Unit 6 – MODELING WITH LINEAR FUNCTIONS

PREREQUISITE SKILLS:

- students should understand how to plot points on the coordinate plane
- students should understand how to write and solve linear equations in two variables
- students should understand function notation
- students should be able to graph linear equations
- students should be able to create linear equations in slope-intercept form

VOCABULARY:

1. independent variable: a variable in an equation that may have its value freely chosen without considering values of any other variable
2. dependent variable: a variable whose value depends upon independent variables
3. transformation: is a process which changes the position and possibly the size and orientation of a shape
4. translation: moving a shape by sliding it without rotating or flipping it. Every point of the object moves a fixed distance in a given direction
5. dilation: a transformation that changes the size of the figure
6. scatter plot: a graph of plotted points that show the relationship between two sets of data
7. correlation: describes a relationship between two data sets
8. positive correlation: a relationship between variables where as one variable increases, so does the other
9. negative correlation: a relationship between variables where as one variable increases, the other variable decreases
10. no correlation: no relationship between two variables
11. correlation coefficient: it is a number that describes the strength and direction of the linear relationship between two variables
12. line of best fit: the line that best fits the data on a scatter plot
13. least squares regression: a method for finding a best fitting line that summarizes the relationship between the two variables
14. residual: it is the difference between the observed y value from the scatter plot and the predicted y value from the regression equation
15. residual plot: it is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis

SKILLS:

- model linear relationships
- transform graphs of linear functions
- create and interpret scatter plots
- use lines of best fit to make predictions
- assess the fit of the linear model

STANDARDS:

F.IF.C.9-1 Compare properties of two linear, quadratic, and/or exponential functions each represented in a different way (algebraically, graphically, numerically in tables, or by verbal descriptions). For example, given a graph of one quadratic function and an algebraic expression for another, say which has the larger maximum.

F.LE.B.5 Interpret the parameters in a linear or exponential function in terms of a context. *(Modeling Standard)

A.CED.A.2-1 Create linear, exponential, and quadratic equations in two or more variables to represent relationships between quantities and graph equations on coordinate axes with labels and scales. Limit exponentials to have integer inputs only. *(Modeling Standard)

F.BF.B.3-1 Identify the effect on the graph of replacing f(x) by f(x) + k, kf(x), f(kx), and f(x + k) for specific values of k (both positive and negative) and find the value of k given the graphs. Experiment with cases and illustrate an explanation of the effects on the graph using technology. Include linear, exponential, quadratic, and absolute value functions.

8.SP.A.1 Construct and interpret scatter plots for bivariate measurement data to investigate patterns of association between two quantities. Describe patterns such as clustering, outliers, positive or negative association, linear association, and nonlinear association.

8.SP.A.3 Use the equation of a linear model to solve problems in the context of bivariate measurement data, interpreting the slope and intercept.

S.ID.B.6a Fit a function to the data; use functions fitted to data to solve problems in the context of the data. Use given functions or choose a function suggested by the context. Emphasize linear, quadratic, and exponential models. *(Modeling Standard)

S.ID.B.6c Fit a linear function for a scatter plot that suggests a linear association. *(Modeling Standard)

S.ID.C.7 Interpret the slope (rate of change) and the intercept (constant term) of a linear model in the context of the data. *(Modeling Standard)

8.SP.A.2 Know that straight lines are widely used to model relationships between two quantitative variables. For scatter plots that suggest a linear association, informally fit a straight line, and informally assess the model fit by judging the closeness of the data points to the line.

S.ID.C.8 Compute (using technology) and interpret the correlation coefficient of a linear fit. *(Modeling Standard)

S.ID.C.9 Distinguish between correlation and causation. *(Modeling Standard)

LEARNING TARGETS:

- 6.1 To compare and interpret the slope and y-intercept of linear functions.
- 6.2 To transform the graph of linear functions.
- 6.3 To find a linear model for a set of bivariate data.
- 6.4 To use technology to find lines of best fit.
- 6.5 To use residuals to determine how well lines of best fit model the data.
- 6.6 To identify correlations between data sets.

BIG IDEAS:

Relationships among quantities can be represented using tables, graphs, verbal descriptions and inequalities. It is important to compare linear functions in various forms to see how they are different or similar and assess their properties. It is necessary to recognize these differences and draw conclusions about these models in real-world situations.

Scatter plots are useful for visualizing a trend or association in data. It is common for an association to exist between data sets. Sometimes the association is linear and sometimes the association is non-linear. Correlation measures how strongly two numerical variables are linearly associated. A common error when interpreting paired data is confusing correlation and causation.

Notes, Examples and Exam Questions

Units 6.1, 6.2 To compare and interpret the slope and y-intercept of linear functions and To transform the graph of linear functions.

Graphs are used to represent relationships between two variables. They help us understand relationships and generalize beyond the data supplied. In functions, a value of *y* or *f(x)* is assigned to each value of *x*. The primary relationship in a function is that the *y*-value depends on the *x*-value. For this reason, the *x*-axis is called the **independent axis** and the *y*-axis is called the **dependent axis.**

The relationship can also be seen in terms of *slope,* as the rate of change between the *dependent y*-axis and the *independent x*-axis: $\dfrac{\text{dependent } \Delta y}{\text{independent } \Delta x}$. In dealing with real situations, it becomes necessary to think of the *x*-axis as some independent variable, such as time, and to think of the *y*-axis as a dependent variable, such as height or distance. Whenever it is clear that one variable depends on the other, the **dependent variable** should be the *y* and the **independent variable** should be the *x*.

<u>**Reading the graph of a situation:**</u> When considering the graph of a situation, remember:

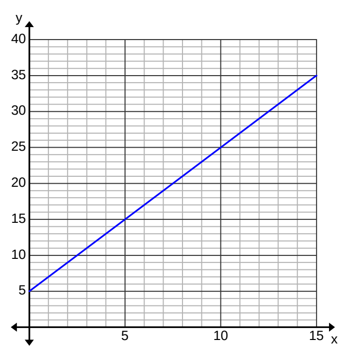
- Questions that involve rates (time per room painted, daily charge) or speeds (miles per hour, births per year) will usually require information about the slope.
- Questions that involve starting times, opening deposits, beginning locations, or similar initial conditions are usually asking questions about the *y*-intercept.

Ex 1: Deluxe Limousine Service charges $5 for the initial pickup of a passenger and then $2 per mile. Make a graph of this situation and describe the slope and *y*-intercept.

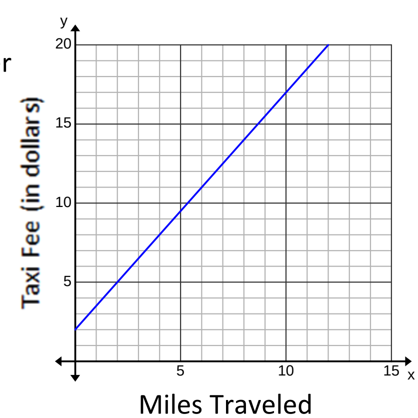
Step 1: Determine the independent and dependent variables. Note that the fee for the ride *depends* on the number of miles traveled. Therefore, the fee will be represented by the variable, *y*, and distance will be represented by the variable, *x*.

Step 2: Create the model by determining slope and *y*-intercept. The change per mile is the fraction that represents the change in price per change in miles travelled, so the slope equals $\frac{\$2}{1\text{ mile}}$ or 1. The *y*-intercept is the starting amount, which is $5.00. Therefore, the model is: $y = \$2x + \$5$.

Step 3: Graph the function.



Ex 2: The graph represents the billing structure for Friendly Taxi. What is the initial pickup fee and charge per mile for this service?



