

A Vector Approach to 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.

Simple Linear Regression

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 \vec{x} and \vec{y} , with

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

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

$$\vec{y} = a\vec{1} + b\vec{x}$$

is well defined.

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

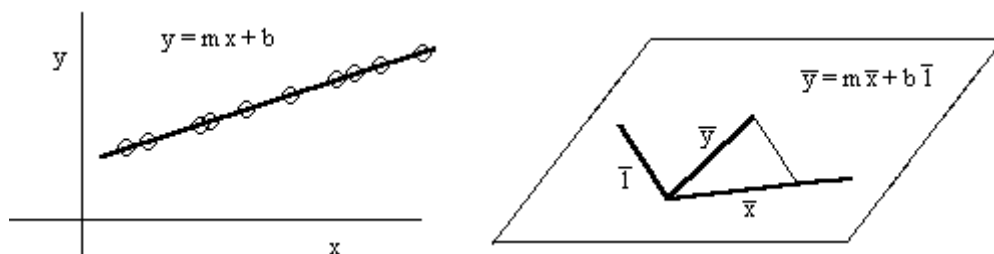


Figure 1: Data lie exactly on the line $y = a + bx$

Of course, in a real data situation, we cannot expect the equation $\vec{y} = a\vec{1} + b\vec{x}$ to hold. If there is any error at all, then $\vec{y} \neq a\vec{1} + b\vec{x}$, that is, the vector \vec{y} is not in the plane of \vec{x} and $\vec{1}$.

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 $\bar{y} = a\bar{1} + b\bar{x}$ is the best linear model for the data.

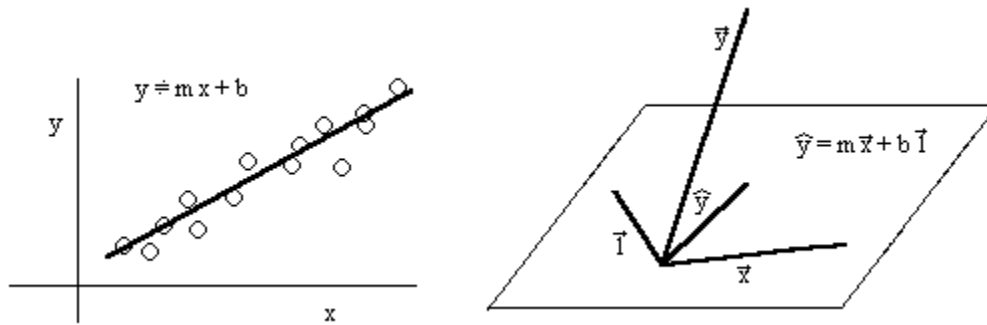


Figure 2: Data do not lie exactly on the line $y = a + bx$

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.

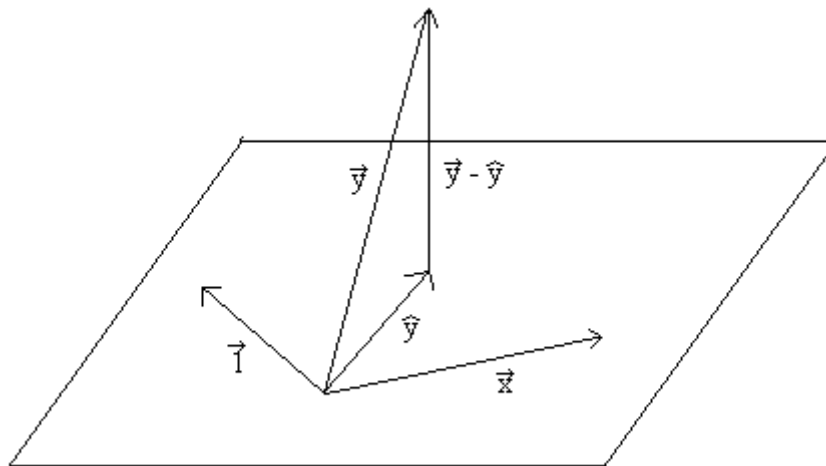


Figure 3: The Geometry of Linear Regression

The vector $\vec{r} = \bar{y} - \hat{y} = \bar{y} - (a\bar{1} + b\bar{x})$, 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 3 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 \vec{x} and $\vec{1}$, we know that the dot products $\vec{r} \cdot \vec{x}$ and $\vec{r} \cdot \vec{1}$ are zero. If

