



AN EXTENDED EXPLORATION

LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

OVERVIEW

Is learning the key to earning? Does going to school pay off? In this extended exploration, you use a large data set from the U.S. Bureau of the Census to examine ways in which education and earnings may be related. Technology is used to fit lines to data, and you learn how to interpret the resulting linear models, called regression lines. You can explore further by finding evidence to support or disprove conjectures, examining questions raised by the analysis, and posing your own questions.

In this exploration, you will

- analyze U.S. Census data
- use regression lines to summarize data
- make conjectures about the relationship between education and earnings in the United States
- find evidence to support or refute your conjectures
- examine the distinction between correlation and causation

Using U.S. Census Data

Does more education mean higher earnings? The answer may seem obvious. We may reasonably expect that having more education gives access to higher-paying jobs. Is this indeed the case? We explore how a social scientist might start to answer these questions using a random sample from U.S. Census data. Our data set, called FAM1000, provides information on 1000 individuals and their families.

The Bureau of the Census, as mandated by the Constitution, conducts a nationwide census every 10 years. To collect more up-to-date information, the Census Bureau also conducts a monthly survey of American households for the Bureau of Labor Statistics called the Current Population Survey, or CPS. The CPS is the largest survey taken between census years. The CPS is based on data collected each month from approximately 50,000 households. Questions are asked about race, education, housing, number of people in the household, earnings, and employment status.¹ The March surveys are the most extensive. Our sample of census data, FAM1000, was extracted from the March 2006 Current Population Survey. It contains information about 1000 individuals randomly chosen from those 16 and older who worked at least 1 week in 2005.



You can use the FAM1000 data and the related software, called *FAM1000 Census Graphs*, to follow the discussion in the text and/or conduct your own case study. The full FAM1000 data set is in the Excel file FAM1000, and condensed versions are in the graph link files FAM1000 A to H. These data files and related software can be downloaded from the web at www.wiley.com/college/kimeclark or on your class Wiley Plus site. The software provides easy-to-use interactive tools for analyzing the FAM1000 data.

Table 1 on page 147 is a *data dictionary* with short definitions for each data category. Think of the data as a large array of rows and columns of facts. Each row represents all the information obtained from one particular respondent about his or her family. Each column contains the coded answers of all the respondents to one particular question. Try deciphering the information contained in the first row of the 1000 rows in the FAM1000 data set in Table 2.

age	sex	region	cencity	marstat	famsize	edu	occup	hrswork	wkswork	yrft	pearnings	ptotinc	faminc	race	hispanic
42	2	4	1	0	3	12	4	40	52	1	\$25,000	\$25,559	\$46,814	5	0

Table 2

Referring to the data dictionary, we learn that the respondent is a 42-year-old female who lives in a city somewhere in the West. She is married with two other people in her family and she has a high school degree. In 2005 she worked in sales or a related occupation for 40 hours a week, 52 weeks of the year. Her personal earnings from work were \$25,000, but her personal total income, was \$25,559. That means she had an additional source of unearned income, such as dividends from stocks or bonds, or interest on a savings account. Her family income totaled \$46,814. She did not identify with any of the racial or Hispanic categories listed on the census form.

¹The results of the survey are used to estimate numerous economic and demographic variables, such as the size of the labor force, the employment rate, earnings, and education levels. The results are widely quoted in the popular press and are published monthly in *The Monthly Labor Review* and *Employment and Earnings*, irregularly in *Current Population Reports* and *Special Labor Force Reports*, and yearly in *Statistical Abstract of the United States* and *The Economic Report of the President*.

Data Dictionary for March 2006 Current Population Survey

Variable	Definition	Unit of Measurement (code and allowable range)	Variable	Definition	Unit of Measurement (code and allowable range)
age	Age	Range is 16 to 85	occup	Occupation group of respondent	7 = Construction and extraction occupations 8 = Installation, maintenance, and repair occupations 9 = Production occupations 10 = Transportation and material moving occupations 11 = Armed Forces
sex	Sex	1 = Male 2 = Female	hrswork	Usual hours worked per week	Range is 1 to 99
region	Census region	1 = Northeast 2 = Midwest 3 = South 4 = West	wkswork	Weeks worked in 2005	Range is 1 to 52
cencty	Residence location	1 = Metropolitan 2 = Nonmetropolitan 3 = Not Identified	yrft	Employed full-time year-round	1 = full-time year-round 2 = part-time year-round 3 = full-time part of the year 4 = part-time part of the year
marstat	Marital status	0 = Presently married 1 = Presently not married	earnings	Total personal earnings from work	Range is \$0 to \$650,000
famsize	Family size	Range is 1 to 39	ptotinc	Personal total income	Range is negative \$1,000,000 to \$10,000,000
educ	Years of education	8 = 8 or fewer years of education 10 = No high school degree, 9–12 years of education 12 = High school degree 13 = Some college 14 = Associate's degree 16 = Bachelor's degree 18 = Master's degree 20 = Doctorate 22 = Professional degree (e.g., MD)	faminc	Family income	Range is negative \$400,000 to \$24,000,000
occup	Occupation group of respondent	0 = Not in universe, or children 1 = Management, business, and financial occupations 2 = Professional and related occupations 3 = Service occupations 4 = Sales and related occupations 5 = Office and administrative support occupations 6 = Farming, fishing, and forestry occupations	race	Race of respondent	1 = White 2 = Black 3 = American Indian, Alaskan Native Only 4 = Asian or Pacific Islander 5 = Other
			hispanic	Hispanic heritage	0 = Not in universe 1 = Mexican 2 = Puerto Rican 3 = Cuban 4 = Central/South American 5 = Other Spanish

Table 1

Source: Adapted from the U.S. Bureau of the Census, Current Population Survey, March 2006, by Jie Chen, Computing Services, University of Massachusetts, Boston.

Summarizing the Data: Regression Lines

Is There a Relationship between Education and Earnings?

In the physical sciences, the relationship among variables is often quite direct; if you hang a weight on a spring, it is clear, even if the exact relationship is not known, that the amount the spring stretches is definitely dependent on the heaviness of the weight. Further, it is reasonably clear that the weight is the *only* important variable; the temperature and the phase of the moon, for example, can safely be neglected.

In the social and life sciences it is usually difficult to tell whether one variable truly depends on another. For example, it is certainly plausible that a person's earnings depend in part on how much formal education he or she has had, since we may suspect that having more education gives access to higher-paying jobs, but many other factors also play a role. Some of these factors, such as the person's age or type of work, are measured in the FAM1000 data set; others, such as family background or good luck, may not have been measured or even be measurable. Despite this complexity, we attempt to determine as much as we can by first looking at the relationship between earnings and education alone.