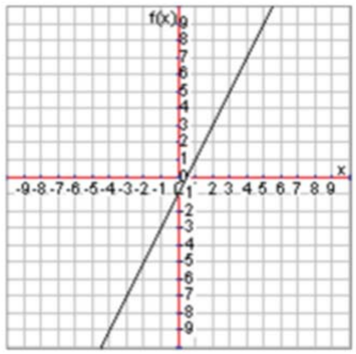
The initial pickup fee is the charge before any driving is done, or in other words, when *x* is zero. The point on the graph with an *x* value of zero is (0, 2). Therefore, the pickup fee is $2.00.

The charge per mile is the change in price per change in miles, or the slope. Two points on the line are: (0, 2) and (4, 8), so $\frac{8-2}{4-0} = \frac{6}{4} = \frac{3\text{ dollars}}{2\text{ miles}} = \$1.50$ per mile.

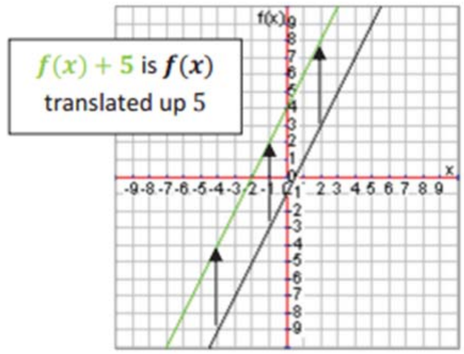
Ex 3: Let $f(x) = 2x - 1$ and transform it to *g(x)* where $g(x) = f(x) + 5$. What would happen?

$f(x) + 5 = 2x - 1 + 5 = 2x + 4$, so $g(x) = 2x + 4$. Now the *y*-intercept is 4 instead of -1 which means that the function has moved up vertically 5 units. This is a vertical translation of 5 units.

Original: $f(x) = 2x - 1$



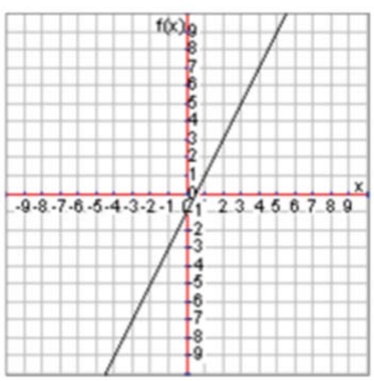
Transformed: $f(x) + 5 = 2x + 4$



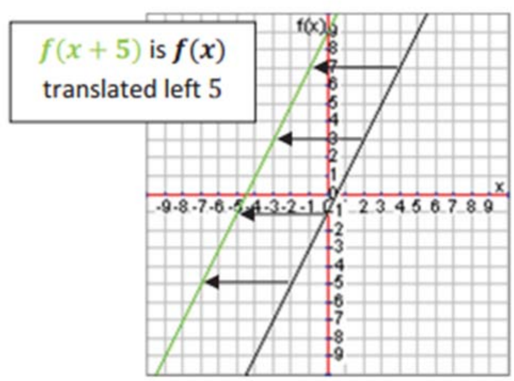
Ex 4: Let $f(x) = 2x - 1$ and transform it to *g(x)* where $g(x) = f(x+5)$. What would happen?

$f(x+5) = 2(x+5) - 1 = 2x + 10 - 1 = 2x + 9$, so $g(x) = 2x + 9$. It looks like this translated (moved) the function up by ten, which is true. However, there is a better way to think about this. Notice that an input of 0 in *f(x+5)* now looks like an input of 5 in *f(x)*. In other words, what was an input of five is now an input of zero. So, a better way to think about this is that it translated the function five units to the left.

Original: $f(x) = 2x - 1$

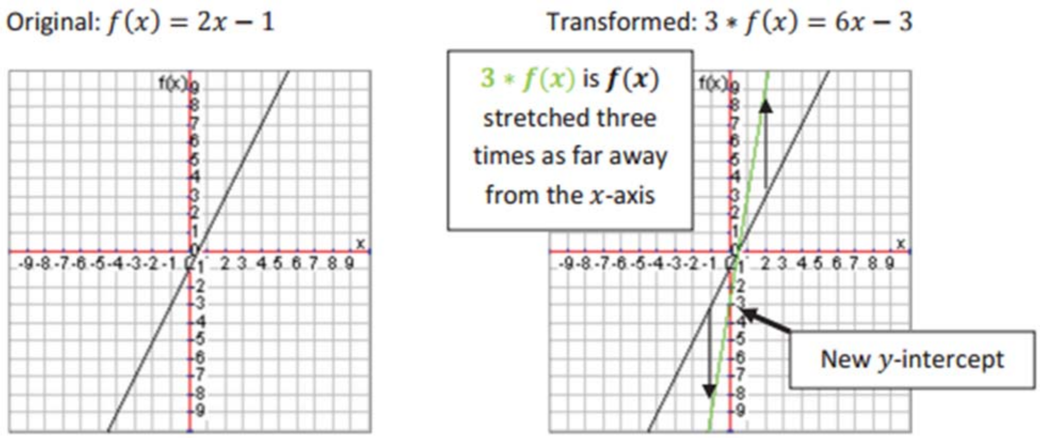


Transformed: $f(x+5) = 2x + 9$



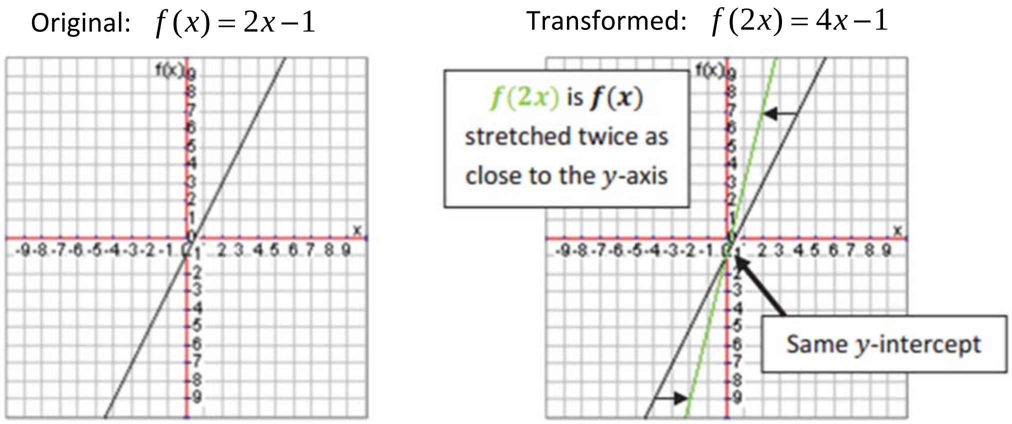
Ex 5: Let $f(x) = 2x - 1$ and transform it to *g(x)* where $g(x) = 3f(x)$. What would happen?

$3f(x) = 3(2x - 1) = 6x - 3$, so $g(x) = 6x - 3$. The output of the transformed function will be three times as big as the original function. This is a dilation transformation. We basically have stretched the graph away from the *x*-axis. Note that the *x*-intercept stayed in place, but the *y*-intercept changed. Since the value of the function at the *x*-intercept is zero, zero times anything is still zero. So that point doesn't move.



Ex 6: Let $f(x) = 2x - 1$ and transform it to *g(x)* where $g(x) = f(2x)$. What would happen?

$f(2x) = 2(2x) - 1 = 4x - 1$, so $g(x) = 4x - 1$. Here again we have a dilation – we basically have stretched the graph closer to the *y*-axis this time. Notice that the *y*-intercept stayed in place, but the *x*-intercept changed.



