cases, so the two lines are almost always different. For the regression line of Y on X, the $(Y - \hat{Y})$ residuals are the focus; these are the vertical distances of the data points from the line, as in Figures 12.2 and 12.3. One of these is displayed red in the left panel of Figure 12.8. The Y on X regression line minimizes $s_{Y.X}$, the SD of the vertical, Y residuals.

By contrast, for the regression line of X on Y, the $(X - \hat{X})$ residuals are the focus, and these are the horizontal distances of the data points from the X on Y regression line. One of these is displayed red in the right panel. The X on Y regression line minimizes $s_{X.Y}$, the SD of those horizontal, X residuals.

The regression line of X on Y minimizes $s_{X.Y}$, the SD of the X residuals.

12.13 a. When do the two regression lines have similar slopes?

b. When are the two regression lines the same? What is their slope?

c. When are the two regression lines most different in slope? Explain.

? Quiz 12.2

1. Before using regression it is important to see the scatterplot. What should you be asking as you inspect it?
   a. Is the relationship linear, because otherwise linear regression should not be used?
   b. Are there outliers, which can have a very strong influence on the regression equation?
   c. Both a and b.
2. Regression is easy with standardized scores (z scores). Given a Z_X, you can calculate the prediction $Z_{\hat{Y}}$ by simply _____.
3. By regression convention, β is the _____ of the standardized regression line; it is equal to _____.

4. Which of the following is true about r^2?
   a. r^2 is calculated simply by multiplying r by itself ($r \times r$).
   b. r^2 reflects the strength of the correlation, and can vary from 0 to 1.
   c. r^2 represents the proportion of variance of Y accounted for by knowing X.
   d. All of the above.
5. If X and Y are correlated, $r = -.4$, then X accounts for ____% of the variance of Y. If the correlation is $r = .1$, X accounts for only ____% of the variance of Y. If X accounts for 64% of the variance of Y, then $r =$ ____ or ____.
6. The regression line for predicting Y from X minimizes the SD of the $(Y - \hat{Y}) / (X - \hat{X})$ residuals, whereas the regression line for predicting X from Y minimizes the SD of the $(Y - \hat{Y}) / (X - \hat{X})$ residuals.

ASSUMPTIONS UNDERLYING REGRESSION PREDICTIONS

For a $\hat{Y}$ estimate to make sense as a prediction, we need to assume random sampling of Y values at any particular value of X.

For the simple linear regression we're discussing, we're assuming both X and Y have interval scaling. In addition, for $\hat{Y}$ estimates to be usable as predictions for new cases, we need to assume random sampling of Y values at any particular value of X. For an example, recall the imaginary Hot Earth Awareness Test, the HEAT. Suppose we're interested in the increase in HEAT scores for students in successive years at your college. Figure 12.9 is a scatterplot of a random sample with $N = 20$ at each year level. Because Y is randomly sampled at each X value, we are justified in using the regression line of Y on X, as shown in the figure, for making predictions. It's not a problem that the X values—1, 2, 3, and 4—were chosen for convenience and are not a random sample.

In addition, we should make predictions only for cases from the same population as the data set. Informally, we're assuming that the relationship observed in the data set holds also for that particular case. For our example, we could use the regression line to make predictions for students at your college,

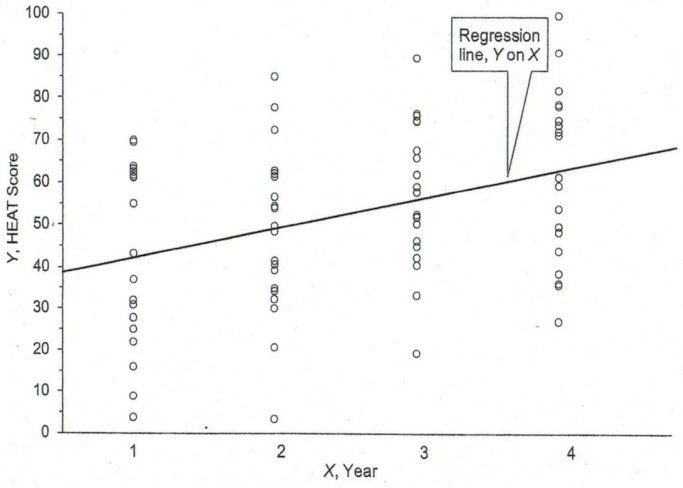


Figure 12.9. Fictitious HEAT scores for samples of $N = 20$ students from years 1, 2, 3, and 4 in college. The predictor variable, X, is year, and the predicted, Y, is HEAT score. The regression line is calculated for the data set of 80 points.

but not for other students—unless we made the additional assumption that students at your college are similar to some wider population of students, at least in relation to HEAT scores.

Any prediction from a regression line requires the assumption that the particular case comes from the same population as the data set.

What if X for the individual case lies outside the range of X in the data set? That suggests that the individual case doesn't comes from the same population—in other words, the assumption that the relationship in the data set applies to the individual case is likely to be suspect, or even silly. A regression line doesn't necessarily apply beyond the range of X values in the original data. It's easy to find crazy examples of regression estimates beyond the original range of X that make no sense. For example, if your friend proudly reports that, over the last two months he has lost 1 kg, you can helpfully say "Great, at that rate in about ten years you'll weigh nothing!"

Be especially cautious about any estimate for X beyond the range of X in the data set.

To discourage possibly unjustified extrapolation—trying to extend the relationship beyond the ranges of the data—ESCI doesn't display the regression line or calculate a $\hat{Y}$ value for X beyond its range in the data set. When using regression, just as any other time we're working with data, it's essential to keep thinking about the meaning of what we're doing. Does it make sense?

Now I'll summarize this brief section on assumptions.

Regression prediction. For a $\hat{Y}$ estimate to make sense as a prediction for a new case, we need to assume:

1. random sampling of Y values at any particular value of X;
2. that the new case comes from the same population as the data. In particular, X for the new case should lie within or close to the range of X in the data; we should be very cautious if it lies much beyond that range.

INFERENCE, AND MEASURING THE UNCERTAINTY OF PREDICTION

At last some CIs, in fact two of them. In this section I'll discuss, first, a CI on the regression slope, b, and second, CI curves for the whole regression line. Then I'll discuss a prediction interval for individual values of Y. There will be cool pictures, with curvy lines.

A Confidence Interval on b, the Regression Slope

We use b, the slope of the regression line, to estimate the population slope, which I'll call $b_{population}$. (I'd like to call it β, but unfortunately β has another meaning.) Now I want a CI on b. Figure 12.3 shows at red 12 that the regression equation for Thomason 1 is $\hat{Y} = 4.22 + 0.78 \times X$. Therefore $b = 0.78$ is our estimate of the population slope, which tells us that one extra LSAT point at pretest (that's X) is likely to be associated, on average, with an increase of around 0.8 points on the LSAT at posttest (which is Y).

Just below the equation is the CI on b, which is [0.50, 1.06], our interval estimate of $b_{population}$. We can interpret that CI as we do any other CI: Most likely the slope of the regression line in the population lies in the interval, with values around 0.78 the most likely. Also, our interval is one from a dance, and may be red—it may miss the true value, although probably not by much. The CI is so far from zero that the p value for testing the null hypothesis that $b_{population} = 0$ would be tiny. Figure 12.6 reports the regression analysis and shows near red 12 that indeed $p = 0$, to three decimal places. If you wished to report a p value, you would report $p < .001$.

The CI on b, the slope of the regression line in the data set, is the interval estimate for $b_{population}$, the regression slope in the population.

For a given b, what would you expect to happen to the CI if N were much larger? Much smaller?

I hope you agree those are very easy questions. As usual, other things being the same, larger N gives us shorter CIs, and thus a more precise estimate of $b_{population}$.