We start with a scatter plot of education and personal earnings from the FAM1000 data set. If we hypothesize that earnings depend on education, then the convention is to graph education on the horizontal axis. Each ordered pair of data values gives a point with coordinates of the form

(education, personal earnings)

Take a moment to examine the scatter plot in Figure 1. Each point refers to respondents with a particular level of education and income and has two coordinates. The first coordinate gives the years of education past grade 8 (So zero represents an

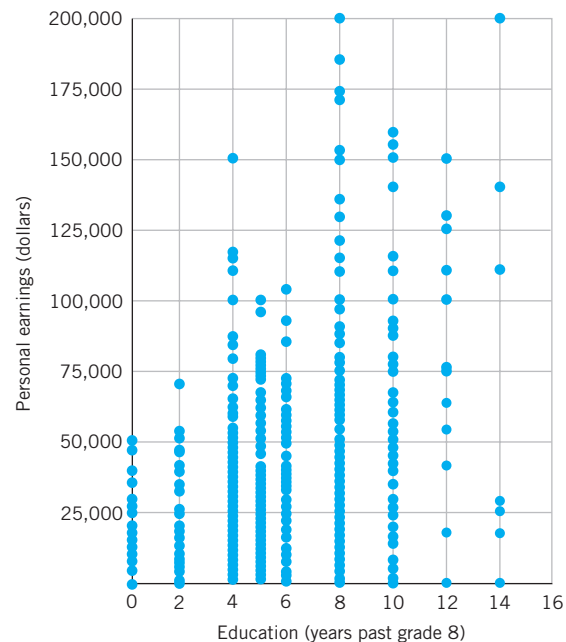


Figure 1 Personal earnings vs. education. In attempting to pick a reasonable scale to display personal earnings on this graph, the vertical axis was cropped at \$200,000, which meant excluding from the display the points for three individuals who each earned more than \$200,000 in wages.

eighth-grade education or less) and the second gives the personal earnings. For example, the points at the very top of the graph represent individuals who make \$200,000 in personal wages. One of these points represents those with 16 years (8 years past eighth grade) of education and the other, those with 22 years (14 years past eighth grade) of education. The coordinates for these points are (8, \$200,000) and (14, \$200,000). We can refer back to the original data set to find out more information about these points, as well as the outliers that are not shown on the graph. Note that some dots represent more than one individual (i.e., for a particular level of education there might be many people with the same income).

How might we think of the relationship between these two variables? Clearly, personal earnings are not a function of education in the mathematical sense since people who have the same amount of education earn widely different amounts. The scatter plot obviously fails the vertical line test.

But suppose that, to form a simple description of these data, we were to insist on finding a simple functional description. And suppose we insist that this simple relationship be a linear function. In Chapter 2, we informally fit linear functions to data. A formal mathematical procedure called *regression analysis* lets us determine what linear function is the “best” approximation to the data; the resulting “best-fit” line is called a *regression line* and is similar in spirit to reporting only the mean of a set of single-variable data, rather than the entire data set. It can be a useful and powerful method of summarizing a set of data.

We can measure how well a line represents a data set by summing the vertical distances squared between the line and the data points. The regression line is the line that makes this sum as small as possible. The calculations necessary to compute this line are tedious, although not difficult, and are easily carried out by computer software and graphing calculators. We can use the *FAM1000 Census Graphs* or Excel program to find regression lines. The techniques for determining regression lines are beyond the scope of this course.

In Figure 2, we show the FAM1000 data set along with a regression line determined from the data points. The equation of the line is

$$\text{personal earnings} = 4188 + 5693 \cdot \text{yrs. of educ. past grade 8}$$

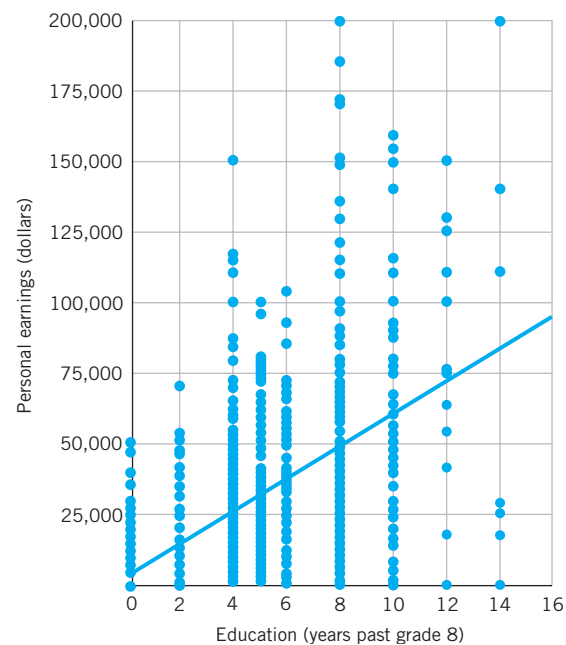


Figure 2 Regression line for personal earnings vs. education.



If you are interested in a standard technique for generating regression lines, a summary of the method used in the course software is provided in the reading “Linear Regression Summary.”

150 AN EXTENDED EXPLORATION: LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

Since personal earnings are in dollars, the units for the term 4188 must be dollars, and the units for 5693 must be dollars per year of education.

This line is certainly more concise than the original set of data points. Looking at the graph, you may judge with your eyes to what extent the line is a good description of the original data set. From the equation, the vertical intercept is 4188. Thus, this model predicts that individual personal earnings for those with an eighth-grade education or less will be \$4188. The number 5693 represents the slope of the regression line, or the average rate of change of personal earnings with respect to years of education. Thus, this model predicts that for each additional year of education, individual personal earnings increase by \$5693.

We emphasize that, although we can construct an approximate linear model for any data set, this does not mean that we really believe that the data are truly represented by a linear relationship. In the same way, we may report the median to summarize a set of data, without believing that the data values are at the median. In both cases, there are features of the original data set that may or may not be important and that we do not report.

The data points are widely scattered about the line, for reasons that are clearly not captured by the linear model. We can eliminate the clutter by grouping together all people with the same years of education and plotting the median personal earnings of each group. The result is the graph in Figure 3, which includes a regression line for the new data.

For each year of education, only a single median earnings point has been graphed. For instance, the point corresponding to 12 years of education past grade 8 has a vertical value of approximately \$75,000; hence the median of the personal earnings of everyone in the FAM1000 data with 20 years of education is about \$75,000. The pattern is now clearer: An upward trend to the right is more obvious in this graph.

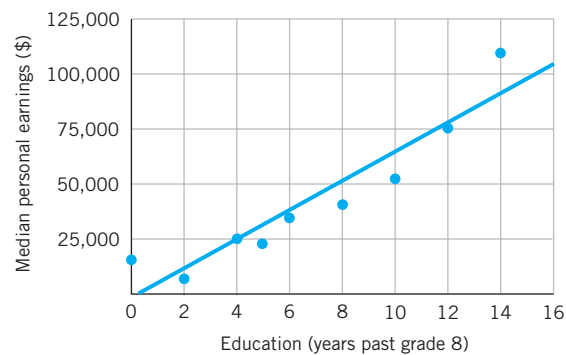


Figure 3 Median personal earnings vs. education.

Every time we construct a simplified representation of an original data set, we should ask ourselves what information has been suppressed. In Figure 3 we have suppressed the spread of data in the vertical direction. For example, there are only 36 people with an eighth-grade education or less but 175 with 16 years of education. Yet each of these sets is represented by a single point.

We can fit a line to the graph in Figure 3 using the same method of linear regression. The equation of this straight line is

$$\text{median personal earnings} = -2237 + 6592 \cdot \text{yrs. of educ. past grade 8}$$

Here, -2237 is the vertical intercept and 6592 is the slope or rate of change of median personal earnings with respect to education. This model predicts that for each additional year of education, median personal earnings increase by \$6592.

Note that this linear model predicts *median personal earnings for the group*, not personal earnings for an individual. The vertical intercept of this line is negative even though all earnings in the original data set are positive. The linear model is clearly inaccurate for those with an eighth-grade education or less.

Figure 4 shows two regression lines: One represents the fit to the medians and the other represents the fit to all of the data. Both of these straight lines are reasonable answers to the question “What straight line best describes the relationship between education and earnings?” and the difference between them indicates the uncertainty in answering such a question. We may argue that the benefit in earnings for each year of education is \$5693, or \$6592, or something between these values.