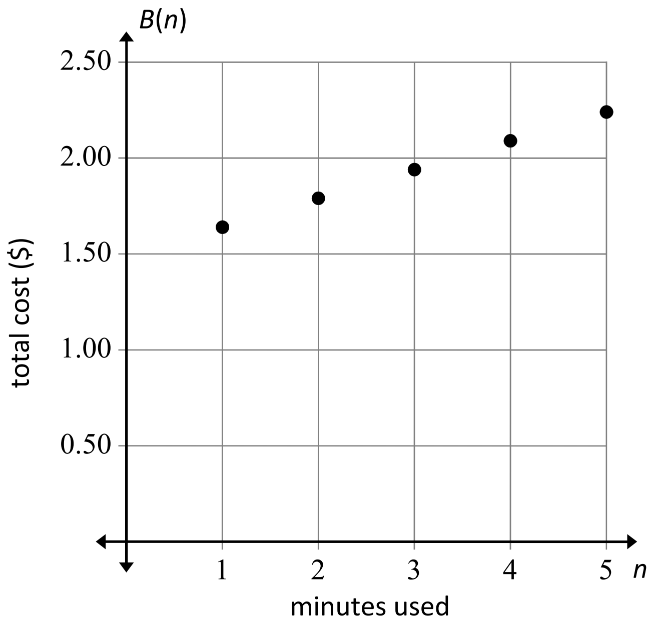
Sample Exam Questions

For questions 1-3, use the scenario below.

A phone call using a prepaid card consists of a fixed fee to place the call plus an additional fee for each minute of the call.

The cost of an *n*-minute phone call with a card from Company A is *A*(*n*) = \$0.99 + \$0.25*n*, where *n* is a positive integer.

The cost of an *n*-minute phone call with a card from Company B is shown in the graph below.



1. The per minute fee for Company B is greater than Company A.

(A) True

(B) False

Ans: B

2. The fixed fee for Company B is greater than Company A.

(A) True

(B) False

Ans: A

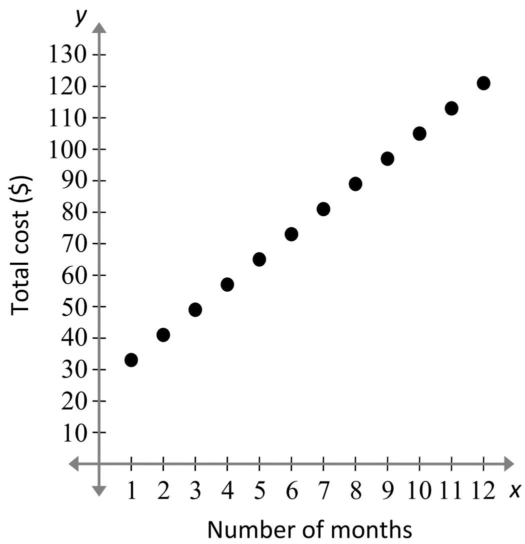
3. A call using Company B will always cost more than the same length call using Company A.

(A) True

(B) False

Ans: B

4. **An online music service charges a $25 start-up fee plus $8 per month for unlimited downloads. The graph illustrates the total cost of a membership for a given number of months.**



What would happen to the graph if the start-up fee changed from $25 to $32?

(A) The slope would increase by $7/month.

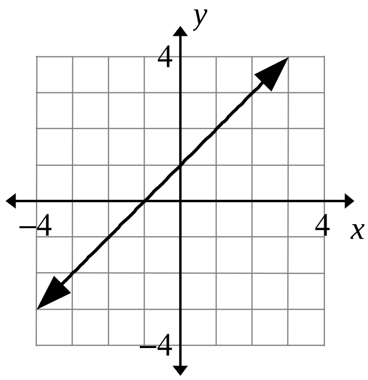
(B) The slope would decrease by $7/month.

(C) The graph would translate up $7.

(D) The graph would translate down $7.

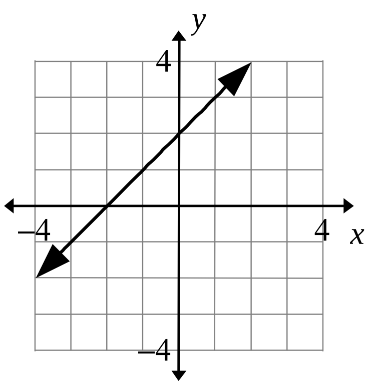
Ans: B

5. The graph shows the linear function $y = f(x)$.

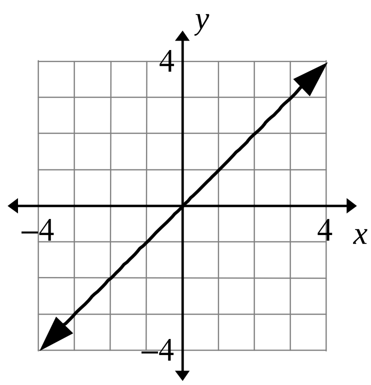


Which graph shows $y = f(x) + 1$?

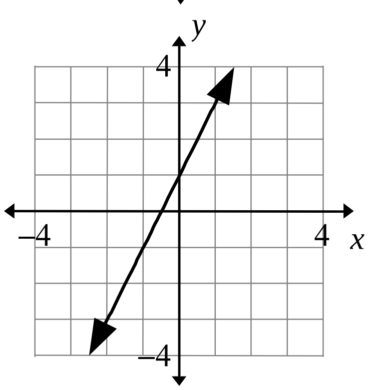
(A)



(B)



(C)



Ans: A

Units 6.3, 6.4 To find a linear model for a set of bivariate data and To use technology to find lines of best fit.

Scatter Plot: a graph of ordered pairs that show a possible relationship between two data sets.

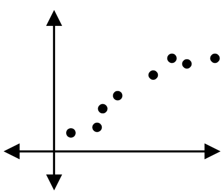
Correlation: the relationship between quantities as shown in a scatter plot

Best-Fitting Line: a line that best fits the pattern of the data in a scatter plot

> *If data have positive correlation, the slope of the best-fitting line will be positive. If data have negative correlation, the slope of the best-fitting line will be negative. If data have relatively no correlation, drawing a best-fitting line is not appropriate.

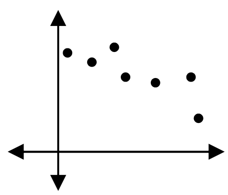
*There are three types of linear correlation:

Positive Correlation



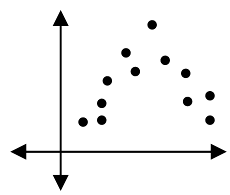
*as *x* increases, *y* also increases

Negative Correlation



*as *x* increases, *y* decreases

Relatively No Linear Correlation



*there is no linear relationship between *x* and *y*

Constructing a Scatter Plot by hand

- Decide which variable is the independent variable (x) and which is the dependent variable (y).
- Construct and label both axes, using reasonable scales for the values. (Use the “break” symbol, if necessary.)
- Plot the data points (ordered pairs)
- Give the scatter plot a title.

Constructing a Scatter Plot on a graphing calculator

- Enter the data values for the independent variable (x) in L_1.
- Enter the data values for the dependent variable (y) in L_2.
- Check that there is the same number of values in both lists.
- Press 

.
- Press 1, going into the first plot. The cursor should be over On. Press ENTER to turn STAT PLOT 1 on.
- Using the arrows and cursor, make sure the first graph type is selected, L_1, and L_2 are the lists to be graphed (unless your data is in other lists), and choose your mark.
- Adjust the window so all the values will appear on the graph. The best window is usually achieved by pressing 

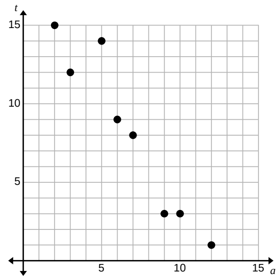
and selecting 9: ZoomStat.