Now for a real example. Zaval et al. (2015) asked a sample of 244 U.S. online respondents about environmental issues, especially climate change, and their willingness to take action for the sake of future generations. Respondents rated agreement with statements like "I feel a sense of responsibility to future generations" and "It is important to me to leave a positive legacy." Responses were combined to give a "Legacy Motivation" scale, range 1 to 6, where higher values indicate a greater concern about what legacy one leaves to future generations. Respondents were later offered \$10, and asked to nominate some proportion to donate to an environmental cause; they would keep the remainder for themselves. To what extent did people who were more concerned about future generations choose to donate more? The regression of amount donated (Y) against Legacy Motivation score (X) had slope of b = \$0.73 per Legacy Motivation scale unit [0.35, 1.12]. That value of b means an increase of one point on the Legacy Motivation scale, which ranges from 1 to 6, gives, on average, an increased donation of 73 cents from the \$10, with CI from 35 to 112 cents. Yes, they donated more, but that doesn't look like a large effect. Zaval et al. used that result to guide further research, including investigation of ways to increase legacy motivation.

esci **12.14** In Scatterplots, scroll right and click at red 19 to load the BodyWellFM data, the full data set for N = 106 women and men as in Figure 11.1. (Note that the axes extend beyond the scale ranges of the two measures, which are 1–5 and 1–7. If you know about Excel you can click near the edge of the figure to select it, then edit the scale ranges for the X and Y axes.) Click at red 9 then 10 to display the regression line for Well-being (Y) on Body Satisfaction (X).

a. What is the regression equation? State b and explain what it tells us.
b. State and interpret the CI on b.
c. Click near red 12 to show the p value. Explain and interpret.
d. Click near red 7 to show the p value for testing the null hypothesis that the population correlation is zero. How do the two p values compare? Does this seem reasonable?
e. Is it reasonable to apply regression to this data set? What assumptions are we making, for what calculations?

Exercise 12.14 referred to the p values for b near red 12, and r near red 7. The first assesses the null hypothesis that $b_{population}$, the regression slope in the population, is zero. The second assesses the null hypothesis that ρ, the population correlation, is zero. The two p values are calculated using different formulas, but should be the same, or very close to the same. This makes sense because zero correlation, as in Panel A of Figure 12.4, corresponds to a horizontal regression line, meaning the slope is zero. Regression slope is zero if, and only if, correlation is zero. Assessing the two null hypotheses should give the same answer, in particular the same p value.

Confidence Intervals for the Mean of Y, at every X

Our second CI is not a couple of values in square brackets, but two curves

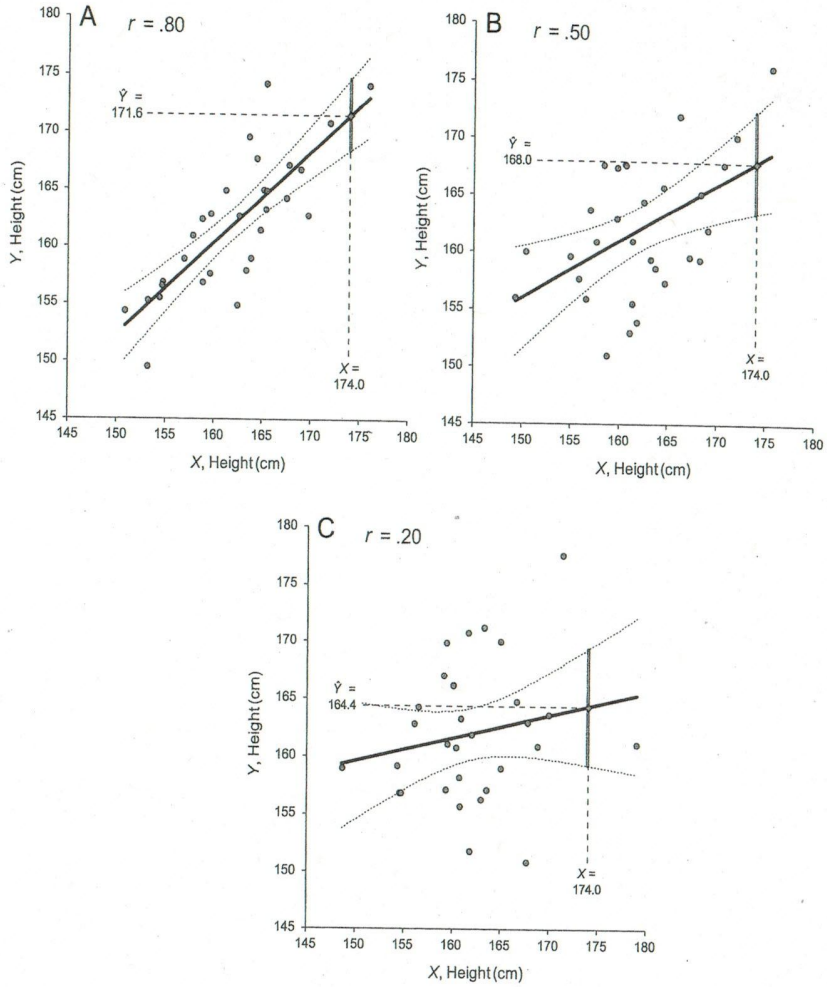


Figure 12.10. Same as Figure 12.7, but with curved dashed lines indicating, at every value of X, the lower and upper limits of the CI for the population mean of Y. The $\hat{Y}$ value, as given by the regression line for X = 174, is marked in each scatterplot, and the heavy red vertical line is the CI on this $\hat{Y}$ value. This CI is the vertical extent between the dashed curves and is the CI for the population mean of Y at X = 174 cm.

with the addition of dashed curves above and below the regression line. These curves are sometimes called "the CI on the regression line", but that's vague wording. Let's focus on a particular value, X = 174, as marked in the scatterplots. Recall Figure 12.9, which illustrates a sample extending vertically above an X value, the sample coming from a population of Y values at that X. That's also the case in Figure 12.10. We assume that vertically above X = 174 on the graph (technically, above a very short interval close to 174) there's an arbitrarily large population of Y values. There are such populations across the full range of X values.

The figure marks the $\hat{Y}$ values given by the regression lines in the three panels, all for X = 174 cm. In each panel, the heavy red vertical line between

the two dashed curves is the CI on the $\hat{Y}$ value, meaning the CI for the mean of the Y population at that X value. In Panel A, for example, the regression line gives $\hat{Y} = 171.6$ cm. The population is the heights (Y values) of an assumed very large number of women who have an identical twin with height very close to $X = 174$ cm. Our estimate of the mean of that population is our estimate of Maria's height—given that her identical twin is 174 cm tall. When I used Scatterplots to make Panel A, I clicked at red 11 to display the dashed CI curves, and saw that the CI on $\hat{Y}$, at $X = 174$, is [168.5, 174.7]. So we estimate Maria's height as 171.6 cm [168.5, 174.7].

The curved CI lines mark the lower and upper limits of the CI for the mean of the Y population at each X value.

Why are the CI lines curved? Think of a dance of the regression lines. Imagine taking repeated samples of size $N = 30$ pairs of identical twin women. Display the regression line for each. Sampling variability would cause those lines to bounce around: up and down a bit, and also rotated either way a bit. The dashed CI lines give us a rough idea of the extent of bouncing around. Because the bouncing includes rotation, the dashed CI lines need to be further apart for more extreme values of X. Putting it another way, CIs will be longer at X values further away from M_X.

Across the three panels in Figure 12.10, the regression line slopes decrease from A to C, because r decreases from .8 to .2. Now consider the dashed CI lines. Does it make sense that they are wider apart and more strongly curved in C than A? Given the greater scatter of points corresponding to $r = .2$ in Panel C, it's reasonable that the dance of the regression lines would show more energetic bouncing—larger jumps up and down, larger rotations—than for $r = .8$ in Panel A. So the CI lines need to be further apart and more curved in C than A.

12.15 If Maria's best woman friend is 174 cm tall, Panel C estimates Maria's height as $\hat{Y} = 164.4$ cm [159.3, 169.5].

a. Explain what that CI refers to.
b. Compare that CI with [168.5, 174.7], the CI stated above for Panel A. Explain the difference.

12.16 a. What proportion of the variance in a woman's height (Y) is accounted for by the variance in her identical twin's height? What proportion by her best woman friend's height?
b. What are the corresponding proportions for variance of the residuals?
c. What aspects of the display in Panels A and C reflect all those proportions?

