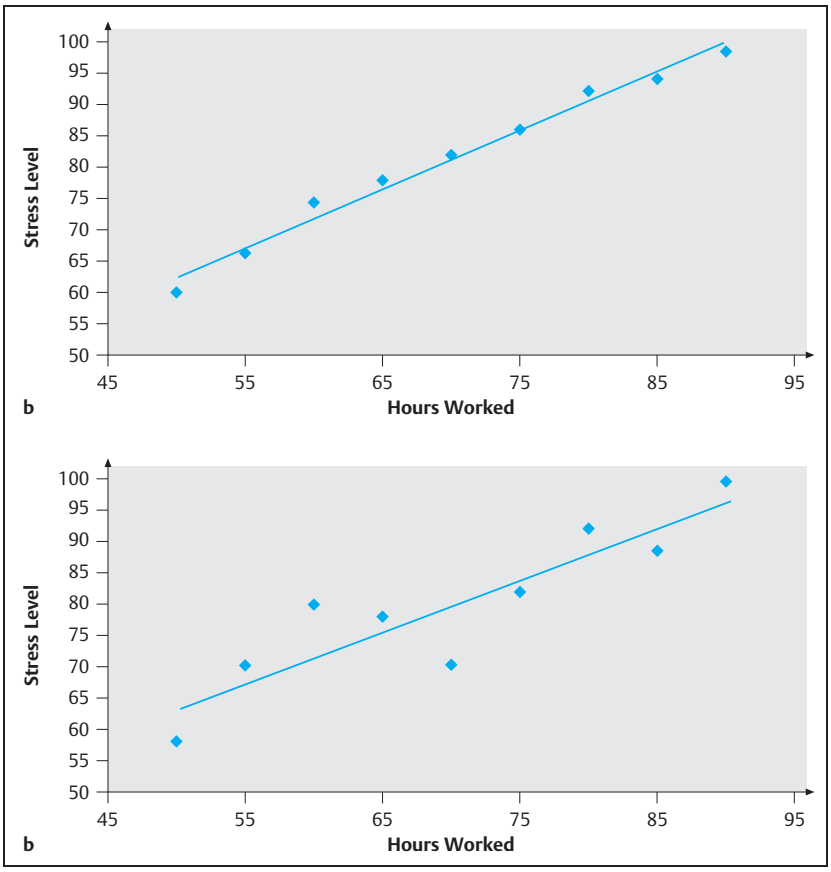

Fig. 44.3 **(a)** Sample graph of medical residents' hours worked versus stress levels showing points lying closer to the fitted line, representing a stronger relationship between the two variables. **(b)** Sample graph of medical residents' hours worked versus stress levels showing points scattered more widely around the fitted line, representing a weaker relationship between the two variables.

sured in hours per week could be converted to minutes per week – a systematic change in our data points that would alter the slope, but have no effect on the relationship between the two variables.

So how do we measure the strength of a relationship between two continuous variables? It is actually quite simple. The closer the data points fall in a straight line to form a linear relationship, the stronger the relationship between the variables.[1] Consider **Fig. 44.3**.

Looking at the two graphs in **Fig. 44.3**, it is apparent that both graphs represent a positive relationship between stress levels and hours worked. That is, increase in hours worked also leads to increases in stress levels. However, the points in **Fig. 44.3a** lie much closer to the fitted line, whereas those in **Fig. 44.3b** are scattered more widely above and below the fitted line. In **Fig. 44.3a**, the fitted line (which is a function of x) better captures the variance in y. In other words, by knowing the value of x (hours worked), we are better able to predict the value of y (stress level).

How exactly is the line of best fit created to capture the greatest amount of variance of the data points? The process is known as *linear regression analysis*.

Linear Regression Analysis

Consider the graph below (**Fig. 44.4**). The data represents the resulting blood loss in femoral neck fracture patients treated with hemiarthroplasty.

The value of y (blood loss) for any given value of x (minutes) can be predicted by analyzing the fitted line. This predicted value of $\hat{y}$ (known as y-hat) can be explained by the basic formula for a straight line,

$$\hat{y} = mx + b \qquad \textbf{Eq. 44.1}$$

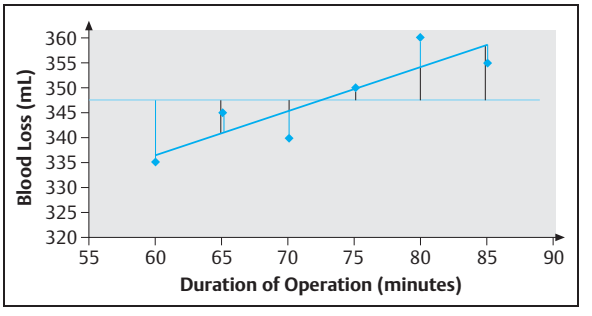

Fig. 44.4 The graph plots operation length versus blood loss in femoral neck fracture patients treated with hemiarthroplasty and depicts the line of best fit.

where m is the slope of the fitted line and b is the y-intercept. However, as the graph depicts, all the points do not lie on the fitted line, and the above formula can only predict the value of blood loss with given error in most cases.