Writing an Equation of a Best-Fitting Line

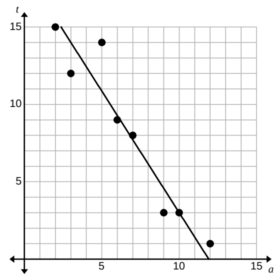
Ex 1: Draw a scatter plot for the age and tail length of some tadpoles. Find the equation of the best-fitting line. Determine any correlation.

Age (days)	5	2	9	7	12	10	3	6
Tail (mm)	14	15	3	8	1	3	12	9

Step One: Draw the scatter plot by plotting the ordered pairs on a coordinate grid. We will use the age in days as the horizontal axis, *a*, and the tail length as the vertical axis, *t*.



Step Two: Sketch the best-fitting line. It should follow the general direction/pattern of the data and come close to all of the data points splitting them even above/below the line.



Step Three: Pick two points on the best-fitting line and write the equation of the line using these two points (you can use points already in the data sets or add new points that fall on the line).

Two points on the line are $(3,14)$ and $(10,3)$. The slope of the line is $\frac{14-3}{3-10}=-\frac{11}{7}$.

Using point-slope form, we have an equation of the line as $t-14=-\frac{11}{7}(a-3)$.

Rewriting in slope-intercept form, we have $t-14=-\frac{11}{7}a+\frac{33}{7} \Rightarrow t=-\frac{11}{7}a+\frac{131}{7}$

Note: Depending on how you drew the best-fitting line, your equation may differ slightly.

Step Four: Determine the type of correlation.

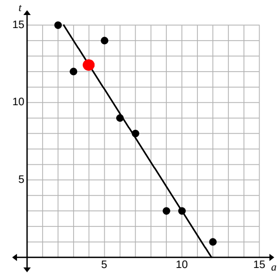
The data have negative correlation because the slope of the best-fitting line is negative.

The points have a fairly strong correlation since they are closer to a straight line than to a total scatter cloud.

Making Predictions with a Best-Fitting Line:

Ex 2: In the tadpole example, predict the tail length of a tadpole that is four days old.

You can estimate using the graph of the best-fitting line:



Looking at the graph, it is somewhere around 12.5 mm long.

We can also use the equation of the best-fitting line to predict the tail length.

$$t = -\frac{11}{7}a + \frac{131}{7} \Rightarrow a = 4$$

$$t = -\frac{11}{7}(4) + \frac{131}{7} = \frac{87}{7} \approx 12.43 \text{ mm long}$$

Ex 3: In the tadpole example, predict how long the tail would be when the tadpole is 7 days old.

Using the equation of the best-fitting line:

$$t = -\frac{11}{7}a + \frac{131}{7}$$

$$t = -\frac{11}{7}(7) + \frac{131}{7}$$

$$t = -\frac{77}{7} + \frac{131}{7}$$

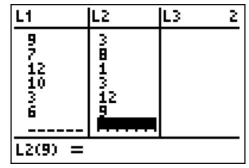
$$t = \frac{54}{7} \approx 7.7mm$$

Note: **Extrapolation** is making a prediction of a *y*-value using an *x*-value outside the range of the sample data. In the examples above, we only have data on tadpoles from age 2 days to 12 days. We would not want to make predictions too far outside of these values because we cannot be confident that the model continues past the given *x* values. We would not use the model to predict the tail length of a tadpole 20 days old. The tail will stop growing at some point.

In the next example, the line of best fit will be determined by the TI-84 calculator using the **Least Squares Linear Regression** Calculator. Out of all possible linear fits, the least squares regression line is the one that has the smallest possible value for the sum of the squares of the residuals. The residuals are the differences between the actual *y* values and the predicted *y* values.

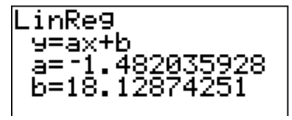
Graphing Calculator Activity: Scatter Plots and Linear Regression

Ex 4: Using the tadpole data, create a scatter plot on the graphing calculator and have it compute the best-fitting line using linear regression.

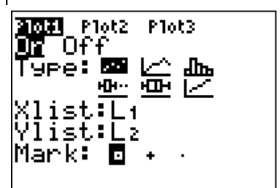


Step One: Enter the data from the table into the Lists L1 and L2.

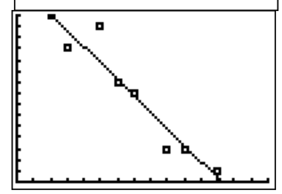
On the home screen, choose STAT then Edit. To clear data from the lists, highlight the name of the list at the top of the column and press CLEAR. Then enter the ages in L1 and the tail lengths in L2.



Step Two: On the home screen, use the following key strokes to calculate the least squares regression line and store it in Y1.



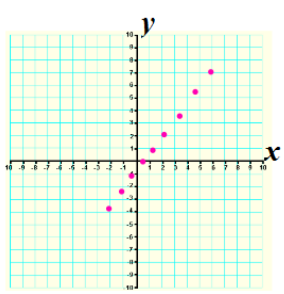
Step Three: Graph the scatter plot with the best-fitting line.



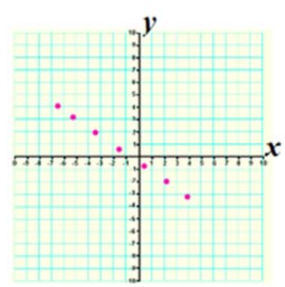
Go to the STAT PLOT menu by pressing 2^{ND} Y=. Choose the options shown in the screen shot. Use the same window as our graph above. Note: The calculator's equation of the line and the one we found earlier are fairly close.

<u>6.5, 6.6 To use residuals to determine how well lines of best fit model the data and To identify correlations between data sets.</u>

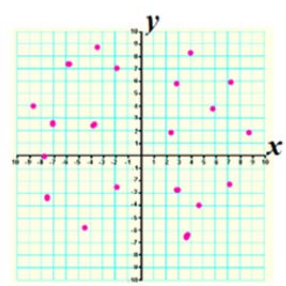
The **correlation coefficient, *r***, is a measure of the strength and direction of the linear relationship between two variables. It is a numerical value that can help us decide if the relationship is strong, moderate or weak. The correlation coefficient is a number between -1 and +1, inclusive. The closer the points are to forming a line, the closer the number is to -1 or +1.



If $r = +1$, then the data points all fall on a perfectly straight line with a positive slope.

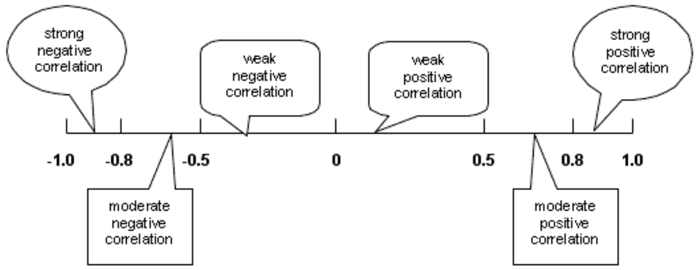


If $r = -1$, then the data points all fall on a perfectly straight line with a negative slope.



The closer r is to 0, the less likely it is that there is a linear relationship between x and y.

Below shows the classification of the correlation coefficient:



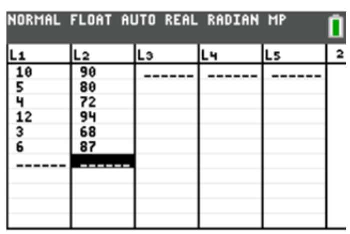
Things to note about the correlation coefficient:

- Graphs with positive slopes have positive *r* values and graphs with negative slopes have negative *r* values.
- *r* is dimensionless (has no units)
- The closer $|r|$ is to 1, the better the data fits the line.
- If *r* is zero, there is no LINEAR correlation. The correlation coefficient only measures linear correlation. Some other correlation (quadratic, logarithmic, power) might exit.
- The correlation coefficient is GREATLY affected by outliers. One ordered pair can raise or lower *r* by a large amount.
- *r* can be found using technology. The formula is very tedious and should not be used.

