

## The "Naturalness" of *Squaring* in Linear Regression

The question most often asked when students begin their study of linear regression and curve fitting is, "why do we minimize the sum of the **squares** of the errors?". Squaring the errors seems like an artificial measure of the total error of the fit. Typically, we fumble around with answers like "we want to make all the errors positive, so positive and negative errors won't negate each other". Then we are faced with explaining why working with squares is simpler than working with absolute values, which accomplish the same task without altering the size of the errors. To understand why the sum of the squares of the errors is the "natural" measure of the fit rather than the artificial measure it appears to be, we need to think about the problem geometrically.

Given a set of  $n$  data points  $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$ , we can think of the data as defining two vectors  $\bar{x}$  and  $\bar{y}$ , with

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \bar{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}.$$

With this interpretation, we re-consider the linear equation  $\bar{y} = m\bar{x} + b$ . This equation now makes no sense, since  $m$  and  $b$  are scalars and  $\bar{x}$  and  $\bar{y}$  are  $n \times 1$  vectors. Implied by the equation is an  $n \times 1$  vector of 1's, which we will call  $\bar{1}$ . Now the equation

$$\bar{y} = m\bar{x} + b\bar{1}$$

is well defined.

If the equation  $\bar{y} = m\bar{x} + b\bar{1}$  is satisfied, then all of the data lie precisely on a line, as shown in Figure 1a. More importantly, we interpret the vector equation  $\bar{y} = m\bar{x} + b\bar{1}$  as saying that vector  $\bar{y}$  lives in the plane defined by  $\bar{x}$  and  $\bar{1}$ .

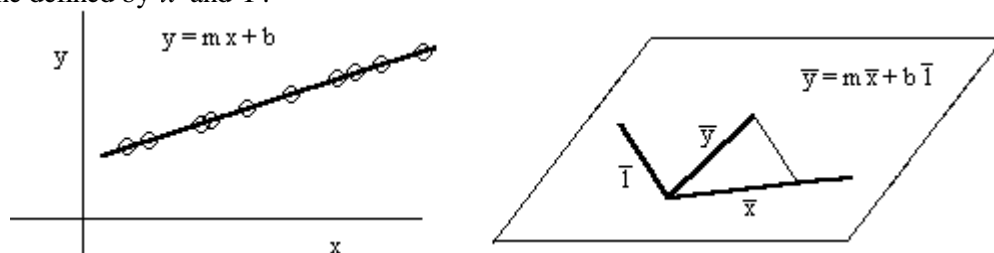


Figure 1a: Data lie exactly on the line  $y = mx + b$

Of course, in a real data situation, we cannot expect the equation  $\bar{y} = m\bar{x} + b\bar{1}$  to hold. If there is any error at all, then  $\bar{y} \neq m\bar{x} + b\bar{1}$ , that is, the vector  $\bar{y}$  is not in the plane of  $\bar{x}$  and  $\bar{1}$ . Nevertheless, we do want to write a linear model to represent the data set. That is, we want to know the  $y$ -values,  $\hat{y}$ , for which  $\hat{y} = m\bar{x} + b\bar{1}$  is the best linear model for the data.

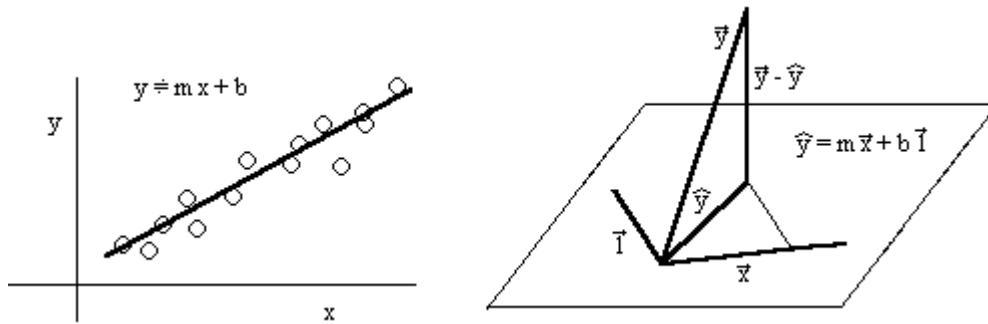


Figure 1b: Data do not lie exactly on the line  $y = mx + b$

The regression question can be stated geometrically as, "of all vectors in the plane of  $\bar{x}$  and  $\bar{1}$ , which is the closest to  $\bar{y}$ ?" We will use this vector as our model  $\hat{y}$ . Figure 2 illustrates the geometry of the regression problem.

