C. Tolmie

# REFIGURING ANTHROPOLOGY

## First Principles Of Probability & Statistics

David Hurst Thomas
American Museum of Natural History

**Waveland Press, Inc.**
Prospect Heights, Illinois

For information about this book, write or call:

Waveland Press, Inc.
P.O. Box 400
Prospect Heights, Illinois 60070
(312) 634-0081

For permission to use copyrighted material, the author is indebted to the following:

FIG. 2.1. By permission of the Trustees of the British Museum (Natural History).
TABLE 2.3. (p. 24) *From Physical Anthropology: An Introduction* by A. J. Kelso. Reprinted by permission of the publisher, J. B. Lippincott Company. Copyright © 1974. (p. 25) Reproduced by permission of the Society for American Archaeology from *Memoirs of the Society for American Archaeology*, Vol. 11, 1956.
FIG. 3.1. From Hulse, Frederick S. *The Human Species: An Introduction to Physical Anthropology*. Copyright © 1963 by Random House, Inc.

FIG. 3.2. From Dozier, Edward P., *The Pueblo Indians of North America*. Copyright © 1970 by Holt, Rinehart and Winston, Inc. Reproduced by permission of Holt, Rinehart and Winston. [This book reissued 1983 by Waveland Press, Inc.]
FIG. 3.4. Reproduced by permission of the Society for American Archaeology from *American Antiquity*, Vol. 35 (4), 1970.
FIG. 3.5. Reproduced by permission of the American Anthropological Association from the *American Anthropologist*, Vol. 73 (3), 1971.
FIG. 13.14. From *Biometry* by Robert R. Sokal and F. James Rohlf. W. H. Freeman and Company. Copyright © 1969.


Second Printing
The 1976 version of this book was entitled *Figuring Anthropology*.

ISBN 0-88133-223-2



Printed in the United States of America.

# 13 Linear Regression

● *Nothing counts, we might say, unless it can be counted.*—G. Rees

## 13.1 THE LINEAR RELATIONSHIP

Throughout the past few chapters, we have assigned rather specific definitions to commonplace terms such as *variable*, *constant*, *random*, *population*, and *significance*. Time has come to consider another Big Word, and that word is *function*. When the value of a random variable $Y$ changes in response to a corresponding change in random variable $X$, then $Y$ is said to be a function of $X$. The nature of this dependency is presently irrelevant. Regardless of whether the dependency is specific or generalized, causal or coincidental, all such relationships can be symbolized as

$$Y = f(X)$$

to be read as "$Y$ is a function of $X$."

Some elementary relationships between random variables were encountered in the discussion of the chi-square statistic, but we must now consider the generalized bivariate relationship in more detail. Specifically, there are two common methods for expressing the relationship between two variables—mathematical equations and graphs.

The most elementary function between two random variables is simply

$$Y = f(X) = X$$

This function tells us that the value of $Y$ must always exactly equal the value of $X$. This simple function is that assumed in tree-ring dating (dendrochronology) for example. A perfect one-to-one relationship exists between the age of a living

tree and the number of *annual rings*: One ring represents one year.

$$\text{number of annual rings} = \text{age in years}$$
$$Y = X$$

Because $Y = f(X) = X$, the number of annual rings ($Y$) is a *function* of $X$, the age of the tree. This function predicts that a ten-year-old tree should have exactly ten growth rings. A 1000-year-old tree must have 1000 rings.

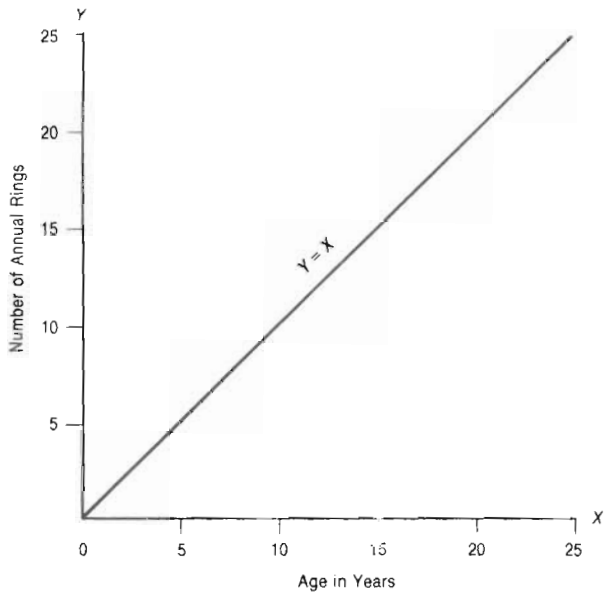


Fig. 13.1 Relationship between age of tree and number of annual growth rings.

Functional relationships can also be graphed. Figure 13.1 shows the graph for the function $Y = X$. The vertical axis (the *ordinate*) generally is taken to denote $Y$ and the horizontal $X$-axis (the *abscissa*) plots the values of the $X$ variable. The $Y$-axis in this case depicts the number of growth rings per tree and the abscissa scales the tree's age in years. The ordinate meets the abscissa at the *origin* of the graph, so the origin of Fig. 13.1 represents zero on both the $X$ and $Y$ scales. Zero age predicts zero annual rings.

The appropriate curve for Fig. 13.1 was found by plotting the various values satisfying the equation $Y = X$.

| When $X$ is: | 0 | 1 | 2 | 4 | 7 | 28 | 99 | ··· |
|---|---|---|---|---|---|---|---|---|
| Then $Y$ is: | 0 | 1 | 2 | 4 | 7 | 28 | 99 | ··· |

This curve is "linear" because all of these potential values can be described by a single straight line. This "curve" commences at the *origin* because both scales truncate at zero; a negative age or a minus count of rings is patently impossible.

Convention dictates that the $X$ variable be called the *predictor* (or *indepen-*

*dent*) variable and that $Y$ be the *predicted* or *dependent* variable. Both terms follow from the general function $Y = f(X)$. Because the $X$ variable can often be controlled in experimental situations, a change in $X$ is said to induce a shift in $Y$. Sometimes these terms reflect a causal sequence in which $X$ is said to cause $Y$, but care must be taken to avoid confusing a *causality* with simple *prediction*. Fire engines are excellent indicators of fires, ambulances associate with automobile accidents, and police officers invariably occur at the scene of a crime.

Age is the independent variable in dendrochronology because we happen to know from plant physiology that age *causes* trees to produce annual rings. Age ($X$) accurately predicts the number of rings ($Y$); in this case, a causal relationship exists. But prediction equations may often be written in the reverse form. The number of rings predicts age. A living tree with ten rings must be exactly ten years of age. This reasoning sets the foundation for the science of dendrochronology.

Tree-ring samples can be counted in living trees by careful use of an increment borer. The tree is not harmed. Edmond Schulman, a dendrochronologist from the University of Arizona, took literally hundreds of borings from a bristlecone forest located at an elevation over 10,000 feet in eastern California. Using the simple function $Y = X$, Schulman discovered the oldest living thing in the world. One bristlecone—lovingly christened *Pine Alpha*—dated back to 2194 B.C. And Pine Alpha still lives! In this case, $Y$ is the tree's age, predicted by $X$, the number of annual rings. Obviously, the decision of which variable is $X$ (the *predictor*) depends strictly upon what one wishes to predict, age or number of rings.

At the risk of repetition, let me underscore once again a canon of statistical inference: Association must never be confused with causality. The predictive relationship implicit in $Y = f(X)$ may represent a true causal linkage, or it may not. The issue is determined by substantive rather than statistical considerations. The common statistical labels *independent* and *dependent* must not be allowed to cloud the causal issue because these terms are often assigned merely for convenience. The choice of independent variables lies with the specific empirical intent or the perspective of the investigator. For this reason, the $X$ variable will be termed the *predictor* variable, to avoid any confusion of true dependence or independence.

Let us now move to a bit more complicated function:

$$Y = f(X) = \beta X$$

where $\beta$ represents any constant.[1] The expression $Y = \beta X$ is another specific example of the general function $Y = f(X)$. In the previous example of a function, $Y = X$, the multiplicative constant ($\beta$) was equal to unity, $\beta = 1$. The function $Y = X$ tells us that an increase in a single unit of $X$ corresponds to an increase in precisely one unit of $Y$. The more general case of $Y = \beta X$ implies a change in one unit of $X$ for every $\beta$ units of change in $Y$. If $Y = 10X$, then one unit change in $X$ produces +10 units of change in $Y$. The constant $\beta$ can likewise be negative, in which case $Y$ decreases with an increase in $X$.

[1] Be careful here not to confuse the regression $\beta$ with the probability of committing a Type II error.

An elementary example of the $Y = \beta X$ function is the rate of exchange between international monetary systems. Although these rates tend to fluctuate daily, the relationship between any two currencies is fixed at any given point in time. On March 4, 1974, the *New York Times* reported the commercial selling rate between currencies of the United States and Spain to be 1.72. Translated into functional notation, this relationship becomes $Y = 1.72X$, where $X$ is the value of the Spanish *peseta* and $Y$ is the value of the U.S. penny (0.01 U.S. dollars). In other words, one U.S. penny is equal to 1.72 pesetas. This relationship is diagrammed in Fig. 13.2. The functional line once again commences at the origin (zero U.S. pennies = zero *pesetas*) and extends indefinitely upward.

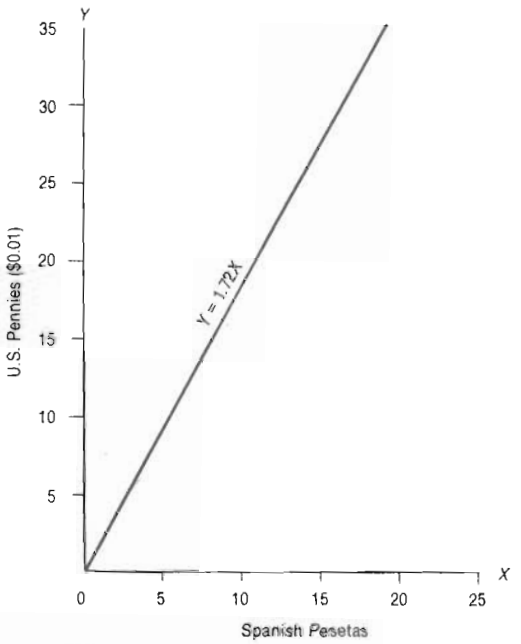


Fig. 13.2 Relationship between U.S. and Spanish currency.

All monetary rates can be expressed in this simple form. Only the absolute value of $\beta$ changes to fit the particular circumstance. Note also how $X$ is arbitrarily assigned to the Spanish currency. There is no causal linkage in any of the currency exchanges; $X$ is simply a convenient point of reference.

Using the multiplicative constant $\beta$ requires us to introduce another new term, the *slope*. The multiplicative constant $\beta$ represents the slope of a line. The slope of the line in Fig. 13.1, for instance, is $\beta = 1$. One unit of change in $X$ creates one unit shift in $Y$. Similarly, in Fig. 13.2, a unit shift in the value of $X$ creates a 1.72 unit shift in $Y$. The magnitude of $Y$ changes more relative to $X$ in the second case because the line is *steeper*. That is, the slope of the line in Fig 13.2 is greater than that of Fig. 13.1. Consider the two equations

$$Y = 1X \qquad Y = 1.72X$$

The only difference between these two functions is the value of the multiplicative constant. In the tree-ring example, $\beta$ must equal unity, while the exchange rate had been fixed at $\beta = 1.72$. So, clearly, *the slope of any line depends only upon the value of* $\beta$. The lines will become steeper as $\beta$ increases. Furthermore, a positive value of $\beta$ means that the function line slopes in a positive direction ($Y$ increasing with $X$); a negative $\beta$ denotes a negative slope ($Y$ decreasing as $X$ increases). Mathematically speaking, the slope of a line is given by the tangent of the angle formed by the function line and the $X$-axis, but this derivation is not important to our purposes. The meaning of $\beta$ is also apparent simply from plotting values of $Y$ for given $X$.