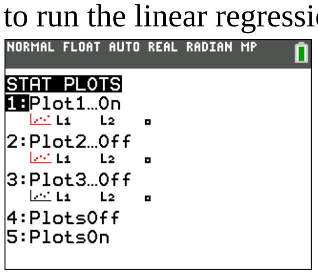
Ex 5: The table shows the amount of study time, in hours, students spend before a test and the test grade. Using technology, find the line of best fit and the correlation coefficient and interpret them in the context of the problem.

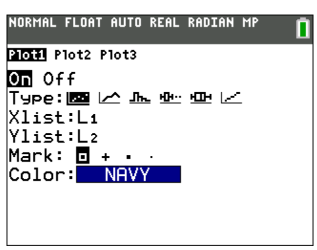
Study Time (in hours)	**Test Grade (in percent)**
10	90
5	80
4	72
12	94
3	68
6	87

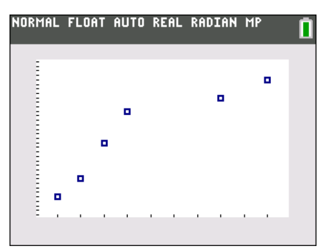
Step 1: Enter the data into L1 and L2. Note that the independent variable (*x*), is study time and will be input into L1. The dependent variable, test grade is input into L2.



Step 2: Draw a scatterplot to see if a linear association exists. Always plot the data first, before running the regression calculation. If no linear pattern is observed in the scatter plot, there is no need to run the linear regression.

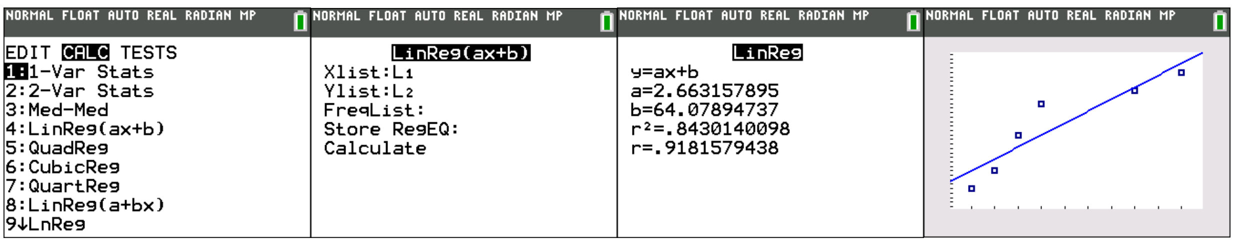






Step 3: Find the LSRL on the calculator. First make sure that the Statistics Diagnostics are on. The Diagnostics must be turned on in order to see the correlation coefficient. If the calculator has the latest version of the operating system (at this writing it is 2.55MP), the diagnostics can be turned on easily. Press the MODE key and arrow down until you see Stat Diagnostics. Make sure that the ON is highlighted. If the calculator does not have the latest version, to turn the calculator diagnostics on, press  to get into the CATALOG menu. Press the down arrow key until you get to "Diagnostics On". Once there, press ENTER twice.

Once the diagnostics have been turned on, calculate the LSRL: 
We will use STAT, CALC 4 although there are two linear regessions on the calculator. STAT, CALC, 8 is not in slope-intercept form, so that is why we will use #4.



Step 4: Interpret the results.

The correlation coefficient is 0.918. This means that there is a strong, positive, linear correlation between study time (in hours) and test grade (in percent).

The equation for the best fitting line is: $\hat{y} = 2.66x + 64.1$. Note the "y hat" – the y with a little hat over it. This is to distinguish it from the actual *y* values. The "y hat" or predicted *y* is the model and can be used to make predictions.

Defining slope and *y*-intercept in context: The *y*-intercept is 64.1. This means that if a student studies zero hours, the predicted test grade value would be 64.1%. The slope is 2.66. This means that as study time increases by one hour, on average, the test grade percentage will increase by 2.66%.

***These definitions of slope and *y*-intercept in context are important.

- *y*-intercept: It is the PREDICTED *y* value when the *x* value is equal to zero. (Sometimes this does not have any context in the problem. For example, if the *x* values were heights, zero would not be defined).

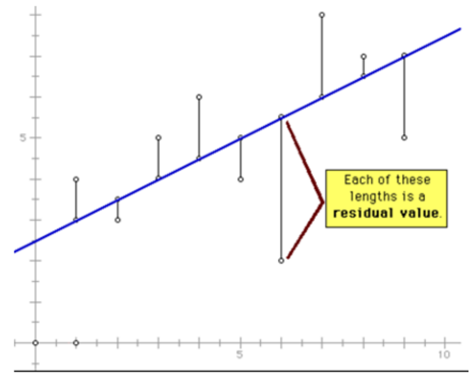
- slope: It is the average change (increase or decrease) in the *y* variable as the *x* variable increases by one unit.

<u>Residuals and residual plots:</u>

The distance between the predicted location of a point on a line and the distance from that line to an actual data point is called a **residual**. To find a particular residual:

residual = actual value - predicted value

residual = $y - \hat{y}$



The vertical distance between the two values is the residual. If the actual *y* value falls below the line, we have a negative residual. If the actual *y* value lies above the line, the residual is positive. The LSRL is the best fitting line because if the positive and negative residuals were added up, the sum would be zero. The positive "errors" and the negative "errors" are balanced out.

Ex 6: The relationship between the selling price of a car (in $1,000) and its age (in years) is estimated from a random sample of cars of a specific model. The model is given by the following equation: $\text{SellingPrice} = -1.182\text{Age} + 24.2$. Define the slope and y-intercept in the context of the problem.

y-intercept: On average, a new car costs about $24,200.00.

slope: For every year that the car gets older, on average, the selling price drops by $1,182.00.

Ex 7: Using the model above, predict the selling price of a car 5 years old.

$\text{SellingPrice} = -1.182(5) + 24.2$

$\text{SellingPrice} = 18.29$ Predicted selling price for a 5 year old car: $18, 290.00

Ex 8: Using the model in Ex 6, find the residual for a 3-year old car that sold for $21,000.00.

The data point that we are interested in is (3, 21). For this case, the actual *y* value is 21. First find the predicted *y* value by substituting 3 into the model and then use the formula to find the residual.

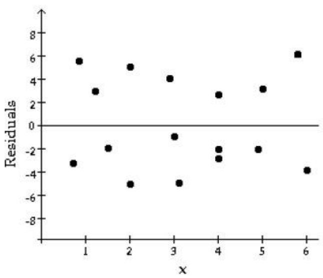
$\hat{y} = -1.182(3) + 24.2 = 20.654 \Rightarrow$ residual = $y - \hat{y}$

residual = $21 - 20.654 = 0.346$

Since the *y* variable is in the thousands, the residual for this data point is $346.00. The car's selling price was $346.00 above the predicted selling price using this model. A positive residual means the point lies above the LSRL.

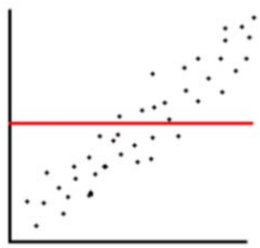
A **residual plot** is a scatter plot that shows the residuals on the vertical axis and the independent variable on the horizontal axis. The residual plot is used to assess if a linear model is appropriate. If the points in a **residual plot** are randomly scattered around the horizontal axis, a linear regression model is appropriate for the data. If there is curvature or a patter in the residual plot, a non-linear model would be more appropriate. The residual plot can be graphed on a graphing calculator and then assessed. When you run a regression on the calculator, the residuals are automatically created and input into a calculator-generated list named RESID. By plotting the *x* values along with the RESID values, we get the residual plot.