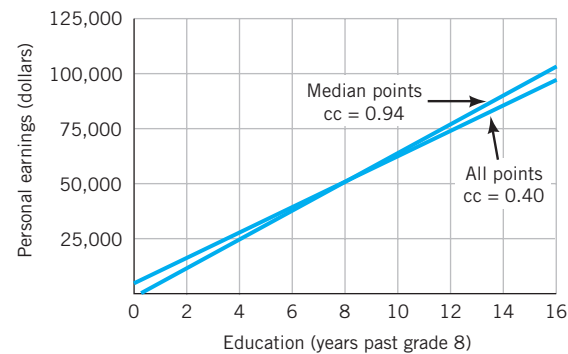


Figure 4 Regression lines for personal earnings vs. education.

Regression Line: How Good a Fit?

Once we have determined a line that approximates our data, we must ask, “How good a fit is our regression line?” To help answer this, statisticians calculate a quantity called the *correlation coefficient*. This number can be computed by statistical software, and we have included it on our graphs and labeled it “cc.”² The correlation coefficient is always between -1 (negative association with no scatter; the data points fit exactly on a line with a negative slope) and 1 (positive association with no scatter; the data points fit exactly on a line with a positive slope). The closer the absolute value of the correlation coefficient is to 1 , the better the fit and the stronger the linear association between the variables.

A small correlation coefficient (with absolute value close to zero) indicates that the variables do not depend linearly on each other. This may be because there is no relationship between them, or because there is a relationship that is something more complicated than linear. In future chapters we discuss many possible nonlinear functional relationships.

There is no definitive answer to the question of when a correlation coefficient is “good enough” to say that the linear regression line is a good fit to the data. A fit to the graph of medians (or means) generally gives a higher correlation coefficient than a fit to the original data set because the scatter has been smoothed out. (See the cc’s in Figure 4.) When in doubt, plot all the data along with the linear model and use your best judgment. The correlation coefficient is only a tool that may help you decide among different possible models or interpretations.



The programs R1–R4 and R7 in *Linear Regression* can help you visualize the links among scatter plots, best-fit lines, and correlation coefficients.



The reading “The Correlation Coefficient” explains how to calculate and interpret the correlation coefficient.

²We use the label “cc” for the correlation coefficient in the text and software to minimize confusion. In a statistics course the correlation coefficient is usually referred to as *Pearson’s r* or just *r*.

152 AN EXTENDED EXPLORATION: LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

EXAMPLE 1 Interpreting the correlation coefficient

Figure 5 contains the data and regression line for *mean* personal earnings vs. education.

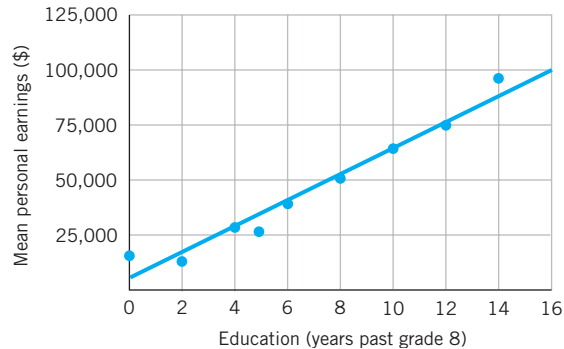


Figure 5 Mean personal earnings vs. education.

The equation of the regression line is

$$\text{mean personal earnings} = 5562 + 5890 \cdot \text{yrs. of educ. past grade 8}$$

$$cc = 0.98.$$

- Interpret the slope of this new regression line.
- Compare this regression line with that for median personal earnings vs. education previously cited in the text:

$$\text{median personal earnings} = -2237 + 6592 \cdot \text{yrs. of educ. past grade 8}$$

$$cc = 0.94$$

What do these equations predict for median and mean personal earnings for 12 years of education (high school)? For 18 years of education (master's degree)?

SOLUTION

- The slope of 5890 suggests that each additional year of education corresponds to a \$5890 yearly increase in mean personal earnings.
- Both equations have *cc*'s close to 1 (one is 0.98, the other 0.94), so both lines are good fits. Using the regression line for means, 12 years of education (or 4 years past grade 8), we get:

$$\begin{aligned} \text{mean personal earnings} &= \$5562 + \$5890(4) \\ &= \$29,122 \end{aligned}$$

$$\text{Rounding to nearest ten} = \$29,120$$

Using the regression line for means for 18 years of education (or 10 years past grade 8), we get:

$$\begin{aligned} \text{mean personal earnings} &= \$5562 + \$5890(10) \\ &= \$64,462 \end{aligned}$$

$$\text{Rounding to nearest ten} = \$64,460$$

Using the regression lines for medians and rounding to the nearest ten, 12 years of education corresponds to \$24,130 and 18 years to \$63,680.

Comparing predictions from the mean and median regression lines, we see that the regression line for means predicts higher earnings than the median regression line. This is reasonable since in general, when using income measures, the mean will often exceed the median since medians are not susceptible to outliers.

EXAMPLE 2 Examine the four scatter plots in Figure 6.

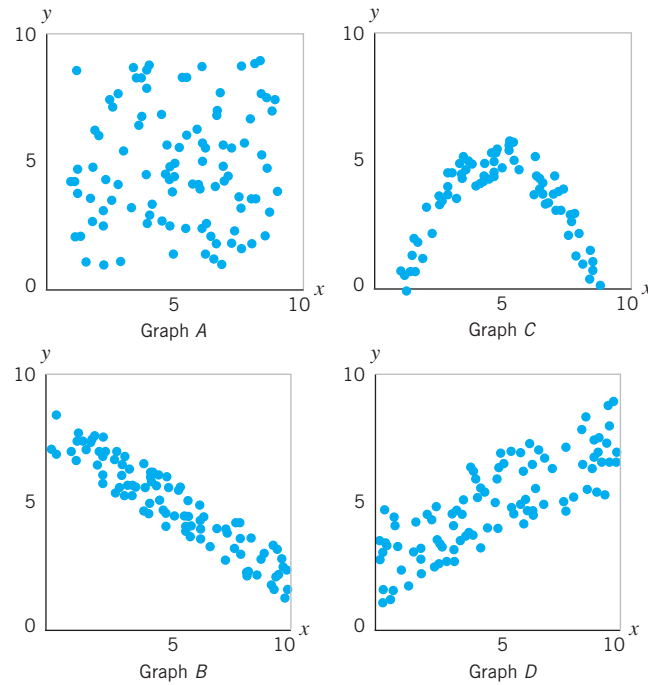


Figure 6 Four scatter plots.

- Which graph shows a positive linear correlation between x and y ? A negative linear correlation? Zero correlation?
- Which graph shows the closest linear correlation between x and y ?

SOLUTION

- Graph D shows a positive linear correlation between x and y (when one variable increases, the other increases). Graph B shows a negative linear correlation (when one variable increases, the other decreases). Both graphs A and C show zero correlation between x and y . Even though graph C shows a pattern in the relationship between x and y , the pattern is not linear.
- Graph B . The correlation coefficient of its regression line would be close to -1 , almost a perfect (negative) correlation.

Interpreting Regression Lines: Correlation versus Causation

One is tempted to conclude that increased education *causes* increased earnings. This may be true, but the model we have used does not offer conclusive proof. This model can show how strong or weak a relationship exists between variables but does not answer the question “Why are the variables related?” We need to be cautious in how we interpret our findings.

Regression lines show *correlation*, not *causation*. We say that two events are correlated when there is a statistical link. If we find a regression line with a correlation coefficient that is close to 1 in absolute value, a strong relationship is suggested. In our previous example, education is positively correlated with personal earnings. If education increases, personal earnings increase. Yet this does not prove that education causes an increase in personal earnings. The reverse might be true; that is, an increase in personal earnings might cause an increase in education. The correlation may be due

154 AN EXTENDED EXPLORATION: LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

to other factors altogether. It might occur purely by chance or be jointly caused by yet another variable. Perhaps both educational opportunities and earning levels are strongly affected by parental education or a history of family wealth. Thus a third variable, such as parental socioeconomic status, may better account for both more education and higher earnings. We call such a variable that may be affecting the results a *hidden variable*.