The lines of Figs. 13.1 and 13.2 pass through the origins of their respective graphs. This is reasonable: Trees of zero age have no annual rings, and all monetary rates of exchange must commence with zero money. But one final statistical phrase is required to completely generalize this discussion of the linear relationship. Several lines are graphed in Fig. 13.3, each of which shares a slope of $\beta = 1$. These lines are all parallel, and differ only in their position relative to the axes. Only line $A$ passes through the origin. The equations for all

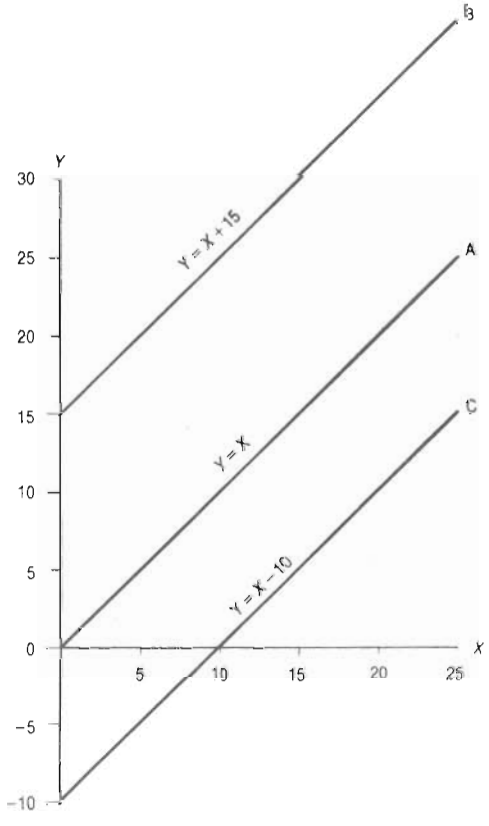


Fig. 13.3 Equations with identical slope but different $Y$-intercepts.

other lines involve a new constant. For line *B*, this constant is $\alpha = 15$ and for line *C*, the constant takes the value of $\alpha = -10$. This new term is called the *additive* constant and is symbolized by $\alpha$.[2] The general formula for all linear relationships can now be given:

$$Y = f(X) = \alpha + \beta X$$

The term $\alpha$ is also known as the *Y-intercept*, since its value is the precise point at which the function line intercepts the *Y*-axis. When $X = 0$, then $Y = \alpha$. Only when $\alpha = 0$ will the line intersect the origin; note that $\alpha = 0$ in Figs. 13.1 and 13.2.

The additive constant can be illustrated by *Bergmann's rule*, that biological principle which relates an animal's size to the temperature of the habitat (see Birdsell 1972:465–467). Bergmann's rule predicts that polar animals should have a greater body size than animals living near the equator. In general, a larger body mass will tend to retard heat loss in colder climates, so less energy is expended by larger bodies. This relationship is expressed in Fig. 13.4. For

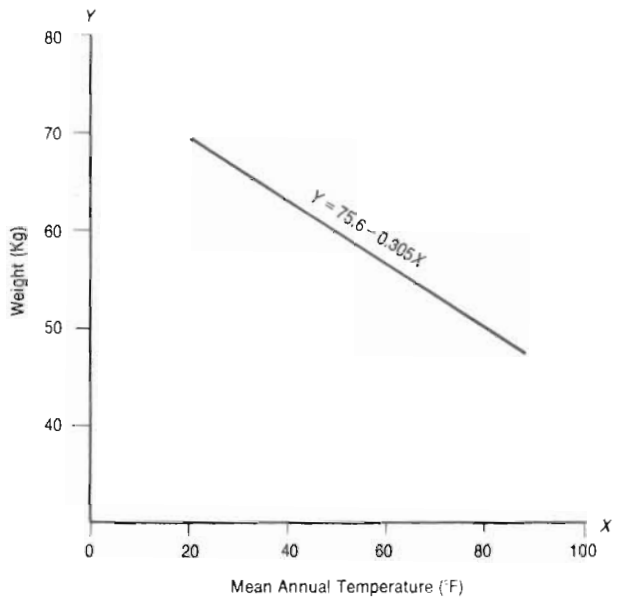


Fig. 13.4 Graph illustrating Bergmann's rule (after Roberts 1953).

those readers who seem comforted by the belief that man has little in common with other animals, Fig. 13.4 might come as something of a shock. This graph has been derived from *human* populations throughout the world, and size clearly decreases with an increase in temperature. Bergmann's rule predicts human size as well as that of the lower beasts. This expression can be

[2]Once again, do not confuse the regression $\alpha$ with the probability of committing a Type I error

summarized as

$$Y = 75.6 - 0.305X$$

where $X$ is measured in degrees Fahrenheit and human weight ($Y$) is given in kilograms. This function tells us a great deal about the interrelationship between temperature and size. For every additional degree of temperature, the average weight of a human population can be expected to decrease about 0.305 kg. Bergmann's rule indicates that environmental factors tend to operate on mankind regardless of culture. Even in a mild climate such as southern California, between 80 and 90 percent of the food consumed is required to maintain a body temperature of 98.6F; quite obviously, the situation is much worse in an Arctic environment. Despite the fact that igloos are heated to a balmy 75 degrees, the true limiting factor upon human populations appears to be the −60°F temperatures encountered outside (Birdsell 1972:467).

To summarize, any *linear* relationship can be described by the simple equation

$$Y = \alpha + \beta X \tag{13.1}$$

where $X$ is the *predictor* variable, $Y$ is the *predicted* variable, $\alpha$ is the $Y$-intercept, and $\beta$ is the slope of the line.

## 13.2 LEAST SQUARES REGRESSION (MODEL I REGRESSION)

With the formal properties of the linear relationship at hand, we arrive at the major topic of this chapter—the concept of *regression.* The actual word "regression" sometimes causes a bit of confusion, but this distraction is unnecessary once one realizes the genesis of the concept. The pioneering effort on the study of linear relationships was made by Sir Francis Galton, a nineteenth-century scholar of rather amazing breadth. Galton was an accomplished statistician, whose early studies of heredity were among the vanguard of pre-Mendelian genetics. He was also a prominent anthropologist (in the original sense of the term), contributing to such diverse fields as dermatoglyphics (fingerprinting), anthropometry, evolution, and eugenics. (Sir Francis is, incidentally, the same Galton whose infamous "problem" has bemused anthropologists for the past 80 years, as discussed in Chapter 15.)

Of immediate interest is Galton's paper entitled "REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE," published by *The Journal of the Anthropological Institute of Great Britain and Ireland* in 1885. In this classic paper, Galton advanced his "Law of Regression." He hypothesized as a result of genetic experiments upon peas that "offspring did not tend to resemble their parent seeds in size—but to be always more mediocre than they—to be smaller than the parents if the parents were large; to be larger than the parents, if the parents were very small" (1885:246). Galton felt that offspring tend to "regress" toward the population average; hence his title "Regression towards Mediocrity. . . ." To generalize this relationship, Sir Francis compiled hundreds of measurements of human stature and plotted these points on the familiar $X$-$Y$ coordinate axis, similar to those already considered. Parent's stature was plotted

against the stature of the offspring, and a marked linear relationship emerged. Galton called the line describing this positive relationship a *line of regression* because it demonstrated how offspring "regressed" toward mediocrity. Statisticians subsequently modified Galton's idea to apply to all lines predicting values of one random variable ($Y$) from knowledge of the other random variable ($X$). Therefore, the line in Fig. 13.4 is a *regression line* because it predicts human size ($Y$), given mean annual temperature ($X$). Similarly, the lines in Section 13.1 predicted a tree's age (given the number of annual tree rings) and the U.S. dollar equivalent to any particular sum of Spanish *pesetas*. Later in this chapter we will even be able to rather accurately predict temperature by counting the number of chirps from crickets. But it is first necessary to examine just how regression lines are computed.

No explanation was offered in Section 13.1 as to how the regression lines were derived. The lines were simply offered as inalterable truth. Closer examination shows that actually two rather different kinds of regressions were considered. The tree-ring example is an *exact* fit, plotted without error. A ten-year-old tree must have exactly ten growth rings, not nine or eleven or any other number. The relationship between the U.S. dollar and the Spanish *peseta* is likewise exact, without any inherent error.

But the equations for Bergmann's rule represent a rather different sort of regression. The line in Fig. 13.4 is not an exact relationship at all, but rather an *estimate* roughly describing some data points. This particular graph was derived by D. F. Roberts of the Anthropology Laboratory at Oxford University. Roberts first surveyed the anthropometric literature and then selected a series of 116 societies from around the world (Roberts 1953). The relationship between these body weights and the mean annual temperature was plotted point by point on the coordinate system shown in Fig. 13.5. Each symbol in Fig. 13.5, represents one society. This method of graphical representation in which $N$ pairs of values for $X$ and $Y$ are arranged into a coordinate system is called a *scatter diagram*, or simply a *scattergram*, so these points represent the actual data relevant to Bergmann's rule. The problem now becomes how to describe these 116 independent points by one simple line.

Unfortunately, any number of lines could be drawn through the points swarming about Fig. 13.5. If ten people were asked to "eyeball" a line to describe the points, ten different regression lines would undoubtedly result. But ten lines describing one phenomenon are not succinct summaries of data, and the problem becomes: How to choose?

A regression line is a linear function, and the direction of that line is completely determined by the values of $\alpha$ and $\beta$ in the equation $Y = \alpha + \beta X$. *Thus, the problem of fitting a line to scattergram points reduces simply to determining values for $\alpha$ and $\beta$ such that the N points lie as closely as possible to the regression line.* Remember that regression lines predict the values of $Y$, given values of $X$. Then this equation must be derived from a scattergram, and the resulting predictions are subject to error precisely because the points tend to scatter. One way to minimize this error would be to draw a line such that exactly half the points would fall above the line and half below. The errors could then be said to "cancel out." But this definition is still unsuitable because many such lines exist and would bisect a swarm of points. A truly satisfactory line of regression must be unique.

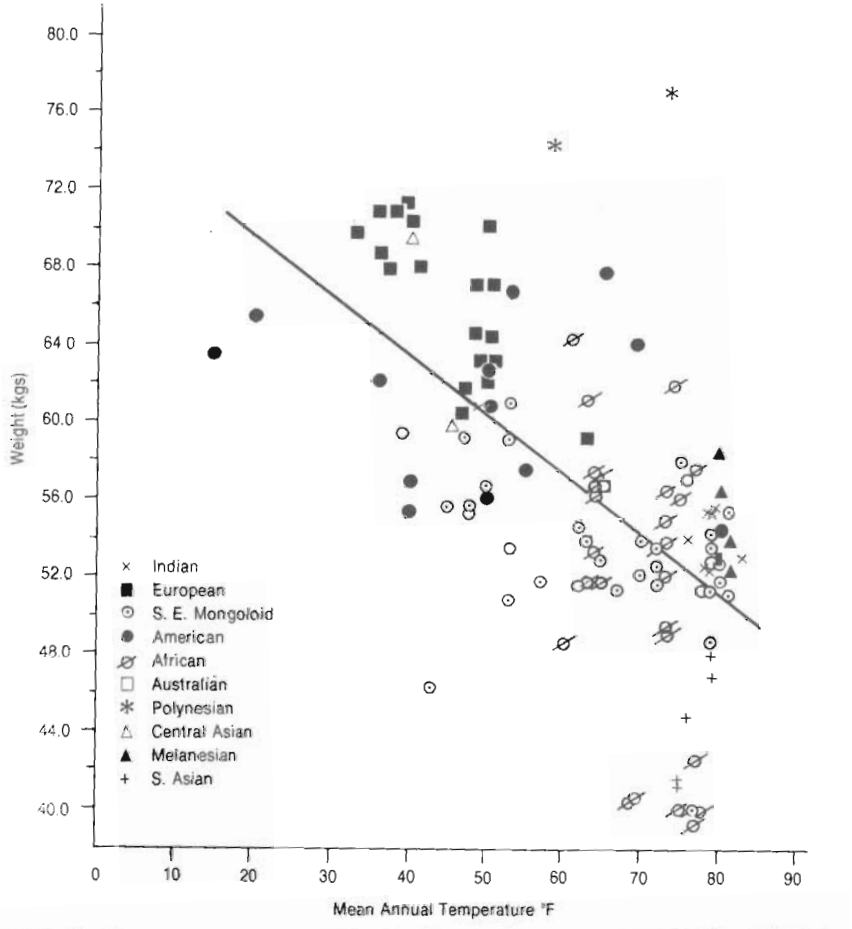


Fig. 13.5 Scattergram used to generate equation for Bergmann's rule in Fig. 13.4 (after Roberts 1953: fig. 7).

In practice, the most suitable fit for a regression line is given by the *criterion of least squares*. Simply defined, the least squares fit places a line such that the *sum of squares of the vertical deviations from this line is minimized*. Consider the graph in Fig. 13.6. Each point has two coordinates, the specific value of random variable $X$ (denoted by $X_i$) and the specific value of the random variable $Y$ (called $Y_i$). The duty of the regression line is to estimate $Y_i$, given $X_i$. There is no error associated with $X_i$ because this figure is arbitrarily selected. Given $X_i$, find $Y_i$. Hence, the total error of estimate in least squares linear regression relates only to the random variable $Y$. This is an important point.

