Figure 13.4. SIMPLE LINEAR REGRESSION: Describing Relationships

Program 8 in: *Against All Odds: Inside Statistics*

In Programs 6 and 7 we looked at the change or growth in a variable over time. This program concerns relationships between two variables, expecially between two quantitative variables. A *quantitative variable* takes numerical values for which arithmetic operations make sense. Other variables are *categorical variables* that record into which of two or more groups an observation falls. In many of the examples we will study, changes in a variable **X** are thought to explain or even cause changes in a second variable **Y**. In such examples, **X** is called an *explanatory variable* and **Y** is called a *response variable*.

A *scatterplot* is a plot of observations on quantitative variables **X** and **Y** as points in the plane. The explanatory variable, if any, is always plotted on the horizontal axis of a scatterplot. The video contains several examples of scatterplots. One shows the relationship between the number **Y** of manatees killed by boats and the number **X** of motorboats registered in Florida for the years 1977-87. Another shows the fluoride content **X** of water supplies and the number **Y** of dental cavities for children in 21 U.S. cities. A third scatterplot pictures the lean body weight **X** and and resting metabolic rate **Y** for 12 women in a study of obesity. In each of these examples, **X** helps explain **Y**. Plotting points with different symbols allows us to see the effect of a categorical variable in a scatterplot. In the video, small and large cities are distinguished by different coloured symbols (green dots, red dots) in the flouride example [See below and overleaf, page 13.28].

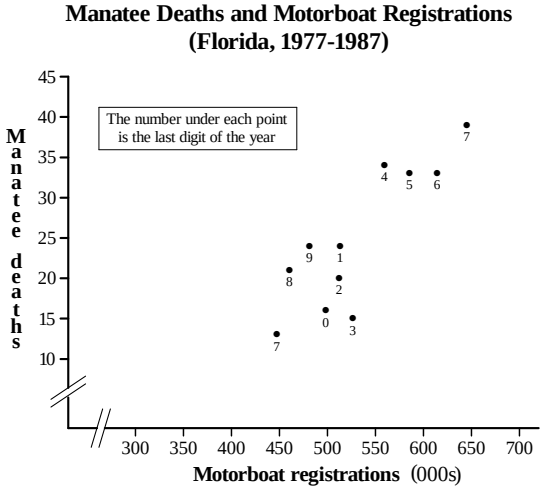
In examining a scatterplot, look for an overall pattern showing the form, direction, and strength of the relationship, and then for outliers or other deviations from this pattern. *Linear* (straight-line) relationships are an important form. If the relationship has a clear direction, we speak of *positive association* when both variables increase together and of *negative association* when an increase in one tends to accompany a *de*crease in the other.

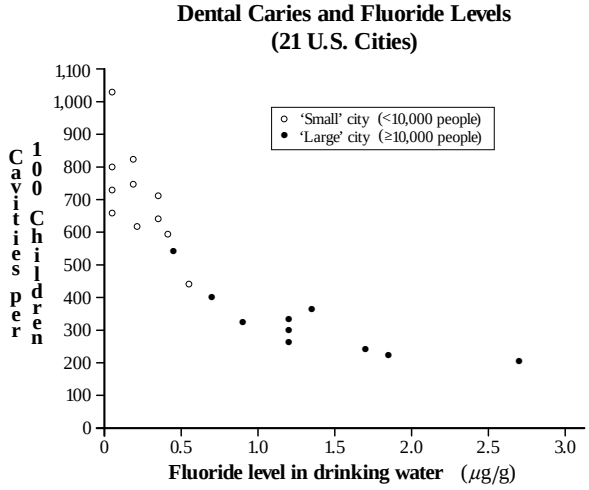
Smoothing a scatterplot by a *median trace* helps reveal the nature of the dependence of **Y** on **X**. The median trace is found by slicing the scatterplot vertically like a loaf of bread, calculating the median of **Y** within each slice, and plotting these medians connected by straight lines. In the video, the negative association between draft number **Y** in the 1970 draft lottery and birth date **X** becomes clear from a median trace (see also pp. 113, 114 of the text third edition).

When a scatterplot suggests that the dependence of **Y** on **X** can be summarized by a straight line, the *least squares regression line* of **Y** on **X** can be calculated. This is the most important calculation in this program. The video calculates the estimated least squares regression line of metabolic rate **Y** on lean body weight **X** for a group of subjects. This fitted line can be used to predict **Y** for a given value of **X**.