An example of a good residual plot:
Note the "random scatter"

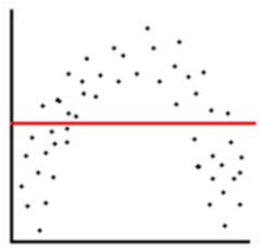


Examples of bad residual plots:

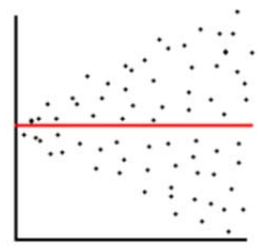
Linear Pattern



Curvature



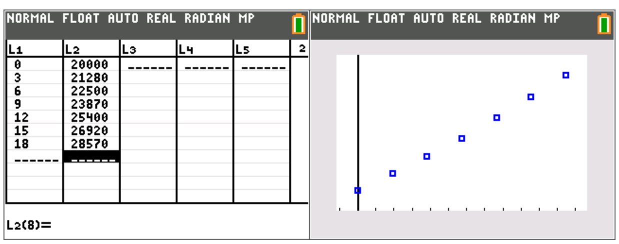
Fan Pattern



Ex 9: The approximate population growth of a town as a function of time starting at year 0 is shown in the table. Determine if a linear model is appropriate for this data.

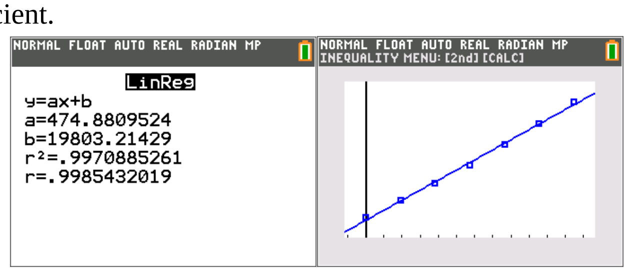
Year	Population
0	20,000
3	21,280
6	22,500
9	23,870
12	25,400
15	26,920
18	28,570

Step One: Create a scatter plot. It looks somewhat linear.



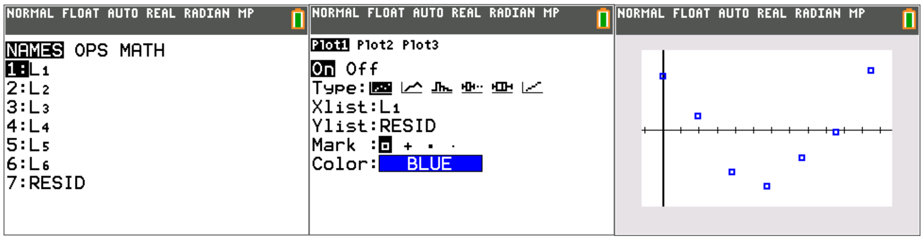
Step Two: Find the LSRL and the correlation coefficient.

The LSRL is $\hat{y} = 474.88x + 19803.2$ and the correlation coefficient is 0.9985. This states that there is a very strong, positive, linear relationship between year and population.



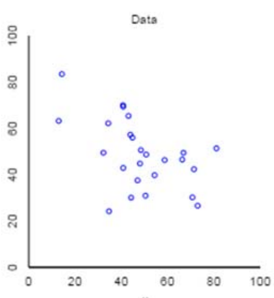
After running STAT, CALC, 4 and finding the LSRL, the calculator has created a list of the residuals. To plot them press 2nd, Y= to get into the Stat Plot Menu. Change the Ylist from L2 to RESID by doing the following: When the cursor is on L2, hit 2nd, STAT to get to the list names. Find RESID in the list and once on the name, hit enter. Plot the new scatter plot by hitting ZOOM 9.

Step Three: Plot the residual plot and assess the appropriateness.

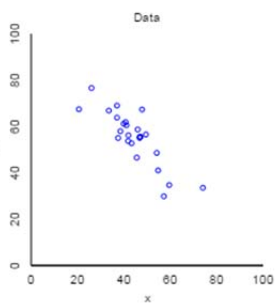


Note that the residuals have a pattern. There is definite curvature in the residuals. This tells us that the linear model is NOT appropriate and the LSRL should not be used to make predictions. The best fit for this data is actually exponential. This problem could be revisited in Unit 8.

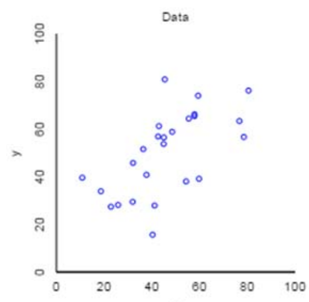
Ex 10: Guess the correlation coefficient of the scatterplots below.



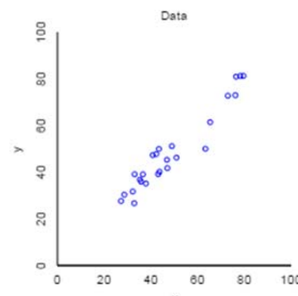
Actual: -0.508



Actual: -0.841



Actual: 0.611

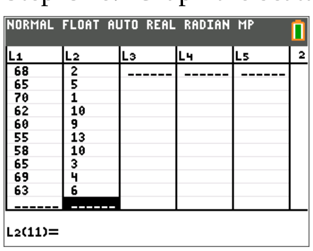


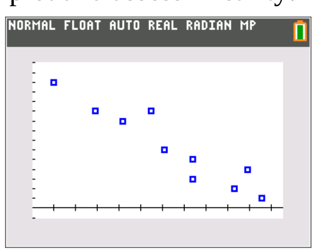
Actual: 0.962

Ex 11: In recent years, physicians have used the so-called diving reflex to reduce abnormally rapid heartbeats in humans by submerging the patient's face in cold water while the breath is held. A research physician conducted an experiment to investigate the effects of various cold temperatures on the pulse rates of ten small children. The results are presented below. Create a scatterplot and determine the graph of the line of best fit. Find the correlation coefficient and asses the residual plot.

Child	Temperature of Water (Fahrenheit)	Reduction in Pulse beats/min
1	68	2
2	65	5
3	70	1
4	62	10
5	60	9
6	55	13
7	58	10
8	65	3
9	69	4
10	63	6

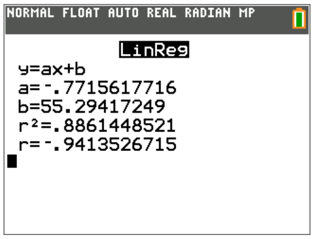
Step One: Graph the scatter plot and assess linearity.





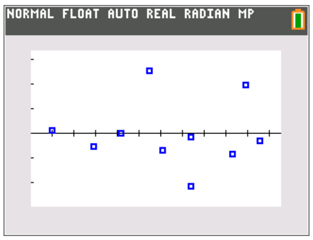
The graph appears linear with a negative relationship.

Step Two: Find the model and assess *r*.



The regression equation is $\hat{y} = -0.77x + 55.3$. The correlation coefficient is -0.941 which tells us there is a strong, negative linear correlation between temperature of the water and reduction in pulse.

Step Three: Plot the residual plot and assess.



The residual plot shows random scatter. There is no apparent curvature or pattern which tells us that the linear model is appropriate. We can use the model to predict observations.

Step Four: Use the model to predict the reduction in pulse for a temperature of 57 degrees.

$\hat{y} = -0.77(57) + 55.3 = 11.41$. We would predict that at 57 degrees, there would be an 11.4 beats/minute reduction in pulse rate.

Step Five: Would you feel comfortable using the model to predict for 80 degrees?

No, we should not use the model to make this prediction. The temperature of 80 degrees lies outside of our data set. This would be extrapolation. We do not know if the linear model will fit this data point.