A glance at Fig. 13.6 reveals that there must actually be *two values* of the random variable $Y$ associated with a given $X_i$. First there are the actual observed values of $Y$. These are the $Y_i$ which comprise the empirical data, such as the 116 societies plotted on Fig. 13.5. But there is also a second meaning of $Y$ implied in all scattergrams, and that is the value of $Y$ *estimated* by the least squares

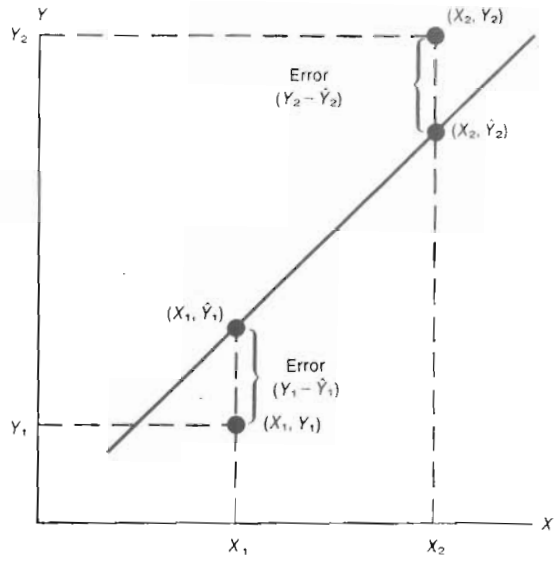


Fig. 13.6

line. Let us call these estimated values the $\hat{Y}_i$. The estimated $\hat{Y}_i$ are computed from the least squares regression equation $Y = \alpha + \beta$ (we will see in a moment how to find these values for $\alpha$ and $\beta$). Because the $\hat{Y}_i$ are given by the regression equation, it follows that all $\hat{Y}_i$ must lie directly upon the least squares line. To reiterate: The *observed* data are represented by the $Y_i$ of a scattergram. The *expected* values of $Y$, all of which lie directly on the regression line, are denoted by $\hat{Y}_i$.

The two distinct sets of coordinates are plotted for each datum point on Fig. 13.6. The first set represents the actual *observed* values, $(X_i, Y_i)$. Also present is the estimated value of each point, predicted by the regression line. The coordinates of the estimated position are $(X_i, \hat{Y}_i)$. Only the $Y_i$ values have been estimated (by $\hat{Y}_i$); the $X_i$ are known and hence error-free. The accuracy of estimation for the least squares line of regression can be judged by the distance between the observed and expected positions, given by $|Y_i - \hat{Y}_i|$. If the regression line (a prediction) passed directly through every observed point, then no error is involved because $\Sigma|Y_i - \hat{Y}_i| = 0$.

But few estimates are that accurate. Most data will have points lying some distance from the regression line. The error for the first datum point on Fig. 13.6 is given by $|Y_1 - \hat{Y}_1|$. This is a measure of how far the actual point lies away from its expected location in the least squares line. The error for the second point is $|Y_2 - \hat{Y}_2|$ and that for the $N$th point is $|Y_N - \hat{Y}_N|$. The least squares method places a line of regression such that the *sum of squares* of the differences between observed and expected values of $Y$ is kept at the smallest possible level. This is the "least squares" criterion:

$$\Sigma(Y_i - \hat{Y}_i)^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \cdots + (Y_N - \hat{Y}_N)^2$$
$$= \textit{minimum}$$

Because all the $X_i$ are "fixed" by arbitrary decision, the total error of estimation is restricted to the vertical ($Y$) dimension. The least squares fit minimizes these vertical distances between the swarm of points and the regression line describing them.

The issue now becomes relatively straightforward: Given a swarm of points (that is, $N$ observed pairs of $Y_i$ and $X_i$), find the constants $\alpha$ and $\beta$ such that $\Sigma(Y_i - \hat{Y}_i)^2$ is a minimum. This problem is solved only by methods beyond the present scope—the formula is derived in Hays (1973:622–623) and Cramér (1946:271–272)—and $\beta$ is given by the following formula:

$$\beta = \frac{\Sigma(X - \mu_X)(Y - \mu_Y)}{\Sigma(X - \mu_X)^2}$$

$$= \frac{\Sigma XY - N\mu_X\mu_Y}{\Sigma X^2 - N\mu_X^2} \tag{13.2}$$

As before, the constant $\beta$ represents the slope of the regression line. Equation (13.2) might look somewhat forbidding at first, but closer inspection reveals that the computations are really quite straightforward. Only a few readily determined values are necessary: $N$ (the number of pairs), $\mu_X$, and $\mu_Y$ (the means of both variables), $\Sigma X^2$ and $\Sigma XY$ (the sum of the cross products). (This expression is actually a *computing* formula, similar to those introduced earlier for finding the variance.)

Once $\beta$ has been computed, it remains only to find $\alpha$. Although not discussed here, the derivation of least squares regression stipulates that the line must always pass through the means of both dimensions, $\mu_X$ and $\mu_Y$. (These means are, of course, computed from the observed datum points rather than the $\hat{Y}_i$, the estimated values of $Y$.) By substituting the mean values into Formula (13.1) and solving this formula for the constant $\alpha$,

$$\alpha = \mu_Y - \beta\mu_X \tag{13.3}$$

Constants $\alpha$ and $\beta$ now define the least squares estimate of the regression equation. The resulting line of regression is the "best fit" because the squared deviations for $Y_i$ from $\hat{Y}_i$ have been minimized. Whenever the least squares method is used, it is customary to denote the regression equation as

$$\hat{Y} = \alpha + \beta X \tag{13.4}$$

The circumflex indicates that values of $\hat{Y}_i$ are *estimated*, but not known. These new methods are illustrated by a simplified example.

We know that body weight increases with height, but what is the exact nature of this relationship? The students in a small physical anthropology seminar were grouped by height into 2 inch intervals. One student was then randomly selected from each group and measured. In this manner, a simple sample was obtained for weight within each arbitrary height increment.

These measurements are plotted on Fig. 13.7. The $X$ variates are graphed on the abscissa as usual. $X$ is "fixed" in this case because each 2 inch height class has been purposely selected rather than randomly sampled. There is no sampling error on $X$ because students were simply assigned to the correct group. The weights become the $Y$ variates in this study, and are plotted on the vertical axis. Each $Y_i$ is a random sample of weight from within a particular

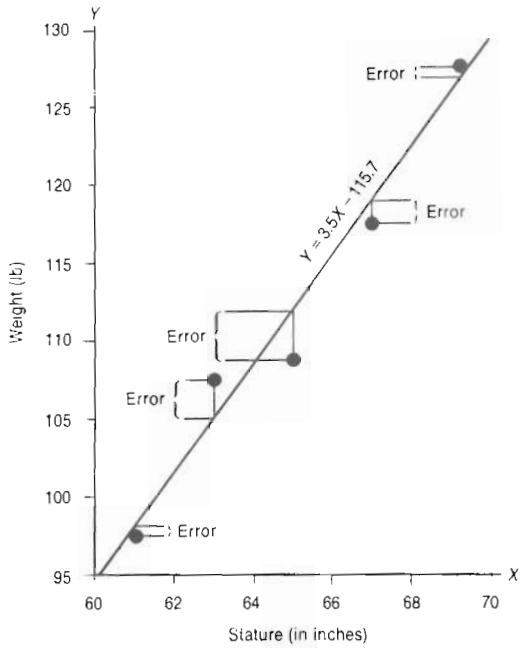


Fig. 13.7

height class. The problem now is to fit a line describing the relationship between these five datum points. The equation of this line can then be used as a loose analogy to estimate the unknown weights for fossil material.

Five quantities are necessary for the least squares method of regression: $\Sigma X$, $\Sigma X^2$, $\Sigma Y$, $\Sigma XY$, and $N$. These values are found in Table 13.1, so the constant $\beta$ is found from Expression (13.2):

$$\beta = \frac{36{,}475 - 5(65)(111.8)}{21{,}165 - 5(65^2)} = +3.5$$

$\alpha$ is found from Formula (13.3):

$$\alpha = 111.8 - 3.5(65) = -115.7$$

**TABLE 13.1**

| Height Class, inches | $X$, Class Midpoint inches | $Y$, Weight, lb | $XY$ | $X_2$ |
|---|---|---|---|---|
| 60–62 | 61 | 98 | 5,978 | 3,721 |
| 62–64 | 63 | 107 | 6,741 | 3,969 |
| 64–66 | 65 | 109 | 7,085 | 4,225 |
| 66–68 | 67 | 117 | 7,839 | 4,489 |
| 68–70 | 69 | 128 | 8,832 | 4,761 |
| | $\Sigma X = 325$ | $\Sigma Y = 559$ | 36,475 | 21,165 |

$\mu_X = 325/5 = 65.0$ in.; $\mu_Y = 559/5 = 111.8$ lb.; $N = 5$.

The final regression equation describing these data is given by Expression (13.4):

$$\begin{aligned}\hat{Y} &= \alpha + \beta X \\ &= -115.7 + 3.5X \\ &= 3.5X - 115.7\end{aligned}$$

This line can now be fitted to the datum points by substituting some arbitrarily selected sample values. Take the hypothetical value of $X_i = 60$. By substituting into the regression Equation (13.4), we find that when $X_i$ is 60 in., then $\hat{Y}_i = 94.3$ lb. This point must lie on the regression line. Similarly, when $X_i = 70$ in., the least squares equation tells us that $\hat{Y}_i = 129.3$ lb. We now have two points which must lie on the line of regression. Because two points always define a line, the new regression line describing the population of five datum points can be drawn on Fig. 13.7.

To summarize: The least squares criterion is a method of computing values of $\alpha$ and $\beta$. This is simple statistical description. The least squares regression equation, $\hat{Y} = \alpha + \beta X$, is thus equivalent to other descriptive measures such as the mean, the median, or the variance. As do all descriptive statistics, the data being described may represent either a sample or a population. If the data constitute a statistical population, then the descriptive summary is called a *parameter*. If the data are sampled from a sample, then the descriptive measure is a *statistic*. No statistical inference has taken place so far.

**Example 13.1**

The table below presents some blood pressure data from a sample of American Indians of the Trio and Wajana tribes of Surinam. These figures were collected by Glanville and Geerdink in 1967 and 1968 on the Upper Courantyne, Lawa, and Tapanahony rivers where missions had recently been established (Glanville and Geerdink 1972).

Find the regression equation which best describes this statistical population.

| Age Group (midpoint), $X$ | Diastolic Blood Pressure, $Y$ | $X^2$ | $XY$ | $Y^2$ |
|---|---|---|---|---|
| 5 | 60 | 25 | 300 | 3,600 |
| 7 | 63 | 49 | 441 | 3,969 |
| 9 | 69 | 81 | 621 | 4,761 |
| 11 | 74 | 121 | 814 | 5,476 |
| 13 | 75 | 169 | 975 | 5,625 |
| 15 | 71 | 225 | 1,065 | 5,041 |
| 17 | 77 | 289 | 1,309 | 5,929 |
| 19 | 85 | 361 | 1,615 | 7,225 |
| 21 | 78 | 441 | 1,638 | 6,084 |
| 117 | 652 | 1,761 | 8,778 | 47,710 |

$\mu_X = 13.0$ years; $\mu_Y = 72.4$ mm; $N = 9$.

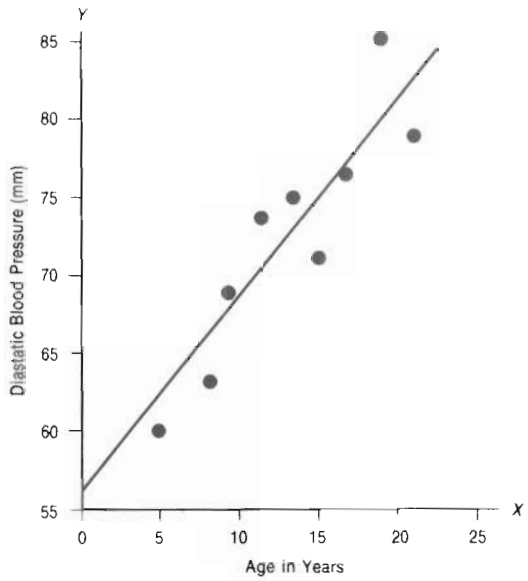


Fig. 13.8

The first step in all problems of linear regression is to plot the scattergram (Fig. 13.8). This helps to determine whether a linear solution is applicable. The data in this case appear to fall in roughly linear fashion, so computation of the regression equation can be attempted.