$$\vec{r} \cdot \vec{1} = 0,$$

then

$$(\bar{y} - \hat{y}) \cdot \vec{1} = (\bar{y} - a\vec{1} - b\vec{x}) \cdot \vec{1} = 0.$$

This gives the linear equation

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

Similarly, if

$$(\bar{y} - \hat{y}) \cdot \vec{x} = 0,$$

then

$$(\bar{y} - a\vec{1} - b\vec{x}) \cdot \vec{x} = 0.$$

This generates the linear equation

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

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

$$\bar{y} \cdot \vec{1} = a(\vec{1} \cdot \vec{1}) + b(\vec{x} \cdot \vec{1})$$

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

Solving this system, we find that

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

Evaluating the dot products generates the more conventional forms

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

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

$$y = 3.2 + 2x.$$

From the vector form of the equation, we find that

$$\hat{y} = 3.2 \cdot \vec{1} + 2 \cdot \vec{x} = \langle 7.2, 9.2, 11.2, 13.2, 15.2 \rangle$$

and the residual vector is

$$\vec{y} - \hat{y} = \langle 7, 10, 10, 14, 15 \rangle - \langle 7.2, 9.2, 11.2, 13.2, 15.2 \rangle = \langle -0.2, 0.8, -1.2, 0.8, -0.2 \rangle.$$

Correlation Coefficient

The correlation coefficient, r , can also be related to the vectors in Figure 2. One form of the

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

all the average of the y -values. The vector \bar{y} can be found by projecting the vector \vec{y} onto the $\vec{1}$ vector. All that is necessary, though, is to recognize that the vector \bar{y} is in the direction of $\vec{1}$.

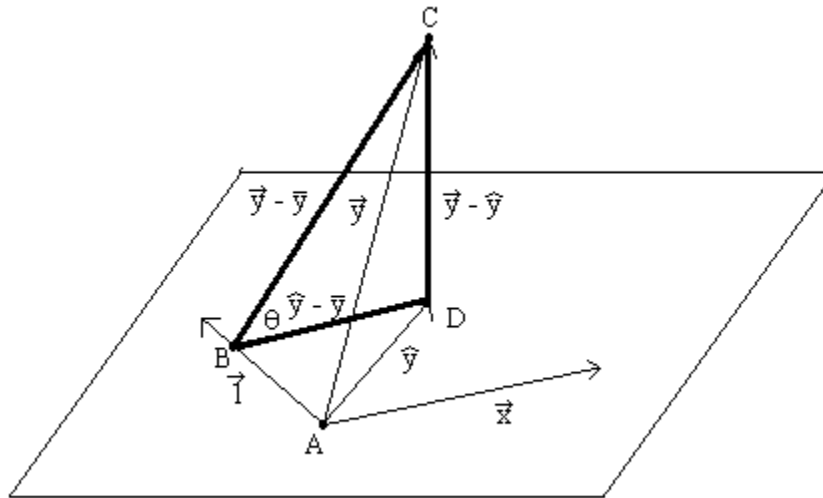


Figure 4: The correlation triangle

In Figure 4, the vector from A to B is vector $\vec{1}$. So, the vector from B to C, then, is vector

$\vec{y} - \bar{y}$ and the vector from B to D is $\hat{y} - \bar{y}$. The length of $\vec{y} - \bar{y}$ is $|\vec{y} - \bar{y}| = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}$ and

of $\hat{y} - \bar{y}$ is $|\hat{y} - \bar{y}| = \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$. These two vectors form the side and hypotenuse of a right

triangle, so the ratio of these two lengths is $\cos(\theta) = \frac{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$. But this is just $\sqrt{R^2} = r$.