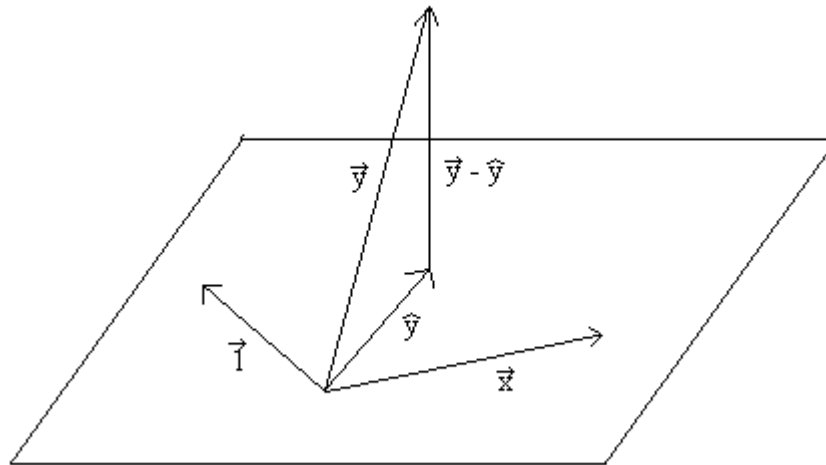


Figure 2: The Geometry of Linear Regression

The vector  $\vec{r} = \bar{y} - \hat{y} = \bar{y} - (m\bar{x} + b\bar{1})$ , known as the residual vector, is the difference in the actual vector  $\bar{y}$  and the model vector  $\hat{y}$ . The residual vector represents the error in the approximation of  $\bar{y}$  by  $\hat{y}$ . We want  $\hat{y}$  to be as close (literally) to  $\bar{y}$  as possible. To achieve the best fit, we want to minimize the length of  $\vec{r}$ . Remember, the length of a vector is found by the generalized Pythagorean Theorem,  $\|\vec{r}\| = \sqrt{r_1^2 + r_2^2 + r_3^2 + \dots + r_n^2}$ . To minimize the length of this vector, you must minimize the sum of the **squares** of its components. What is at first viewed as an artificial device to make terms positive is actually the "natural" (in the sense of Pythagorean) measure of the length of a vector!

Figure 1 contains all of the information required to compute the least squares components as well. Minimizing the length of  $\vec{r}$  requires that  $\vec{r}$  be perpendicular to the plane of  $\bar{x}$  and  $\bar{1}$ . Since  $\vec{r}$  is perpendicular to both  $\bar{x}$  and  $\bar{1}$ , we know that the dot products  $\vec{r} \cdot \bar{x}$  and  $\vec{r} \cdot \bar{1}$  are zero. If  $(\bar{y} - \hat{y}) \cdot \bar{x} = (\bar{y} - m\bar{x} - b\bar{1}) \cdot \bar{x} = 0$ , then  $(\bar{y} - \hat{y}) \cdot \bar{x} = (\bar{y} - m\bar{x} - b\bar{1}) \cdot \bar{x} = 0$ . This generates the linear equation

$$\bar{y} \cdot \bar{x} = m(\bar{x} \cdot \bar{x}) + b(\bar{1} \cdot \bar{x}).$$

Similarly, if  $\vec{r} \cdot \vec{1} = 0$ , then  $(\vec{y} - \hat{y}) \cdot \vec{1} = (\vec{y} - m\vec{x} - b\vec{1}) \cdot \vec{1} = 0$ . This gives the linear equation

$$\vec{y} \cdot \vec{1} = m(\vec{x} \cdot \vec{1}) + b(\vec{1} \cdot \vec{1}).$$

To determine the values of  $m$  and  $b$  that minimize the size of the residual vector, we only have to solve a system of two linear equations in two unknowns,

$$\begin{aligned}\vec{y} \cdot \vec{x} &= m(\vec{x} \cdot \vec{x}) + b(\vec{1} \cdot \vec{x}) \\ \vec{y} \cdot \vec{1} &= m(\vec{x} \cdot \vec{1}) + b(\vec{1} \cdot \vec{1}).\end{aligned}$$

Solving this system, we find that

$$m = \frac{(\vec{y} \cdot \vec{x})(\vec{1} \cdot \vec{1}) - (\vec{y} \cdot \vec{1})(\vec{1} \cdot \vec{x})}{(\vec{x} \cdot \vec{x})(\vec{1} \cdot \vec{1}) - (\vec{x} \cdot \vec{1})(\vec{1} \cdot \vec{x})} \quad \text{and} \quad b = \frac{(\vec{y} \cdot \vec{1})(\vec{x} \cdot \vec{x}) - (\vec{y} \cdot \vec{x})(\vec{x} \cdot \vec{1})}{(\vec{x} \cdot \vec{x})(\vec{1} \cdot \vec{1}) - (\vec{x} \cdot \vec{1})(\vec{1} \cdot \vec{x})}.$$

Evaluating the dot products generates the more conventional forms

$$m = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \quad \text{and} \quad b = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}.$$

For the data set (2, 7), (3, 10), (4, 10), (5, 14) and (6, 15) we have  $\vec{x} = \langle 2, 3, 4, 5, 6 \rangle$  and  $\vec{y} = \langle 7, 10, 10, 14, 15 \rangle$ , with

$$\sum x_i = 20, \quad \sum y_i = 56, \quad \sum x_i y_i = 244, \quad \sum x_i^2 = 90, \quad \left( \sum x_i \right)^2 = 400 \quad \text{and} \quad n = 5.$$

Substituting, we have  $m = \frac{5(244) - (20)(56)}{5(90) - 400} = 2$  and  $b = \frac{(56)(90) - (244)(20)}{5(90) - 400} = 3.2$  and so the least squares linear fit is

$$y = 2x + 3.2.$$

From the vector form of the equation, we find that  $\hat{y} = 2 \cdot \vec{x} + 3.2 \cdot \vec{1} = \langle 7.2, 9.2, 11.2, 13.2, 15.2 \rangle$ .