The multiplicative constant is found to be

$$\beta = \frac{8778 - 9(13.0)(72.4)}{1761 - 9(13.0)^2} = \frac{307.2}{240}$$

$$= 1.28$$

By substitution, the $Y$-intercept is

$$\alpha = 72.4 - 1.28(13.0) = 55.76$$

The regression equation for the relationship is thus

$$\hat{Y} = 55.8 + 1.28X$$

This line can now be plotted by solving the equation for several arbitrary values of $X$.

| When $X$ is | $\cdots$ | then $\hat{Y}$ must be |
|---|---|---|
| 6 | | 63.4 |
| 10 | | 68.6 |
| 20 | | 81.40 |

This line has been added to the scattergram. Note that the line of regression must pass through both $\mu_X$ and $\mu_Y$.

**Example 13.2**

Use Kroeber's data (Kroeber 1925: 891) on California Indian populations (Table 3.1) to find the line of regression best describing the depopulation in California between 1835 and 1860.

In this case, *time* is the *predictor* variable ($X$) and *population* is the *predicted* variable ($Y$). The scattergram indicates that these data plot on a relatively straight line, but the standard regression procedures are complicated by the large numbers involved (Fig. 13.9). Each variable will be

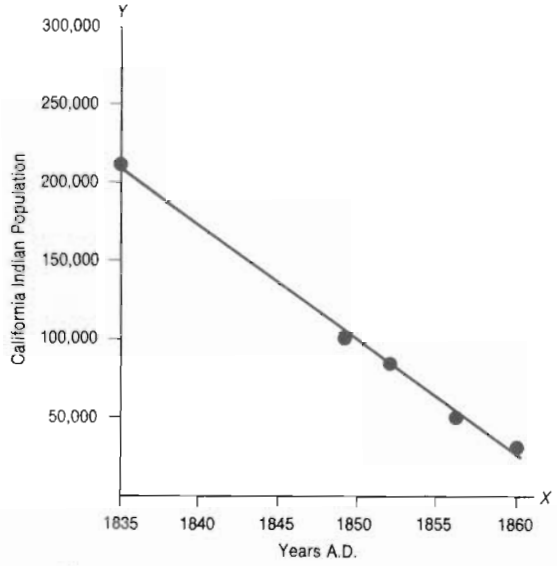


Fig. 13.9 (Data from Kroeber 1925: 891).

*coded* for easier computations. *Time* can be coded by subtracting 1800 from each variate; since time is relatively distributed, this coding simply takes A.D. 1800 as point zero rather than the year A.D./B.C. Population will be coded as $0.0001Y$. The computations are as follows:

| Time | | Population | | | |
|---|---|---|---|---|---|
| Raw Data | Coded Data, $X$ | Raw Data | Coded Data, $\hat{Y}$ | $X^2$ | $XY$ |
| 1,835 | 35 | 210,000 | 21.0 | 1,225 | 735.0 |
| 1,849 | 49 | 100,000 | 10.0 | 2,401 | 490.0 |
| 1,852 | 52 | 85,000 | 8.5 | 2,704 | 442.0 |
| 1,856 | 56 | 50,000 | 5.0 | 3,136 | 280.0 |
| 1,860 | 60 | 35,000 | 3.5 | 3,600 | 210.0 |
| | 252 | | 48.0 | 13,066 | 2157.0 |

$\mu_X = 50.4$ years; $\mu_Y = 9.6$ people; $N = 5$.

The regression constants are computed as usual

$$\beta = \frac{2157 - 5(50.4)(9.6)}{13066 - 5(50.4)^2} = -0.718$$

$$\alpha = 9.6 + 0.718(50.4) = 45.79$$

Thus, the *coded* regression equation is

$$\hat{Y} = 45.79 - 0.718X$$

Sample values can now be computed, decoded by reversing the coding procedure, and the regression line plotted.

| When (coded) $X$ is ··· | then (coded) $Y$ must be | Decoded $X$ | Decoded $\hat{Y}$ |
|---|---|---|---|
| 40 | 17.07 | 1840 | 170,700 |
| 50 | 9.89 | 1850 | 96,900 |
| 55 | 6.30 | 1855 | 63,000 |

The resulting line of regression has been plotted on the scattergram.

## 13.3 ESTIMATING THE ERRORS OF REGRESSION FOR POPULATIONS

So far we have simply assumed that a linear relationship exists between the random variable $X$ and $Y$. But the least squares procedure can be applied to any array of $N$ points, whether or not a linear relationship holds; so, it becomes necessary to determine whether or not the regression equation is meaningful to a specific set of data. Consider the following example.

Clinical researchers have developed a number of methods to determine a child's age, based strictly upon skeletal evidence. One common source of data is the *Greulich–Pyle Atlas*, which presents standardized X-rays of wrist and hand ossification. The X-rays of any living child can be assigned a "skeletal age" by comparison with the standards from this Atlas. Data on 52 subadult males were collected by the Denver Child Research Council in order to assess the accuracy of skeletal age, and the least squares method was used to fit a line to these points. Consider these $N = 52$ observations to be a statistical population. Are the expected values, the $\hat{Y}_i$, on a straight line?

The first important measure of linearity in a population is known as the *standard error of estimate*. So far, we have explicitly assumed that no errors are involved on the $X$ random variable. All the errors of estimation are due to deviations in a vertical ($Y$) direction. The chronological ages plotted on Fig. 13.10 qualify under this model because there are no errors in finding chronological age; we know this from birth records. The total deviation between the $Y$ points and the regression line must be due to errors involved in reading the wrist X-rays. This error is given by $\Sigma(Y_i - \hat{Y}_i)^2$. The *average error* for each point is found simply by dividing the summed squared deviations by $N = 52$, the number

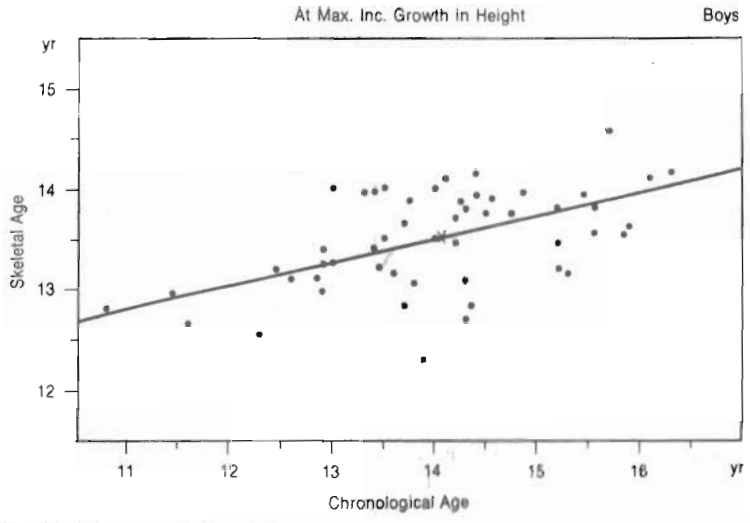


Fig. 13.10 Scattergram of skeletal age against chronological age at the time of maximum increment of growth for 52 boys (after Maresh 1971: fig. 5).

**TABLE 13.2 Computations for Fig. 13.10.**

$\Sigma X = 730.20$ years; $\Sigma X^2 = 10{,}334.98$ years$^2$
$\Sigma Y = 701.60$ years; $\Sigma Y^2 = 9{,}480.44$ years$^2$
$\Sigma XY = 9{,}870.39$ years$^2$; $N = 52$

$$\mu_X = \frac{730.20}{52} = 14.0423 \text{ years}$$

$$\mu_Y = \frac{701.60}{52} = 13.4923 \text{ years}$$

$$\beta = \frac{9870.39 - 52(14.0423)(13.4923)}{10{,}334.98 - 52(14.0423)^2} = 0.225$$

$$\alpha = 13.4923 - 0.225(14.0423) = 10.33 \text{ years}$$

of independent variates in the population. This new index of average deviation is known as the (*population*) *mean squared error of estimate*, denoted by $\sigma^2_{Y \cdot X}$:

$$\sigma^2_{Y \cdot X} = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{N} \tag{13.5}$$

The mean squared error indicates the variance for $Y$, given $X$. The mean squared error of estimate for Fig. 13.10 is $\sigma^2_{Y \cdot X} = 0.194$ years$^2$. This measure denotes the degree of variation between the actual population of $N = 52$ points and the least square estimates of the regression line, $\hat{Y} = 10.33 + 0.225X$. If all the population points were to fall exactly on the regression line, then the relationship would be perfectly linear and $\sigma^2_{Y \cdot X} = 0.0$. The larger the mean squared error, the greater is the deviation from linearity. A strong analogy exists between the mean square

error of estimate and the population variance: Whereas $\sigma^2$ accounts for the variability about a single *point* ($\mu$), $\sigma^2_{Y \cdot X}$ considers the variability about a single *line*, determined by $\hat{Y} = \alpha + \beta$.

Unfortunately, $\sigma^2_{Y \cdot X}$ is expressed in *squared units*, such as years$^2$, cm$^2$, or grams$^2$. This shortcoming, encountered with the population variance, is remedied by taking the square root of the parameter. Hence, the (*population*) *standard error of estimate* is defined as

$$\sigma_{Y \cdot X} = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{N}} \tag{13.6}$$

The meaning of $\sigma_{Y \cdot X}$ is really quite close to that of the population standard deviation. The predictions resulting from a least squares regression line with a small $\sigma_{Y \cdot X}$ will be relatively accurate. That is, the $\hat{Y}_i$, which lie on a straight line, satisfactorily describe the observed $Y_i$. A large standard error of estimate warns that the relationship is only weakly linear, and hence description by a straight line lacks accuracy. When $\sigma_{Y \cdot X}$ is small, a knowledge of any $X_i$ tells us a great deal about the corresponding value of $Y_i$. When $\sigma_{Y \cdot X} = 0$, then $\hat{Y}_i = Y_i$, and the relationship between $X$ and $Y$ is perfectly linear.

As with the standard deviation, a useful shortcut computing formula simplifies calculation of the population standard error of estimate for regression. Not only are the computations involved in Equation (13.6) too laborious, but the numerous subtractions tend to introduce considerable errors of rounding. The following computational formula is always preferred for finding the standard error of estimate:

$$\sigma_{Y \cdot X} = \sqrt{\frac{\Sigma Y^2 - \alpha \Sigma Y - \beta \Sigma XY}{N}} \tag{13.7}$$

The only new quantity required by the computing formula is $\Sigma Y^2$; all the remaining sums are required to find the regression constants $\alpha$ and $\beta$. The regression of skeletal age on chronological age is found to have a population standard error of estimate of

$$\sigma_{Y \cdot X} = \sqrt{\frac{9480.44 - 10.33(701.60) - 0.225(9870.39)}{52}}$$

$$= 0.482 \text{ year}$$

Measuring goodness of fit for regression can be approached in an alternative method. The average skeletal age of the population of $N = 52$ pre-adults plotted in Fig. 13.8 is known to be $\mu_Y = 13.49$ years. The variability about $\mu_Y$ is customarily denoted by the population standard deviation, as

$$\sigma_Y = \sqrt{\frac{\Sigma(Y_i - \mu_Y)^2}{N}}$$

Using computing Formula (4.18), we find that the population standard deviation for skeletal age is

$$\sigma_Y = \sqrt{\frac{\Sigma Y_i^2 - [(\Sigma Y_i)^2/N]}{N}} = \sqrt{\frac{9480.44 - (701.6^2/52)}{52}}$$

$$= 0.523 \text{ year}$$

Two estimates of variability in the $Y_i$ are thus available: The population standard error of estimate, $\sigma_{Y \cdot X} = 0.482$ year, and the population standard deviation, $\sigma_Y = 0.515$ year. It is clear that the least squares regression line provides a more accurate estimator of the $Y_i$ than does the mere standard deviation of $Y$. The *improvement in estimation* of the linear estimate over the point estimate can be expressed as the ratio of the mean squared standard error of estimate, $\sigma^2_{Y \cdot X}$, to the population variance, $\sigma_Y^2$:

$$k^2 = \frac{\sigma^2_{Y \cdot X}}{\sigma_Y^2} \tag{13.8}$$

The expression $k^2$ is known as the *Coefficient of Nondetermination*, and its meaning should be obvious: $k^2$ represents the proportion of variability in the $Y$ variable which remains *unexplained* after the nature of the $X$ and $Y$ articulation has been assessed by $\sigma^2_{Y \cdot X}$. That is, $k^2$ tells us how much variability the regression equation does not "account for." When $k^2$ is equal to zero (that is, $\sigma_{Y \cdot X} = 0.0$), then a perfect correspondence between $X$ and $Y$ must exist because all the variability in $Y$ can be accounted for by a knowledge of $X$. This means that $\hat{Y}_i = Y_i$ for all $i$. Conversely, when $k^2$ equals unity, no relationship exists at all: $X$ tells us exactly nothing about $Y$. The Coefficient of Nondetermination for the example of skeletal versus chronological age is

$$k^2 = \frac{0.482^2}{0.523^2} = \frac{0.2323}{0.2652} = 0.849$$

This value of $k^2$ tells us that the regression of the $Y_i$ on $X$ *fails to account* for about 85 percent of the total variability known to exist in $Y_i$. We could also reverse this coin and concentrate upon the amount of variance *explained* by regression, as expressed by the *Coefficient of Determination* ($\rho^2$):

$$\rho^2 = 1 - k^2 = 1 - \frac{\sigma^2_{Y \cdot X}}{\sigma_Y^2} \tag{13.9}$$

The Coefficient of Determination, denoted by the Greek letter $\rho^2$, is merely the complement of $k^2$. The above value of $k^2 = 0.849$ must have $\rho^2 = 1.0 - 0.849 = 0.151$. The Coefficient of Determination indicates that the regression equation accounts for only about 15 percent of the variability in skeletal age. The choice between $k^2$ and $\rho^2$ reflects only one's philosophy: optimistic (percent explained by $\rho^2$) or pessimistic (percent unexplained by $k^2$).

