Linear Regressions Summary

KEY TERMS

- Least Squares Method
- centroid
- regression line
- interpolation
- extrapolation
- correlation
- correlation coefficient
- coefficient of determination
- causation
- necessary condition
- sufficient condition
- common response
- confounding variable

LESSON 1

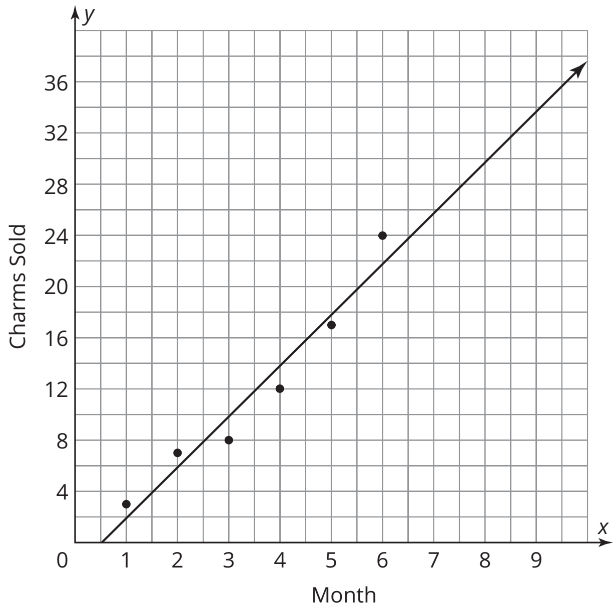
Like a Glove

Technology calculates the line of best fit of data on a scatter plot using the **Least Squares Method**. The line includes the **centroid**, or the point whose *x* value is the mean of all *x* values and whose *y* value is the mean of all *y* values. The **regression line** has the smallest possible vertical distance from each given data point to the line. The sum of the squares of these distances is at a minimum with the regression line.

If there is a linear association between the independent and dependent variables, a linear regression can be used to make predictions within the data set. Using a linear regression to make predictions within the data set is called **interpolation**. To make predictions outside the data set is called **extrapolation**.

For example, consider the situation of Nina selling charms to her classmates. The table records the sales of her charms over the months since she began selling them.

Month	Charms Sold
1	3
2	7
3	8
4	12
5	17
6	24



The regression line modeling the situation is graphed on the scatter plot shown.

The linear regression equation is $y = 3.97x - 2.07$.

Using the equation to interpolate, Nina should sell about 14 charms in the fourth month.

$$y = 3.97(4) - 2.07$$
$$= 13.81$$

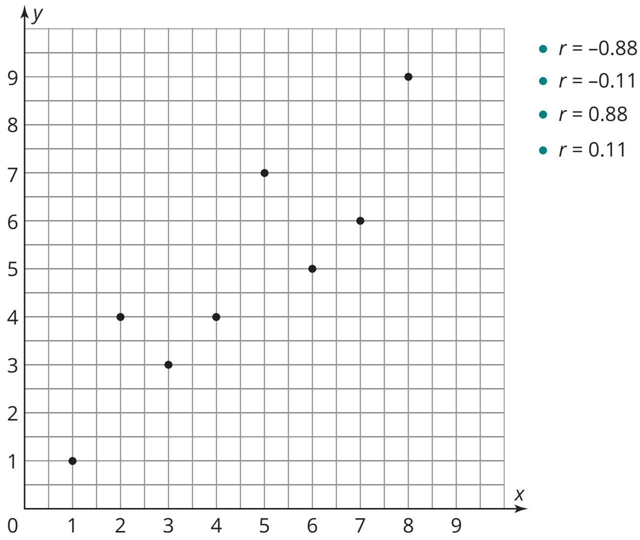
Using the equation to extrapolate, Nina should sell about 30 charms in the eighth month.

$$y = 3.97(8) - 2.07$$
$$= 29.69$$

LESSON **2**

Gotta Keep It Correlatin'

The measure of how well a regression fits a set of data is called **correlation**. When dealing with regression equations, the variable r is used to represent a value called the **correlation coefficient**. The correlation coefficient indicates how close the data are to the graph of the regression equation. The correlation coefficient falls between −1 and 0 if the data show a negative association or between 0 and 1 if the data show a positive association. The closer the r-value is to 1 or −1, the stronger the relationship is between the two. The **coefficient of determination**, r^2, measures how well the regression line fits the data. It represents the percentage of variation of the observed values of the data points from their predicted values.



For example, consider the possible *r*-values for a linear regression given for the data graphed in the scatter plot.

The data has a positive correlation. Because of this, *r*-value must be positive. Also, the data are fairly close to forming a straight line, so of the choices, $r = 0.88$ would be the most accurate. Technology can be used to verify the correlation coefficient. The coefficient of determination for this data set is 0.7744.

When interpreting the correlation between two variables, you are looking at the association between the variables. While an association may exist, that does not mean there is causation between the variables. **Causation** is when one event causes a second event. A correlation is a **necessary condition** for causation, but a correlation is not a **sufficient condition** for causation. Correlation may be due to a **common response**, which is when another reason may cause the same result, or a **confounding variable**, which is when other variables are either unknown or unobserved.

For example, consider an experiment conducted by a group of college students that found that more class absences correlated to rainy days. The group concluded that rain causes students to be sick. However, this correlation does not imply causation. Rain is neither a necessary condition (because students can get sick on days it does not rain) nor a sufficient condition (because not every student who is absent is necessarily sick) for students being sick.