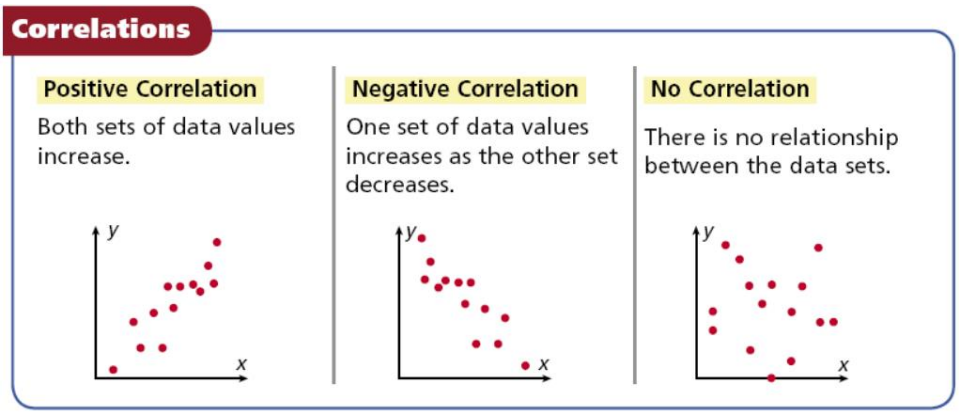
❖ Correlation

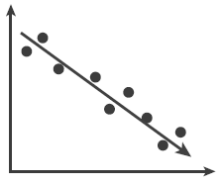
Two-variable data is a collection of paired variable values, such as a series of measurements of air temperature at different times of day. One method of visualizing two-variable data is called a **scatter plot**: a graph of points with one variable plotted along each axis. A recognizable pattern in the arrangement of points suggests a mathematical relationship between the variables.

Correlation is a measure of the strength and direction of the relationship between two variables. The correlation is positive if both variables tend to increase together, negative if one decreases while the other increases, and we say there is "no correlation" if the change in the two variables appears to be unrelated.

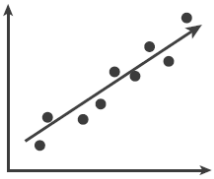


❖ Correlation Coefficient

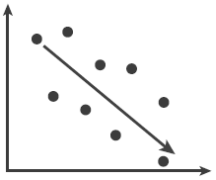
One way to quantify the correlation of a data set is with the **correlation coefficient**, denoted by r. The correlation coefficient varies from -1 to 1, with the sign of r corresponding to the type of correlation (positive or negative). Strongly correlated data points look more like points that lie in a straight line, and have values of r closer to 1 or -1. Weakly correlated data will have values closer to 0.



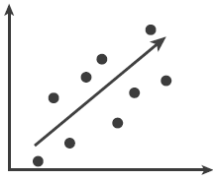
STRONG NEGATIVE CORRELATION
points lie close to a line with a negative slope
r is close to -1



STRONG POSITIVE CORRELATION
points lie close to a line with a positive slope
r is close to 1

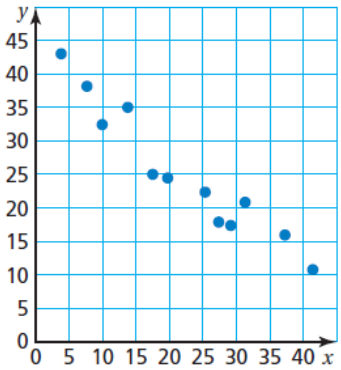


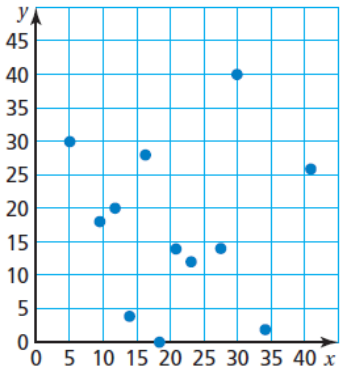
WEAK NEGATIVE CORRELATION
points loosely follow a lie with a negative slope
r is between 0 & -1

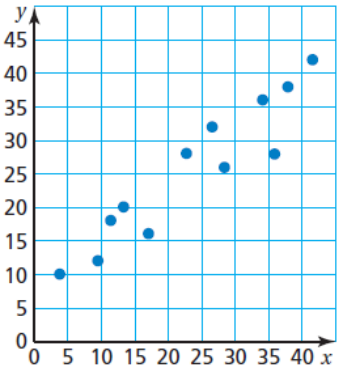


WEAK POSITIVE CORRELATION
points loosely follow a lie with a positive slope
r is between 0 & 1

➢ Describe the correlation and use the scatter plot to estimate the value of r. Indicate whether r is closer to -1, - 0.5, 0, 0.5, or 1.