One further index of interest is the square root of the Coefficient of Determination:

$$\rho = \sqrt{\rho^2} = \sqrt{1 - \frac{\sigma^2_{Y \cdot X}}{\sigma_Y^2}} \tag{13.10}$$

The symbol $\rho$ is also known as the *population correlation coefficient*. This measure, an extremely important index of the bivariate relationships between two normally distributed populations, is used in practice considerably more often than either coefficients of determination or nondetermination. So important is rho that Chapter 14 is devoted to discussion of the correlation coefficient.

## 13.4 LEAST SQUARES REGRESSION AS STATISTICAL INFERENCE

Hubert Blalock (1972: 364) calls regression lines the *laws* of social science. All else being equal, a knowledge of $X$ sufficiently predicts the behavior of $Y$. But anthropologists are accustomed to accepting some rather loosely constructed "laws," not only because of the crude measurements, but also because of the general variability of human behavior. Most social scientists freely acknowledge that while the laws of the physical scientists are quite exact, social and behavioral laws must always remain less precise, more fluid, only actuarial in nature. It seems that the laws of social science have a built-in degree of variability. So, if social scientists indeed seek the "underlying laws" governing human behavior, and if regression equations can serve as one expression of these laws, *how then is the sampling variability in regression to be handled*?

Regression lines have been treated simply as descriptive devices used to summarize populations of bivariate points. The equation $\hat{Y} = 7179.6X - 13{,}189{,}132$ *describes* the quantitative depopulation of California Indians between 1835 and 1860; the expression $\hat{Y} = 3.5X - 115.7$ *describes* the relationship between stature and weight in a certain physical anthropology seminar; the equation $\hat{Y} = 10.33 + 0.225X$ *describes* the relationship between chronological age and skeletal age in 52 subadults from the Denver area. Used in this manner, regression is a tool which allows anthropologists to fit descriptive lines to known populations of points.

Accordingly, the indices derived in Section 13.3 were all based upon *populations* of variates, and Greek letters were used to denote the population parameters. But regression fulfills a more critical niche in anthropology than does mere description. Regression lines computed for *samples of variates* and the least squares equations used in statistical inference to predict an unknown parameter seem an observed sample. Regression equations have functioned to this point only in the limited capacity of descriptive devices. Regression can now be treated as a tool for inferential statistics.

Figure 13.11 illustrates the sampling distribution for the regression equation. For every value of $X$ there exists a *distribution* of $Y_i$ values. The regression equation estimates one single value ($\hat{Y}_i$) of these values. The prediction $\hat{Y}$ occurs on the regression line directly above the preselected value on the $X$-axis. But the actual observed values of the $Y$ variable—denoted simply by $Y_i$—could lie anywhere on the vertical axis above the given $X$ value, as illustrated in Fig. 13.9. The predicted value $Y_i$ resulting from the least squares equation is thus taken as an estimator of the theoretical mean of the distribution of $Y_i$ at point $X$. A different normal distribution holds for every interval along the $X$-axis. The scatter of such points depends not only upon the value of $\hat{Y}_i$ (the estimated mean above $X_i$) but also upon the shape of the population about the line of regression. If the population of points lies close to the line, then the observed sample of points should also land rather close to the line of regression.

Section 13.3 introduced the concept of the population standard error of estimate due to regression. The average error was determined for every $Y_i$ in a population by dividing the sum of squared deviations by the number of points. $\sigma_{Y \cdot X}$ is a population parameter, applicable whenever an entire population of

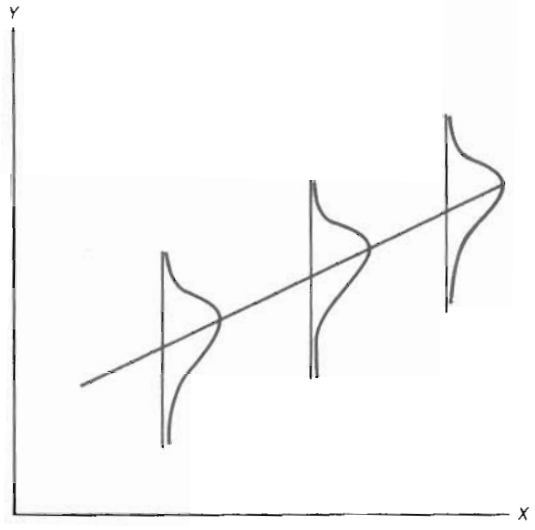


Fig. 13.11

variates can be observed, but when this complete population is only partly visible—when this population has been *sampled*—then the parameter $\sigma_{Y \cdot X}$ can be *estimated* only by an appropriate sample statistic. Let us denote the (*sample*) *standard error of estimate* as $S_{Y \cdot X}$:

$$S_{Y \cdot X} = (\text{estimate of } \sigma_{Y \cdot X}) = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-2}} \tag{13.11}$$

$S_{Y \cdot X}$ is a sample statistic which functions as do other statistics considered earlier. The sample standard deviation, for example, was used as an estimator of the population standard deviation whenever only a sample of size $n < N$ was available. The population standard deviation, $\sigma_X$, was computed with a denominator of simply $N$, thereby providing average variability for each variate within the population. But a slight modification in $S_X$ was computed with a denominator of $(n-k)$, where $k$ denoted the number of degrees of freedom lost in the act of computing the standard deviation. Because $\bar{X}$ was required before $S_X$ could be computed, exactly one degree of freedom was lost, so $S_X$ was computed with a denominator of $(n-1)$.

A parallel situation holds when the standard error is estimated for regression. The population parameter was computed strictly as the average squared deviation per variate pair, with a denominator of $N$. But to obtain an estimate of $\sigma_{Y \cdot X}$, one must consider the number of degrees of freedom lost due to computations. In this case, $k = 2$, which means that two previously computed quantities are necessary: the regression coefficients $\alpha$ and $\beta$. So it is that the denominator of Expression (13.11) contains $(n-2)$, whereas the denominator of $\sigma_{Y \cdot X}$ is simply $N$.

The population standard error of estimate was computed earlier to be $\sigma_{Y \cdot X} = \sqrt{12.074/52} = 0.482$ year. $\sigma_{Y \cdot X}$ is a parameter describing a population of

52 variates. If the 52 subjects had been considered to be a *sample* of size $n = 52$ (rather than a *population* of size $N = 52$), then the sample standard error of estimation would properly be

$$S_{Y \cdot X} = \sqrt{\frac{12.074}{52 - 2}} = 0.491 \text{ year}$$

The difference in value between parameter and statistic in this case is virtually nil because of the large sample size. But in the blood pressure example (Example 13.1) with $N = 9$, the difference is rather large: $S_{Y \cdot X} - \sigma_{Y \cdot X} = 4.117 - 3.631 = 0.486$ mm. This discrepancy, apparent in all small samples, causes certain sampling difficulties which will be subsequently considered.

As with most estimators of variability, the standard error of regression is based upon deviations about means. But a *computing formula* is available which enables computation of standard error of estimate without the necessity of finding each individual deviation:

$$S_{Y \cdot X} = \sqrt{\frac{\Sigma Y^2 - a \Sigma Y - b \Sigma XY}{n - 2}} \tag{13.12}$$

This formula is generally preferable to (13.11) because errors of rounding are minimized.

The *confidence limits* about a specific regression prediction are given as

$$\text{confidence limits} = \hat{Y} \pm tS_{\hat{Y}} \tag{13.13}$$

where

$$S_{\hat{Y}} = S_{Y \cdot X} \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{(\Sigma X_i^2 - (\Sigma X_i)^2/n)}}$$

for a specific value of $X$ and $(n - 2)$ degrees of freedom. The general format of Expression (13.13) should be familiar: Confidence limits are defined as some critical region on either side of some mean, this interval being defined by $t$ and $S_{\hat{Y}}$. The $t$-value is determined as before from Table A.4 by the appropriate level of significance (95 percent confidence interval, 99 percent confidence interval, etc.) and also by the number of degrees of freedom, in this case $(n - 2)$ degrees of freedom. The measure of variability, $S_{\hat{Y}}$, is the direct analog of $S_{\bar{X}}$ which was used in earlier confidence limits computations. Just as $S_{\bar{X}}$ is the standard error of the mean $\bar{X}$, so is $S_{\hat{Y}}$ the standard error of the estimate $\hat{Y}$. Remember that these confidence limits apply only to the specified value of $X$ and not to the entire regression equation.

We are now in a position to use the least squares estimate as an inferential statistic. Let us return to the case of blood pressure among the Surinamese (Example 13.1). Remember that the informants were considered earlier to constitute a population of $N = 9$ variates. The relationship between age ($X$) and blood pressure ($Y$) was described by the standard regression equation:

$$Y = \alpha + \beta X = 55.8 + 1.28X$$

Suppose now that these same informants are taken to represent a sample of size $n = 9$, which was randomly selected from the biological population of all

Surinamese. (To accomplish such sampling, it would be necessary to independently draw a single informant from each age group.) The population parameters are unknown in this situation, and must be estimated by sample statistics. To keep this new distinction straight, it is necessary to redefine the regression equation. The regression coefficients $\alpha$ and $\beta$ are now unknown parameters, which must be estimated by sample statistics, which we denote as $a$ and $b$. The *least squares regression equation for samples* is thus

$$\hat{Y} = a + bX \tag{13.14}$$

The new statistics $a$ and $b$ are defined identically to $\alpha$ and $\beta$ except that all parameters have been replaced by statistics:

$$b = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\Sigma(X - \bar{X})^2} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2} \tag{13.15}$$

$$a = \bar{Y} - b\bar{X} \tag{13.16}$$

The symbolism of regression might seem a bit excessive, but this is really necessary in order to keep the descriptive functions of least squares regression from its inferential function.

The least squares equation describing the *sample* of $n = 9$ Surinam informants is therefore

$$\hat{Y} = a + bX = 55.8 + 1.28X$$

The numbers here remain unchanged from Example 13.1, although now $a$ is taken to estimate $\alpha$ and $b$ estimates $\beta$.

How is this regression expression used as an inferential statistic? Suppose that we wish to predict the blood pressure of a ten-year-old Surinamese. The regression equation for $X = 10$ yields $\hat{Y}$:

$$\hat{Y} = 55.8 + 1.28(10) = 68.6 \text{ mm}$$

Of course nobody should expect that all ten-year-old Surinam informants will have exactly this blood pressure, but $\hat{Y}$ serves as an estimate for the average blood pressure of ten-year-olds. The sample standard error of regression was computed previously to be $S_{Y \cdot X} = 4.117$ mm, so the 95 percent confidence limits are given by Formula (13.13):

$$\text{confidence limits} = \hat{Y} \pm t_{0.05} S_{\hat{Y}}$$

$$= 68.6 \pm 2.365(4.117)\sqrt{\frac{1}{9} + \frac{(10.0 - 13.0)^2}{1761 - 117^2/9}}$$

$$= 68.6 \pm 2.365(1.587)$$

$$= 68.6 \pm 3.753 \text{ mm}$$

The value of $t_{0.05}$ was found from Table A.4 listed under $(9 - 2) = 7$ degrees of freedom. At the 95 percent confidence level, we expect a randomly selected ten-year-old Surinam male to have a blood pressure reading greater than 64.85 mm but less than 72.35.

