Let's do it! 13.9 Is There an Error?

Identify and explain the error in each of the following situations. If there is no error, write *no error*.

(a) Our study shows that the correlation between the number of hours a child spends watching television per week and the child's reading level is $r = -1.04$.

(b) We found that the correlation between a voter's political party and income is $r = 0.38$.

Interactive Statistics, 2nd Ed.
Aliaga / Gunderson, Prentice Hall © 2003

Given that the scatterplot shows a linear relationship is plausible, a strong correlation, a value of r close to $+1$ or -1, tells us that the two variables are closely associated in a linear way. A positive correlation implies that they increase together, while a negative correlation implies that as one variable increases the other tends to decrease. We cannot stress enough the importance of examining the scatterplot along with the reporting of a correlation. Correlations can be misleading. Any study dealing with relationships must be on the lookout for possible confounding variables. Even a very strong correlation does not imply that changes in the explanatory variable will actually cause changes in the response variable. Outliers can also have quite an impact on correlations. The next examples illustrate these cautionary notes regarding correlation.

❖ EXAMPLE 13.7 Watch Out for Third Variables

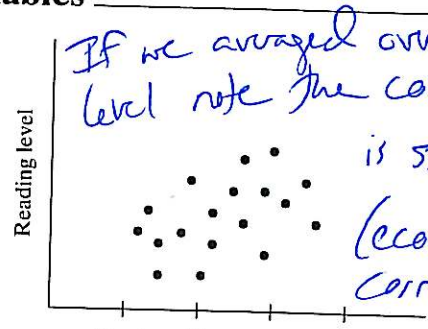
Consider the accompanying scatterplot. The correlation is positive, indicating that an increase in the number of TV hours watched by children corresponds to an increase in the reading level of the children. What is going on here?

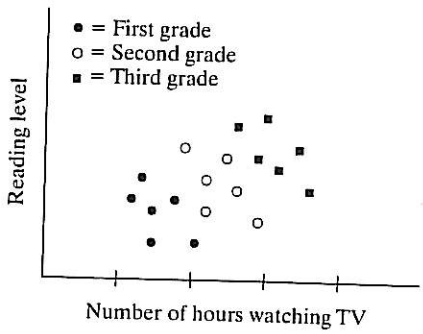
Details of the data tell you that the children were of different ages, ranging from first grade through third grade.

The lower scatterplot at the right shows you the data for the three separate grades, denoted by the different plotting symbols. If you examine the data for each grade separately, you see that the correlation coefficient within each grade is negative, not positive. The variable *grade* is a third variable that is *confounded* with the response, reading level.

The comparison between the two variables is complicated because the subjects are at different grade levels. *Aggregating the results* across the three grades is not appropriate. This is an example of Simpson's paradox, which we will see again in Chapter 14.

If we averaged over grade level note the correlation is stronger (ecological correlation)





❖

(f) If every woman dated a man exactly 3 inches taller than she, what would be the correlation between male and female heights?

(g) Compute the slope of the regression line of male height on female height.

(h) Suppose that the heights of the men were measured in centimeters (cm) rather than in inches, but that the heights of the women remained in inches. (There are 2.54 cm to an inch.)

(i) Now what would be the correlation between male and female heights?

(ii) What would be the new slope of the regression line of male height on female height?

13.32 A random sample of seven households was obtained, and information on their income and food expenditures for the past month was collected. The data (in hundreds of dollars) are given.

Income ($100s) ($x$)	22	32	16	37	12	27	17
Food Expend. ($100s) ($y$)	7	8	5	10	4	6	6

(a) Make a scatterplot of the data, find the least squares regression line, and superimpose it on your scatterplot.

(b) Should you predict the food expenditure for a household with an income of $5200? Explain.

(c) How much would food expenditure change for a $100 increase in income?

(d) Find the residual for the third observation ($x = 16, y = 5$).

(e) Is the following statement true or false? "The correlation coefficient between income and food expenditure is approximately $0.93." Explain.

13.8 Regression Effect*

- Will a special reading program improve the reading levels for students with extremely low reading levels?
- Will carrying a "good luck" charm improve the batting average of baseball players experiencing an early season slump?
- Will an extra review section for students whose Exam 1 scores were the lowest lead to higher scores on Exam 2?

Although these interventions may have some effect, another possible explanation, due to nonrandom sampling of the subjects, is called the **regression effect**.

Definition: In nearly all test–retest situations, the lowest group on the first test will on average show some improvement on the second test—and the highest group on the first test will on average perform worse. This is the **regression effect**.