Prediction Intervals for Individual Values of *Y*

We've been estimating the *mean* of Y for a particular X. If Maria's identical twin is 174 cm tall, Panel A estimates Maria's height as $\hat{Y} = 171.6$ cm [168.5, 174.7]. The CI tells us that the *mean* of the population of women with a 174 cm tall identical twin is, most likely, between about 168 and 175 cm. However, individual women have heights that are scattered below and above that mean. To get some idea of how widely scattered, we need the *prediction interval* for the height of an individual woman in that population.

The prediction interval for individual values of Y at a particular X value reflects both uncertainty in the estimate of mean Y, and the spread of

This is our final uncertainty topic—considering the prediction of individual values of Y at a particular X. We need a prediction interval, not a CI, because CIs estimate population parameters such as means, but here we focus on individual values, not a parameter. In other words, we are interested in individual

the previous subsection. We can expect the prediction interval for an individual Y value at $X = 174$ cm to be long because it must reflect:

1. uncertainty in the population mean of Y at this X, as quantified by the CI we discussed in the previous subsection, plus also
2. the spread of individual values, as indicated by the SD of that population.

Prediction intervals for an individual case are often long, even when CIs on b and $\hat{Y}$ are short. Individual cases may be widely spread.

Figure 12.11 is the same as Figure 12.10, but with the addition of two more dashed lines in each panel, to mark the lower and upper limits of the prediction interval for individual values of Y, at each X value.

Figure 12.11 shows that the prediction intervals are indeed long. In each panel the prediction interval for individual values of Y, when $X = 174$ cm, is marked by the fine green vertical line between the two open diamonds. For

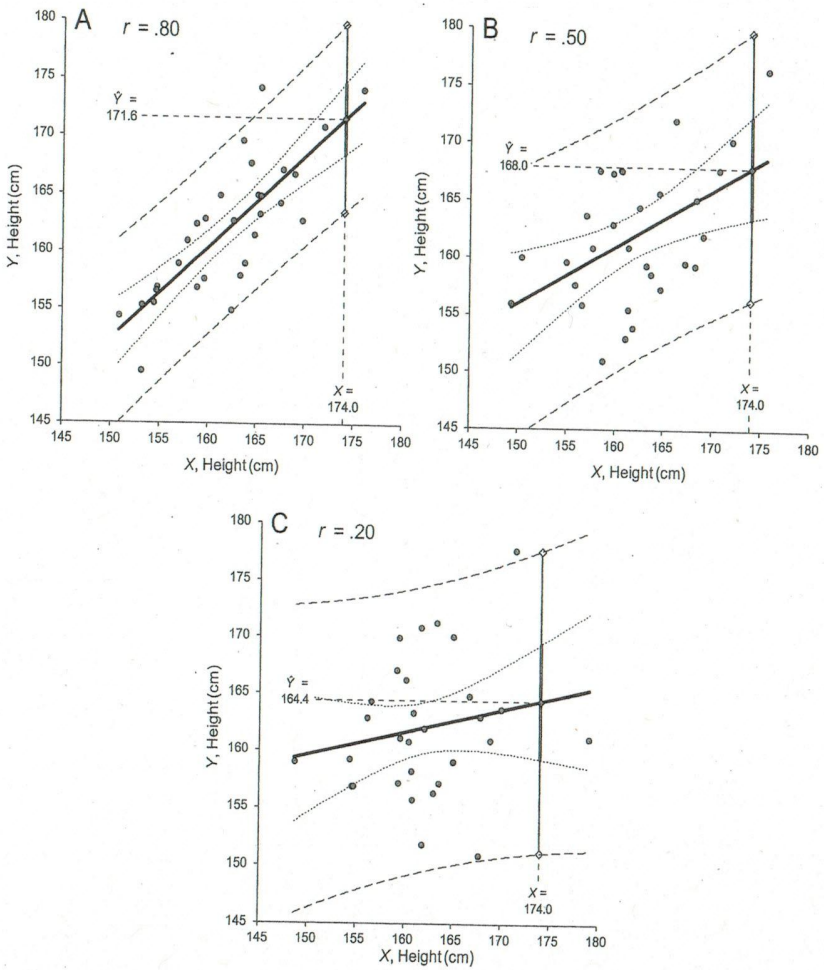


Figure 12.11. Same as Figure 12.10, but with display also of longer-dashed lines indicating, at every value of X, the lower and upper limits of the prediction interval for individual values of Y. The fine green vertical line between the two open green

Panel A, the prediction interval is (163.5, 179.7). Despite the large correlation, $r = .8$, the heights of individual women who have an identical twin of height $X = 174$ cm are widely spread. We shouldn't be surprised if Maria herself is, perhaps, 165 cm, or even 181 cm tall. That's a sobering but important lesson: When considering regression, keep in mind that individual cases are likely to be widely spread, even if we get nice short CIs for slopes and means.

Do long prediction intervals mean that regression predictions are useless? In Panel A, eyeball the prediction intervals for $X = 152$ cm and $X = 174$ cm. Each is long, but they are very different, with only modest overlap. They could, therefore, give us useful predictions for individuals, even though the long intervals remind us that individual values are widely spread. However, doing the same for Panel C gives prediction intervals that are largely overlapping, so, for small r, and, therefore, a regression line not far from horizontal, prediction intervals for individual Y values may be of limited practical value.

Here's a thought experiment: Suppose Figure 12.11 was based on three samples of $N = 1{,}000$, rather than $N = 30$, and that r was, once again, .8, .5, and .2 for the three panels. Consider the CI curves for mean Y, as also displayed in Figure 12.10. How do you think those pairs of curves may be different with the much larger samples? Now consider the prediction interval curves. How do you think they may be different from those in Figure 12.11?

Thinking, thinking, discussing...

Think about CIs for a population parameter, and the spread of individual values in a population, which is relevant for prediction intervals. I'll discuss this thought experiment shortly.

12.17 In Exercise 12.15, I stated that Panel C estimates Maria's height as $\hat{Y} = 164.4$ cm [159.3, 169.5]. The prediction interval is (151.1, 177.7). Interpret those two intervals and explain the difference.

12.18 In **Scatterplots**, load the Thomason 3 data and turn on the regression line and the CI curves for the mean of Y, the HCTA posttest.

a. Display a chosen X value and the corresponding $\hat{Y}$ value. Select $X = 72$. Note $\hat{Y}$ and its CI, for this X value. Interpret.

b. Display the prediction interval curves. Note the prediction interval for $X = 72$ and interpret.

Back to that thought experiment. With $N = 1{,}000$ the CI for mean Y at some X would be much shorter, and so the CI curves much closer to the regression line than in Figures 12.10 and 12.11. The prediction interval is shortened a little by that shortening of the CI—that's component 1 I described above. However, the second component is the spread of individual values in the population of Y values at the particular X and this is not influenced by the size of the sample we take. Therefore, prediction intervals are only a little shorter with $N = 1{,}000$, and the longer-dashed prediction interval curves only a little closer to the regression line.

ASSUMPTIONS UNDERLYING REGRESSION CIS AND PREDICTION INTERVALS

To calculate regression CIs, we need to assume Y is normally distributed in the population, and that the variance of Y is homogeneous for all X.

In an earlier section I said that regression predictions require the assumption of random sampling of Y values at any particular value of X. For regression CIs we need more. We need to assume, in addition, that Y is normally distributed in the population and that the variance of Y is homogeneous for all X. To illustrate, I'll use Figure 12.12, which is the same as Figure 12.9 but displays also the population of Y (HEAT scores) at each X. We need to assume those populations are normally distributed, and that all have the same standard deviation, meaning that we need homogeneity of variance of Y for all X. I generated the data in the two figures by sampling randomly from the populations shown, which all do have the same standard deviation, so the assumptions needed for regression CIs are met.

Note that, for regression CIs, we need to make assumptions only for the predicted variable, Y. Just as for regression prediction, the X values can be chosen for convenience and need not be a sample. Contrast with the requirement for bivariate normality when we calculate CIs on *correlation* r, as we discussed in Chapter 11. For regression CIs, however, it's not a problem if X and Y *do* come from a bivariate normal population, because that gives us what we need: Y that's sampled from a normal population and variance of Y that's homogeneous for all X.