The correlation coefficient is the cosine of the angle between $\bar{y} - \bar{y}$ and $\hat{y} - \bar{y}$. The value of the cosine varies from -1 to 1 , as expected for r . In our example,

$$\cos(\theta) = \frac{(\hat{y} - \bar{y}) \cdot (\bar{y} - \bar{y})}{|\hat{y} - \bar{y}| |\bar{y} - \bar{y}|} = \frac{40}{\sqrt{40} \sqrt{42.8}} \approx 0.9667.$$

References:

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

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

Appendix B:

Multivariate Regression

In simple linear regression, to find the values of the scalars a and b , we solved the system of equations

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

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

This system can be written as a single matrix equation given below:

$$\begin{bmatrix} \bar{y} \cdot \bar{1} \\ \bar{y} \cdot \bar{x} \end{bmatrix}_{2 \times 1} = \begin{bmatrix} \bar{1} \cdot \bar{1} & \bar{x} \cdot \bar{1} \\ \bar{1} \cdot \bar{x} & \bar{x} \cdot \bar{x} \end{bmatrix}_{2 \times 2} \begin{bmatrix} a \\ b \end{bmatrix}_{2 \times 1} \quad (1)$$

where we define the two matrices X and Y using the three known vectors $\bar{1}$, \bar{x} , and \bar{y} , with matrix X being an $n \times 2$ matrix containing the explanatory vectors $\bar{1}$ and \bar{x} ,

$$X_{n \times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \text{ and } Y \text{ being the } n \times 1 \text{ matrix containing vector } \bar{y}, Y_{n \times 1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}. \text{ We can}$$

rewrite the matrix equation (1) above in terms of X and Y .

The left side of equation (1) is the product of X^T and Y ,

$$\begin{bmatrix} \bar{y} \cdot \bar{1} \\ \bar{y} \cdot \bar{x} \end{bmatrix}_{2 \times 1} = (X^T Y),$$

and the right side of equation (1) is

$$\begin{bmatrix} \bar{1} \cdot \bar{1} & \bar{x} \cdot \bar{1} \\ \bar{1} \cdot \bar{x} & \bar{x} \cdot \bar{x} \end{bmatrix}_{2 \times 2} \begin{bmatrix} a \\ b \end{bmatrix}_{2 \times 1} = (X^T X) \begin{bmatrix} a \\ b \end{bmatrix}_{2 \times 1} = (X^T X) \beta,$$

where β is the matrix of coefficients a and b .

Solving for β , we find that $\beta = (X^T X)^{-1} (X^T Y)$. This matrix equation will perform all forms of polynomial and multiple regression.

Given a set of n data points

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), \dots, (x_n, y_n),$$

we have $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ and $Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}$. For example, (2,7), (3,10), (4,10), (5,14), and

(6,15) is represented by $X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \end{bmatrix}$ and $Y = \begin{bmatrix} 7 \\ 10 \\ 10 \\ 14 \\ 15 \end{bmatrix}$, with $(X^T X)^{-1} (X^T Y) = \begin{bmatrix} 3.2 \\ 2 \end{bmatrix}$, so our

regression equation is $\hat{y} = 3.2 + 2x$. The matrix product $X\beta$ will give the fitted values and $Y - X\beta$ the residuals.

