

Fitting a Least Squares Line to Data

A widely used method of finding a linear model for a set of ordered pairs is the method of least squares. Although the equation of the least squares line can be found without using calculus, calculus is an efficient way to find the slope m and y -intercept b of the least squares line. The principle of the least squares line is that the sum of the squares of the residuals should be minimized. If we write an expression for the sum of the squares of the residuals for a given set of data, we can then use the derivative to help us find the minimum value of this expression. Since the equation of the least squares line contains two unknown quantities, slope and intercept, whose values we must determine, our expression for the sum of the squares of the residuals will include two unknowns.

We will develop the method used to determine the slope and intercept of a least squares line using sample data gathered in an experiment. In the experiment, weights were suspended from a spring, causing the spring to stretch. Table 3.33 gives the weight suspended from the spring (in decagrams) and the resultant length of the spring (in centimeters).

Weight (in dg)	1	2	3	4	5
Length (in cm)	6	7	9	10	11

Table 3.33 Weights and lengths of spring

The scatter plot of length versus weight in Figure 3.34 shows that a line is an appropriate model for this data set. The residuals are represented graphically by the vertical distances between the data points and the linear model. Residual values can be calculated by subtracting the y -coordinate of a point on the linear model from the y -coordinate of a corresponding data point. The points on the linear model $y = mx + b$ have coordinates $(1, m + b)$, $(2, 2m + b)$, $(3, 3m + b)$, $(4, 4m + b)$ and $(5, 5m + b)$. The residuals associated with each of the five data points are $6 - (m + b)$, $7 - (2m + b)$, $9 - (3m + b)$, $10 - (4m + b)$, and $11 - (5m + b)$. Therefore, an expression for S , the sum of the squares of the residuals,

is given by $S = [6 - (m + b)]^2 + [7 - (2m + b)]^2 + [9 - (3m + b)]^2 + [10 - (4m + b)]^2 + [11 - (5m + b)]^2$

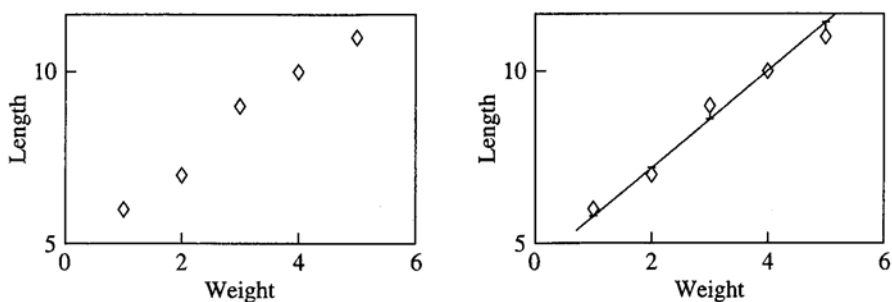


Figure 3.34 Scatter plot of length versus weight and a linear model

Exercises for Investigation

1. Treat m as a constant and find the derivative of S with respect to b .
2. Treat b as a constant and find the derivative of S with respect to m .
3. If S is to be minimized, both of the derivatives from problems 1 and 2 must be equal to zero. Find the values of m and b so that both of these derivatives are equal to zero.

Taking the derivative of a function with respect to one independent variable while treating the other variables as if they are constants produces derivatives that are called partial derivatives. The partial

derivative of S with respect to b is written $\frac{\partial S}{\partial b}$ which means m is treated as a constant. (The symbol ∂

is the lowercase Greek letter delta; Δ is the uppercase delta.) The partial derivative of S with respect

to m is written $\frac{\partial S}{\partial m}$ which means b is treated as a constant. In general, the concept of partial

derivatives is more complicated than we have made it seem here and is usually covered in a multivariable calculus course.

4. Suppose we replace the specific data points for the spring with the general data points (x_1, y_1) (x_2, y_2) (x_3, y_3) (x_4, y_4) and (x_5, y_5) . Repeat the process used in problems 1-3 to find formulas for m and b in terms of the x_i 's and the y_i 's, where i is an integer that ranges in value from 1 to 5. *Summation notation* is useful in these expressions. For example, in summation notation,

the sum of the x_i 's is written as $\sum_{i=1}^5 x_i$ and the sum of the y_i 's is written as $\sum_{i=1}^5 y_i$. (The symbol Σ

is the uppercase Greek letter *sigma*, which in this case represents *sum*.) Here, i is the index and the numbers 1 and 5 indicate that integer values from 1 to 5, inclusive, are to be used in place of i in the

summand. That is, $\sum_{i=1}^5 x_i = x_1 + x_2 + x_3 + x_4 + x_5$.

5. Now that we have found expressions for the slope and intercept of the least squares line for any data set of 5 points, we can generalize the results for any size set of data. Assume we have N points with coordinates (x_i, y_i) where i is an integer from 1 to N . Solve for m and b in terms of the x_i 's and the y_i 's and write the expressions using summation notation.

6. The data set shown in Table 3.35 gives the data that students gathered to test the hypothesis that a person's arm span and height are equal. Use the formulas found for m and b in problem 5 to determine the least squares linear model for this data. Compare the results with the least squares line determined by a calculator or computer

Arm span (in cm)	Height (in cm)
157	171
164	168
177	178
162	167
193	181
183	186
158	164
175	175
156	159

Table 3.35 Ann span and height data

7. The method for finding the equation of the least squares line can be extended to finding a least squares quadratic model. Suppose we want to find a model of the form $y = ax^2 + bx + c$ for a set of data. Write an expression for the sum of the squared residuals S in terms of the constants a , b , and c . Use partial derivatives $\frac{\partial S}{\partial a}$, $\frac{\partial S}{\partial b}$ and $\frac{\partial S}{\partial c}$ to find values for a , b , and c that minimize S . Your answers should be in terms of the data points (x_i, y_i) where i is an integer from 1 to N . It may help to use the fact that $\frac{d}{dx} \left\{ \sum_{i=1}^N [f(x_i)]^2 \right\} = \sum_{i=1}^N \frac{d}{dx} [f(x_i)]^2$.

(Hint: It is not necessary to expand $[f(x_i)]^2$ before taking the derivative.)