Name: _____ Date: _____

LINEAR REGRESSION AND LINES OF BEST FIT
COMMON CORE ALGEBRA II

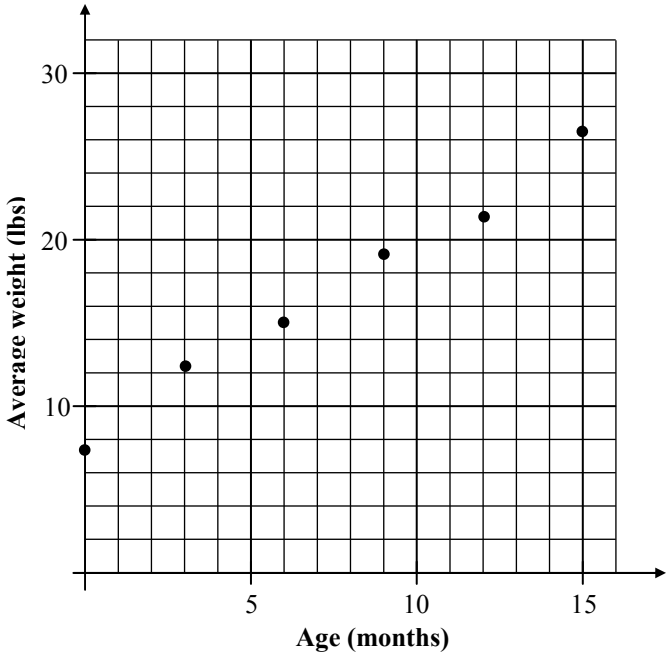
Oftentimes in science, a mathematical relationship between two variables is desired for predictive purposes. In the real world, the relationship between two variables is not always a perfect one, thus we often look for the "best" curve that can fit the data. Today we will review how to do this with a linear function.

***Exercise* #1:** A pediatrician would like to determine the relationship between infant female weights versus age. The pediatrician studies 100 newborn girls and finds their average weight at the end of 3 month intervals. The data is shown below and graphed on the scatter plot.

Age (months)	0	3	6	9	12	15
Average Weight (pounds)	7.2	12.2	15.1	19.4	21.5	26.3

(a) Using a ruler, draw a line that you think best fits this data. As a general guideline, try to draw it such that there are as many data points above the line as below it.

(b) By picking two points that are on the line (not necessarily data points), determine the equation of your best fit line. Round your coefficients to the nearest *tenth*.



(c) Using the linear regression command on your calculator, find the equation of the best fit line

(d) Use your calculator to determine the **linear correlation coefficient.** Round to the nearest *thousandth*. How can you interpret this value in terms of the variation in weight due to age?

***Exercise* #2:** Using the equation that your calculator produced in Exercise #1, predict the weight of a baby girl after 10 months. Round your answer to the nearest tenth of a pound.

The use of a model to predict outputs when the input is within the range of the known data is called **interpolation**. Interpolation tends to be fairly accurate.

***Exercise* #3:** Using the equation that your calculator produced in Exercise #1, predict the weight of a baby girl after 2 years. Round your answer to the nearest tenth of a pound.

The use of a model to predict outputs when the input is outside of the range of the known input data is called **extrapolation**. Models are most helpful when they can be used to extrapolate, but tend to be less accurate.

***Exercise* #4:** Biologists are trying to create a least-squares regression equation (another name for best fit line) relating the length of steelhead salmon to their weight. Seven salmon were measured and weighed with the data given below.

Length (inches)	22	24	28	34	39	42	48
Weight (pounds)	3.43	4.46	7.08	14.21	22.19	31.22	35.67

(a) Determine the least-squares regression equation, in the form $y = ax + b$, for this data. Round all coefficients to the nearest *hundredth*.

(b) Using your equation from part (a), determine the expected weight of a salmon that is 30 inches long.

(c) Using your equation from part (a), determine the expected weight of a salmon that is 52 inches long.

(d) In which part, (b) or (c), did you use interpolation and in which part did you use extrapolation? Explain.

Name: _____ Date: _____

LINEAR REGRESSION AND LINES OF BEST FIT
COMMON CORE ALGEBRA II HOMEWORK

FLUENCY

1. Which of the following linear equations would best fit the data set shown below?

(1) $y = 2.4x + 18.7$ (3) $y = -1.6x + 27.2$

(2) $y = -0.8x + 18.1$ (4) $y = 1.9x - 15.6$

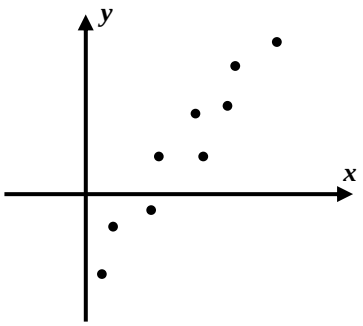
x	2	5	9	15
y	26	17	12	4

_____

2. A scatter plot is shown below. Which of the following *could* be the equation of the best fit line for the data set?

(1) $y = 1.8x - 3.2$ (3) $y = -2.9x + 8.3$

(2) $y = -3.5x - 12.4$ (4) $y = 6.5x + 3.9$



_____

3. A line of best fit was created for a data set that only included values of x on the interval $12 \le x \le 52$. For which of the following values of x would using this model represent extrapolation?

(1) $x = 26$ (3) $x = 14$

(2) $x = 50$ (4) $x = 6$

_____

4. Which of the following is true about the line of best fit for the data set given in roster form below?

(1) It has a positive slope and negative y-intercept.
(2) It has both a positive slope and y-intercept.
(3) It has both a negative slope and y-intercept.
(4) It has a negative slope and positive y-intercept.

$\{(0,-3),(2,4),(6,10),(15,12)\}$

_____

APPLICATIONS

5. An agronomist is studying the height of a corn plants as a function of the number of days since the corn germinated (appeared above the ground). Based on the following data, use your calculator to determine the best fit line in $y = ax + b$ form. Round all coefficients to the nearest *tenth*.

Time, x (days)	3	8	12	20	28	32	40
Height, y (inches)	2.5	4.5	6.2	9.3	12.9	14.4	16.8

6. Heavier cars typically get worse gas mileage (their miles per gallon) than lighter cars. The table below gives the weight versus the highway gas mileage for seven vehicles.

Vehicle Weight (thousands of pounds)	2.5	2.9	3.1	3.0	4.2	6.6	3.4
Gas Mileage (miles per gallon)	34	36	31	29	23	12	26

(a) Determine the best fit linear equation, in $y = ax + b$ form, for this data set. Round all coefficients to the nearest tenth.

(b) Using your model from part (a), determine the gas mileage, to the nearest mile per gallon, for a vehicle that weighs 3500 pounds.

(c) Is the prediction you made in (b) and example of interpolation or extrapolation? Explain.

(d) What is the value of the correlation coefficient to the nearest *hundredth*? Why is it negative?

7. The superintendent of the Clarksville Central School District is attempting to predict the growth in student population in the coming years. The table below gives the population for her district for selected years.

Year	1990	1992	1995	1997	2002	2005
District Population	3520	3605	3771	3860	4135	4285

(a) Find the equation for the line of best fit, in $y = ax + b$ form, where x represents the years *since* 1990 and y represents the district's population. Round all coefficients to the nearest *hundredth*.

(b) Use your model from part (a) to predict the district's population in the year 2020. Round your answer to the nearest whole number.

(c) What are the units of the slope of this linear model?

(d) What does the slope of this model represent? Think about your answer to part (c).