The true value of y is expressed by the following formula:

$$y = mx + b + \varepsilon \qquad \textbf{Eq. 44.2}$$

The error term, ε, explains the variance of the points from the fitted line due to factors, other than operation length, that influence blood loss. The error term is also known as the *residual variance*.[2] The value of the error term for each point on the above graph is the vertical distance between the actual point and the fitted line, as spanned by the green line. **Figure 44.5** demonstrates the error terms associated with the first three data points in **Fig. 44.4** with a magnified view.

The goal of regression analysis is to find values for the slope and y-intercept that will yield a line of best fit that explains the variance in y to the highest degree possible. Essentially, we need a line that lies as close as possible to all the data points, so that the residual variance is minimized. This introduces the concept of linear regression analysis. Formally stated, linear regression analysis is the creation of a line of best fit that minimizes the sum of squares of all the error terms (*sum of squares residual*). This is also known as *least-squares regression analysis*. The product of this analysis, the fitted line, is known as the least-squares regression line.[1,2] It is important to note that the error terms are not summed directly because some are positive values while others are negative, which will cancel each other if summed directly. Rather, regression analysis sums the square of each error term so that all values are positive.

Jargon Simplified: Linear Regression Analysis and Least-Squares Regression Line

Linear regression analysis: A statistical approach for measuring the linear relationship between two continuous variables by creating a line of best fit, which best captures the variance in the dependent variable.

Least-squares regression line: The line of best fit that is produced by linear regression analysis. It is defined by the general formula, $y = mx + b$.

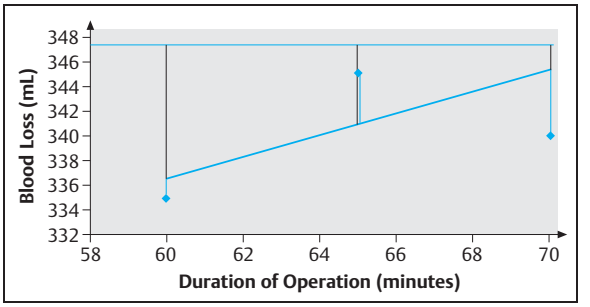

Fig. 44.5 The graph plots operation length versus blood loss in femoral neck fracture patients treated with hemiarthroplasty and depicts the line of best fit.

Mathematically, linear regression analysis minimizes the following expression:

$$\text{Sum of squares residual} = \Sigma\,(y - \hat{y})^2 \qquad \textbf{Eq. 44.3}$$

For **Fig. 44.4**, the equation of the regression analysis is, $y = 0.8857x + 283.29$. Thus, the slope of the fitted line is 0.89, indicating that for every additional minute of surgery there is an increased blood loss of 0.89 mL. Although the y-intercept is 283 mL, it should be understood that this value is meaningless – an operation of zero minutes cannot correspond with a blood loss of 283 mL. It is important to note that although the equation produced by regression analysis enables us to project values beyond our given data points; it must be done cautiously to ensure that such projections are meaningful and truthful.[1] Although the math is not displayed, our regression analysis corresponds with a sum of squares residual of 94.3.

Coefficient of Determination and Correlation Coefficient

Regression analysis provides values for the slope and y-intercept, which maximize the variance in y captured by the line of best fit. However, we have yet to discuss how to measure the magnitude of variance that is actually captured by this line. For that, it is essential to introduce the *coefficient of determination*. The coefficient of determination, expressed as r^2, measures the percentage of variation in y that is accounted for by the least-squares regression. It can also be viewed as the percentage of total variation in the dependent variable that is explained by the independent variable.[2]

$$r^2 = \frac{\textit{variance of } \hat{y}}{\textit{variance of } y} \qquad \textbf{Eq. 44.4}$$

or

$$r^2 = \frac{\textit{sum of squares regression}}{\textit{sum of squares regression} + \textit{sum of squares residual}} \qquad \textbf{Eq. 44.5}$$

The numerator in Eq. 44.4, the variance of $\hat{y}$, is the sum of squares of $\hat{y} - \bar{y}$ – the difference between the fitted point that corresponds to each of the actual data points and the mean value for the independent variable. This is a measure of the variance that the independent variable y would