To consider whether the assumptions needed for regression CIs are justified, consider the nature of Y: Is it reasonable to assume it's sampled from a normally distributed population? We also need homogeneity of variance of Y for all X, but this can be difficult to assess. It's common to make that assumption unless there are strong reasons against, or strong indications in the data that the standard deviation of Y changes across the range of X.

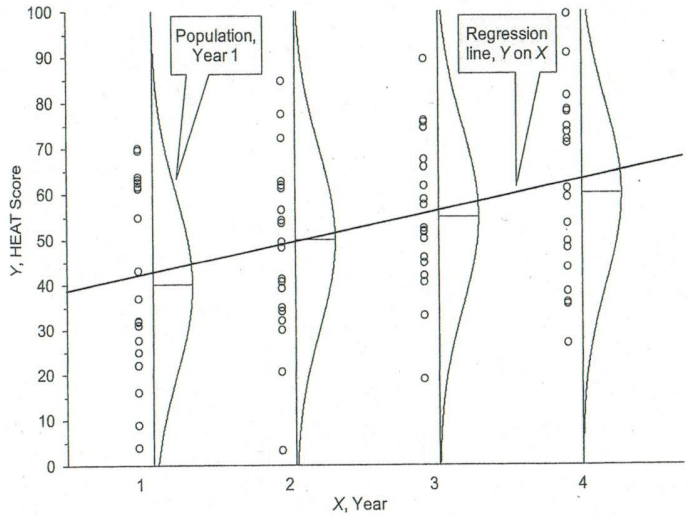


Figure 12.12. Same as Figure 12.9, but showing also the population of Y (HEAT scores) at each year, assumed to be normally distributed. The short horizontal lines mark population means.

I've said those assumptions are needed for "regression CIs", but I should be more specific. They are needed for the CI on b, the CI for the mean of Y at a particular X, as in Figure 12.10, and also for the prediction interval for individual values of Y at a particular X, as in Figure 12.11. For the latter two intervals we also need the assumption that the case with the particular X value, for which we calculate the interval, is from the same population as the data.

Now I'll summarize both the previous section on assumptions and this section.

Regression prediction, from previous section. For a $\hat{Y}$ estimate to make sense as a prediction for a new case, we need to assume:

1. random sampling of Y values at any particular value of X;
2. that the new case comes from the same population as the data. In particular, X for the new case should lie within or close to the range of X in the data—we should be very cautious if it lies much beyond that range.

CI on regression slope, b. To calculate such a CI, we need Assumption 1 above and also need to assume:

3. that the population of Y values at any particular value of X is normally distributed;
4. homogeneity of variance of Y across all values of X.

CI for the mean of Y at a particular X. To calculate such a CI, as in Figure 12.10, we need:

- all of Assumptions 1, 2, 3, and 4 above.

Prediction interval for individual values of Y. To calculate such an interval for a particular value of X, as in Figure 12.11, we again need:

- all of Assumptions 1, 2, 3, and 4 above.

THREE INTERVALS: WHICH SHOULD I USE?

At this point you are entitled to feel that there's lots going on that's different, but sufficiently similar to be confusing. I hope it helps if I use an example to summarize. It's vital to know the range of options you have, and to be able to choose which you need. Let's go back to Daniel, who scored $X = 3.0$ for Body Satisfaction. We have four main options. If you wish to follow along, go to **Scatterplots**, scroll right and click at red 21 to load the BodyWellM data set for the $N = 47$ males as in Figure 12.1.

1. *Find the regression equation and use it to make a $\hat{Y}$ estimate.* I'm interested in the research question of how X and Y relate, so I want the regression equation for Y on X. I'm willing to assume random sampling of Y, so I clicked at red 9 and saw

Regression line, Y on X.

$$\hat{Y} = 1.66 + 0.93 \times X \qquad (12.14)$$

The regression slope in the sample is $b = 0.93$, meaning that, on average, 1 unit increase in Body Satisfaction corresponds to a 0.93 unit increase in Well-being. (Not necessarily causal!)

What's our estimate of Daniel's Well-being (Y) score? I'll assume Daniel is from the same student population as the data. I could use Equation 12.14 to calculate $\hat{Y}$ for Daniel's $X = 3.0$, but instead I clicked three checkboxes at red 10 and used the slider to set $X = 3.0$ and find $\hat{Y} = 4.44$ as our point estimate for Daniel.

2. *Find the confidence interval on b.* My research interest is in X and Y *in the population*, so the regression line is a good start, but as well as the point estimate, $b = 0.93$, of the population slope, I want the interval estimate. This is shown near red 12 as [0.48, 1.38], a long interval because our sample isn't large. This CI relies on the reasonable assumption that we have random sampling of Y from a normally distributed population with homogeneous variance across all X. We estimate $b_{population} = 0.93$ [0.48, 1.38].
3. *Find the confidence interval for mean Y at a particular X.* For $X = 3.0$, we found $\hat{Y} = 4.44$, our point estimate of mean Y of the population of all students who have $X = 3.0$. I clicked at red 11 and saw two dotted curves and a vertical red CI, as in Figure 12.10. The CI is [3.99, 4.89], which is our interval estimate of mean Y when $X = 3.0$. To conclude, our estimate of Daniel's Well-being score is $\hat{Y} = 4.44$ [3.99, 4.89].
4. *Find the prediction interval for individual values of Y at a particular X.* That CI might mislead us into thinking we have a fairly precise idea of Daniel's Well-being score. A quick glance at the large scatter in Figure 12.1 should remind us that individual points are widely spread. For a more realistic idea of Daniel's Y score we need the prediction interval for Y, when $X = 3.0$. I clicked lower below red 11 and saw that interval to be (2.60, 6.28). I also saw the curved, dashed lines and vertical green interval between diamonds, as in Figure 12.11. The long prediction interval reflects both the uncertainty in estimating mean Y at $X = 3.0$ (the CI we found in 3 above) and the large vertical spread of individual values of Y for students with $X = 3.0$. The Y score of our particular subject, Daniel, is most likely around 4 or 5, but could be anywhere in the prediction interval (2.60, 6.28), or even a small distance beyond.

The regression equation and CI on b tell about the relation of X and Y in the population. $\hat{Y}$ and its CI tell about all cases having a stated X value. The prediction interval tells about individual Y values for a stated X value.

So, which should you choose? That depends on what you want. To study the relationship of X and Y, you probably want the regression equation and CI on b (use 1 and 2 above). If you want to know about all students with $X = 3.0$, use 3. If you want the best guidance for one particular student, for example Daniel, use 4.

REGRESSION TO THE MEAN: A POSSIBLY STRANGE NATURAL PHENOMENON

Whenever a correlation is not 1 or −1, *regression to the mean* occurs. It's a natural and inevitable phenomenon, so it's worth discussing, although it can seem strange. Consider our estimates of Maria's height. In Panel A of Figure 12.10, correlation is .8, the regression line is steep, and $X = 174$ cm gives an estimate of $\hat{Y} = 171.6$ cm for Maria. In Panel C, correlation is .2, the regression line is not far from horizontal, and $X = 174$ cm gives an estimate of $\hat{Y} = 164.4$ cm.

Whenever the correlation of X and Y is not 1 or −1, a $\hat{Y}$ estimate is closer to the mean than the X value it was estimated from. This is ***regression to***

Small r indicates regression line near horizontal and much regression to the mean. Large r indicates steep regression line and little regression to the mean.

Both $\hat{Y}$ estimates have shifted, or *regressed*, closer to the mean than X, and the larger the correlation, the steeper the regression line and the smaller the shift toward the mean.

Figure 12.13 shows this in a different way, the curved red arrows indicating how the $\hat{Y}$ values are closer to the mean of 162 cm than the X values from which they were calculated. For $r = .8$, the shift is small—from 174 down to 171.6 cm, but for $r = .2$ the shift is larger. The red arrow on the left shows a third case: If $r = .5$, $X = 156$ cm gives $\hat{Y} = 159$ cm, which is also shifted closer to the mean, and by an intermediate amount.

