THE RATIO OF AVERAGES AND THE AVERAGE OF RATIOS AS "BEST" ESTIMATES IN LINEAR REGRESSION*

By D. S. Robson BU-24-M November, 1951

It is postulated that the random variable Y is linearly related to the nonrandom variable X. If, for example, the nonrandom variable X were age of a pine seedling, measured in days after planting of the seed, and the random variable Y were height of a seedling, measured in millimeters, the postulate would read: the average height of pine seedlings, measured in mm., is proportional to age, measured in days after planting - or this may be reworded to say: on the average, the height of pine seedlings is directly proportional to age measured in days from planting. Although the postulate may in reality be false when applied to this particular example, i.e., average height may not actually show a constant increase with age, it is certainly feasible to say that the average height of seedling on the day of planting of the seeds (X=0) is zero, and for the purposes of this discussion we shall suppose that the average height of seedlings does show a constant increase from day zero onward. Thus, if the average height of seedlings X days after planting is βX mm. then on the (X+1)'th day the average height will have increased β mm., from βX to $\beta(X+1)$. Notice, in particular, that our postulate does not state that <u>all</u> seedlings are of height βX mm. on the X'th day after planting; our postulate involves the considerably more reasonable statement that the <u>average</u> height of plants on day X is βX mm. If all plants were of height βX on day X and all plants were of height $\beta(X+1)$ on day X+1 no problem would arise in estimating the constant β; one would simply observe the height on those two days and note the difference. That difference would be β mm. In practice, of course, one

---

* See Sections 6.2 and 6.3 in Snedecor's "Statistical Methods."

finds natural variation among seedlings, and even if we happened to observe one plant of height βX mm. on day X it does not follow from our postulate that this particular plant must attain a height of exactly $\beta(X+1)$ mm. on day $(X+1)$.

Let us now formulate this argument mathematically. We have said that the average height of seedlings on day X is directly proportional to X, and we have called the unknown proportionality factor β. We have also observed that the height Y of any particular seedling on day X need not equal the exact mean $(=\beta X)$ but may deviate from βX by a quantity which we shall now call ε (epsilon), i.e.,

$$\varepsilon = Y - \beta X \tag{1}$$

The expression (1) is generally written in the form

$$Y = \beta X + \varepsilon \tag{2}$$

and in this form is called a "linear model." Since the average value of Y (= height of seedling) on day X is βX it must follow that the average value of ε $(=Y-\beta X)$ on day X is zero. Another property of the deviate ε follows by noting that if the variance of Y (=height of seedling) about the mean height $(=\beta X)$ on day X is defined to be the average value of the squared deviation from the mean then [variance of Y on day X] = [average value of $(Y-\beta X)^2$] = [average value of ε^2].

Suppose now that we have n observations on the random variable Y (= height of seedlings); we shall denote the first observation by Y_1, the second by Y_2, etc. - or, in general, we denote the i'th observation by Y_i. Similarly, we say that the first observation was made on day X_1, the second on day X_2, etc. - or, in general, the i'th observation was made on day X_i. The postulate, expressed by (2), then says Y_1 is an observation from a population which has a

mean of βX_1 and that Y_1 deviates from this mean by a quantity ε_1. The variance about the mean βX of this population from which Y_1 was drawn is the average value of the squared deviations from the mean, i.e., the average value of ε_1^2.* This variance will be denoted by σ_1^2 ; similarly, the average value of ε_2^2 will be denoted by σ_2^2, and, in general, the average value of ε_i^2 will be called σ_i^2.

The general problem of estimating the unknown quantity β, given a set of n observations of the form

$$Y_i = \beta X_i + \varepsilon_i \quad , \qquad i=1, \ldots, n \tag{3}$$

has been solved for the case where $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$ and the n observations are independent. From the data we calculate an estimate b of β which has the property that the sum of the squared deviations of the Y_i's from their estimated means, the bX_i's, is a minimum, i.e., b is the number such that

$$\sum_{i=1}^{n} (Y_i - bX_i)^2 = \text{a minimum} . \tag{4}$$

The number b which satisfies (4) is called the "least squares estimate" of β. It is an unbiased estimate of β and is, moreover, the best unbiased estimate of β in the sense that the variance of the least squares estimate is smaller than the variance of any other unbiased estimate. The quantity b is determined as follows:

