Teaching Simple Linear Regression Using the Casio fx-7000G Calculator

Michael B. Fiske
The Ohio State University
Department of Educational Theory and Practice

In traditional elementary statistics courses, the teaching of simple linear regression by the method of least squares is complicated by inadequate student facility with arithmetic and their lack of understanding of slope. Textbooks used in these courses either devote minimal comments to linear regression (Moore, 1979) or substantial development to the fundamental concepts of slope and intercept (Freedman, Pisani, & Purves, 1978; Mosteller, Fienberg, & Rourke, 1983). The ready availability of calculators with linear regression modes and graphic displays both simplifies the calculations and enhances the opportunity for student understanding through wide exploration of linear regression models. This paper describes the linear regression functions available on the Casio fx-7000G calculator and discusses three teaching modules for exploring data sets in an elementary statistics course.

The Casio fx-7000G calculator uses the technique of least squares linear regression with the model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. The estimators of the regression coefficients, b_0 (constant term) and b_1 (slope), are calculated by the formula:

$$b_0 = \frac{\Sigma y_i - b_1 \Sigma x_i}{n}$$

and

$$b_1 = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma (x_i)^2 - (\Sigma x_i)^2} .$$

Once the calculator is in the linear regression graphics mode and a satisfactory range for the data is set, data are entered in the form: independent variable, dependent variable (x_i, y_i). As each data item is entered, it is displayed on the graphics screen. The user can check that the correct number of items (n) have been entered and correct any incorrectly entered data. A drawback is that corrected points remain in uncorrected form on the graphics display. After entering the data, the user has the option of immediately graphing the estimated regression line or of obtaining the estimators of the regression coefficients by keying the appropriate function keys. If desired, the following values are obtainable on single keys: the averages $\bar{x}$ and $\bar{y}$; the sums Σx_i, Σy_i; the sums of the squares $\Sigma (x_i)^2$, $\Sigma (y_i)^2$; the

sum of the cross product terms $\sum x_i y_i$; the biased and unbiased standard deviations $s_{x,n}$, $s_{x,n-1}$, $s_{y,n}$, $s_{y,n-1}$; and the coefficient of correlation, r.

The first teaching module uses data collected by Doll (Freedman et al., 1978) on the per capita consumption of cigarettes in 11 countries in 1930 and on the male death rate from lung cancer in the same countries in 1950. The module introduces students to the method of data entry and correction on the calculator and assists them in choosing and then entering an appropriate viewing range for the data set (0 to 1500 cigarettes per capita and 0 to 500 deaths per million men). Before proceeding to a regression model, students are asked to examine the scatterplot for any relationship they see between cigarette consumption and death rate from lung cancer. In addition, they are asked to discuss why different time frames were used for the data and why only men were studied. A simple keying sequence produces the graph of the estimated regression line (Figure 1). The horizontal axis is per capita cigarette consumption, and the vertical axis is death rate per million. Students are asked to examine the graph of the line to see how well it fits the data, identifying any points that lie far above or below the line. The estimated intercept (b_0) and slope (b_1) are obtained using function keys, resulting in the following estimated regression equation:

Deaths per million men from lung cancer = 67.56 + 0.2284 x (per capita cigarette consumption).

Predicted death rates based on the regression equation are introduced through the use of the $\hat{y}$ function. In order to see the effect of an outlying point on the regression line, the point for the United States (1300, 200) is removed and the new regression line plotted (Figure 2). Students can compare the suitability of each regression equation and be led to discuss why the United States had a substantially lower death rate.

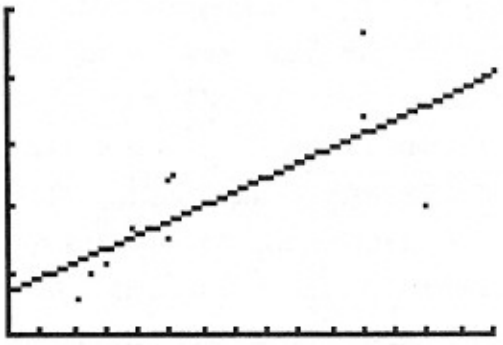

Figure 1

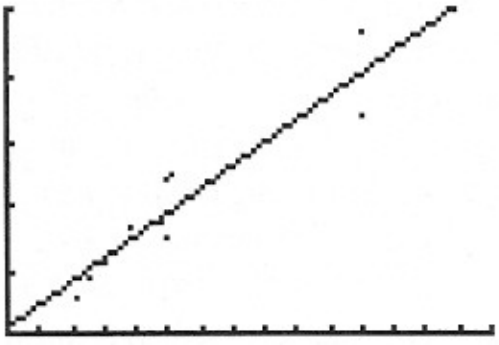

Figure 2

The second module uses data from the Olympic 100 m and 800 m running events to demonstrate how a logarithmic transformation of the dependent variable (running time) allows a comparison of the percent of decrease in running time between the two events (Cleveland, 1985). After entering the data in the form (Olympiad - 1, logarithm of running time), students use the estimated regression equations to predict future winning times. The graphs of the two regression lines (Figures 3 and 4) show that the percent of decrease in running times is almost the same for each event; that is, the slopes of the lines are nearly the same. The horizontal axes are the years (1896 to 1984, scale 12 years), and the vertical axes are the logarithms of the time (100 m run: 2.24 to 2.46; scale .1; 800 m run: 4.6 to 4.85, scale .05).