Here's an example we might consider strange: Recall Exercises 12.10 and 12.11 about Susan and her female relatives. Here I'll discuss the pattern without using formulas. The correlation is positive (and not 1) for the heights of Susan and her daughter. (Studies of women's and daughters' heights typically report correlations near .5, the value I stated earlier, but the pattern I'm discussing doesn't depend on any particular value of r). Susan is 174 cm tall, well above the mean, so we expect her daughter's height to regress downward toward the mean. We expect her granddaughter's height when adult to regress even further down toward the mean, and so on for further generations. Might this suggest that, after a few generations, every woman is pretty much of average height? That's why this phenomenon has sometimes been considered a paradox.

Now consider going backward in time, rather than forward. The correlation for the heights of Susan and her mother is also positive (and not 1). Therefore, we expect Susan's mother's height to regress down toward the mean, her grandmother's height even further down, and so on. All those statements going forward and backward in time are actually true in the world. Regression to the mean may seem strange, but it captures something important about how correlation works in practice.

Note that causation plays no part in this discussion, which depends only on r. We used regression in Figure 12.10 to estimate Maria's height without

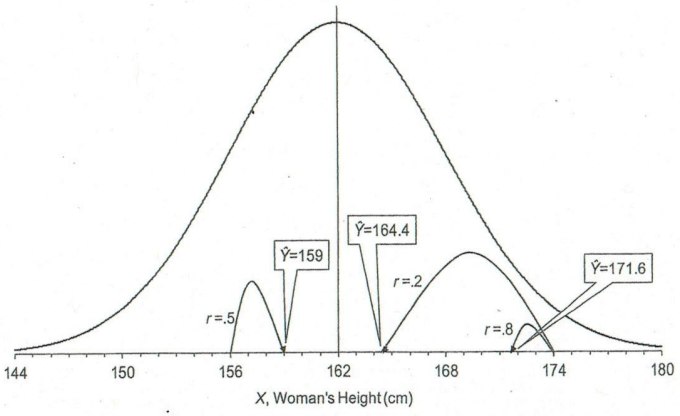


Figure 12.13. Normal distribution of women's heights, with mean 162 cm and standard deviation 6 cm. The three curved red arrows illustrate regression predictions of height ($\hat{Y}$ values) for three different situations. The shortest arrow illustrates $X = 174$ cm and $r = .8$; the longest illustrates the same X and $r = .2$; and the arrow on the left illustrates $X = 156$ and $r = .5$. The red arrows highlight the extent to which the $\hat{Y}$ estimate regresses toward the mean in each case: Larger r gives less regression—as we'd

any mention of causation—only the correlations mattered. For a particular r, we'll see the same pattern of regression to the mean, whatever X and Y are measuring and regardless of any causation or lack of causation.

Regression to the Mean and the Variability of Y

A $\hat{Y}$ estimate is regressed toward the mean (unless $|r| = 1$), but this does not imply any decrease in the variability of Y.

The key issue is that we've been discussing point estimates of the *mean* of Y, which don't tell us anything about the *variability* of Y. Consider an extreme case: Suppose the correlation between the heights of women who are neighbors is $r = 0$. If I tell you the height of Maria's woman neighbor, what's your estimate of Maria's height? Yes, 162 cm, the mean, and the regression line is horizontal. What's the extent of regression to the mean? It's 100%. However, we expect the standard deviation of the heights of women neighbors (including Maria) to be the same as that of all women, 6 cm. A large amount of regression to the mean doesn't reduce the amount of variability of Y. Our particular Maria could be short or tall. Even if our best height estimate for Susan's great-great-great-granddaughter is almost totally regressed to the mean, that doesn't imply any reduction in the standard deviation for that younger generation of women. Same for her great-great-great-grandmother. There is no paradox.

In other words, regression estimates the mean, but individual cases are likely to show great variability. That's why the prediction interval for individual values of Y is long, as in Figure 12.11. Our estimate for Susan's daughter is 168 cm, but a particular daughter may be shorter or taller than 168 cm, perhaps even taller than Susan. My family provides two examples: I'm 192 cm tall, well above average. Regression would lead me to expect sons who are shorter than me. However, both my sons are distinctly taller than me. Unlikely, but in our case true. And, in case you are wondering, my wife is not a giant!

Regression to the mean may be tricky to think about, but it happens in the world all the time. Let's consider some examples.

Everyday Regression to the Mean

Suppose you are very pleased with your basketball (or golf, or favorite computer game) score today, one of the best you have achieved. What's your prediction for next time you play? Your scores for two successive times you play are probably positively correlated, but there's almost certainly some variability in how you perform, in which case r is less than 1. Therefore, if today's score is high, your next score will, on average, be regressed toward the mean, and therefore lower than today's. If you score a personal best one day, but do less well the next day, you don't need to search for any complex reasons, because the decrease may have been caused by nothing more than the boring old natural phenomenon of regression to the mean. Conversely, after a really bad day, be encouraged that next time you are likely to do better, merely because of regression to the mean. But no guarantees.

Regression to the mean occurs often in the world, whenever two variables are less than perfectly correlated and one is used to make an estimate for the other.

When any variables X and Y are correlated less than perfectly, meaning r is not 1 or −1, then, given a value of X, the mean Y we expect is less extreme, less distant from the mean, somewhat regressed toward the mean, compared with that given value of X. The higher the correlation, the less the regression to the mean, but only if $|r| = 1$ is there no regression.

When learning to fly, one of the greatest challenges is achieving a smooth landing, and student pilots typically spend many hours making practice

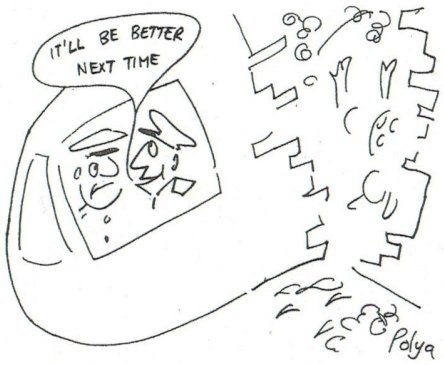
landings. An instructor who had studied psychology and knew about positive reinforcement and punishment decided that, to help her students improve, she would systematically give a few words of praise whenever a student achieved a particularly smooth landing, and a few harsh words after any very bumpy landing. She kept careful records and was disappointed to find that, after a good landing and praise, a student would, on average, make a less good landing. However, after receiving harsh words for a poor landing, a student was likely to make a better landing. She concluded that praise didn't work, but punishment—her harsh words—did. Were her conclusions justified? Does psychology have it wrong about praise and positive reinforcement?

That story is worth considering and discussing, especially if you have studied what psychology says about positive reinforcement. I'll come back to it shortly, but take a break and call a friend to discuss the story. Is it plausible?

12.19 A large basketball club fields many teams. Each week they give a special mention to the three players scoring the most points that week. The committee observes that only rarely does a player receive a special mention for two weeks in a row. They fear that receiving a special mention somehow puts players off their game, so the scheme is discontinued. Was that decision justified? Explain.

12.20 I dare not tell my friends, but I believe that a headache will go away if I wear a pink hat, face north, and wiggle my toes for one minute. I tried it secretly for my most recent five headaches, and every time my headache was distinctly less severe an hour later. Can I expect my headache treatment to make me famous? Explain.

Back to the flying instructor. We hope that the student's landings gradually improve, but, from landing to landing, there's almost certainly considerable variation in quality. In other words, there's less than a perfect correlation from one landing to the next. Therefore, a pattern, on average, of less good landings after particularly smooth landings, can be explained by regression to the mean. Seeing improvement, on average, after particularly rocky landings can be explained the same way. In other words, because of variation in the quality of successive landings, the correlation is less than perfect and so regression to the mean occurs. We'd therefore expect the observed pattern, even if the praise and admonishment had no effect.



There's an additional interesting twist to this story. Consider what the instructor experiences. Regression to the mean leads to the instructor being punished (student likely to land less well) for giving praise, but rewarded (a better landing) for giving punishment—the harsh words. It's an unfortunate fact that we go through life tending to be punished after we give reward, but rewarded after we give admonishment. How sad! Perhaps, to compensate, we should all try hard to limit our criticisms of other people, and find reasons for giving positive comments?