b is the number which minimizes

$$\sum_{i=1}^{n} (Y_i - bX_i)^2$$

$$= \sum_{i=1}^{n} (Y_i^2 - 2bX_iY_i + b^2X_i^2)$$

$$= \sum_{i=1}^{n} Y_i^2 - 2b \sum_{i=1}^{n} X_iY_i + b^2 \sum_{i=1}^{n} X_i^2$$

---

* On day X_1 we observed the height Y_1 of a particular plant, and corresponding to this particular Y_1 there is a particular ε_1. By "average value of ε_1^2" we mean the average of all ε_1^2 for all plants on day X_1. Symbolically, the "average value of ε_1^2" is written $E(\varepsilon_1^2)$.

now multiply by $\dfrac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2} = 1$ to get

$$\frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2} \sum_{i=1}^{n} Y_i^2 - 2b \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n} X_i^2} \sum_{i=1}^{n} X_i Y_i + b^2 \frac{\Sigma X_i^2}{\Sigma X_i^2} \sum_{i=1}^{n} X_i^2$$

now add $0 = \dfrac{(\sum_{i=1}^{n} X_i Y_i)^2}{\sum_{i=1}^{n} X_i^2} - \dfrac{(\sum_{i=1}^{n} X_i Y_i)^2}{\sum_{i=1}^{n} X_i^2}$ to get

$$= \frac{1}{\sum_{i=1}^{n} X_i^2} \left\{ \sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} X_i Y_i)^2 + [(\sum_{i=1}^{n} X_i Y_i)^2 - 2b \sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} X_i Y_i + b^2 (\sum_{i=1}^{n} X_i^2)^2] \right\}$$

$$= \frac{1}{\sum_{i=1}^{n} X_i^2} \left\{ \Sigma X_i^2 \Sigma Y_i^2 - (\Sigma X_i Y_i)^2 + [(\Sigma X_i Y_i) - b \Sigma X_i^2]^2 \right\}$$

and now, since

$$\sum_{i=1}^{n} (Y_i - bX_i)^2 = \frac{1}{\sum_{i=1}^{n} X_i^2} \left\{ \sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} X_i Y_i)^2 + [(\Sigma X_i Y_i) - b \sum_{i=1}^{n} X_i^2]^2 \right\}$$

the problem of determining the number b which minimizes

$$\sum_{i=1}^{n} (Y_i - bX_i)^2$$

has been resolved into the problem of finding the b which minimizes

$$\frac{1}{\sum_{i=1}^{n} X_i^2} \left\{ \left(\sum_{i=1}^{n} X_i^2 \right) \left(\sum_{i=1}^{n} Y_i^2 \right) - \left(\sum_{i=1}^{n} X_i Y_i \right)^2 + \left[\left(\sum_{i=1}^{n} X_i Y_i \right) - b \sum_{i=1}^{n} X_i^2 \right]^2 \right\} .$$

We note that the value of the number b effects only the squared term

$$\left[\left(\sum_{i=1}^{n} X_i Y_i \right) - b \sum_{i=1}^{n} X_i^2 \right]^2 .$$

This term, being a square, is always positive, and so the smallest value that it can attain is zero. Hence, if

$$\left[(\Sigma X_i Y_i) - b \, \Sigma X_i^2 \right] = 0$$

then

$$\sum_{i=1}^{n} (Y_i - bX_i)^2 = \text{a minimum} = \frac{1}{\sum_{i=1}^{n} X_i^2} \left\{ \left(\sum_{i=1}^{n} X_i^2 \right) \left(\sum_{i=1}^{n} Y_i^2 \right) - \left(\sum_{i=1}^{n} X_i Y_i \right)^2 + 0 \right\}$$

Hence, the number b which minimizes $\sum_{i=1}^{n} (Y_i - b X_i)^2$ is determined by solving the equation

$$\left(\sum_{i=1}^{n} X_i Y_i \right) - b \sum_{i=1}^{n} X_i^2 = 0$$

$$\text{or} \quad \left(\sum_{i=1}^{n} X_i Y_i \right) = b \sum_{i=1}^{n} X_i^2$$

$$\text{or} \quad b = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} . \tag{5}$$