Quadratic and Other Models

If we want to fit a quadratic regression to these data, simply alter the X matrix. To fit $\hat{y} = a + bx + cx^2$, matrix X will have a column of 1's (for a), a column of x 's (for b),

and a column of x^2 's (for c). In this case $X = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \end{bmatrix}$ and $Y = \begin{bmatrix} 7 \\ 10 \\ 10 \\ 14 \\ 15 \end{bmatrix}$, with

$(X^T X)^{-1} (X^T Y) = \begin{bmatrix} 3.2 \\ 2 \\ 0 \end{bmatrix}$. In this case we get the same linear equation, which is strange,

but that's what happened. If we wanted to find a model of the form $\hat{y} = ax + bx^3$, then we form matrix X with a column of x 's and a column of x^3 's. In this case,

$X = \begin{bmatrix} 2 & 8 \\ 3 & 27 \\ 4 & 64 \\ 5 & 125 \\ 6 & 216 \end{bmatrix}$ and $Y = \begin{bmatrix} 7 \\ 10 \\ 10 \\ 14 \\ 15 \end{bmatrix}$, with $(X^T X)^{-1} (X^T Y) = \begin{bmatrix} 3.255 \\ -0.0215 \end{bmatrix}$, and the model is

$$\hat{y} = 3.255x - 0.0215x^3.$$

To perform a multivariate fit, create the X matrix so that it has the form of the model you want to use. If we have the ordered triples (x, y, z) and we want to fit $z = a + bx + cy$,

the X matrix will be an $n \times 3$ matrix, with a column of 1's, and column of x 's and a column of y 's. For example, if we have the data

$$(1, 2, 5), (2, 1, 3), (2, 2, 7), (3, 5, 10), (4, 3, 12), \text{ and } (5, 5, 19),$$

we create the matrices

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 5 \\ 1 & 4 & 3 \\ 1 & 5 & 5 \end{bmatrix} \text{ and } Y = \begin{bmatrix} 5 \\ 3 \\ 7 \\ 10 \\ 12 \\ 19 \end{bmatrix}, \text{ with } (X^T X)^{-1} (X^T Y) = \begin{bmatrix} -1.679 \\ 2.698 \\ 1.123 \end{bmatrix},$$

and the model will be $\hat{z} = -1.679 + 2.698x + 1.123y$. If we want the fit

$$z = ax + by + cxy + dy^2, \text{ matrix } X \text{ would be } n \times 4 \text{ with } X = \begin{bmatrix} 1 & 2 & 2 & 4 \\ 2 & 1 & 2 & 1 \\ 2 & 2 & 4 & 4 \\ 3 & 5 & 15 & 25 \\ 4 & 3 & 12 & 9 \\ 5 & 5 & 25 & 25 \end{bmatrix} \text{ and}$$

$$Y = \begin{bmatrix} 5 \\ 3 \\ 7 \\ 10 \\ 12 \\ 19 \end{bmatrix}, \text{ with } (X^T X)^{-1} (X^T Y) = \begin{bmatrix} -1.076 \\ 3.871 \\ 1.096 \\ -0.901 \end{bmatrix}, \text{ and the model would be}$$

$$z = -1.076x + 3.871y + 1.096xy - 0.901y^2.$$

Finally, for the temperature-insulation data, we want a model of the form

$$\hat{G}as = a + bTemp + cInsulation + d(Temp \cdot Insulation).$$

So we create the matrices

$$X_{n \times 4} = \begin{bmatrix} 1 & Temp_1 & 0 & 0 \\ 1 & Temp_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & Temp_{n-1} & 1 & Temp_{n-1} \\ 1 & Temp_n & 1 & Temp_n \end{bmatrix}, Y_{n \times 1} = \begin{bmatrix} Gas_1 \\ Gas_2 \\ Gas_3 \\ \vdots \\ Gas_n \end{bmatrix} \text{ and } (X^T X)^{-1} (X^T Y) = \beta_{4 \times 1} = \begin{bmatrix} 6.85 \\ -0.393 \\ -2.26 \\ 0.144 \end{bmatrix}$$

The model, then, is $\hat{G}as = 6.85 - 0.393Temp - 2.26Insulation + 0.144(Temp \cdot Insulation)$.

If $Insulation = 0$

then we have $\hat{G}as = 6.85 - 0.393Temp$

and if $Insulation = 1$

we have $\hat{G}as = 4.59 - 0.249Temp$.

References:

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

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.