Figure 7 shows a clear correlation between the number of radios and the proportion of insane people in England between 1924 and 1937. (People were required to have a license to own a radio.)

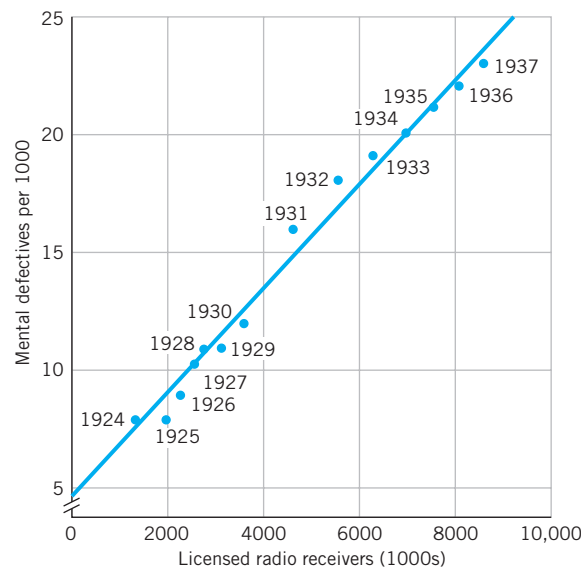


Figure 7 A curious correlation?

Source: E. A. Tufte, *Data Analysis for Politics and Policy*, p. 90.
Copyright © 1974 by Prentice Hall, Inc., Upper Saddle River, NJ.
Reprinted with permission.

Are you convinced that radios cause insanity? Or are both variables just increasing with the years? We tend to accept as reasonable the argument that an increase in education causes an increase in personal earnings, because the results seem intuitively possible and they match our preconceptions. But we balk when asked to believe that an increase in radios causes an increase in insanity. Yet the arguments are based on the same sort of statistical reasoning. The flaw in the reasoning is that statistics can show only that events occur together or are correlated, but *statistics can never prove that one event causes another*. Any time you are tempted to jump to the conclusion that one event causes another because they are correlated, think about the radios in England!

Raising More Questions

When a strong link is found between variables, often the next step is to raise questions whose answers may provide more insight into the nature of the relationship. How can the evidence be strengthened? What if we used the mean instead of the median, restricted ourselves to year-round full-time workers, or used other income measures, such as total personal income or total family income? Will the relationship between education and income still hold? Are there other variables that affect the relationship?

Do Earnings Depend on Age?

We started our exploration by looking at how earnings depend on education, because it seems natural that more education might lead to more earnings. But it is equally plausible that a person's income might depend on his or her age. People may generally earn more as they advance through their working careers, but their earnings usually drop when they eventually retire. We can examine the FAM1000 data to look for evidence to support this hypothesis.

It's hard to see much when we plot all the data points. This time we use *mean* personal earnings and plot it versus age in Figure 8. The graph seems to suggest that up until about age 50, as age increases, mean personal earnings increase. After age 50, as people move into middle age and retirement, mean personal earnings tend to decrease. So age does seem to affect personal earnings, in a way that is roughly consistent with our intuition. But the relationship appears to be nonlinear, and so linear regression may not be a very effective tool to explore this dependence.

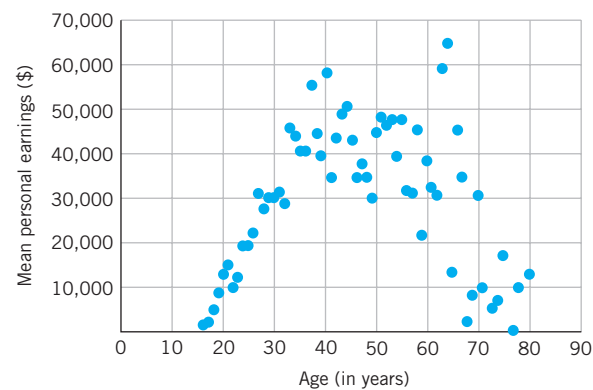


Figure 8 Mean personal wages vs. age.

The FAM1000 data set contains internal relationships that are not obvious on a first analysis. We might, for example, also investigate the relationship between education and age. Age may be acting as a hidden variable influencing the relationship between education and income.

There are a few simple ways to attempt to minimize the effect of age. For example, we can restrict our analysis to individuals who are all roughly the same age. This sample still would include a very diverse collection of people. More sophisticated strategies involve statistical techniques such as *multivariable analysis*, a topic beyond the scope of this course.

Do Earnings Depend upon Gender?

We can continue to look for relationships in the FAM1000 data set by using some of the other variables to sort the data in different ways. For example, we can look at whether the relationship between earnings and education is different for men and women. One way to do this is to compute the mean personal earnings for each year of education for men and women separately and restrict ourselves to those who work full-time year-round.

If we put the data for men and women on the same graph (Figure 9), it is easier to make comparisons. We can see in Figure 9 that for those working full-time year-round, the mean personal earnings of men are consistently higher than the mean personal earnings of women. We can also examine the best-fit lines for mean personal earnings versus education for men and for women shown in Figure 10.

156 AN EXTENDED EXPLORATION: LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

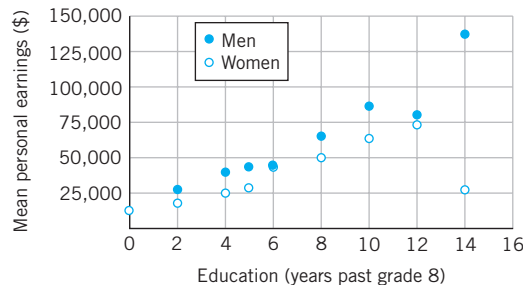


Figure 9 Mean personal earnings vs. education for women and for men (working full-time year-round).

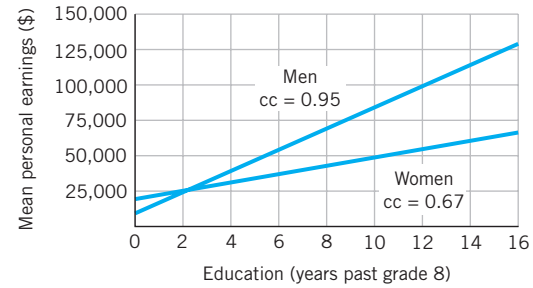


Figure 10 Regression lines for mean personal earnings vs. education for women and for men (working full-time year-round).

The linear model for mean personal earnings for men working full-time is given by

$$P_{\text{men}} = 9965 + 7423E$$

where E = years of education past grade 8 and P_{men} = mean personal earnings for men. The correlation coefficient is 0.95. The rate of change of mean personal earnings with respect to education is approximately \$7423/year. For males in this set, the mean personal earnings increase by roughly \$7423 for each additional year of education.

For women working full-time the comparable linear model is

$$P_{\text{women}} = 19,190 + 3000E$$

where E = years of education past grade 8 and P_{women} = mean personal earnings for women. The correlation coefficient is 0.67. As you might predict from the relative status of men and women in the U.S. workforce, the rate of change for women is much lower. For women in this sample, the model predicts that the mean personal earnings increase by only \$3,000 for each additional year of education. In Figure 10, we can see that the regression line for men is steeper than the one for women when plotted on the same grid. In Figure 9, the mean personal earnings for any particular number of years of education are consistently lower for women than for men. The disparity in mean personal earnings between men and women is most dramatic for those with 14 years of education beyond grade 8 or 22 years of education. Although the vertical intercept of the regression line for men's personal earnings is below that for women, after 2 years of education beyond grade 8 the regression line for men lies above that for women.