Select another age, say, $X = 6$ years. The regression prediction is $\hat{Y} =$

63.44 mm, with the following 95 percent confidence interval:

$$\text{confidence limits} = 63.4 \pm 2.365(4.117)\sqrt{\frac{1}{9} + \frac{(6.0 - 13.0)^2}{1761 - 117^2/9}}$$
$$= 63.4 \pm 2.365(2.311)$$
$$= 63.4 \pm 5.466 \text{ mm}.$$

Similarly, the 95 percent confidence interval for $X = 20$ years is

$$\text{confidence limits} = 81.4 \pm 2.365(4.117)\sqrt{\frac{1}{9} + \frac{(20.0 - 13.0)^2}{1761 - 117^2/9}}$$
$$= 81.4 \pm 5.466 \text{ mm}$$

Note that these confidence intervals about every $Y_i$ are *specific for a given* $X$. The confidence interval for $\hat{Y} = 68.6$ mm (corresponding to $X = 10$) are $\pm 3.753$ mm, while those for an age of $X = 6$ are $\pm 5.466$ mm. The closer the samples fall to the mean of $X$ (in this case, $\bar{X} = 13.0$ years), the smaller will be the confidence interval about $\hat{Y}$. That is, the predictions become less accurate as the given $X$ deviates from $\bar{X}$. Note further that because $X = 6$ and $X = 20$ are equidistant from the mean age of $\bar{X} = 13.0$ years, the confidence intervals are identical.

These three confidence intervals have been plotted on Fig. 13.12. The enclosed

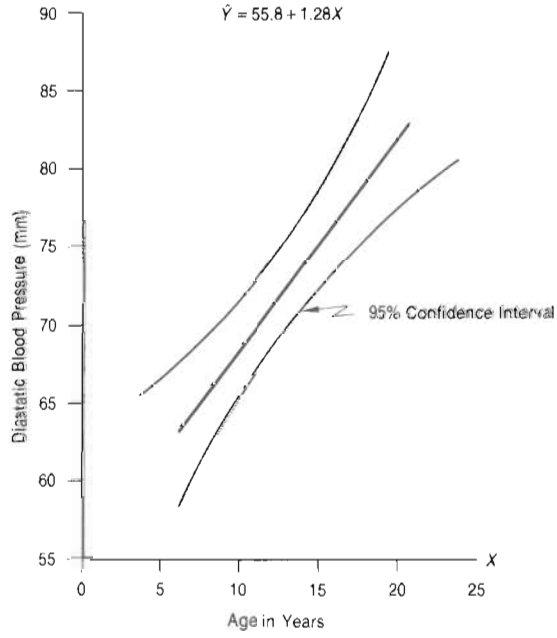


Fig. 13.12 Regression of blood pressure and age for Surinam informants (data from Glanville and Geerdink 1972: table 2).

area represents the approximate 95 percent confidence band for the total regression equation. This band does not parallel the line of regression; rather it pinches in toward the mean and fans out as one deviates from the mean because predictions about the mean of $X$ are known to be more accurate than those some distance from $\bar{X}$. Furthermore, note that the confidence band about the regression line does not extend beyond the observed range of the sample variates.

**Example 13.3**

Determine the standard error of estimate for the stature-weight data in Fig. 13.7.

The sample standard error of estimate can be found using Formula (13.12), a method which entails actually computing the deviations between observed and predicted variates.

| $X$ | $Y$ | $\hat{Y}$ $\hat{Y} = a + bX$ | $\lvert Y - \hat{Y} \rvert$ | $\lvert Y - \hat{Y} \rvert^2$ |
|---|---|---|---|---|
| 61 | 98 | 3.5(61) − 115.7 = 97.8 | 0.2 | 0.04 |
| 63 | 107 | 3.5(63) − 115.7 = 104.8 | 2.2 | 4.84 |
| 65 | 109 | 3.5(65) − 115.7 = 111.8 | 2.8 | 7.84 |
| 67 | 117 | 3.5(67) − 115.7 = 118.8 | 1.8 | 3.24 |
| 69 | 128 | 3.5(69) − 115.7 = 125.8 | 2.2 | 4.84 |
| | | | | 20.80 |

Substituting into (13.11),

$$S_{Y \cdot X} = \sqrt{\frac{20.80}{3}} = 2.63 \text{ lb}$$

The computing formula allows much easier computation, however, because the sums involved are the same as those used in Table 13.1 to find $a$ and $b$.

$$S_{Y \cdot X} = \sqrt{\frac{63{,}007 - (-115.7)559 - 3.5(36{,}475)}{3}}$$

$$= \sqrt{\frac{20.80}{3}} = 2.63 \text{ lb}$$

The only new quantity required is $Y^2$. Whenever tables of computations are framed, one should usually include a column for $Y^2$. Even though $Y^2$ is not directly involved in finding the regression equation, this sum is handy when assessing errors of regression.

## 13.5 ASSUMPTIONS OF LEAST SQUARES REGRESSION

The least squares method has been used for two distinct purposes in this chapter: Description and inference. As long as the objective is simply to describe a swarm of points, the assumptions need not trouble us. It is not necessary to assume anything about the form of the distribution or the variability of the $Y_i$ over the $X_i$, or even to worry about the level of measurement implied (Hays 1973: 636). All a least squares description does is treat $N$ distinct cases *as if* they were linear. The regression equation describes the population of points in terms of their tendency to associate in a linear fashion. This description is tenable only for the exact $N$ points considered.

But when the least squares method is used to generate inferences about unknown values of $\hat{Y}$, then some important assumptions are required:

1. *The predictor variable, X, is measured without error.* The levels of $X_i$ are to be arbitrarily selected beforehand by the investigator and do not result from any sort of sampling operation. There is only one special situation, however (the so-called *Berkson case*), which permits a special sort of error to creep into the $X_i$. As long as the errors on $X$ are strictly resultant from inaccuracies of measurement or the lack of suitable experimental precision, then least squares regression can still be valid. That is, if informants are selected for age, a certain amount of error might result among nonliterate subjects who are truly ignorant of their age. Or when students are grouped into classes of increasing stature, there may be some small error due to using a 1 meter tape. Errors of this kind can be permitted only as long as their magnitude is totally unrelated to the magnitude of the variate (see Sokal and Rohlf 1969: 482–483). This sort of error occurs on Fig. 13.5. With only this exception, random fluctuation of $X$ is not permitted in least squares procedure.
2. *The samples along the regression line are homoscedastic.* Each of the normally distributed populations of $Y_i$ above each $X$ must have the same variance. That is, the $\sigma_{Y \cdot X}$ for all $X$ are assumed to be equal.
3. *A linear relationship must exist between X and Y* (or a suitable transformation must be applied; discussed in Chapter 14).
4. *Both X and Y are measurable on at least an interval scale.*
5. *The line of regression applies only within the observed range of the X.*
6. *The $Y_i$ for any given level of X must be independently and randomly selected from a normally distributed population above X* (see Fig. 13.11).

## 13.6 LEAST SQUARES REGRESSION THROUGH THE ORIGIN

Situations sometimes demand that specific regression lines must commence, or pass through, the origin of the coordinate graph, and the previous example (Section 13.1) relating age to the number of annual growth rings serves as a case in point. The discussion was presented *as if* one year will always produce *precisely* one ring, but this assumption holds true only in the long run. Specific trees are known to fail to add rings in some years, or the rings are too indistinct

for detection by dendrochronologists. In trees such as the bristlecone pine, several trunks often exist. Some of these trunks might be dormant at any one time, thereby failing to add rings over sometimes lengthy intervals. But despite the exigencies which introduce error into the simple $Y = X$ relationship, the regression line should still logically commence at the origin of the graph and then progress outward (as in Fig. 13.1). When the age is zero, there must be zero tree rings, regardless of what random errors are introduced by climatic and physiological factors.

It becomes useful in many such cases to fit a special line of regression which automatically begins at the origin (0, 0). This is a limited case of the general regression situation $Y = a + bX$, in which the $Y$-intercept is a priori defined to be $a = 0.0$. That is, this line of regression is required to intercept the $Y$-axis at point zero. So, the formula for the *regression through the origin* reduces simply to

$$\hat{Y} = b''X \tag{13.17}$$

where $b'' = \Sigma XY/\Sigma X^2$. The superscript $b''$ distinguishes the slope computed from Expression (13.15) for the common slope $b$.

This simplified least squares estimate can be illustrated in the case of germinating plant seeds. When planted, a seed has zero height, and time of growth is also zero. The stalk then progresses steadily through both time and height, in its inexorable climb upward. Height measurements were taken at odd intervals on a single stalk of corn which was growing near Davis, California (Table 13.3). A regression line can be fitted to these data, using the standard $\hat{Y} = a + bX$ method (see Table 13.3), with the following result:

$$\hat{Y} = 15.3 \text{ cm} + 8.23X$$

where $Y$ is measured in centimeters and $X$ is age in weeks. A series of sample points can be readily generated from this expression, and the line has been fitted to the actual data in Fig. 13.11. But note that this conventional approach to regression produces a line intersecting the $Y$-axis at $a = 15.3$ cm. Even the rankest urbanite must surely recognize that no hybrid, regardless of its vigor, could possibly begin growth with a height of 15 cm! Although the least squares fit has adequately minimized $\Sigma(Y_i - \hat{Y}_i)^2$, the resulting equation produces substantively ridiculous results.

A more appropriate regression line would commence at the origin, thereby denoting zero height to a newly planted corn kernel. Taking the $Y$-intercept a priori to be $a = 0.0$, the regression Formula (13.17) is readily applicable. From the data in Table 13.3, the proper slope is found to be

$$b'' = \frac{\Sigma XY}{\Sigma X^2} = \frac{14{,}600}{1593} = 9.17$$

The new regression equation is thus

$$\hat{Y} = b''X = 9.17X$$

Sample values have been computed from this new equation and plotted in Fig. 13.13. The two lines of regression are nearly parallel, the difference in slope amounting to only $(b'' - b) = 9.17 - 8.23 = 0.94$ cm increase per week. But the

**TABLE 13.3 Observations on one stalk of corn near Davis, California.**

| Age, weeks $X$ | Height, cm $Y$ | $X^2$ | $XY$ |
|---|---|---|---|
| 4 | 56 | 16 | 224 |
| 8 | 81 | 64 | 648 |
| 12 | 110 | 144 | 1,320 |
| 14 | 130 | 196 | 1,820 |
| 17 | 150 | 289 | 2,550 |
| 20 | 172 | 400 | 3,440 |
| 22 | 209 | 484 | 4,598 |
| 97 | 908 | 1,593 | 14,600 |

$\bar{X} = 13.9$ weeks; $\bar{Y} = 129.7$ cm; $n = 7$.

*Regression by standard least squares methods*:

$$b = \frac{14{,}600 - 7(13.9)(129.7)}{1593 - 7(13.9)^2} = 8.23$$

$$a = 129.7 - 8.23(13.9) = 15.3$$

Then

$$\hat{Y} = 15.3 + 8.23X$$

*Regression through the origin*:

$$b = \frac{14{,}600}{1593} = 9.17X$$

$$\hat{Y} = 9.17X$$

15.3 cm difference in the $Y$-intercept causes a rather wide difference between the actual paths of the lines.

| | $X$, Age in weeks | | | |
|---|---|---|---|---|
| $Y$, Height in cm | 5 | 10 | 15 | 25 |
| $\hat{Y} = 15.3 + 8.23X$ | 56.5 | 97.6 | 138.8 | 221.1 |
| $\hat{Y} = 9.17X$ | 49.9 | 91.7 | 137.6 | 229.3 |
| Difference | 6.5 | 5.9 | 1.2 | 8.2 |

The difference between the two methods of regression can also be illustrated by comparing the sample values from each equation. Both regression lines pass through the sample means, so the discrepancy between predictions becomes worse as the samples diverge from the mean values.