Step Six: What is the residual for the point (68, 2)?

$$\hat{y} = -0.77(68) + 55.3 = 2.94 \quad \Rightarrow \quad \text{resid} = 2 - 2.94 = -0.94$$

The residual value is -0.94. This means that the point would fall below the LSRL and the model would OVERPREDICT.

<u>Correlation versus Causation</u>

In cases where we see strong correlation, we sometimes assume that the *x*-values, or independent variables, cause the *y*-values, or dependent variables. Just because two variables are strongly correlated, we cannot assume that one causes the other. A correlation may be a mere coincidence or the values of both variables might be caused by the values of a third variable that is unknown to us.

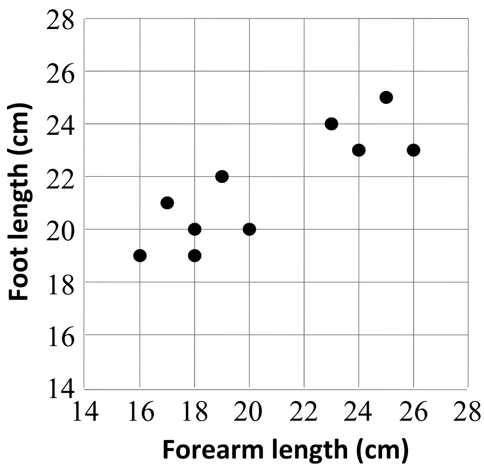
Example: If we find a high correlation between sales of a particular product and arrests for a particular crime, we cannot say that one causes the other.

Example: If we find that the frequency of a parent taking a child to the library is strongly correlated to the child's interest in books, we may be correct in assuming the frequency of library visits is caused by a child's interest in books.

Example: In the start of the 20th century, it was noted that there was a strong positive correlation between the number of radios and the number of people in insane asylums. Does this mean that listening to the radio will cause you to go crazy? No, there is another third variable – population – that is affecting them both. There was a population explosion so more radios were being purchased and more people needed asylums.

Sample Exam Questions

1. The scatterplot below represents the forearm lengths and foot lengths of 10 people.



Based on a linear model of the data, which is the best prediction for the length of a person's foot if his/her forearm length is 21 centimeters?

(A) 19 cm

(B) 20 cm

(C) 22 cm

(D) 24 cm

Ans: C

2. A scatterplot is made of a city's population over time. The equation of the line of best fit is $\hat{p} = 629t + 150{,}000$ where $\hat{p}$ is the city's predicted population size and t is the number of years since 2000. What is the meaning of the slope of this line?

(A) In 2000, the city's population was about 629 people.

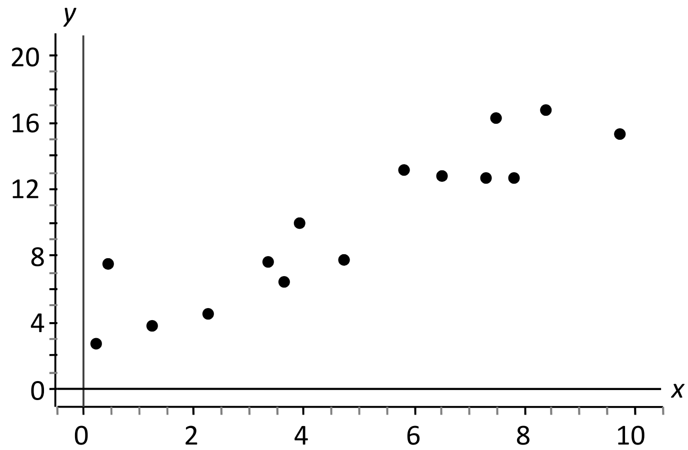
(B) In 2000, the city's population was about 150,000 people.

(C) The city's population increases by about 629 people each year.

(D) The city's population increases by about 150,000 people each year.

Ans: C

3. The line of best fit for the scatterplot below is $\hat{y} = 1.4x + 2.9$



Predict y when $x = 6$.

(A) 2.2

(B) 10.5

(C) 11.3

(D) 18.8

Ans: C

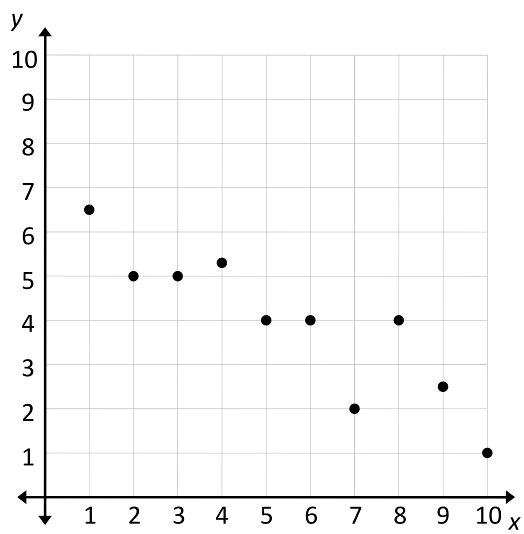
4. Which equation best describes fits the data shown in the scatterplot?

(A) $y = -\frac{3}{5}x + 7$

(B) $y = -\frac{1}{3}x + 8$

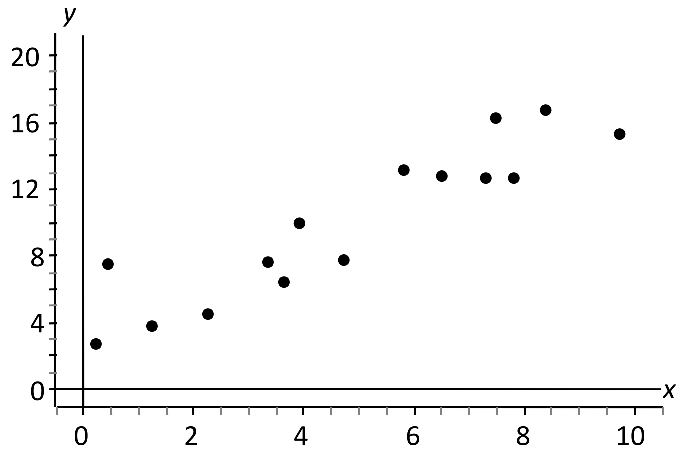
(C) $y = x + 8$

(D) $y = 4$



Ans: A

5. The line of best fit for the scatterplot below is $\hat{y} = 1.4x + 2.9$



What is the residual for the point (4, 10)?

(A) –1.5

(B) 1.5

(C) 8.5

(D) 10

Ans: B

6. The equation $\hat{y} = 31.4 - 0.12x$, gives the predicted population $\hat{y}$ of a city (in thousands) *x* years after 1975. What is meaning of the *y*-intercept?

(A) In 1975, the city's population was about 120 people.

(B) In 1975, the city's population was about 31,400 people.

(C) The city's population decreases by about 120 people each year.

(D) The city's population decreases by about 31,400 people each year.

Ans: B

7. The data below comes from a scatterplot.

x	2	3	4	5	6	7	8	8	8	9	10	10
y	2	8	4	1	10	4	6	10	2	7	3	9

Which best describes the linear relationship between *x* and *y*?

(A) Weak or no correlation

(B) Strong positive correlation

(C) Strong negative correlation

Ans: A

For questions 8-10, evaluate the truth of each statement about the correlation coefficient *r*.

8. A value of *r* near zero indicates there is a weak linear relationship between *x* and *y*.

(A) True

(B) False

Ans: A

9. A value of *r* = –0.5 indicates a weaker linear relationship between *x* and *y* than a value of *r* = 0.5.

(A) True

(B) False

Ans: B

10. A value of *r* = 1 indicates that there is a cause-and-effect relationship between *x* and *y*.

(A) True

(B) False

Ans: B

11. The equation $\hat{P} = -9.50m + 509$ gives the predicted price $\hat{P}$ of a particular style of television *m* months after the style first became available. What is the meaning of the *P*-intercept?