*This is an interesting, but optional, topic.

This regression effect can also be described in terms of a "skill" and "luck" model for test scores. We can think of an individual's test score as being equal to the actual skill level of that individual plus some contribution due to luck. A common model for the luck component is to assume that luck averages out to zero.

$$\textbf{Test 1} = \textbf{Actual Skill Level} + \textbf{Luck for Test 1}$$
$$\textbf{Test 2} = \textbf{Actual Skill Level} + \textbf{Luck for Test 2}$$

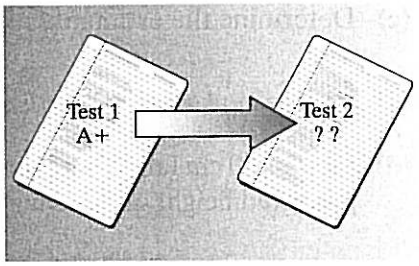
A very high score on Test 1 is generally due to a high skill level and some good luck. Since the luck component, also referred to as *chance* or *random variation*, has an expected value of zero, we would *expect* a high score on Test 2, due to the common high skill level, but not quite as high as Test 1.

We present the idea of the regression effect in this chapter because it arises in situations when the two measures are associated. In this case, it is possible to express the expected response (expected value of y) at a given value of x in terms of this correlation. Starting with the least squares regression equation, and doing a bit of algebra, we can reexpress the least squares regression line as follows:

$$\frac{\text{expected } Y \text{ (for given } x) - \text{ average of } Y}{\text{standard deviation of } Y} = (r)\frac{x - \text{ average of } X}{\text{standard deviation of } X}.$$

Since the correlation is generally a fraction, $-1 < r < 1$, the quantity on the left side is smaller, in absolute value, than the quantity to the right of the correlation. For example, if the value of x is three standard deviations above the average of X and the correlation is 0.6, then the expected value of Y for this value of x is only $1.8(=(0.6)3)$ standard deviations above the average of Y—also extreme, but not as extreme as the value of x.

❖ EXAMPLE 13.12 Test Scores



Consider two test scores you obtain in a course. Suppose that you score three standard deviations higher than the average on the first test. What do you expect your score to be on the subsequent test? Your score on the first test is substantially higher than the average—an indication of your effort, ability, and perhaps some good luck. Your effort and ability may still provide you with an above-average performance on the second test, but the contribution from *luck* may not be the same.

The presence of a common factor contributing to both test scores implies the correlation between the test scores is positive. And the regression effect says we *expect* a lower performance on the second test, relative to the standards on that test—but this is expected, on average, and does not apply to every individual. If we wanted to learn whether or not positive reinforcement works, we must use this "treatment" on a random sample of students—not just those whose performance was exceptionally high on the first test. ❖

❖ EXAMPLE 13.13 Alzheimer's Disease

Alzheimer's is a disease that generally fluctuates quite a bit from day to day—random fluctuations. When testing various drug treatments, it is very difficult to separate what portion of the "improvement" is due to the regression effect, what portion is due to the random fluctuations, and what portion is attributed, if any, to the actual treatment.

What is it called when the effects of various factors cannot be separated? The factors are said to be confounded. ❖

Be Careful: Two Issues Regarding the Regression Effect

1. It is a statement about what we would expect on average—it says nothing about variance. The variation in the measurements is not decreasing.
2. It is a statement about what we would expect on average—it says nothing about what will occur for an individual.

Let's do it! 13.16 Heights of Mothers and Daughters

In a study on the relationship between the heights of mothers and daughters, a regression model was proposed with x, the explanatory variable, the height of the mother, and y, the response variable, the height of the daughter. Two hundred mother–daughter pairs were selected and their heights were measured in centimeters. The following statistics were calculated:

$$\bar{x} = 150, \quad s_X = 10, \quad \bar{y} = 150, \quad s_Y = 10, \quad r = 0.8.$$

(a) What are the units of measurement for $\bar{y}$, s_Y, and r?

(b) If we measure heights in feet instead of centimeters, will the value of the correlation change?

(c) Determine the estimated regression line.

(d) Jane is 140 cm tall. What is her standardized height—that is, what is her z-score? Give the predicted height of her daughter, first in centimeters, and then also in standardized units.

(e) Repeat (d) for Mary of height 150 and Betty of height 160.

(f) The predicted heights of the three daughters have been pulled toward the mean. This is why the regression effect is also referred to as *regression toward the mean*. Do you think the variance of heights of all daughters is less than the variance of the heights of all mothers? Explain.