Each prediction method possesses certain advantages. The initial equation of least squares is more accurate, providing the exact minimum value of the total error of estimate. So, when predicting a single occurrence of $Y$ from an $X$—How tall will the corn be in 15 weeks?—the standard least squares method will provide superior results. But, while introducing a somewhat higher total error of

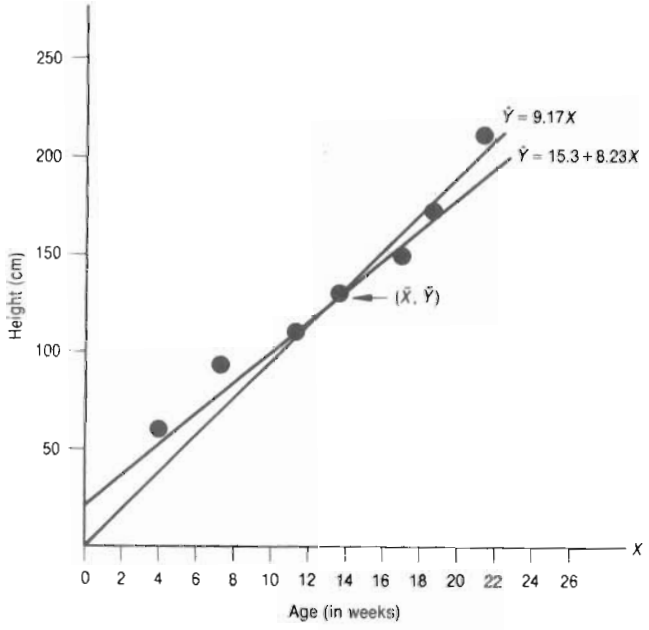


Fig. 13.13 Regression of corn height and age, plotted by least squares and through-the-origin

estimate, the regression through the origin produces a graphic solution which is more acceptable in terms of overall substantive implications. The decision as to which method is best will depend only upon the situation at hand rather than upon abstract mathematical properties.

## 13.7 LEAST SQUARES REGRESSION OF *Y* ON *X* VERSUS REGRESSION OF *X* ON *Y*

Care has been taken to restrict discussion to the "regression of $Y$ on $X$." The $X$ values have been taken as the *predictor* variates, fixed at predetermined values. $Y$ has been the *predicted* variable, presumably randomly selected from the population of points above each $X_i$. Thus, both the mode of sampling and the substantive predictive interest serve to distinguish $X$ from $Y$ in the regression model considered so far. But circumstances will arise occasionally in which it becomes necessary to reverse the relationship and attempt to predict values of $X$, given the $Y_i$. A new regression equation would thus be involved:

$$\hat{X} = c + dY \tag{13.18}$$

These new regression coefficients correspond to the $X$-intercept and slope, except that the quantity $\Sigma(X_i - \hat{X}_i)^2$ has been minimized.

When the resulting line for Equation (13.18) is plotted, it will almost never correspond exactly to the line produced by $\hat{Y} = a + bX$. In fact, as long as error is

present on *either* variable, different lines will always result. Only when all points lie exactly on their lines of regression will the two lines coincide. That is, only when $\Sigma(X_i - \hat{X}_i)^2 = \Sigma(Y_i - \hat{Y}_i)^2 = 0$ will a single regression equation apply for both $X$ on $Y$ and $Y$ on $X$. Extreme caution is in order whenever one attempts to tamper with the predictor-predicted relationship in regression. Sokal and Rohlf (1969:446–448) have discussed methods for predicting $X$ from a given $Y$ including a method to compute a confidence interval for this inverse prediction. But more often, however, the assumptions of least squares regression will not be satisfied in such cases, and a second regression method will prove a more effective means of prediction.

## 13.8 MODEL II REGRESSION

The regression method discussed so far has been based upon the assumptions listed in Section 13.5. Least squares methods are especially well suited for controlled experimental conditions in which the predictor values—the $X_i$—can be artificially "fixed" by the investigator. Errors on the $X$ variable are thereby eliminated. Now we must consider a second method of fitting a regression line, called *Model II*. Under *Model II the X variable is no longer fixed.* The $X_i$ are randomly selected in a manner identical to selection of the $Y_i$ variates. In fact, the assignment of the labels "predictor" and "predicted" in Model II regression is merely for convenience because no intrinsic difference is necessary to regress $X$ on $Y$ or $Y$ on $X$ under the Model II method of regression.

When using the least squares procedure for estimation, it was necessary to assume that the $Y_i$ were normally distributed for every $X_i$. No assumptions were necessary regarding $X$ except that the level of measurement was interval scale. The sampling distribution of Model I (least squares regression, Section 13.2) was depicted in Fig. 13.11. But Model II methods of regression consider $X$ and $Y$ to be equivalent variables. The $Y_i$ must be normally distributed above the $X_i$ and the $X_i$ must likewise be normally distributed across the $Y$. Because both $X$ and $Y$ are assumed to be independent and normally distributed, the population distribution is known as *bivariate normal.* To visualize this configuration, it is necessary to conceive of a very large number of datum points stacked about the intersection of $\mu_X$ and $\mu_Y$. Of course not all points will land exactly on the two means, so random errors will distribute the points farther and farther from this intersection. The greater these random errors, the more dispersed are the datum points. The computer-generated diagram in Fig. 13.14(a) shows that, in three dimensions, the bivariate normal distribution is bell-shaped. As long as the $X$ and $Y$ are measured on identical scales, the bivariate normal mound tends to be symmetrical. When differing scales of measurement are involved, or whenever correlation between $X$ and $Y$ is strong, the bell-shape becomes distorted into a more elliptical shape [see Fig. 13.14(b)]. Both the Model II regression and the correlation coefficient $r$ (to be considered in detail in Chapter 14) assume that the $X_i$ and $Y_i$ are randomly sampled from such a bivariate normal population distribution.

Least squares models can, of course, be used to fit a line describing a bivariate normal distribution, but because of the errors on $X$, inferences from that line are not valid. Probably the best technique for fitting a line to random

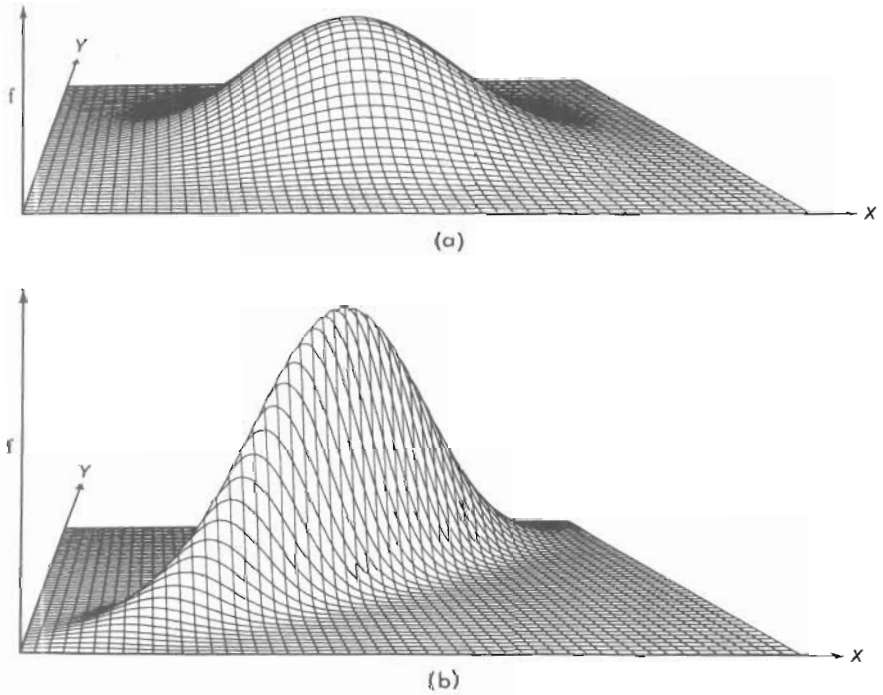


Fig. 13.14 Computer generated frequency diagrams illustrating the bivariate normal distribution. The parametric correlation between variables $X$ and $Y$ in Fig. 13.14(a) is equal to zero, while in Fig. 13.14(b), $\rho = 0.9$ (after Sokal and Rohlf 1969: figs. 15.1 and 15.2).

samples of $X$ and $Y$ is *Bartlett's Three-Group Method*. The estimates of the parametric regression coefficients are quite accurate, but an even greater advantage is that Bartlett's method is extremely easy to compute.

Originally introduced by Bartlett in 1949, this regression technique involves a few simple steps:

1. Rank the variates into descending order on one of the variables; since both variables have been randomly sampled, assignment of $X$ and $Y$ is arbitrary.
2. Divide the ordered array into thirds. Should the number of $(X_i, Y_i)$ pairs not be a multiple of 3, then arrange the categories so that the first and third groups are of the same size.
3. Compute the grand means $\bar{X}$ and $\bar{Y}$ as usual, and then also find the subgroup means for the first and third groups ($\bar{X}_1$, $\bar{Y}_1$, $\bar{X}_3$, $\bar{Y}_3$).
4. The slope of the Model II regression equation is given by

$$b' = \frac{\bar{Y}_3 - \bar{Y}_1}{\bar{X}_3 - \bar{X}_1} \tag{13.19}$$

Following Sokal and Rohlf (1969), the regression coefficients for Model II will be denoted as $a'$ and $b'$ to distinguish them from their Model I counterparts.

5. The *Y*-intercept is given by

$$a' = \bar{Y} - b'\bar{X} \tag{13.20}$$

The resulting equation provides the "best fit" when both variables have been randomly sampled. This is not a least squares regression, so special methods are required for significance testing and confidence interval computations. The following example illustrates Bartlett's method of regression.

Like many anthropologists, I spend a good deal of my time doing fieldwork, which generally entails camping in some fairly remote spot for months on end. One summer, when I was excavating Gatecliff Shelter in central Nevada, I found myself lolling about the evening campfire, listening to the crickets and watching the sagebrush grow. The crickets reminded me that I read somewhere that crickets chirp in response to the temperature: The colder the night, the less the crickets chirp, until they finally refuse to sound off at all in freezing temperatures. I mentioned this astounding little piece of trivia to my crew members, and they must have been as desperate for entertainment as I was, because we all commenced counting cricket chirps to see how cold it was. Of course none of us knew the magic formula for converting chirps to temperature, so we decided to derive our own formula.[3] We performed the counts off and on throughout the night and the following morning, and our results are tabulated in Table 13.4.

**TABLE 13.4 Field Data for cricket chirps from Monitor Valley, central Nevada.**

| Number of Chirps per Minute $X$ | Temperature, °F $Y$ | |
|---|---|---|
| 47 | 50 | |
| 61 | 56 | $\bar{X}_1 = \frac{180}{3} = 60.00$ |
| 72 | 57 | $\bar{Y}_1 = \frac{163}{3} = 54.33$ |
| 78 | 57 | |
| 93 | 63 | |
| 106 | 67 | |
| 110 | 68 | |
| 122 | 73 | $\bar{X}_3 = \frac{364}{3} = 121.33$ |
| 132 | 70 | $\bar{Y}_3 = \frac{211}{3} = 70.33$ |

$\bar{X} = 821/9 = 91.22$; $\bar{Y} = 561/9 = 62.33$.

[3]For the uninitiated, counting cricket chirps is not as easy as it might sound. Some of the critters are boldly sounding off for all to hear, but others—perhaps the younger crickets—are squeaky and difficult to hear. We finally devised a method to reduce the error in our counts: three of us counted the same chirps for 15 sec, but the episode was recorded only if two of us agreed on our count.

The problem was this: Given the number of cricket chirps per minute ($X$), find the equation which will predict temperature in degrees Fahrenheit ($Y$). In this case, both the $Y_i$ and the $X_i$ have been sampled rather than arbitrarily selected; thus, the least squares approximation is invalidated. As long as the populations of cricket chirps and temperature are distributed in bivariate normal fashion, Bartlett's method of regression is appropriate. The first step in finding the magic formula for using the cricket's "vocal thermometer" is to reorder the data into ascending order of $X$, and divide this array into thirds. The necessary means are computed in Table 13.4, so $a'$ and $b'$ are computed as follows:

$$b' = \frac{\bar{Y}_3 - \bar{Y}_1}{\bar{X}_3 - \bar{X}_1} = \frac{70.33 - 54.33}{121.33 - 60.00} = 0.2609$$

$$a' = \bar{Y} - b'\bar{X} = 62.33 - (0.26)91.22 = 38.61$$

Rounding the computed values of $a'$ and $b'$ to quantities more suitable for field usage, the final prediction equation is thus

$$\hat{Y} = 39 + 0.26X$$

A few sample points must be computed in order to fit this line to the scattergram:

| | | | |
|---|---|---|---|
| Chirps per minute ($X$): | 50 | 100 | 150 |
| Temperature °F ($Y$): | 52 | 65 | 78 |

The data and the best fit line appear in Fig. 13.15. Note that the line determined by Bartlett's Best Fit Method not only passes through the subgroup means, but also through the grand means $\bar{X}$ and $\bar{Y}$.