The correlation coefficient  $r$  also has a geometric interpretation. One form of the computational

formula for  $r^2$  is  $r^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$ , where  $\bar{y}$  is the vector whose elements are all the average of the  $y$ -

values. The correlation coefficient  $r$  can be found by taking the square root of this ratio,

$$r = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}}{\sqrt{\sum_{i=1}^n (\bar{y}_i - \bar{y}_i)^2}}. \text{ The vector } \bar{y} \text{ can be found by projecting vector } \hat{y} \text{ onto the } \bar{1} \text{ vector. To see this,}$$

consider the length of the projection using the vector formula  $\cos(\theta) = \frac{|\bar{u}|}{|\bar{y}|}$ . Solving for  $|\bar{u}|$  we have

$$|\bar{u}| = |\bar{y}| \cdot \frac{\bar{1} \cdot \bar{y}}{|\bar{1}| |\bar{y}|} = \frac{\bar{1} \cdot \bar{y}}{|\bar{1}|}. \text{ In our example, this is } |\bar{u}| = \frac{1 \cdot 7 + 1 \cdot 10 + 1 \cdot 10 + 1 \cdot 14 + 1 \cdot 15}{\sqrt{5}} = \frac{56}{\sqrt{5}}. \text{ The vector}$$

in the direction of  $\bar{1}$  with a length of  $\frac{56}{\sqrt{5}}$  is the vector  $\bar{y} = \langle \frac{56}{5}, \frac{56}{5}, \frac{56}{5}, \frac{56}{5}, \frac{56}{5} \rangle$ . So, in Figure 3, the vector from A to B is vector  $\bar{y}$ , the vector from B to C, then is  $\bar{y} - \hat{y}$ , and the vector from B to D is  $\hat{y} - \bar{y}$ .

The lengths of  $\bar{y} - \hat{y}$  and  $\hat{y} - \bar{y}$ , respectively, are  $\sqrt{\sum_{i=1}^n (\bar{y}_i - \hat{y}_i)^2}$  and  $\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}$ , so the ratio of these two given by  $r$  is just the cosine of the angle between the two vectors,  $\bar{y} - \hat{y}$  and  $\hat{y} - \bar{y}$ .

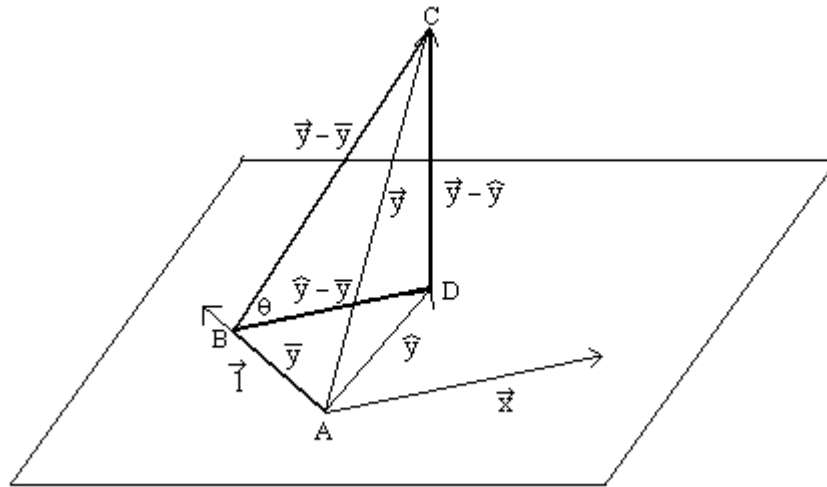


Figure 3: The correlation coefficient and angle  $\theta$

$$\text{In this example } \cos\theta = \frac{(\hat{y} - \bar{y}) \cdot (\bar{y} - \hat{y})}{|\hat{y} - \bar{y}| |\bar{y} - \hat{y}|} = \frac{40}{\sqrt{40} \cdot \sqrt{42.8}} = 0.966736489. \text{ The values of the cosine vary}$$

between -1 and 1, with negative values representing angle greater than 90 degrees. Moreover, the shorter the residual vector  $\bar{y} - \hat{y}$ , the smaller the angle, and the closer  $\hat{y}$  is to  $\bar{y}$ .

#### References:

Johnson, Richard A. and Wichern, Dean W., *Applied Multivariate Statistical Analysis, 3rd*, Prentice Hall, 1992.

Saville, David J. and Wood, Graham R., *Statistical Methods, The Geometric Approach*, Springer-Verlag, 1991.

Box, George, William Hunter, and J. Stuart Hunter, *Statistics for Experimenters*, John Wiley and Sons, New York, 1978.