Here's a final point about regression to the mean. It always occurs, unless $|r| = 1$, but to know the direction of regression of the $\hat{Y}$ estimate we need to know whether the X value is below or above the mean. If you tell me your basketball score and I have no idea whether it's below or above your typical or mean score, I can't say whether your score tomorrow is likely to be higher or lower. Only if you tell me that it's a particularly poor or good score—for you—can I say that tomorrow you are likely to do better, or worse, respectively, because of regression to the mean. Of course that's an average estimate, and on a particular tomorrow your score could jump either way.

Always keep in mind regression to the mean as a possible explanation—it happens in all sorts of common situations and sometimes has surprising consequences. Think of the everyday meaning of the word "regression". Does our topic in this final section of the chapter suggest why the statistical technique we've been discussing was given that name?

It's almost time, at last, for take-home messages. To help you write yours, you might think back to some of the pictures we've discussed: a scatterplot with a regression line, the short vertical lines that are the residuals, three scatterplots and regression lines for estimating Maria's height, dashed CI curves, and even more widely spaced curves for prediction intervals. Reflect again on Susan and her relatives, and on the comments you might make to a student pilot practicing landings.

? Quiz 12.3

1. When making a prediction about a new case,
   a. the new case must be from the same population sampled to generate the regression equation.
   b. it must make sense to assume random sampling of Y at any level of X.
   c. we should be especially cautious of making predictions for any X value outside the range of the sample used to generate the regression equation.
   d. All of the above.
2. Consider the assumption that X and Y have a bivariate normal distribution in the population. For each purpose below, state whether that assumption is or is not necessary.
   a. To calculate r
   b. To calculate the CI on r
   c. To calculate prediction $\hat{Y}$ for a particular value of X
   d. To calculate the CI on b
   e. To calculate the prediction interval for Y at a particular value of X
3. Ange is using regression to use GRE scores to predict graduate school success. For each purpose listed below, which of the following intervals should she use?
   (i) CI on b, (ii) CI for mean of Y at X, (iii) prediction interval for individual values of Y at a specific X.
   a. Ange wants to describe the typical level of success expected for students who score 160 on the GRE.
   b. Ange is helping to evaluate a particular applicant who has scored 169 on the GRE; she wants to express the likely range of possibilities for this applicant.
   c. Ange wants to express the degree to which GRE scores are related to graduate school success.
4. For the different intervals above, which will change the least as larger sample sizes are collected? Why?
5. If the absolute value of the correlation of two variables is less than ____, the estimate of the predicted variable will be closer to / further from the mean than the value of the predictor variable used to make the estimate. This phenomenon is called _____ _____.

6. The couple next door sometimes have loud arguments, which I find disturbing. Whenever they have an especially loud exchange I concentrate hard and try to send them a mental message of calm. It seems to work—usually, within half an hour they are much quieter. Does this show I have wonderful powers of the mind? Explain.

12.21 An advertisement encourages people to enroll for further training by declaring that every extra year of education increases annual income by an average of $7,000.

a. Assuming that figure comes from a regression analysis, explain what it's telling us.
b. What assumptions are implicit in the encouragement of the ad? Are they reasonable?

12.22 I told you about my belief that wearing a pink hat, and so forth, fixes any headache I have.

a. Suppose I always follow that ritual whenever I have a headache. Is my belief in the effectiveness of the ritual likely to weaken or strengthen over time? Explain.
b. You might regard my belief in that ritual as superstition. (Of course, I don't.) Describe some other superstition that might be perpetuated in the same way.
c. Tell me about a belief of yours that's perpetuated in the same way. You don't need to answer this publicly.

12.23 You have developed a computer game to help teach spelling. A group of children who scored more than one standard deviation below their age norm on a spelling test played your game for two sessions. They took the test again a week later and obtained a distinctly higher average score.

a. Should you be encouraged? Explain.
b. Suggest a better design of study to assess your game.

12.24 If you wish, revise your take-home messages.

Reporting Your Work

Regression provides a chance to quantitatively explore the relationship between two variables. Reporting it is similar to reporting correlation, but focuses on the regression equation. Your research report should typically include these elements:

- Whether the regression analysis is planned or exploratory, unless this is already clear
- Basic descriptive statistics for both the X and Y variables
- A scatterplot with the regression line
- The regression equation, including the CI for the slope
- If desired, the p value for the slope, usually against the null hypothesis that $b_{population} = 0$
- If desired, the standardized slope, β, and its CI; For regression with only one predictor variable, this is the same as r and its CI, so you do not need to report both
- Discussion of the regression equation that focuses on the slope and its CI; specifically, consider how the slope informs thinking about how changes in the X variable relate to changes in the Y variable—but be careful to avoid causal language
- A discussion of the degree to which the assumptions for regression are met; when assumptions are violated, be sure your interpretation is suitably tentative

Reporting regression is similar to reporting correlation but focuses on the regression equation.

```
As planned, we used linear regression analysis to predict
well-being scores from body satisfaction scores (b = 0.82,
95% CI [0.51, 1.12], a = 2.10). This CI is fairly long but
indicates that each unit of additional body satisfaction is
associated with a half up to a full unit of additional well-
being (1-7 scale). Relative to other predictors of well-being
this is a strong relationship. One caution in this analysis is
the restricted range of the body satisfaction scores, as all
participants scored at least 2 on this 1-5 scale.
```

Often regression analysis is reported simply to provide a quantitative sense of how the X and Y variables are related. Clearly, though, an additional benefit of regression analysis is the ability to make predictions about general trends or even specific individuals.

- If you are making predictions only about general trends in the Y variable (what mean Y is likely for various values of X), then include the CI for the mean of Y for each value of X of interest (and include these curves in the scatterplot).
- If you are making predictions about specific individuals, then include Y prediction intervals (and include the prediction interval lines in the scatterplot).

For example:

```
Based on this regression analysis, participants with a body
image score of 2 would be predicted to have an average well-
being of 3.74, 95% CI [3.21, 4.26]. This indicates that the
average well-being in this group is likely to be moderate.

Based on this regression analysis, a participant with a body
image score of 2 would be predicted to have a well-being score
of 3.74, 95% PI (1.62, 5.85). This is a very long prediction
interval, but it indicates that a participant at this low-
end score for body satisfaction is very unlikely to be at the
highest levels for well-being.
```


Take-Home Messages

- Like correlation, the regression line is an expression of the linear component of the relationship of X and Y, but neither may provide the full story of that relationship. In neither case is there necessarily any causation.
- The regression line for Y on X minimizes the SD of the $(Y - \hat{Y})$ residuals. It passes through the means point (M_x, M_y) and has slope $b = r \times (s_y/s_x)$, which is r times the slope of the $r = 1$ line.
- The standardized regression line of Z_y on Z_x passes through (0, 0) and has slope r.
- The proportion of s^2_Y, the total variance of Y, that can be accounted for by X is r^2; the remaining proportion is $(1 - r^2)$, which is $s^2_{Y.X}$, the variance of the residuals.
- To use a $\hat{Y}$ estimate for a new case with a particular value of X, we need to assume Y is randomly sampled and that the new case comes from the same population as the data set. Calculating any CI, or a prediction interval, requires in addition the assumptions that Y comes from a normal population and that the variance of Y is homogeneous for all X.
- The CI on b, the sample regression slope, is the interval estimate for $b_{population}$, the population regression slope.
- Curved CI lines mark the lower and upper limits of the CI for the mean of the Y population at each X value.
- The prediction interval for individual values of Y at a particular X value is usually long, because it reflects both uncertainty in the estimate of mean Y at that value of X, and the spread of individual Y values.
- Unless two variables are perfectly correlated, an estimate for the predicted variable shows *regression to the mean*, a natural phenomenon occurring frequently in the world

End-of-Chapter Exercises

1) It probably comes as no surprise to learn that friendly people tend to have more close friends than unfriendly people. To what extent can you predict how many friends someone has just from knowing how friendly they are? To investigate, 64 psychology majors completed a questionnaire measuring their friendliness and their number of close friends. The two variables were moderately correlated, $r = .34$, 95% CI [.10, .54]. Table 12.1 shows descriptive statistics.