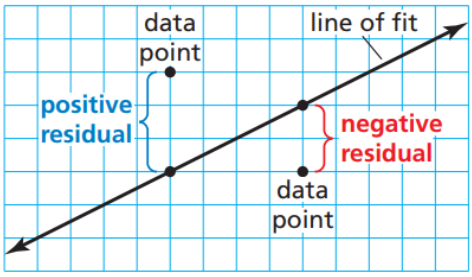






❖ 9.3 & 9.4 ~ Creating & Using Residual Plots

A **residual** is the difference of the y-value of a data point and the corresponding y-value found using the line of fit. A residual can be positive, negative, or zero.



A scatter plot of the residuals shows how well a model fits a data set. If the model is a good fit, then the absolute values of the residuals are relatively small, and the residual points will be more or less evenly dispersed about the horizontal axis. If the model is not a good fit, then the residual points will form some type of pattern that suggests the data are not linear. Wildly scattered residual points suggest that the data might have no correlation.

Residuals help to determine if a curve (shape) is appropriate for the data.

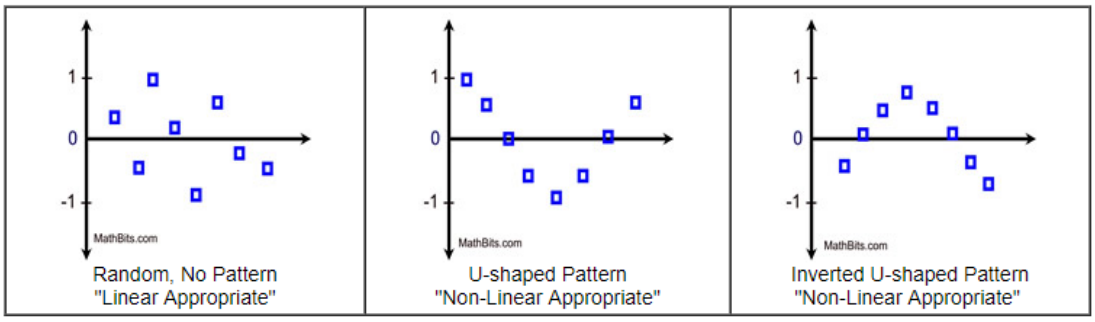
LINEAR VS. NON-LINEAR

Linear associations are the most popular statistical relationships since they are easy to read and interpret. We will spend the majority of our time working with linear relationships, and residuals can tell us when we have an appropriate linear model.

When you look at your scatter plot, and you are unsure if the shape (curve) you chose for your regression equation will create the best model, a residual plot will help you make a decision as to whether the model you chose will, or will not, be an appropriate *linear* model.

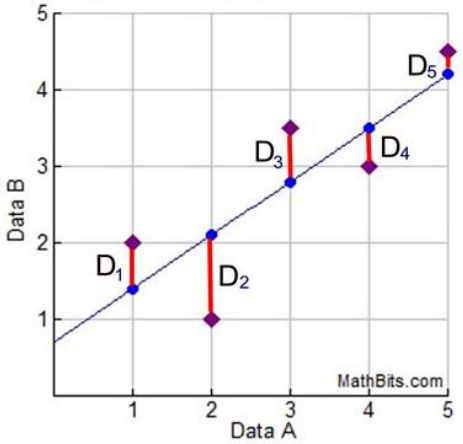
A residual plot is a scatter plot that shows the residuals on the vertical axis and the independent variable on the horizontal axis. The plot will help you to decide on whether a linear model is appropriate for your data.

Appropriate linear model: when plots are randomly placed, above and below x-axis ($y = 0$).
Appropriate non-linear model: when plots follow a pattern, resembling a curve.



When a pattern is observed in a residual plot, a linear regression model is probably **NOT** appropriate for your data.

The residuals are the red line segments, referenced by the letter "D" (for distance), vertically connecting the scatter plot points to the corrdinating points on the linear regression line.



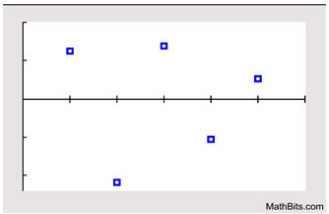
◆ Scatter Plot Points:
{(1,2), (2,1), (3,3½), (4,3), (5,4)}

● Regression Points
{(1,1.4), (2,2.1), (3,2.8), (4,3.5), (5,4.2)}

The Red Line Segments:
The red line segments represent the distances between the y-values of the actual scatter plot points, and the y-values of the regression equation at those points.

