Virginia Journal of Science
Volume 44, Number 4
Winter 1993

A Statistical Result Derived from Mechanical Equilibrium

J. N. Boyd and **P. N. Raychowdhury**,
Department of Mathematical Sciences,
Virginia Commonwealth University,
Richmond, VA 23284-2014

ABSTRACT

The statistical formula SST = SSM + SSE is central to the ideas of linear regression. If values (x) of a random variable X and corresponding values (y) of a second random variable Y are assumed to have the linear relationship $y = b_1x + b_0$ as obtained by the least-squares fit, then the statistical formula partitions the total variation of the observed values of y from their mean (SST) into two summands which represent the variation of the predicted values of y from the mean (SSM) and the variation of the observed values from the corresponding values predicted by the regression (SSE).

INTRODUCTION

Let us suppose that we have n data points (x_i, y_i), $i = 1,2,3, ..., n$. The first coordinate x is taken to be a value of an independent random variable X and y is taken to be the corresponding value of a dependent or response random variable Y. The line ℓ which has equation

$$y = b_1x + b_0 \tag{1}$$

and minimizes the sum

$$S = \sum (y_i - b_1 x_i - b_0)^2 \tag{2}$$

is the least-squares regression line for the data. Since sums will always run from 1 to n over all data points, we omit the limits of summation in our notation.

Let $\hat{y}_i = b_1 x_i + b_0$ denote the value of the response variable predicted by the model when $x = x_i$. In general, $\hat{y}_i \neq y_i$ and we denote the difference by $r_i = y_i - \hat{y}_i$ for each i. This difference is called the i-th residual for the model and sum 2 can be rewritten as

$$S = \sum r_i^2 \ . \tag{3}$$

For a further bit of notation, we let $\bar{y} = \sum y_i / n$ denote the mean of the observed values of y.

THE STATISTICAL FORMULA

Let $SST = \sum (y_i - \bar{y})^2$, $SSM = \sum (\hat{y}_i - \bar{y})^2$, and $SSE = \sum (y_i - \hat{y}_i)^2$. The symbols SST, SSM, and SSE stand for "total sum of squares," "model sum of squares," and "error sum of squares." The formula which we wish to derive is a significant one:

$$SST = SSM + SSE. \tag{4}$$

Its derivation is not generally given in introductory statistics texts or in texts devoted to the methods of statistics.

AN EQUIVALENT PROBLEM

Let us begin our derivation by noting that $y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$ and that $(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$. After summing both sides of the last equation from 1 to n, it becomes clear that our problem is equivalent to showing that

$$T = \sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \tag{5}$$

is zero.

Letting $y_i = \hat{y}_i + r_i$ in equation 5, we can rewrite that equation as

$$T = \sum (\hat{y}_i - \bar{y}) r_i \ . \tag{6}$$

The computations to develop the regression equation $y = b_1x + b_0$ make it clear that $(\bar{x}, \bar{y})$ satisfy the equation where $\bar{x}$ and $\bar{y}$ are the means of the observed x_i and y_i, respectively. Thus equation 6 becomes

$$T = b_1 \sum (x_i - \bar{x}) r_i \ . \tag{7}$$

THE LINE IN MECHANICAL EQUILIBRIUM

Let us imagine that regression line ℓ is pivoted at $(\bar{x}, \bar{y})$ and is free to rotate about the axis perpendicular to the xy-plane at $(\bar{x}, \bar{y})$. The line together with several data points is shown in Figure 1.

If $b_1 = 0$, the derivation of equation 4 is trivial. Therefore, we assume that $b_1 \neq 0$. Then equation 4 is true if and only if $\sum (x_i - \bar{x}) r_i = 0$.

Let us imagine a configuration of n springs above and below line ℓ. The springs may be stretched parallel to the y-axis and the spring constant for each is $k > 0$. Imagine that the i-th spring has one end fixed on the line $x = x_i$ and that the other end is free but constrained to stay on the line $x = x_i$. When the spring is unstretched, this free end is located at (x_i, y_i). Stretch the spring a distance $|r_i|$ and attach the free end to line ℓ. The situation is illustrated in Figure 2. We imagine that at the point of attachment there is a frictionless ring which can slide on ℓ so that the springs remain parallel to the y-axis if the line rotates through a small angle about $(\bar{x}, \bar{y})$.

The elastic potential energy for the configuration of springs is $\sum \frac{k}{2} r_i^2$. Since the line is the least-squares regression line, $\sum r_i^2$ and the elastic potential energy must both be minima, and the line is in stable mechanical equilibrium.