Going deeper

What other variables have we ignored that we may want to consider in a more refined analysis of the impact of gender on personal earnings? We could, for example, consider type of job or amount of work.

- *Type of job.* Do women and men make the same salaries when they hold the same types of jobs? We could compare only people within the same profession and ask whether the same level of education corresponds to the same level of personal earnings for women as for men. There are many more questions, such as: Are there more men than women in higher-paying professions? Do men and women have the same access to higher-paying jobs, given the same level of education?
- *Amount of work.* In our analysis, we used women and men who were working full-time to explore the impact of gender on earnings. Typically, part-time jobs pay less than full-time jobs, and more women hold part-time jobs than men. In addition, there are usually more women than men who are unemployed. What prediction would you make if all people in FAM1000 were used to study the impact of gender on earnings?

How Good Are the Data?

For this exploration earnings were defined and measured in a number of ways. What issues are raised by the way income in general is defined and measured? What groups of people may be undercounted in the U.S. Current Population Survey? What are some of the current controversies about how the U.S. Bureau of the Census collects the census data? You may want to search online and in the library for articles on the controversies surrounding the census. Who else collects data, and how can you determine if the data are reliable? As access to data is made easier and easier in our Internet society, the ability to assess the reliability and validity of the data becomes more and more important.

How Good Is the Analysis?

What other factors might affect earnings that are not covered by our analysis? What are some limitations to using regression lines to summarize data? Are there hidden variables (such as a history of family wealth or parents' socioeconomic status) that may affect both level of educational attainment and higher earnings? You may want to explore other methods of analysis that address some of these questions and could potentially reveal different patterns.

The following readings at www.wiley.com/college/kimeclark are additional resources for examining the relationship between income and educational attainment in the United States.



- The Bureau of Labor Statistics created a website where you can access all the data on earnings and education from the most recent Current Population Survey at <http://data.bls.gov/PDQ/outside.jsp?survey=le>. This site allows you to search according to categories, access historical data, and create graphs.
- U.S. Census Bureau News, *Census Bureau Data Underscore Value of College Degree*. October 2006.
- Bureau of Labor Statistics, *Education Pays*. U.S. Department of Labor. January 2007.
- M. Maier, "Wealth, Income, and Poverty," from *The Data Game: Controversies in Social Science Statistics* (New York: M. E. Sharpe, 1999). Reprinted with permission.
- *Income in the United States: 2002*. Bureau of the Census, Current Population Reports, P60–221. September 2003.
- *The Big Payoff: Educational Attainment and Synthetic Estimates of Work-Life Earnings*. Bureau of the Census, Current Population Reports, P23–210. July 2002.

Exploring on Your Own

Your journey into exploratory data analysis is just beginning. You now have some tools to examine further the complex relationship between education and income. You may want to explore answers to the questions raised above or to your own questions and conjectures. For example, what other variables besides age do you think affect earnings? What other variables besides gender may affect the relationship between education and earnings? How would our analysis change if we used other income measures, such as personal total income or family income?

Working with Partners

You may want to work with a partner so that you can discuss questions that may be worth pursuing and help each other interpret and analyze the findings. In addition, you can compare two regression lines more easily by using two computers or two graphing calculators.

Generating Conjectures

One way to start is to generate conjectures about the effects of other variables or different income measures on the relationship between education and earnings. (See the Data Dictionary in Table 1 for variables included in the FAM1000 Census data.) You can then generate and compare regression lines using the procedures described next.

Procedures for Finding Regression Lines



1. Finding regression lines:

a. Using a computer:

Open “F3: Regression with Multiple Subsets” in *FAM1000 Census Graphs*. This program allows you to find regression lines for education vs. earnings for different income variables and for different groups of people. Select (by clicking on the appropriate box) one of the four income variables: personal earnings, personal hourly wage, personal total income, or family income. Then select at least two regression lines that it would make sense to compare (e.g., men vs. women, white vs. non-white, two or more regions of the country). You should do some browsing through the various regression line options to pick those that are the most interesting. Print out your regression lines (on overhead transparencies if possible) so that you can present your findings to the class.



b. Using graphing calculators and graph link files:

FAM1000 graph link files A to H contain data for generating regression lines for several income variables as a function of years of education. The Graphing Calculator Manual contains descriptions and hardcopy of the files, as well as instructions for downloading and transferring them to TI-83 and TI-84 calculators. Decide on at least two regression lines that are interesting to compare.

2. For each of the regression lines you choose, work together with your partner to record the following information in your notebooks:

The equation of the regression line

What the variables represent

A reasonable domain

The subset of the data the line represents (men? non-whites?)

The correlation coefficient

Whether or not the line is a good fit and why

The slope

Interpretation of the slope (e.g., for each additional year of education, median personal earnings rise by such and such an amount)

Discussion/Analysis

With your partner, explore ways of comparing the two regression lines. What do the correlation coefficients tell you about the strength of these relationships? How do the two slopes compare? Is one group better off? Is that group better off no matter how many years of education they have? What factors might be hidden or not taken into account?

Were your original conjectures supported by your findings? What additional evidence could be used to support your analysis? Are your findings surprising in any way? If so, why?



You may wish to continue researching questions raised by your analysis by returning to the original FAM1000 data or examining additional sources such as the related readings at www.wiley.com/college/kimeclark or the Current Population Survey website at www.census.gov/cps.

Results

Prepare a 60-second summary of your results. Discuss with your partner how to present your findings. What are the limitations of the data? What are the strengths and weaknesses of your analysis? What factors are hidden or not taken into account? What questions are raised?

EXERCISES

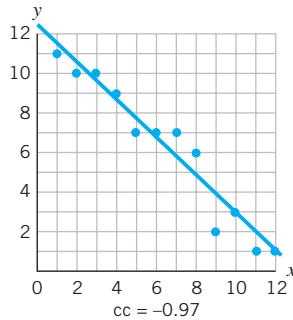
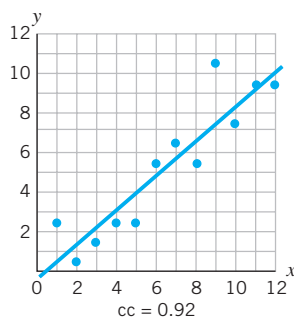
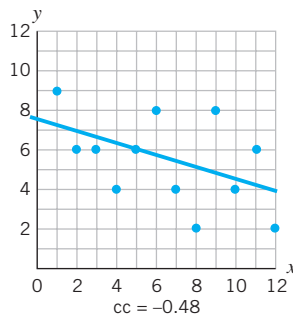
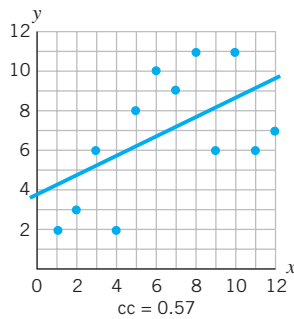
Technology is required for generating scatter plots and regression lines in Exercises 11, 12, and 18.

1. a. Evaluate each of the following:

$$|0.65| \quad |-0.68| \quad |-0.07| \quad |0.70|$$

- b. List the absolute values in part (a) in ascending order from the smallest to the largest.

2. The accompanying figures show regression lines and corresponding correlation coefficients for four different scatter plots. Which of the lines describes the strongest linear relationship between the variables? Which of the lines describes the weakest linear relationship?