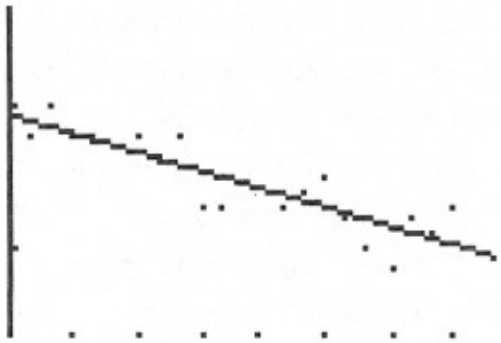

Figure 3 (100 m)

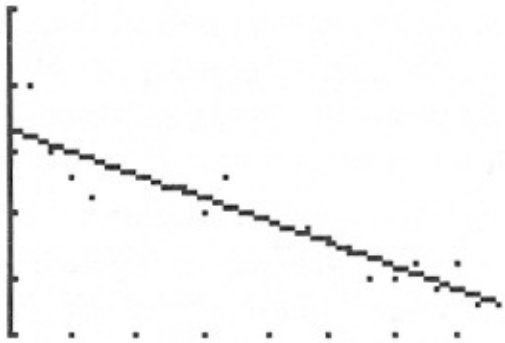

Figure 4 (800 m)

In the third module, the data of interest are infant mortality rates (deaths for children under 1 year of age per 1000 live births) in the state of Massachusetts during the years 1850 to 1970 (Travers, Stout, Swift, & Sextro, 1985). The initial regression line does not fit the data well. Thus, it is suggested that students break the data into two parts, the first 3 and the last 10 decades, fitting two regression lines (Figures 5 and 6). On the horizontal axes are the years, scaled by 10, and on the vertical axes is the infant mortality rate, scaled by 25 and ranging from 0 to 180 deaths per 1000 live births. Students are asked to explain why a downward trend began to occur in 1880. Furthermore, they are asked to explain the predictive value of the regression equations outside the range of the data. Additional explorations involve comparing infant mortality rates for social classes and with other industrialized countries.

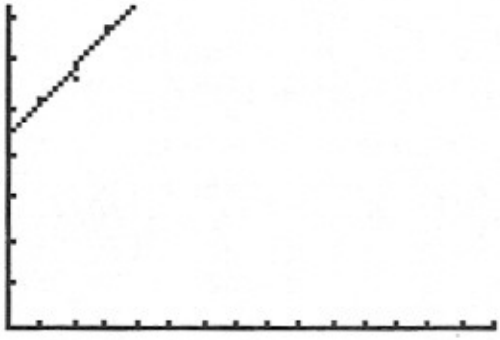

Figure 5

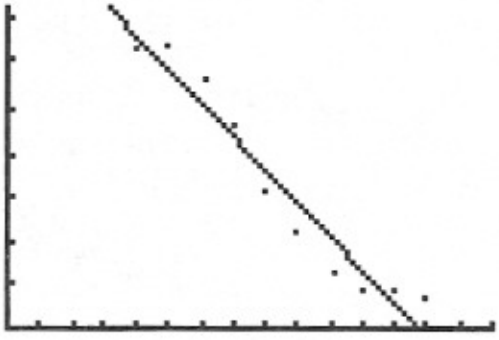

Figure 6

These three modules present the opportunity for expanded exploration within a traditional elementary statistics setting, including transformations of data and piecewise regression. Students are freed from extensive arithmetic calculation and graphing. They have the opportunity to ask and answer interpretive and evaluative questions regarding the data and the estimated regression line. The ease with which the calculator is used opens a new window on the world of statistics for the beginning student.

References

Biehler, R. (1985). Interrelations between computers, statistics, and teaching mathematics. *The influence of computers and informatics on mathematics and its teaching: Supporting papers* (pp. 209-214). Strasbourg: Institut de Recherches sur l'Enseignement des Mathématiques.

Cleveland, W. S. (1985). *The elements of graphing data.* Monterey, CA: Wadsworth.

Freedman, D., Pisani, R., & Purves, R. (1978). *Statistics.* New York: Norton .

Gordon, F. S. (1988). Computer use in teaching statistics. In D. A. Smith, G.J. Porter, L. C. Leinbach, & R. H. Wenger (Eds.), *Computers and mathematics education: The use of computers in undergraduate instruction.* (MAA Notes No. 9, pp. 79-83). Washington, D. C.: The Mathematical Association of America.

Moore, D. S. (1979). *Statistics: Concepts and controversies.* San Francisco: Freeman.

Mosteller, F., Fienberg, S. E., & Rourke, R. E. K. (1983). *Beginning satistics with data analysis..* Reading, MA: Addison-Wesley.

Travers, K. J., Stout, W. F., Swift, J. H., & Sextro, J. (1985). *Using satistics.* Menlo Park, CA: Addison Wesley.