The conditions of equilibrium are that the algebraic sum of forces in the y-direction is zero and that torques about the axis at $(\bar{x}, \bar{y})$ must sum to zero.

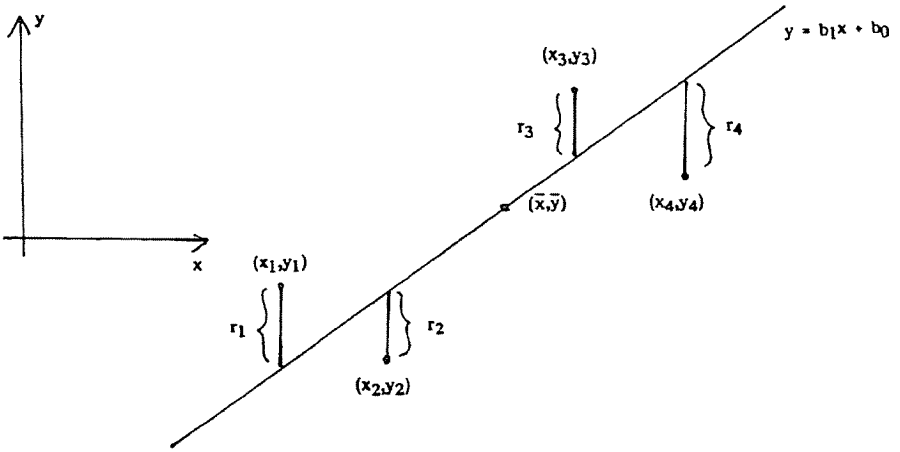

FIGURE 1. The Geometry of Line ℓ.

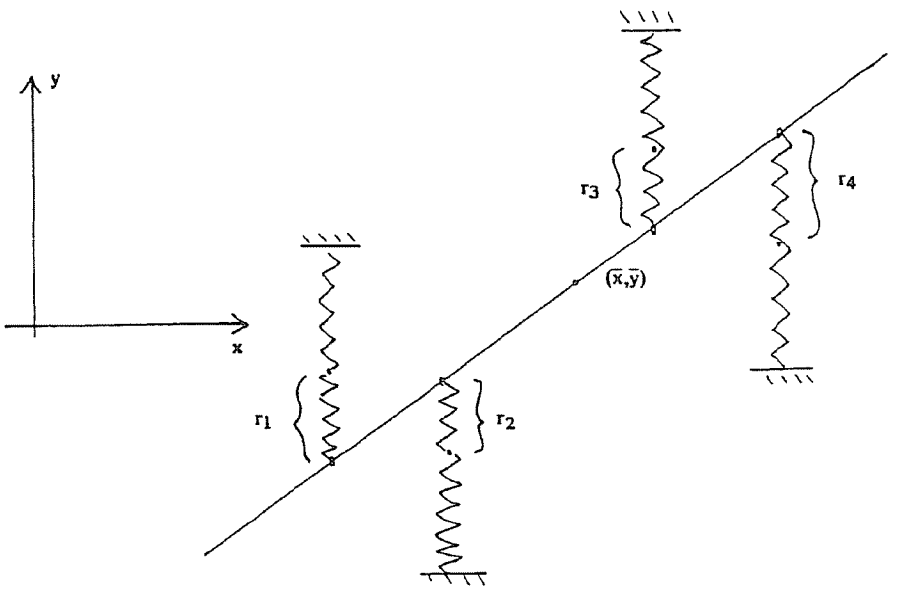

FIGURE 2. Line ℓ in Mechanical Equilibrium.

The first condition tells us that $\sum k\, r_i = 0$ since the force exerted by the i-th spring is kr_i. Thus $\sum r_i = 0$. This result that the residuals must sum to zero is a bonus for us.

The second condition gives us the result we set out to obtain. Since the i-th torque is computed by force times lever arm or $k\, r_i\, (x_i - \bar{x})$, we have $\sum k\, (x_i - \bar{x})\, r_i = 0$. Division by k yields $\sum (x_i - \bar{x})\, r_i = 0$.

Thus SST = SSM + SSE.

CONCLUSION

The physical picture of the least-square regression line is a pleasing one. The line is under tension and is in equilibrium. After small displacements, the line tends to return to its least-square configuration. The model is stable. The statistical meaning of the formula derived is well known (Moore and McCabe, 1993).

LITERATURE CITED

Moore, D. S. and McCabe, G. P. 1993. Introduction to the Practice of Statistics. NewYork: W.H. Freeman and Company, p. 656.