(A) The original price of the television was about $9.50.

(B) The original price of the television was about $509.00.

(C) The price of the television decreases by about $9.50 each month.

(D) The price of the television increases by about $509.00 each month.

Ans: B

For questions 12 and 13, use the following scenario.

A linear model describes the relationship between two variables, *x* and *y*. The correlation coefficient of the linear fit is *r* = –0.9.

12. The slope of the line of best fit is negative.

(A) True

(B) False

Ans: A

13. The linear relationship between *x* and *y* is weak.

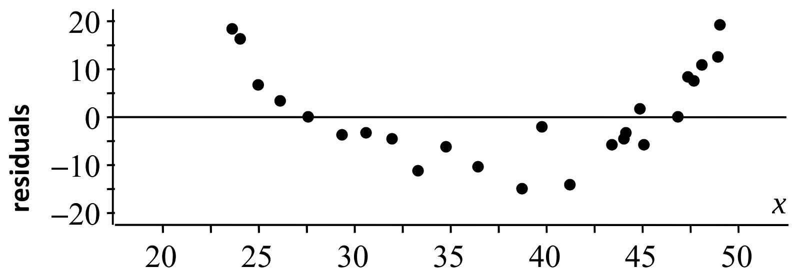
(A) True

(B) False

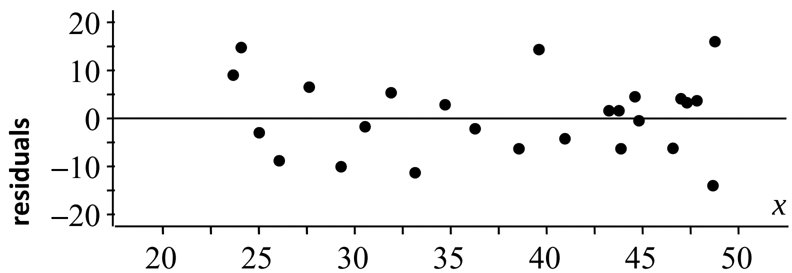
Ans: B

14. Two residual plots are shown below.

Plot I



Plot II



Which residual plot(s) would indicate a linear model is appropriate?

(A) Plot I only

(B) Plot II only

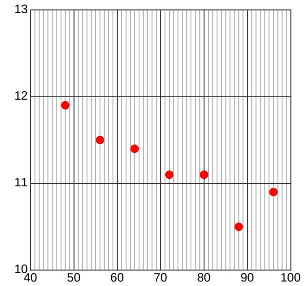
(C) Both Plot I and Plot II

(D) Neither Plot I nor Plot II

Ans: B

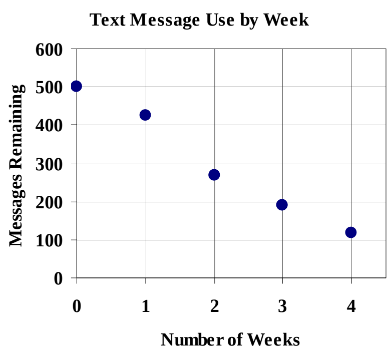
15. **Draw a scatter plot for the Women's Olympic 100-meter times since the year 1900. Find the equation of the best-fitting line. Determine any correlation. Predict the time of the women's 100-meter in the year 2000.**

Year (1900)	48	56	64	72	80	88	96
Time (sec)	11.9	11.5	11.4	11.1	11.1	10.5	10.9



SOLUTION: $t = -0.02y + 12.9$; There is a negative correlation. 10.9 seconds in the year 2000

16. Casey made a scatter plot to show the total number of text messages remaining on her plan throughout the month.



Which linear equation best models Casey's data?

A. $y = 500x - 100$

B. $y = 100x + 500$

C. $y = -x + 500$

D. $y = -100x + 500$ ANS: D

For #17-19, identify the correlation you would expect to see between each pair of data sets.

17. The temperature in Dallas and the number of cars sold in Philadelphia.

A. positive correlation

B. negative correlation

C. no correlation ANS: C

18. The number of members in a family and the size of the family's grocery bill.

    A. positive correlation
    B. negative correlation
    C. no correlation

    ANS: A

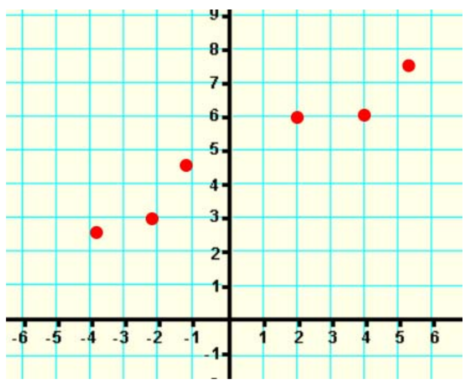
19. The number of times you sharpen your pencil and the length of your pencil.

    A. positive correlation
    B. negative correlation
    C. no correlation

    ANS: B

20. Based on the scatter plot, what would be the best estimation of the correlation coefficient, *r*?

    A. -0.8
    B. 0
    C. 0.4
    D. 0.8



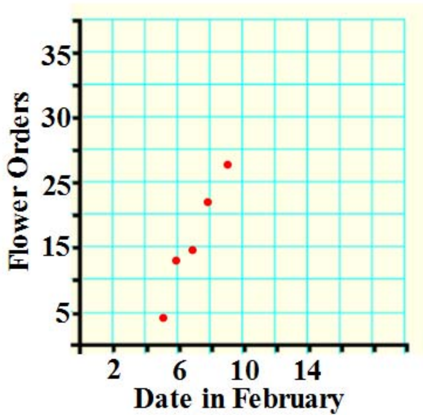
ANS: D

21. Rita had guests for dinner at her house eight times and has recorded the number of guests and the total cost for each meal in the table below. Graph a scatter plot of the data and determine the best-fitting line. Use the best-fitting line to predict the cost of dinner for 11 guests.

Guests	3	4	4	6	6	7	8	8
Cost ($)	30	65	88	90	115	160	150	162

SOLUTION: $c = 23.60g - 28.50$; 11 guests would cost about $231

22. The scatter plot shows the number of orders placed for flowers before Valentine's Day at one shop. Based on this relationship, predict the number of flower orders placed on February 12.



Solution: About 45

23. Which situation best describes a positive correlation?

A. The amount of rainfall on Fridays.

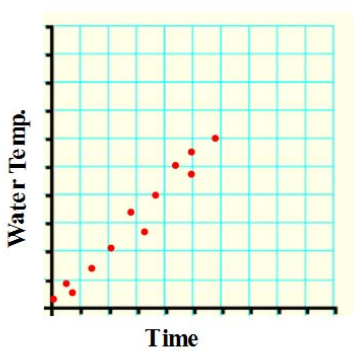
B. The height of a candle and the amount of time it has burned.

C. The price of a pizza and the number of toppings added to it.

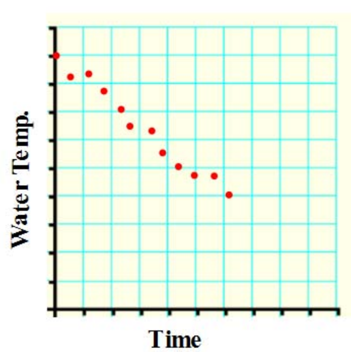
D. The temperature of a cup of hot chocolate and the length of time it sits. ANS: C

Students graphed a scatter plot for the temperature of hot bath water and time if no new water was added. Which graph is incorrect? Explain the error.

a.



b.



SOLUTION: Graph a is incorrect. The graph shows a positive correlation between time and water temperature. Temperature would go down as time passes, so the graph should be showing a negative correlation.