Figure 13.6. SIMPLE LINEAR REGRESSION: Correlation

Program 9 in: *Against All Odds: Inside Statistics*

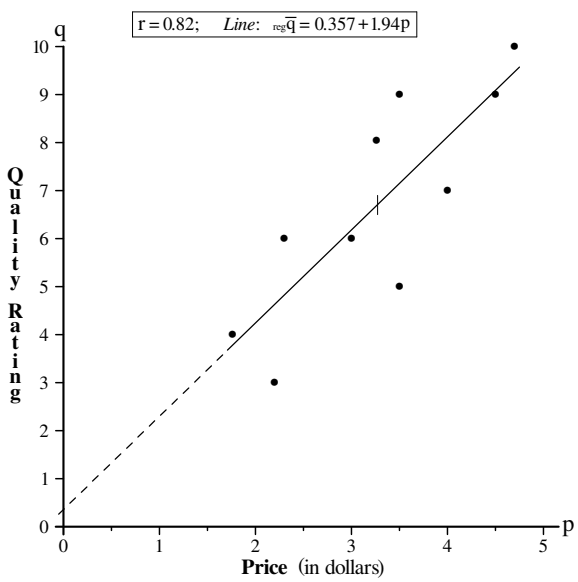
This program contains two main segments. The first introduces the *correlation coefficient* r, an important measure of the strength and direction of the straight-line association between two *quantitative* variates. The video illustrates the use of correlation in describing the results of a study of identical (or monozygotic) twins who have been raised apart. These twins have a strong correlation between physical characteristics and, more surprisingly, a moderately strong correlation between mental and behavioural characteristics. Correlation always satisfies the inequality $-1 \le r \le 1$, and $r = \pm 1$ only in the case of *perfect* linear association. The value of r is *not* affected by changes in the unit of measurement of either variate.

Linear regression signifies a straight-line relationship between an explanatory variate and a response variate. Correlation applies to *any* two quantitative variates; it does *not* require that one be an explanatory variate. When we have an explanatory variate and a reponse variable, however, correlation and regression are closely related. The second part of this Program describes the relationship between correlation and regression. The *squared correlation coefficient* r^2 (also called the *coefficient of determination*) is the fraction of the variation in one variate accounted for (or 'explained' by) least squares regression on the other variate. The estimated regression of **Y** on **X** is the straight line with slope $b_1 = r s_y / s_x$ and passing through the point $(\bar{x}, \bar{y})$, the *point of averages.* In the video, the Coleman Report uses r^2 to describe how characteristics of schools are surprisingly poor predictors of student achievement.

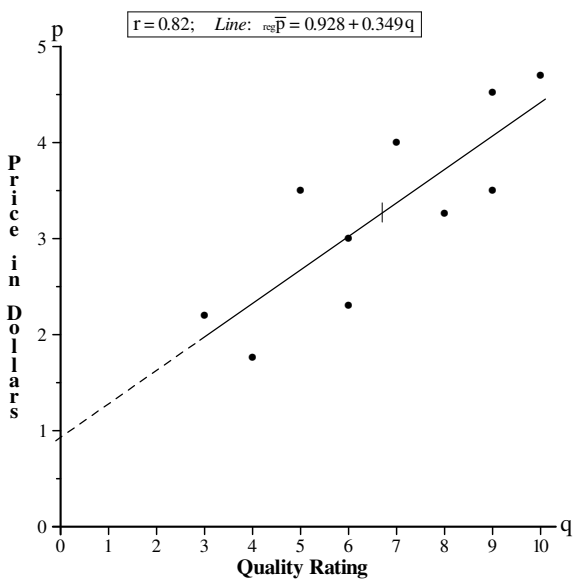
The results of computing a correlation or regression must be interpreted with due attention to the following: the possible effects of lurking variates, the lack of resistance of these procedures, the danger of extrapolation, the fact that correlations based on averages (*e.g.*, ecological correlations) are usually *higher* than those for the corresponding individuals, and an understanding that correlation and linear regression measure only *straight-line* relationships to the possible exclusion of *other* important aspects of the data.

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

Quality and Price of Chocolate Cakes



Price and Quality of Chocolate Cakes



Values quoted in the video for some other correlations are:

∗ Between the heights of identical twins raised *apart*	0.92
∗ Between personality inventories of identical twins raised *apart*	0.49
∗ Between personality inventories of identical twins raised *together*	0.52
∗ Between home run output and strikeouts (in baseball)	0.7

1 From the perspective of a data-based investigating, outline the *weakness(es)* of the data collection aspects, as they are portrayed in the video, of Dr. Amabile's investigation of quality and price of chocolate cakes; give your discussion in point form.

- Indicate briefly how you would *overcome* the weakness(es) you identify.

(*continued overleaf*)

2 What is the implication of the *dashed* parts of the estimated regression lines in the two diagrams overleaf on page 13.37? Explain briefly.