Four regression lines with their correlation coefficients.

3. The following equation represents the best-fit regression line for median personal earnings vs. years of education for the 298 people in the FAM1000 data set who live in the southern region of the United States.

$$S = -2105 + 6139E$$

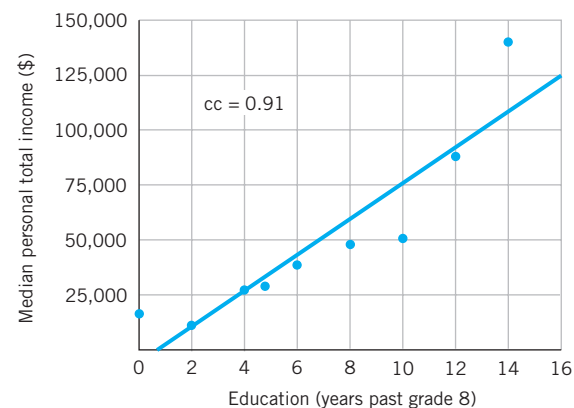
where S = median personal earnings in the South and E = years of education past grade 8, and $cc = 0.72$.

- a. Identify the slope of the regression line, the vertical intercept, and the correlation coefficient.
 b. What does the slope mean in this context?
 c. By what amount does this regression line predict that median personal earnings for those who live in the South change for 1 additional year of education? For 10 additional years of education?

In Exercises 4 to 6, the data analyzed are from the FAM1000 data files and the equations can be generated using *FAM1000 Census Graphs*. Here, the income measure is personal total income, which includes personal earnings (from work) and other sources of unearned income, such as interest and dividends on investments.



4. The accompanying graph and regression line show median personal total income vs. years of education past grade 8.



$$\text{Median personal total income} = -3687 + 7994 \cdot \text{yrs. educ. past grade 8}$$

- a. What is the slope of the regression line?
 b. Interpret the slope in this context.
 c. By what amount does this regression line predict that median personal total income changes for 1 additional year of education? For 10 additional years of education?
 d. What features of the data are not well described by the regression line?

160 AN EXTENDED EXPLORATION: LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

5. The following equation represents a best-fit regression line for median personal total income of white males vs. years of education past grade eight:

$$\text{median personal total income}_{\text{white males}} = -8850 + 10,773 \cdot \text{yrs. educ. past grade 8}$$

The correlation coefficient is 0.87 and the sample size is 452 white males.

- What is the rate of change of median personal total income with respect to years of education?
- Generate three points that lie on this regression line. Use two of these points to calculate the slope of the regression line.
- How does this slope relate to your answer to part (a)?
- Sketch the graph.

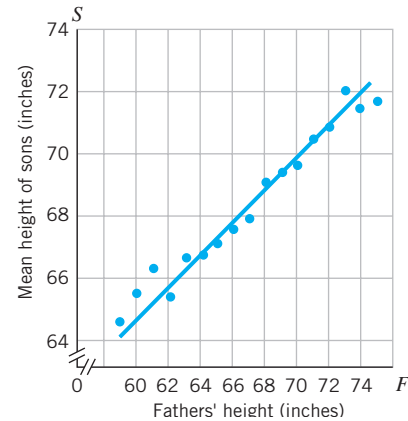
6. From the FAM1000 data, the best-fit regression line for median personal total income of white females vs. years of education past grade 8 is:

$$\text{median personal total income}_{\text{white females}} = 6760 + 4114 \cdot \text{yrs. educ. past grade 8}$$

The correlation coefficient is 0.93 and the sample size is 374 white females.

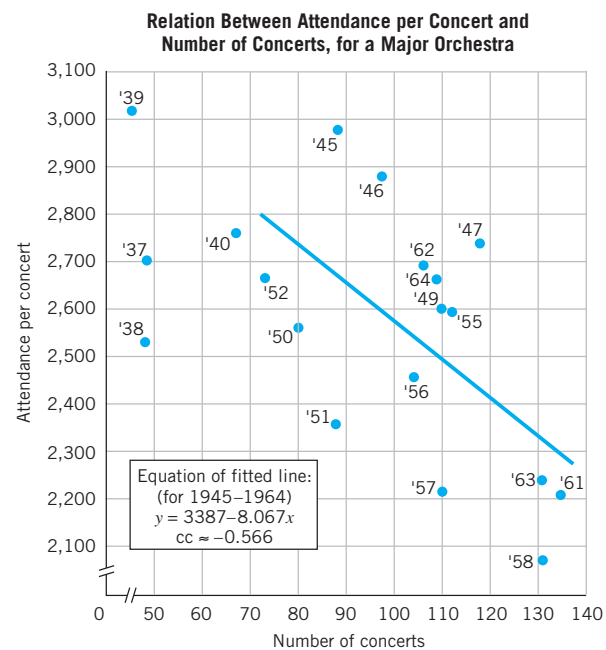
- Interpret the number 4114 in this equation.
- Generate a small table with three points that lie on this regression line. Use two of these points to calculate the slope of the regression line.
- How does this slope relate to your answer to part (a)?
- Sketch the graph.
- Describe some differences between median personal total income vs. education for white females and for white males (see Exercise 5). What are some of the limitations of the model in making this comparison?

7. The term “linear regression” was coined in 1903 by Karl Pearson as part of his efforts to understand the way physical characteristics are passed from generation to generation. He assembled and graphed measurements of the heights of fathers and their fully grown sons from more than a thousand families. The independent variable, F , was the height of the fathers. The dependent variable, S , was the mean height of the sons who all had fathers with the same height. The best-fit line for the data points had a slope of 0.516, which is much less than 1. If, on average, the sons grew to the same height as their fathers, the slope would equal 1. Tall fathers would have tall sons and short fathers would have equally short sons. Instead, the graph shows that whereas the sons of tall fathers are still tall, they are not (on average) as tall as their fathers. Similarly, the sons of short fathers are not as short as their fathers. Pearson termed this *regression*; the heights of sons *regress* back toward the height that is the mean for that population. The equation of this regression line is $S = 33.73 + 0.516F$, where F = height of fathers in inches and S = mean height of sons in inches.



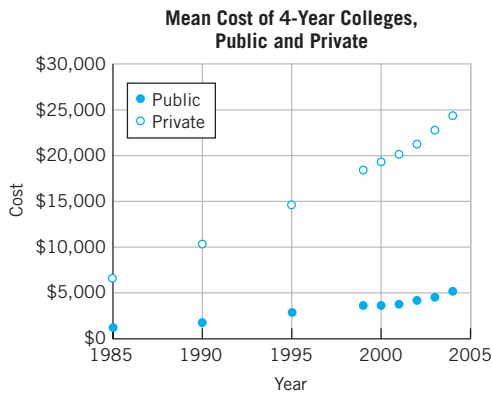
From Snedecor and Cochran, *Statistical Methods*, 8th ed. By permission of the Iowa State University Press. Copyright © 1967.

- Interpret the number 0.516 in this context.
 - Use the regression line to predict the mean height of sons whose fathers are 64 inches tall and of those whose fathers are 73 inches tall.
 - Predict the height of a son who has the same height as his father.
 - If there were over 1000 families, why are there only 17 data points on this graph?
8. The book *Performing Arts—The Economic Dilemma* studied the economics of concerts, operas, and ballets. It included the following scatter plot and corresponding regression line relating attendance per concert to the number of concerts given, for a major orchestra. What do you think were their conclusions?



Source: William Baumol and William G. Bowen, *Performing Arts—The Economic Dilemma*. Reprinted with permission from Twentieth Century Fund.