Now let us consider the problem of estimating β when the condition that $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$ is <u>not</u> fulfilled, i.e., when the variance σ_1^2 about the average height $(=\beta X_1)$ on day X_1 is different from the variance σ_2^2 about the

average height ($=\beta X_2$) on day X_2, and so on. Suppose first that the variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ are known. If, then, we divide each observation by its known variance we obtain a set of transformed observations

$$\frac{Y_1}{\sigma_1} = \frac{X_1}{\sigma_1}\beta + \frac{\varepsilon_1}{\sigma_1}$$

$$\frac{Y_2}{\sigma_2} = \frac{X_2}{\sigma_2}\beta + \frac{\varepsilon_2}{\sigma_2}$$

$$\vdots$$

$$\frac{Y_n}{\sigma_n} = \frac{X_n}{\sigma_n}\beta + \frac{\varepsilon_n}{\sigma_n}$$

which have the property that

1) the variance of $\frac{Y_1}{\sigma_1}$ about its mean of $\beta \frac{X_1}{\sigma_1}$ is the average value of the squared deviation $\left(\frac{Y_1}{\sigma_1} - \beta \frac{X_1}{\sigma_1} \right)^2$, or the average value of $\frac{\varepsilon_1^2}{\sigma_1^2}$. The average value of ε_1^2 is, however, σ_1^2, hence the average value of $\frac{\varepsilon_1^2}{\sigma_1^2}$ is 1.

2) similarly, the variance of $\frac{Y_2}{\sigma_2}$ about its mean of $\beta \frac{X_1}{\sigma_1}$ is the average value of $\frac{\varepsilon_2^2}{\sigma_2^2}$, which, again, is $\frac{\sigma_2^2}{\sigma_2^2} = 1$.

$\vdots$

n) similarly, the variance of $\frac{Y_n}{\sigma_n}$ about its mean of $\beta \frac{X_n}{\sigma_n}$ is 1.

Thus, the transformed variables have a constant variance and hence lend themselves to the least squares procedure described above; in other words, the estimate of β

(6) $$b = \frac{\sum_{i=1}^{n} \frac{X_i}{\sigma_i}\frac{Y_i}{\sigma_i}}{\sum_{i=1}^{n} \left(\frac{X_i}{\sigma_i}\right)^2}$$

is the minimum variance unbiased estimate of β.

In practice, of course, one never knows the variances $\sigma_1^2, \ldots, \sigma_n^2$, but suppose, instead, that one knows simply that the variance σ_i^2, whatever it may be, is directly proportional to X_i, that is, $\sigma_i^2 = kX_i$ where k is an unknown constant. Then if we substitute $\sqrt{kX_i}$ for σ_i in (6) we obtain

(7) $$b = \frac{\sum_{i=1}^{n} \frac{X_i}{\sqrt{kX_i}}\frac{Y_i}{\sqrt{kX_i}}}{\sum_{i=1}^{n} \left(\frac{X_i}{\sqrt{kX_i}}\right)^2} = \frac{\sum_{i=1}^{n} \frac{Y_i}{k}}{\sum_{i=1}^{n} \frac{X_i}{k}} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\bar{Y}}{\bar{X}}$$

which proves that if σ_i^2 is directly proportional to X_i then $b = \frac{\bar{Y}}{\bar{X}}$ is the best unbiased estimate of β. Likewise, if it were known that the standard errors σ_i were proportional to X_i, that is, $\sigma_i = cX_i$ where c is an unknown constant, and if substitute cX_i for σ_i in (6) we obtain

(8) $$b = \frac{\sum_{i=1}^{n} \frac{X_i}{cX_i}\frac{Y_i}{cX_i}}{\sum_{i=1}^{n} \left(\frac{X_i}{cX_i}\right)^2} = \frac{\sum_{i=1}^{n} \frac{Y_i}{c^2X_i}}{\sum_{i=1}^{n} \frac{1}{c^2}} = \frac{\sum_{i=1}^{n} \frac{Y_i}{X_i}}{\sum_{i=1}^{n} (1)} = \frac{\sum_{i=1}^{n} \frac{Y_i}{X_i}}{n}$$

which proves that $b = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{X_i}$ is the minimum variance unbiased estimate of β under these conditions.