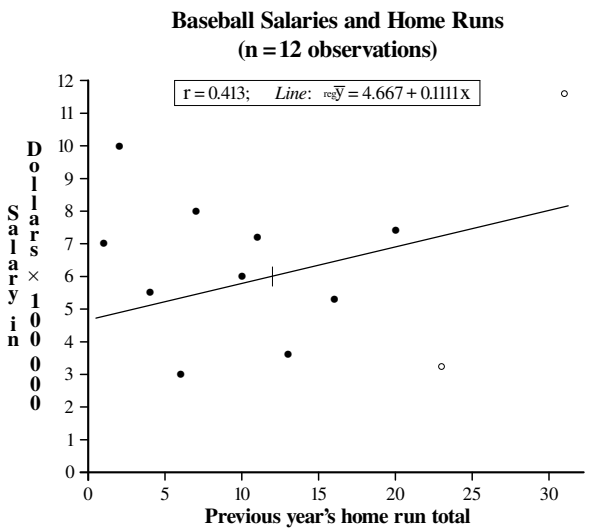
The following data (indexed by $j = 1, 2, \ldots, 12$) are similar to those shown as a scatter diagram in Program 9 for baseball players' salaries *vs* their previous year's home run total:

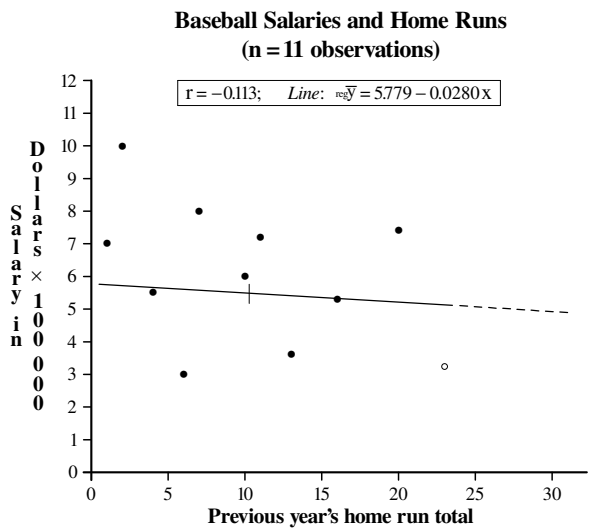
Player number (j)	1	2	3	4	5	6	7	8	9	10	11	12
Home run total (x_j)	1	2	4	6	7	10	11	13	16	20	23	31
Salary in $\$\times 10^5$ (y_j)	7.0	4.1	5.5	3.0	8.0	6.0	7.2	3.6	5.3	7.4	3.3	11.6;

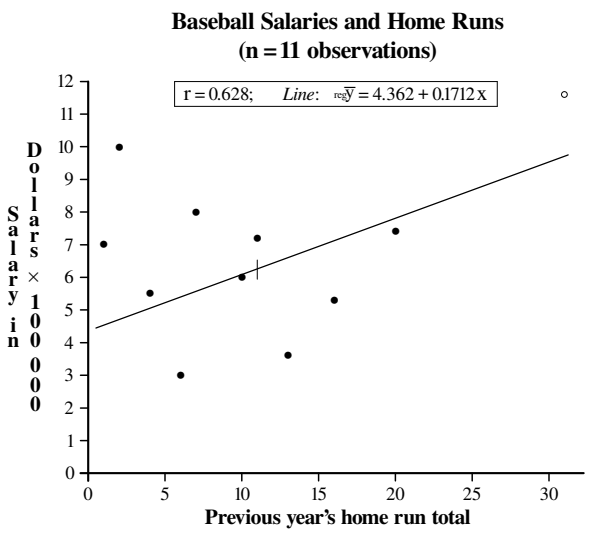
for these $n = 12$ observations:

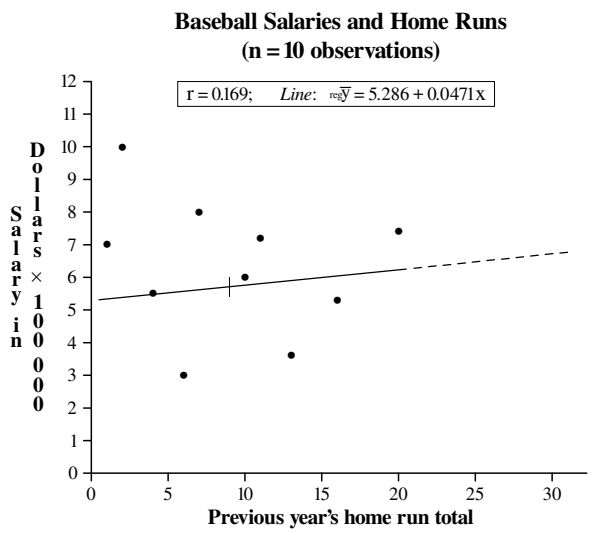
$$\sum_{j=1}^{n} x_j = 144, \qquad \sum_{j=1}^{n} x_j^2 = 2{,}642; \qquad \sum_{j=1}^{n} y_j = 72, \qquad \sum_{j=1}^{n} y_j^2 = 498.16; \qquad \sum_{j=1}^{n} x_j y_j = 965.5.$$

The scatter diagram, correlation, and estimated least squares regression line, based on these data, are shown below; all 12 observations are included for the first diagram, and then either or both of the (anomalous) *12*th and *11*th observations are *ex*cluded for the other three plots. The short vertical line crossing each estimated regression line indicates the point of averages $(\bar{x}, \bar{y})$.









3 During the interview shown in the video, Daniel Seligman of *Fortune Magazine* says: *The players .... are going up there trying to hit that long ball and make big bucks.* Comment briefly on this statement in light of the information presented above.

4 What matter(s) of general relevance to data analysis are illustrated by the information presented above? Explain briefly for each matter you describe.