Table 12.1 Summary Statistics for Friendliness and Numbers of Close Friends for $N = 64$ Students

Friendliness (scale from 1–5), X		Number of close friends (open ended), Y	
M	s	M	s
3.58	0.43	8.33	6.27

a. Think about those values. Does it seem reasonable to assume Y is normally distributed in the population?

b. Before using this information for regression, it would be best to see a scatterplot. What would you be looking for to confirm that regression is appropriate?

c. We'd like to use friendliness to predict number of close friends. What is the slope in the regression equation?

d. What is the intercept in the regression equation?

e. Using the regression equation, how many friends do you predict for someone who is fairly unfriendly ($X = 2$)?

f. Using the regression equation, how many friends do you predict for someone who is very friendly ($X = 5$)?

g. No mathematics, just think: How many friends do you predict for someone of exactly average friendliness ($X = 3.6$)? Why? Use the regression equation to check your intuition.

h. If you find out that your participant of average friendliness actually has 10 friends, what is the residual of prediction?

i. Calculate the regression equation going the other way: using number of close friends to predict friendliness.

j. What if someone reported having 300 close friends? What would their predicted level of friendliness be? Recall that friendliness was measured on a scale from 1 to 5. What's gone wrong using regression?

k. The sample data came from undergraduate psychology majors in the United States. Would it be reasonable to use this data set to make predictions about European psychology majors? About U.S. high school students? About Facebook users?

2) Maybe you're thinking about buying a house after college? Regression can help you hunt for a bargain. Use the book website to download the **Home_Prices** data set. This file contains real-estate listings from 1997 to 2003 in a city in California. Let's explore the extent to which the size of the home (in square meters) predicts the sale price.

a. Use ESCI to create a scatterplot of the relationship between home size (X) and asking price (Y). Does this data set seem suitable for use with regression?

b. To what extent is home size and asking price related in this sample? What is the 95% CI for the relationship between home size and asking price in the population of houses?

c. What is the regression equation for predicting asking price from home size?

d. Use ESCI to show the residuals for home prices—notice that some houses are listed at prices that fall above the regression line and others at prices that fall below. If you are hunting for a bargain, which type of house would you want to look at? Why? What is the house with the largest negative residual?

e. Does a large residual mean that the seller has made a mistake and should adjust the asking price? Why or why not?

f. If a house has a size of 185.8 m², what is the predicted asking price?

g. How well can we predict the mean asking price for houses of 185 m²? Use ESCI to obtain the 95% CI for the mean of Y when $X = 185$.

h. How well would our predictions hold up with a new data set? To investigate, 10 further cases, not included in the regression analysis, were taken from the same population. These are listed in Table 12.2, which is partially completed. For each house, use ESCI to predict price from house size. Record the prediction ($\hat{Y}$), the 95% prediction interval (PI), and calculate the residual ($Y - \hat{Y}$). In the second column from the right, record whether or not the PI includes the asking price. For the first two houses, check you get the same values as shown. Fill in values for the last four houses. In how many of the 10 cases does the PI include Y, the asking price? Calculations to the nearest $1,000 are fine.

Table 12.2 A Further Sample of 10 Houses In a Californian City

Case	Size, X (m²)	$\hat{Y}$ ($1,000)	95% PI ($1,000)		Asking Price, Y ($1,000)	Within PI? (Y/N)	Residual ($1,000)
1	133.8	297	−56	650	149	Y	−148
2	158.0	362	9	716	549	Y	187
3	142.7	321	−32	674	435	Y	114
4	121.7	264	-89	617	299	Y	35
5	203.2	485	132	838	625	Y	140
6	195.1	463	110	817	399	Y	−64
7	140.4				187		
8	197.9				1,290		
9	130.1				265		
10	113.8				199		

i. For a given X, our 95% PIs are really long. Yet we can predict the *mean* of Y for a given X with a very short 95% CI (as in g above). Why the difference? Would an even larger sample size help shorten the PIs substantially? Explain.

j. The data for this regression analysis were collected from 1997 to 2003. In 2015, a realtor is asked to help sell a house in the same city that is 99.1 m². What is the predicted price for this house using the regression equation you have from the 1997–2003 data set? Would it be reasonable to use this prediction in setting a price for this house? Why or why not? Do regression equations have expiration dates? Should they?

3) Happiness may not be important just for the person feeling it; happiness may also promote kind, altruistic behavior. Brethel-Haurwitz et al. (2014) examined this idea by collecting data on U.S. states. A Gallup poll in 2010 was used to measure each state's well-being index, a measure of mean happiness for the state's residents on a scale from 0 to 100. Next, a

kidney donation database for 1999–2010 was used to figure out each state's rate (number of donations per 1,000,000 people) of non-directed kidney donations—that's giving one kidney to a stranger, an extremely generous and altruistic thing to do! You can download the data from the book website (**Altruism_Happiness**).

a. Use ESCI to create a scatterplot of the relationship between a state's 2010 well-being index (X) and rate of kidney donations (Y). Any comments on the scatterplot? Does this data set seem suitable for use with regression?

b. To what extent is a state's well-being index and rate of kidney donation related in this sample? What is the 95% CI for the relationship in the population? What does "population" refer to?

c. What is the regression equation for predicting rate of kidney donation from well-being index?

d. In Table 12.3 record the predictions for four states, in the column labeled $\hat{Y}$ (2010).

Table 12.3 Well-Being Means for Four States, for 2010 and 2013

State	Well_Being_2010	$\hat{Y}$ (2010)	Well_Being_2013	$\hat{Y}$ (2013)
WY	69.2		65.6	
HI	71		68.4	
ND	68		70.4	
NV	64.2		66.6	

e. The table includes updated well-being means, for 2013. Use those to make predictions, and record these in the column labeled $\hat{Y}$ (2013). Compare with the predictions based on the 2010 well-being data. Discuss which we should use.

f. To make predictions using the 2013 well-being means you used the regression equation generated using the 2010 well-being data. Discuss the extent that's reasonable.

g. From 2008 to 2013 Gallup measured state-wide well-being based on six indicators: life evaluations, emotional health, work environment, physical health, healthy behavior, and access to services. In 2014, however, Gallup changed the way it measures state-wide well-being; it now bases scores on five somewhat different indicators: purpose, social life, financial satisfaction, community, and physical health. Can the regression equation you developed be used with 2014 well-being data? Discuss.

Answers to Quizzes

Quiz 12.1

1) Y, X, $\hat{Y}$; 2) slope, intercept; 3) A residual is an error in prediction, calculated as $(Y-\hat{Y})$, which is the difference between the actual and predicted value of Y; 4) most useful is c., $r = -.6$; least useful is a., $r = 0$. The best predictions come from the strongest relationship, but the direction of the relationship does not matter for making good predictions; 5) slope; 6) d.

Quiz 12.2

1) c; 2) multiplying Z_x by r (so $Z_{\hat{Y}} = r \times Z_X$); 3) slope, r; 4) d; 5) 16, 1, .8, −.8; 6) $(Y-\hat{Y})$, $(X-\bar{X})$.

Quiz 12.3

1) d; 2) No, yes, no, no, no (for c–e, less restrictive assumptions are required); 3) a. needs (ii), b. needs (iii), c. needs (i); 4) The prediction interval for individual Y at a particular X will change least because this depends not only on uncertainty in the estimate of mean Y at that value of X, but also on the spread of individual Y values at that level of X, which doesn't change for different N; 5) 1, closer to, regression to the mean; 6) I may have such powers (I suspect you do too), but this observation doesn't give good supporting evidence. Most likely, the noise level is lower 30 minutes after a very loud exchange simply because of variation over time and regression to the mean.

Answers to In-Chapter Exercises

12.2 a. From 0 (horizontal) to 0.87; b. It's large for slope 0 or 0.87 and smaller in between; c. When the slope is 0.78, the same (or almost the same) as the calculated value of b, then $s_{Y.X} = 1.37$, its minimum value, and the line and values change from red to black.

12.3 a. $\hat{Y} = 4.22 + 0.78 \times X$; b. 11.24, small, because r is large and therefore the points are not scattered widely; c. For $X = 0$, $\hat{Y} = 4.22$, but $X = 0$ is way outside the range of X for which we have data, and the relationship of X and Y might not be the same so far beyond that range. A pretest score of 0 might suggest an absent or delinquent student, so $\hat{Y} = 0$ might be a better guess, but that's a guess—not a regression estimate.