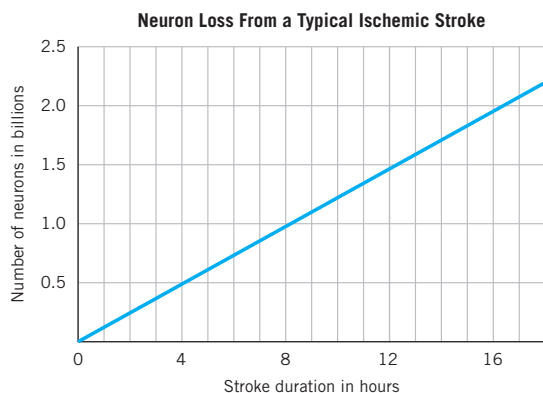
9. (Optional use of technology.) The accompanying graph gives the mean annual cost for tuition and fees at public and private 4-year colleges in the United States since 1985.



Source: U.S. Bureau of the Census, *Statistical Abstract of the United States: 2006*.

It is clear that the cost of higher education is going up, but public education is still less expensive than private. The graph suggests that costs of both public and private education versus time can be roughly represented as straight lines.

- By hand, sketch two lines that best represent the data. Calculate the rate of change of education cost per year for public and for private education by estimating the coordinates of two points that lie on the line, and then estimating the slope.
 - Construct an equation for each of your lines in part (a). (Set 1985 as year 0.) If you are using technology, generate two regression lines from the Excel or graph link data file EDUCOSTS and compare these equations with the ones you constructed.
 - If the costs continue to rise at the same rates for both sorts of schools, what would be the respective costs for public and private education in the year 2010? Does this seem plausible to you? Why or why not?
10. Stroke is the third-leading cause of death in the United States, behind heart disease and cancer. The accompanying graph shows the average neuron loss for a typical ischemic stroke.



Source: American Heart Association.

- Find the slope of the line by estimating the coordinates of two points on the line. Interpret the meaning of the slope in this context.
- Construct an equation for the line, where n = number of neurons in billions and d = duration of stroke in hours.
- The average human forebrain has about 22 billion neurons and the average stroke lasts about 10 hours. Find the percentage of neurons that are lost from a 10-hour stroke.

11. (Requires technology.) The accompanying table shows (for the years 1965 to 2005 and for people 18 and over) the total percentage of cigarette smokers, the percentage of males who are smokers, and the percentage of females who are smokers.



Percentage of Adult Smokers (18 years and older)

Year	Total Population	Males	Females
1965	42.4	51.9	33.9
1974	37.1	43.1	32.1
1979	33.5	37.5	29.9
1983	32.1	35.1	29.5
1985	30.1	32.6	27.9
1987	28.8	31.2	26.5
1988	28.1	30.8	25.7
1990	25.5	28.4	22.8
1991	25.6	28.1	23.5
1992	26.5	28.6	24.6
1993	25.0	27.7	22.5
1994	25.5	28.2	23.1
1995	24.7	27.0	22.6
2000	23.3	25.7	21.0
2003	21.6	24.1	19.2
2005	20.9	23.9	18.1

Source: U.S. Bureau of the Census, *Statistical Abstract of the United States: 2006*.

- Construct a scatter plot of the percentage of all smokers 18 and older vs. time.
 - Calculate the average rate of change from 1965 to 2005.
 - Calculate the average rate of change from 1990 to 2005. Be sure to specify the units in each case.
- On your graph, sketch an approximate regression line. By estimating coordinates of points on your regression line, calculate the average rate of change of the percentage of total smokers with respect to time.
- Using technology, generate a regression line for the percentage of all smokers 18 and older as a function of time. (Set 1965 as year 0.) Record the equation and the correlation coefficient. How good a fit is this regression line to the data? Compare the rate of change for your hand-generated regression line to the rate of change for the technology-generated regression line.
- Using technology, generate and record regression lines (and their associated correlation coefficients) for the percentages of both males and females who are smokers vs. time.
- Write a summary paragraph using the results from your graphs and calculations to describe the trends in smoking from 1965 to 2005.

162 AN EXTENDED EXPLORATION: LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

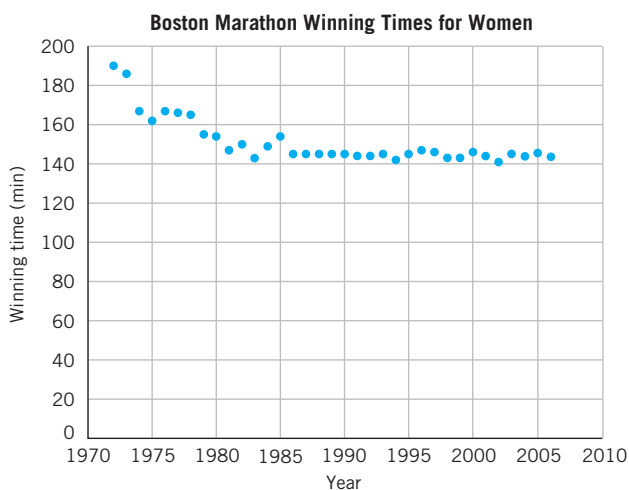
12. (Requires technology.) The accompanying table shows the calories per minute burned by a 154-pound person moving at speeds from 2.5 to 12 miles/hour (mph). (Note: A fast walk is about 5 mph; faster than that is considered jogging or slow running.) Marathons, about 26 miles long, are now run in slightly over 2 hours, so that top distance runners are approaching a speed of 13 mph.



Speed (mph)	Calories per Minute	Speed (mph)	Calories per Minute
2.5	3.0	6.0	12.0
3.0	3.7	7.0	14.0
3.5	4.2	8.0	15.6
4.0	5.5	9.0	17.5
4.5	7.0	10.0	19.6
5.0	8.3	11.0	21.7
5.5	10.1	12.0	24.5

- Plot the data.
- Does it look as if the relationship between speed and calories per minute is linear? If so, generate a linear model. Identify the variables and a reasonable domain for the model, and interpret the slope and vertical intercept. How well does your line fit the data?
- Describe in your own words what the model tells you about the relationship between speed and calories per minute.

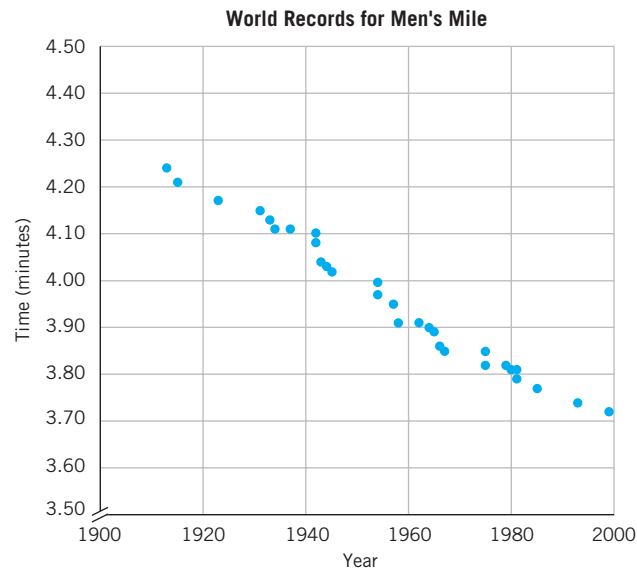
13. (Optional use of technology.) The accompanying graph shows the winning running times in minutes for women in the Boston Marathon.



Source: www.bostonmarathon.org/BostonMarathon/PastChampions.asp.