The fit of a regression line is examined by plotting the *estimated residuals*, or differences between the observed and predicted values of **Y**. Look for *outliers*, which are points with unusually large estimated residuals. Also look for nonlinear patterns and uneven variation about the line. *Influential observations*, individual points that substantially change the regression line, must also be spotted and examined. Influential observations are often outliers in the **X** variable, but may *not* have large residuals. Evidence of the effects of *lurking variables* on **Y** may be provided by plots of y and of the estimated residuals against the time order of the observations.

Copyright © 1985 by Consortium for Mathematics and Its Applications (COMAP), Inc.
Reprinted with permission of W. H. Freeman and Company.





1995-04-20

(*continued overleaf*)

The following information (indexed by $j = 1, 2, \ldots, 12$) refers to the investigation of obesity described in the video:

No. (j)	Weight (x_j)	Rate (y_j)
1	36.1	995
2	54.6	1,425
3	48.5	1,396
4	42.0	1,418
5	50.6	1,502
6	42.0	1,256
7	40.3	1,189
8	33.1	913
9	42.4	1,124
10	34.5	1,052
11	51.1	1,347
12	*41.2*	*1,204*

For these $n = 12$ observations:

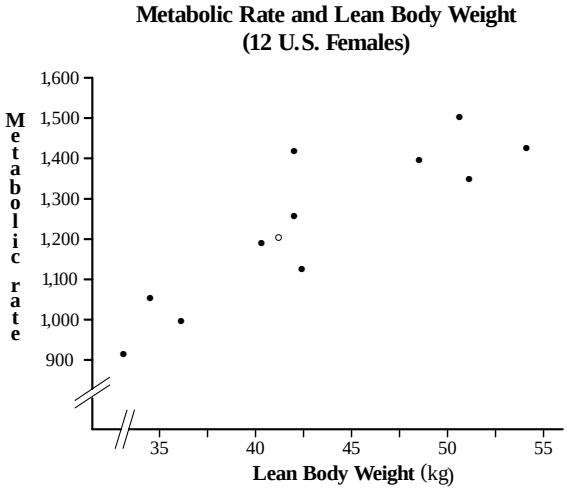
$$\sum_{j=1}^{n} x_j = 516.4,$$

$$\sum_{j=1}^{n} x_j^2 = 22{,}741.34;$$

$$\sum_{j=1}^{n} y_j = 14{,}821,$$

$$\sum_{j=1}^{n} y_j^2 = 8{,}695{,}125;$$

$$\sum_{j=1}^{n} x_j y_j = 650{,}264.8.$$



The expressions for the least squares *estimates* of the slope (b_1) and the intercept (b_0) of the regression of **Y** on **X** are:

$$b_1 = \frac{n\sum_{j=1}^{n} x_j y_j - \left(\sum_{j=1}^{n} x_j\right)\left(\sum_{j=1}^{n} y_j\right)}{n\sum_{i=1}^{n} x_j^2 - \left(\sum_{j=1}^{n} x_j\right)^2} \quad ; \quad b_0 = \bar{y} - b_1\bar{x}$$

[In the video, b_1 and b_0 are denoted $\hat{b}$ and $\hat{a}$, respectively].

1 You are the medical officer of health for a city of 50,000 people, and city council has voted to introduce fluoridation of the city's drinking water. You are familiar with medical evidence that strongly supports the role of fluoride in reducing dental caries, particularly in children, and also with the claims of a link between fluoride in drinking water and increased rates of cancer. Use the right-hand scatter diagram given overleaf on page 13.27 to recommend the *level* of fluoride to be used in your city's drinking water. Briefly justify your choice of level.

2 Using the information provided above, show that the estimated regression of metabolic rate on lean body weight, based on the data for the 12 U.S. females, is: $\hat{y}_{reg} = 201.16 + 24.026x$.

Draw this line on the scatter diagram given above at the right.

- Use this model to *predict* the metabolic rate of a female with a lean body weight of 33 kg.
  - Outline the factors which determine the *accuracy* of this prediction.
- In the video, rounded values of 43 and 1,235 are given for $\bar{x}$ and $\bar{y}$, leading to a value of 201.882 being quoted for b_0. Comment briefly on these values, particularly that of b_0, in light of your own calculation of the estimated least squares regression line.