12.4 b. $\hat{Y} = 11.24$ is our best point estimate of Y when $X = 9$, based on the regression line.

12.5 a, b. All the same.

12.6 a. 12, 10; b. Those points do not lie on the regression line, and therefore their Y values do not equal $\hat{Y} = 11.24$, which is no problem.

12.7 a. The point reduces r from .89 to .74 because it's an outlier; b. Slope changes from 0.78 to 0.76. The line drops, because of the low outlier, but rotates only a little because the new point is close to the mean of X. The line represents the points less well with the outlier added; c. The point (18, 7) is much more influential than (12, 7), making a much larger change to r, because it's more extreme, especially by being extreme on X as well as Y. Points that are outliers on both X and Y have much more influence on r and b.

12.8 a. In both scatterplots, r and the regression slope change greatly. In the left figure the means cross move, but not in the right figure; b. $\beta = r$ and is .89 for the original points and .40 with the additional point. Both regression slopes change accordingly.

12.9 a. Today's temperature, X, and tomorrow's, Y; b. $Z_X = (14 - 20)/4 = -1.5$. $Z_{\hat{Y}} = .6 \times (-1.5) = -0.90$, so prediction is $\hat{Y} = 20 + (-0.90) \times 4 = 16.4$ °C, c. If $r = 0$, predict 20 °C, the mean; d. Closer because regression line is steeper.

12.10 a. 2; b. $Z_{\hat{Y}} = .5 \times 2 = 1$. $\hat{Y} = 168$ cm; c. For $X = 168$ for the daughter, $\hat{Y} = 165$ cm for the granddaughter. $\hat{Y} = 163.5$ cm for the great-granddaughter.

12.11 a. $Z_{\hat{Y}} = .5 \times 2 = 1$; b. $\hat{Y} = 168$ cm is the best estimate of a mother's height, given her daughter (Susan) is $X = 174$ cm tall.

12.12 a. 64%, so 36% of variance in Y is the variance of the residuals, which no doubt fits with Panel A, although variance is hard to eyeball; b. For Panel C, 4% and 96%, again perhaps reasonable.

12.13 a. When r is close to 1 or −1; b. When $r = 1$ or −1, the two regression lines are the same, with slope (s_Y/s_X) or $-(s_Y/s_X)$; c. They are most different when $r = 0$, in which case the Y on X line is horizontal at M_Y, the mean of Y, and the X on Y line is vertical at M_X.

12.14 a. $\hat{Y} = 2.10 + 0.82X$; $b = 0.82$ is the regression slope in the data, meaning the slope of the regression line of Y on X; b. The CI on b is [0.52, 1.12], so the slope in the population could plausibly be anywhere in that fairly long interval; c. At red 12, ESCI shows $p = 0$ (which we'd report as $p < .001$) for the null hypothesis of zero regression slope in the population; d. At red 7, $p = 0$. The two p values are the same, as we'd expect, because if $r = 0$ the slope of the regression line is also 0; e. To calculate the regression line we're assuming random sampling of Y (Well-being), which is probably reasonable. To calculate the CI on b, and the p value at red 12, we need to assume also that Y is normally distributed in the population, with variance that is homogeneous for all X. To calculate the p value at red 7, for correlation, we need the stronger assumption that the data are from a bivariate normal population. All those assumptions are probably reasonable.

12.15 a. The CI tells us that the mean height of all women who have a best woman friend who is 174 cm tall lies, most likely, in that interval; b. The CI for Panel A tells us the same, but for all women who have an identical twin who is 174 cm tall. It's much shorter because the correlation is much higher, so we have much better information about Maria's likely height.

12.16 a. 64%, 4%, and these values reflect the very different slopes of the two regression lines; b. 36%, 96%; c. The differing Y residuals in the two scatterplots—the differing amounts of scatter of the points around the regression line in A and C.

12.17 The CI estimates the mean height of all women with a best woman friend who's 174 cm tall, whereas the much longer prediction interval reflects that uncertainty and in addition the variation in height of all those individual women. It tells us about the spread in the distribution of heights of all women with a best woman friend who's 174 cm tall.

12.18 a. $\hat{Y}$ = 73.9 [72.1, 75.8], so 73.9 is the best point estimate of population mean posttest HCTA when the pretest is 72, and that CI is the interval estimate for that mean; b. The prediction interval is (64.2, 83.7) and tells us about the full distribution of individual posttest scores when pretest = 72.

12.19 Performance week to week no doubt shows some variability, so the correlation is less than 1, and therefore regression to the mean must occur: an extreme performance this week is likely to be followed by a less extreme performance next week. There is no need to invoke any extra reason, such as a player being put off their game. It may be worth bringing the awards back.

12.20 Headache usually varies over time, so regression to the mean could explain why a headache now is, on average, followed by lesser headache in an hour. My ritual could easily be having no effect, unless my belief is strong enough to produce a placebo effect. Alas, I can't expect my headache treatment to make me famous.

12.21 a. The $7,000 is no doubt *b* for the regression line of annual income on years of education, probably based on data for a large sample of people. It implies that, on average, income increases by that amount for an additional year of education; b. The ad assumes causality of income by education, whereas no doubt there are many other variables involved. For example, intelligence, family background, and personal motivation may be, to some extent, causal for both years of education, and income.

12.22 a. By regression to the mean, quite often my headache reduces and so my belief in the ritual could be strengthened; b. Other superstitions that could be reinforced in the same way need to predict an event likely to happen because of regression to the mean. The superstition must relate to an extreme value on *X* leading to a less extreme value on *Y*, where *X* and *Y* are correlated. A belief that it rains more at weekends does not qualify, but a belief that shouting at the clouds on a very wet day causes the following day to be less wet does qualify.

12.23 a. Children selected for scoring poorly are likely to do less poorly, on average, at a second testing, by regression to the mean—so your game may have no effect; b. A better study would randomly allocate your poor spellers either to a group who work with your game, or another group who have some comparison activity, perhaps working with a textbook.

13

Frequencies, Proportions, and Risk

So far in this book we've mainly used effect size measures that rely on interval scaling—including means, Cohen's *d*, and correlations. In this chapter we take a step back to require only nominal scaling and focus on the *proportion*, which is a highly useful ES measure based on frequencies, or simple counts. We'll consider research questions for which a proportion provides the answer and, as we've done so many times with other ES measures, we'll focus on the CI as well as the point estimate. Then we'll discuss research questions where we need the difference between two proportions and the CI on that difference. My first example investigates the possibility of telepathy, which is communication with another person using a psychic power, through "the power of the mind".

Do some people have psychic powers, powers that science doesn't know about? Can some people bend spoons at a distance, or use telepathy to communicate with others? Some claimed psychics give highly convincing demonstrations, which persuade many people that their psychic powers are real. However, stage magicians can also impress us with demonstrations of what look like psychic phenomena. Many people believe psychic powers exist, while many others are convinced that they don't.

Some scientists investigate possible psychic powers, under the label of *parapsychology*. Some studies of telepathy are especially relevant here because they can be analyzed simply by calculating the proportion of trials that were correct. The *proportion* is a simple effect size measure that's calculated from frequencies.

The agenda for this chapter is:

- Psychic powers and research in parapsychology
- Frequencies, nominal scaling, and proportions
- The CI on a proportion
- The difference between two proportions, and the CI on this difference
- Frequency tables analyzed using proportions, and via an alternative approach: the chi-square statistic
- Another application of proportions: risk, and the difference between risks

RESEARCH ON POSSIBLE PSYCHIC POWERS

Mentalists are stage magicians with a particular interest in using their skills to demonstrate what look like psychic phenomena. Many are both convincing and entertaining. Some magicians challenge psychics to demonstrate their claimed powers under conditions allowing scientific scrutiny. For a famous example, search online for "James Randi"—who has long had on offer a cash prize, now $1,000,000, for anyone who can demonstrate psychic powers under agreed scientific conditions. So far, he has not had to pay up. For an entertaining take on psychic powers, see the Woody Allen movie *Magic in the Moonlight*. You can