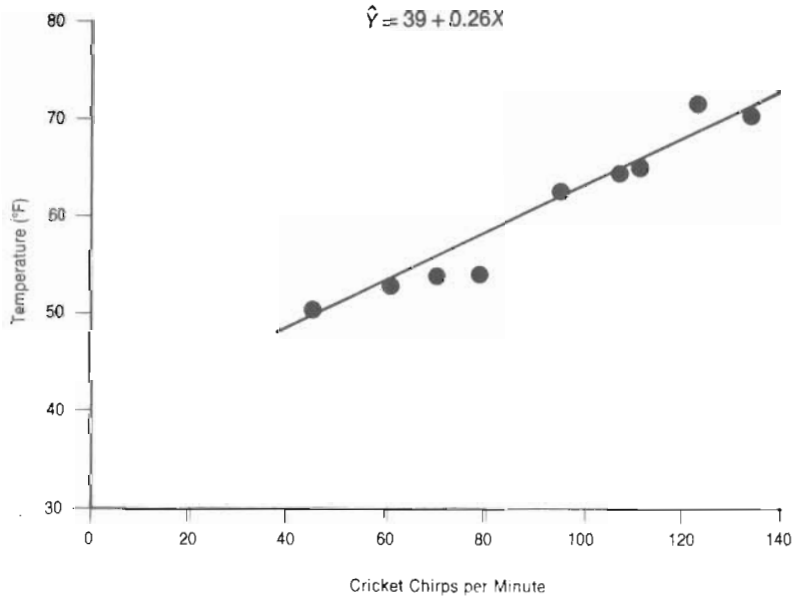


Fig. 13.15 Magic formula for connecting cricket chirps to temperature.

Because Bartlett's method does not assume fixed effects on $X$, the prediction equation for $X$ from $Y$ is simply the reverse of $Y$ from $X$. If we had wished to predict the number of chirps for any particular temperature, the equation would be readily found as follows:

$$b' = \frac{\bar{X}_3 - \bar{X}_1}{\bar{Y}_3 - \bar{Y}_1} = \frac{121.33 - 60.00}{70.33 - 54.33} = 3.833$$
$$a' = \bar{X} - b'\bar{Y} = 91.22 - 3.833(62.33) = -147.7$$
$$\hat{X} = 3.7Y - 148$$

Note that this revised equation will still predict the above sample values.

We noted above that Bartlett's method of finding the best-fit line has the distinct advantage of considerably easier computations than the least squares method. Unfortunately, this computational ease is more than offset by the difficulties in finding the confidence limits to Bartlett's line of regression. The interested reader is referred to discussions in Simpson, Roe, and Lewontin (1960:233–235) and Sokal and Rohlf (1969:483–486) for the appropriate methods.

The standard error of estimate is also invalid when Bartlett's Three-Group Method is used. The best measure of strength of linearity is the correlation coefficient $r$, which is discussed in Chapter 14.

Inference from Model II regression makes the following assumptions[4]:

1. *Variables X and Y are assumed to be in bivariate normal distribution* (see Fig. 13.12).
2. *A linear relationship must exist between X and Y.*
3. *Both X and Y are measurable on at least an interval scale.*
4. *The line of regression applies only within the observed range of the X.*

[4]Some confusion seems to exist about the topic of inference from linear regression, not only among the users of the technique but also among mathematical statisticians as to just which methods of regression are applicable to which empirical situations. The advocates of a *hard line* approach to regression restrict the true least squares method (Model I regression) to cases in which the values of $X$ have been rigidly predetermined, as in laboratory situations or agricultural field experiments. Although a certain degree of error might creep into $X$ through faulty observation, the variable $X$ can still be considered to be "fixed" as long as it has not been *sampled* in the conventional sense. Some less rigorous techniques of regression are available for use whenever the variable $X$ has been sampled rather than selected. Examples of this *hard* approach to regression are found in Sokal and Rohlf (1969), Hays (1973), Bliss (1967), and Dixon and Massey (1969).

A second perspective—I hesitate to term the *soft line*—holds that the mathematically valid distinction between fixed and random effects on $X$ has little relevance to actual application of regression methods. To paraphrase one such advocate, the hard approach makes good mathematical statistics but rather poor science. This view is held by Simpson, Roe, and Lewontin (1960), who argue that the Model I regression techniques can be applied to $X$ variables of any sort (fixed or random), as long as measurement is interval scale or better. The distinction is often made between the predictive (functional) and the descriptive (structural) purposes of regression equations.

Finding myself at the crossroads of this dilemma, I have opted for the hard-line approach to the related topics of regression and correlation. Models I and II regression methods are presented in a format following Sokal and Rohlf (1969), among others. Both extreme positions are represented in the recent literature of anthropology. I personally remain undecided about the efficacy of totally ignoring the mathematical strictures for fixed effects on $X$; and, in addition, once the more rigorous methods of regression have been mastered, then one is in excellent position to select a *hard or soft* posture at will, depending upon the particular applications at hand. That is, once the two models of regression are understood, then one is free to select methods from a position of strength and knowledge rather than by mere dogma. The *soft* position—at least at an introductory level—lacks this flexibility.

**Example 13.4**

While analyzing the archaeological findings at Fort Michilmackinac, Lewis Binford made the interesting observation that kaolin pipestems can be used to date historic archaeological sites. It seems that during the seventeenth and eighteenth centuries, the average diameter of these pipestems decreased in a remarkably consistent fashion. The relationship between site age and stem-bore diameter has been assembled by Heighton and Deagan (1971: fig. 1) for the 12 colonial archaeological sites tabulated below.

| Site | Age, A.D. | Pipestem Diameter; 1/64 in. |
|---|---|---|
| Williamsburg (Coke Garret I) | 1757 | 4.62 |
| Clay Bank | 1695 | 6.11 |
| Tutter's Neck, Va. (Pit A) | 1706 | 5.82 |
| Silver Bluff, S.C. | 1748 | 4.91 |
| Ft. Frederica, Ga. | 1743 | 4.91 |
| Archer Cottage, Yorktown | 1769 | 4.31 |
| Ft. Michilmackinak | 1768 | 4.55 |
| Ft. Michilmackinak Barracks | 1775 | 4.07 |
| Warrasqueoi | 1688 | 6.50 |
| Brunswick Town | 1751 | 4.88 |
| Ft. Necessity | 1754 | 4.4 |
| Spaulding's Lower Store | 1770 | 4.63 |

Find the line of regression which allows age predictions of an archaeological site from its mean stem-bore diameter.

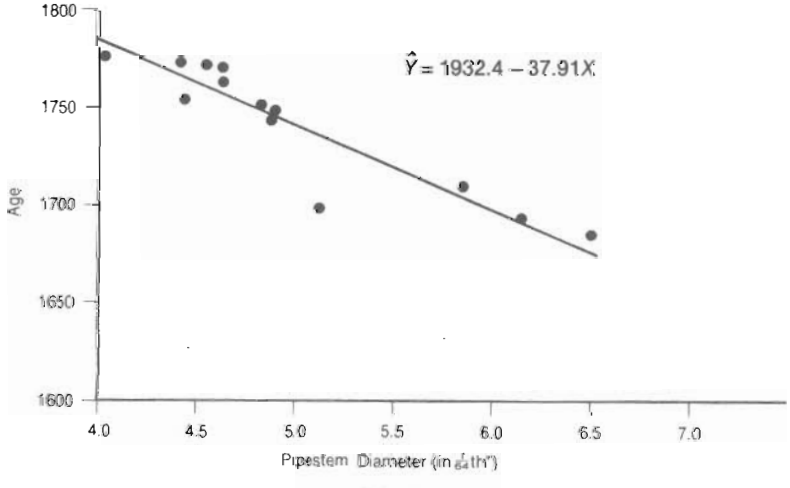


Fig. 13.16

The least squares approximation does not apply to this case because neither variable is "fixed" in a statistical sense. Specifically, we wish to predict the age of archaeological sites ($Y$) from knowledge of the pipestem diameters ($X$), but both $X$ and $Y$ are random variables. The data must first be plotted to determine whether a linear description makes sense. Since these variables appear to be roughly linear in fashion, Bartlett's Model II regression can be used to estimate the best fit.

| $X$ | $Y$ | $X^2$ | $Y^2$ | $XY$ |
|---|---|---|---|---|
| 4.07 | 1,775 | 16.56 | 3,150,625 | 7,224.25 |
| 4.31 | 1,769 | 18.58 | 3,129,361 | 7,624.39 |
| 4.4 | 1,754 | 19.36 | 3,076,516 | 7,717.60 |
| 4.55 | 1,768 | 20.70 | 3,125,824 | 8,044.40 |
| 4.62 | 1,757 | 21.34 | 3,087,049 | 8,117.34 |
| 4.62 | 1,770 | 21.44 | 3,132,900 | 8,134.91 |
| 4.88 | 1,751 | 23.81 | 3,066,001 | 8,544.88 |
| 4.91 | 1,743 | 24.11 | 3,038,049 | 8,558.13 |
| 4.91 | 1,748 | 24.11 | 3,055,504 | 8,582.68 |
| 5.82 | 1,706 | 33.87 | 2,910,436 | 9,928.92 |
| 6.11 | 1,695 | 37.33 | 2,873,025 | 10,356.45 |
| 6.50 | 1,688 | 42.25 | 2,849,344 | 10,972.00 |

The grand mean and group means must be computed:

$$\bar{X} = \frac{59.71}{12} = 4.98 \qquad \bar{Y} = \frac{20,924}{12} = 1743.67$$

$$\bar{X}_1 = \frac{17.33}{4} = 4.33 \qquad \bar{X}_3 = \frac{23.34}{4} = 5.84$$

$$\bar{Y}_1 = \frac{7066}{4} = 1766.50 \qquad \bar{Y}_3 = \frac{6837}{4} = 1709.25$$

The Model II regression equation constants are thus

$$b' = \frac{1709.25 - 1766.50}{5.84 - 4.33} = -37.91$$

$$a' = 1743.67 - (-37.91)(4.98) = 1932.4$$

$$\hat{Y} = 1932.4 - 37.91X$$

The following sample values indicate how this regression equation can be used by the historic archaeologist.

| When the mean bore diameter is · · · | The estimated age of the site is |
|---|---|
| 4.25 | A.D. 1771 |
| 5.00 | A.D. 1743 |
| 6.00 | A.D. 1705 |

## SUGGESTIONS FOR FURTHER READING

Bliss (1967: chapter 13)
Simpson, Roe, and Lewontin (1960: chapter 11)
Sokal and Rohlf (1969: chapter 14)

## EXERCISES

13.1 Given the following data:

| $X$ | $Y$ |
|---|---|
| −2 | 0 |
| 1 | 3 |
| 3 | 7 |
| 4 | 10 |
| 7 | 16 |

(a) Find the equation of the least squares line describing these five points.
(b) Draw the scattergram and fit the regression line.
(c) What is $Y$ when $X = 2$?
(d) Find the population standard error of estimate.

13.2 Given the following data:

| $X$ | $Y$ |
|---|---|
| 2 | 18 |
| 4 | 15 |
| 5 | 12 |
| 9 | 7 |
| 10 | 2 |

(a) Find the least squares equation describing these data.
(b) Draw the scattergram and regression line.
(c) Find the population standard error of estimate.

*13.3 Given the following variates:

| $X$ | $Y$ |
|---|---|
| 10 | 12 |
| 17 | 26 |
| 20 | 42 |
| 16 | 22 |
| 18 | 26 |
| 23 | 45 |
| 29 | 50 |
| 8 | 6 |
| 13 | 20 |

(a) Find the least squares line describing these points.

(b) Use Model II techniques to determine the regression equation.
(c) Plot both equations on a scattergram.
(d) What assumptions are necessary for each method?

13.4 By studying radiographs of the fetus, McKim, Hutchinson, and Gavan (1972) derived the following regression equation projecting prenatal age from the length of the femur in the unborn rhesus monkey:

$$\hat{Y} = 50 - 0.35X$$

where $\hat{Y}$ is "days prior to birth" and $X$ is "femoral length in millimeters."
(a) Graph this equation.
(b) When the femur is 20 mm in length, about how many days prior to birth is the fetus?
(c) How many days prior to birth is a fetus with a 40 mm femur?

*13.5 Return to the data from the Grasshopper Ruin (Exercise 10.7).
(a) Find the equation which allows prediction of hearth size from a knowledge of room size for the later rooms.
(b) If another room from the early rooms is found to be 20 $m^2$, how large would you predict its firehearth to be?
(c) What measure can we use to determine the accuracy of this prediction?