The lengths of the red line segments are called RESIDUALS.

THE RESIDUAL PLOT:



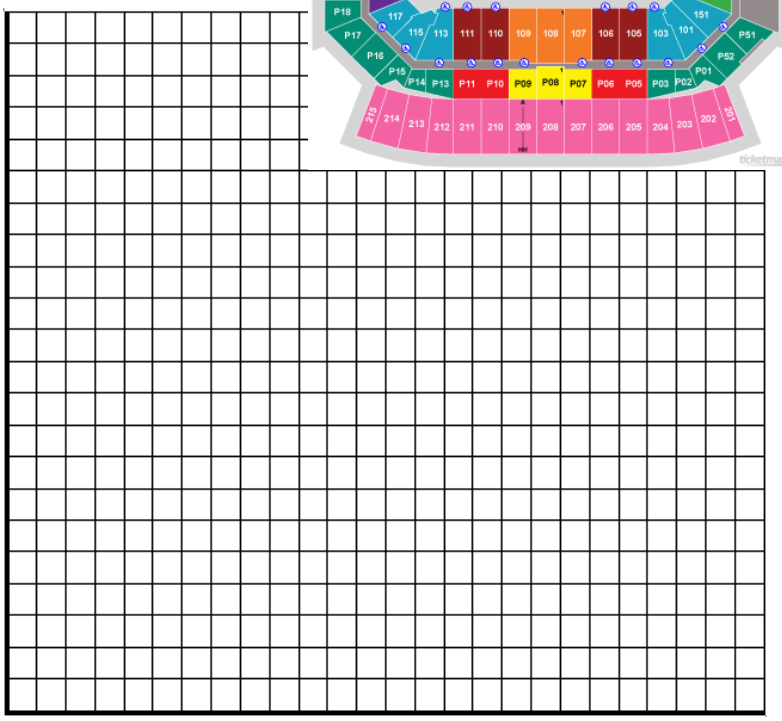
The plots do not follow a pattern. A linear model **IS** appropriate.

- Creating & Using a Residual Plot
  - Is there any relationship between the price of the ticket and the distance from the soccer field?

    Assume I am going to get a seat right at the half-way line (centered) and that row 1 is close to the field and row 50 is further away up in the stands.

    - Make a scatterplot of the data:

ROWS BACK, X	PRICE, Y
2	$205
8	$175
10	$170
15	$105
20	$185
25	$100
30	$65
32	$78
40	$50
48	$20



- What type of relationship is there between two variables? Explain.

- The correlation is: r = -0.919. Describe what -0.919 correlation means in this problem.

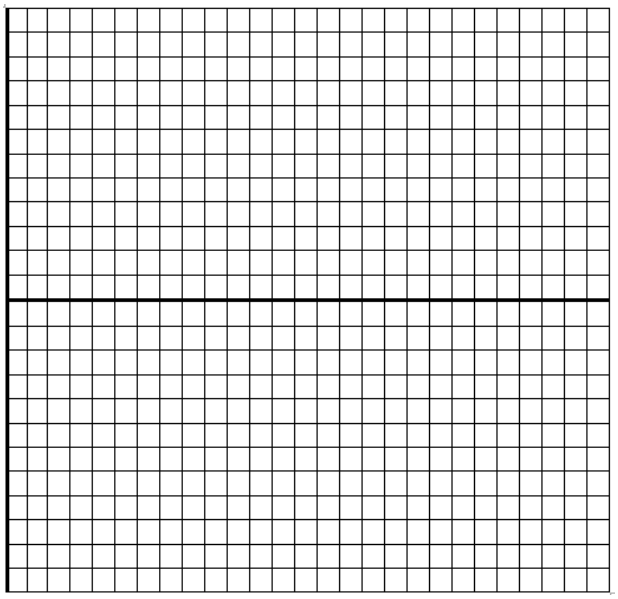
- The least squares regression equation is $y = -4x + 207$. In your opinion, do you think that the line is a good fit or not? Justify your answer.

- Is Row 15 above or below the regression line, how about row 20?

- How many points are above the line of fit and how many are below?

- Use residuals to determine if the least squares regression equation, $y = -4x + 207$, is a good fit for the data. (Make a scatterplot of the residuals.)

x	y	y-VALUE FROM MODEL	RESIDUAL
2	205		
8	175		
10	170		
15	105		
20	185		
25	100		
30	65		
32	78		
40	50		
48	20		



- What does the residual plot tell us in this situation?

- Does this confirm your results about least squares regression equation, $y = -4x + 207$? Is it a good fit? Was your prediction correct? Explain.

- Is there an association between the price of a ticket for an Orlando City game and the number of rows from the field?