- Sketch a line that best approximates the data by hand. Set 1972 as year 0 and compute an equation for this line. Interpret the slope of your line in this context. If using technology, generate a regression line from the data in the Excel or graph link file MARATHON and compare its equation with the equation you computed by hand.
- If the marathon times continue to change at the rate given in your linear model, predict the winning running time for the women's marathon in 2010. Does that seem reasonable? If not, why not?
- The graph seems to flatten out after about 1986. Based on this trend, what would you predict for the winning running time for the women's marathon in 2010? Does this prediction seem more realistic than your previous prediction?
- Write a short paragraph summarizing the trends in the Boston Marathon winning times for women.

14. (Optional use of technology.) The accompanying graph shows the world record times for the men's mile. As of January 2006, the 1999 record still stands. Note that several times the standing world record was broken more than once during a year.



Source: www.runnersworld.com.

- Generate a line that approximates the data (by hand or, if using technology, with the data in the Excel or graph link file MENS MILE. (Set 1910 or 1913 as year 0). If you are using technology, specify the correlation coefficient. Interpret the slope of your line in this context.

- b. If the world record times continue to change at the rate specified in your linear model, predict the record time for the men's mile in 2010. Does your prediction seem reasonable? If not, why not?
 - c. In what year would your linear model predict the world record to be 0 minutes? Since this is impossible, what do you think is a reasonable domain for your model? Describe what you think would happen in the years after those included in your domain.
 - d. Write a short paragraph summarizing the trends in the world record times for the men's mile.
15. The temperature at which water boils is affected by the difference in atmospheric pressure at different altitudes above sea level. The classic cookbook *The Joy of Cooking* by Irma S. Rombauer and Marion Rombauer Becker gives the data in the accompanying table (rounded to the nearest degree) on the boiling temperature of water at different altitudes above sea level.

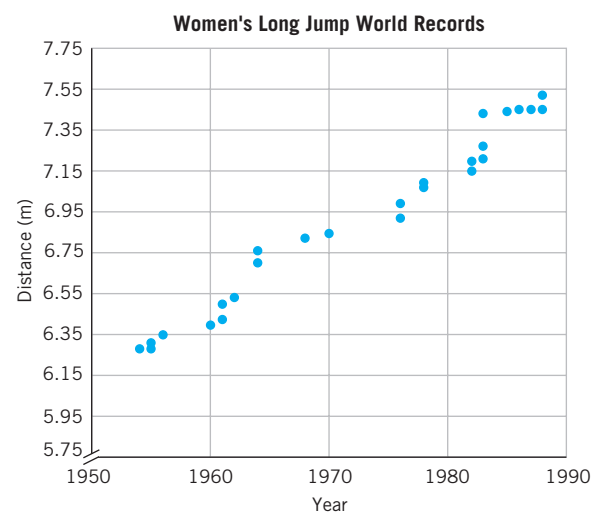
Boiling Temperature of Water

Altitude (ft above sea level)	Temperature Boiling °F
0	212
2,000	208
5,000	203
7,500	198
10,000	194
15,000	185
30,000	158

- a. Use the accompanying table for the following:
 - i. Plot boiling temperature in degrees Fahrenheit, °F, vs. the altitude. Find a formula to describe the boiling temperature of water, in °F, as a function of altitude.
 - ii. According to your formula, what is the temperature at which water will boil where you live? Can you verify this? What other factors could influence the temperature at which water will boil?
- b. The highest point in the United States is Mount McKinley in Alaska, at 20,320 feet above sea level; the lowest point is Death Valley in California, at 285 feet below sea level. You can think of distances below sea level as negative altitudes from sea level. At what temperature in degrees Fahrenheit will water boil in each of these locations according to your formula?

- c. Using your formula, find an altitude at which water can be made to boil at 32°F, the freezing point of water at sea level. At what altitude would your formula predict this would happen? (Note that airplane cabins are pressurized to near sea-level atmospheric pressure conditions in order to avoid unhealthy conditions resulting from high altitude.)

16. (Optional use of technology.) The accompanying graph shows the world distance records for the women's long jump. Several times a new long-jump record was set more than once during a given year.

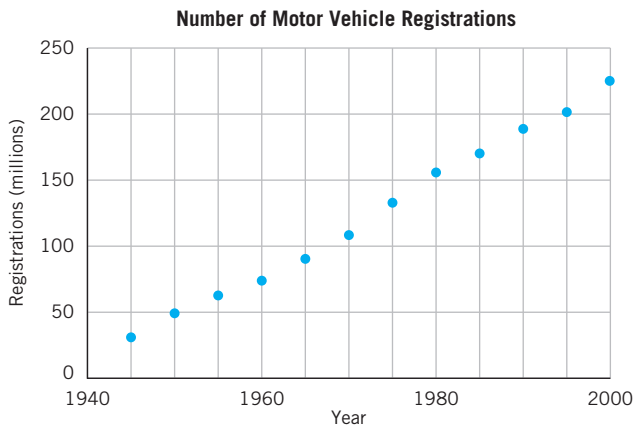


Note: As of January 2006, the 1988 long jump record still stands.
 Source: Data extracted from the website at <http://www.uta.fi/~csmipe/sports/eng/mwr.html>.

- a. Generate a line that approximates the data (by hand or, if using technology, use the data from the Excel or graph link file LONGJUMP. (Set 1950 or 1954 as year 0). Interpret the slope of your line in this context. If you are using technology, specify the correlation coefficient.
- b. If the world record distances continue to change at the rate described in your linear model, predict the world record distance for the women's long jump in the year 2010.
- c. What would your model predict for the record in 1943? How does this compare with the actual 1943 record of 6.25 meters? What do you think would be a reasonable domain for your model? What do you think the data would look like for years outside your specified domain?

164 AN EXTENDED EXPLORATION: LOOKING FOR LINKS BETWEEN EDUCATION AND EARNINGS

17. (Optional use of technology.) The accompanying graph shows the increasing number of motor vehicle registrations (cars and trucks) in the United States.



Source: U.S. Federal Highway Administration, *Highway Statistics*, annual.

- Using 1945 as the base year, find a linear equation that would be a reasonable model for the data. If using technology, use the data in the Excel or graph link file MOTOR.
- Interpret the slope of your line in this context.
- What would your model predict for the number of motor vehicle registrations in 2004? How does this compare with the actual data value of 228.3 million?
- Using your model, how many motor vehicles will be registered in the United States in 2010? Do you think this is a reasonable prediction? Why or why not?

18. (Requires technology.) Examine the following data on U.S. union membership from 1930 to 2004.



U.S. Union Membership, 1930–2004

Year	Labor Force* (thousands)	Union Members† (thousands)	Percentage of Labor Force
1930	29,424	3,401	11.6
1940	32,376	8,717	26.9
1950	45,222	14,267	31.5
1960	54,234	17,049	31.4
1970	70,920	19,381	27.3
1980	90,564	19,843	21.9
1990	103,905	16,740	16.1
2000	120,786	16,258	13.5
2003	122,481	15,800	12.9
2004	123,564	15,472	12.5

*Does not include agricultural employment; from 1985, does not include self-employed or unemployed persons.

†From 1930 to 1980, includes dues-paying members of traditional trade unions, regardless of employment status; from 1985, includes members of employee associations that engage in collective bargaining with employers.

Source: Bureau of Labor Statistics, U.S. Dept. of Labor.

- Graph the percentage of labor force in unions vs. time from 1950 to 2004. Measuring time in years since 1950, find a linear regression formula for these data using technology.
 - When does your formula predict that only 10% of the labor force will be unionized?
 - What data would you want to examine to understand why union membership is declining?
- If a study shows that smoking and lung cancer have a high positive correlation, does this mean that smoking causes lung cancer? Explain your answer.
 - Parental income has been found to have a high positive correlation with their children's academic success. What are two different ways you could interpret this finding?