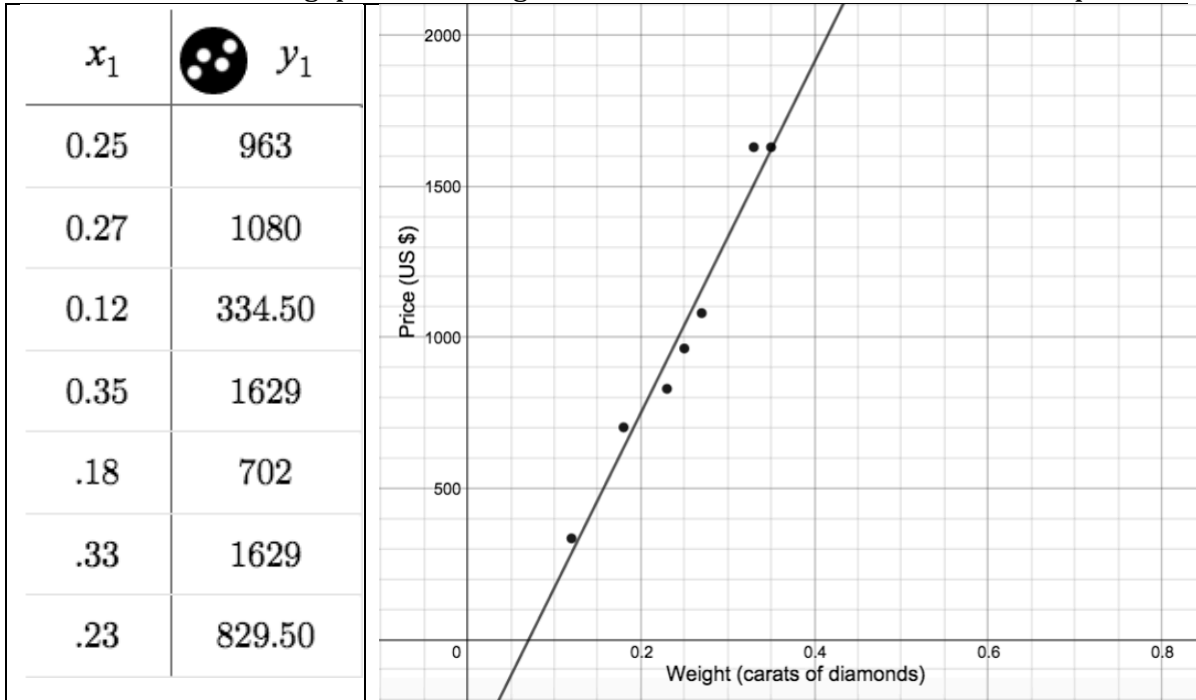


KEY

Linear Regression Practice Problems Name: _____

SECTION ONE

The following Desmos worksheet serves as a model for diamond prices (in US \$) as a function of weight (in carats of diamonds). The table shows data for seven diamonds sold at auction in Singapore. The “regression line” is also shown in the scatterplot.



Create this Desmos worksheet yourself

1. What challenges did you have in creating this sheet?

Various possible answers by student.

2. slope of regression line: **5806 approximately**

3. y-intercept of regression line: **- 411 approximately**

4. correlation coefficient: **0.984**

SECTION TWO

These questions all refer to the data set and model on the first page.

1. For the previous data, write an equation for the “regression line.”

Use $y = mx + b$ form where x represents weight (in carats of diamonds) and y represents price (in US \$).

$$y = 5806x - 411$$

2. Complete the following table correctly for the given data.

(You may ROUND “predicted y ” and “residual” to nearest WHOLE number.)

observed x	observed y	predicted y	residual	residual squared
0.25	963	1040	-77	5929
0.27	1080	1156	-76	5776
0.12	(a) 334.5	(b) 286	(c) 49	(d) 2401
0.35	1629	1621	8	(e) 64
0.18	702	634	68	4624
0.33	1629	(f) 1505	(g) 124	(h) 15376
0.23	829.5	924	-95	8930
(i) sum of squared residuals ?				43100

9025?

43195?

3. Jose found a “least squares line” for the data above that has a “sum of squared residuals” of 53000 .

Do you believe that Jose’s line is a good “least squares line” for this data? **No**

Why or why not? **The best fitting line above has sum of squared residuals of 43100 (43195?) so 53000 is not the smallest sum of squared residuals.**

4. Alyona used the same regression line you wrote in #1 above to predict the price of a 1-carat diamond.

What was her prediction? **$y = 5806(1) - 411 = 5395$**

Would you expect this to be a good prediction? **No**

Why or why not?

This is an EXTRAPOLATION which is much higher than any of the observed x values. Highest “observed x ” is 0.35.

5. Which of the following best describes the correlation between diamond weight and price: (circle your choice) **0.98**

- (a) strong positive (b) weak positive
(c) no correlation (d) weak negative (e) strong negative

6. Justify your answer to #5. (i.e. Why did you choose your answer?)

**A perfect positive correlation is +1.
This value (0.98) is very close to that.**

7. (a) Find \bar{x} , the mean (observed) weight of the diamonds: **0.247**

(b) Find \bar{y} , the mean (observed) price of the diamonds: **1023.857**

(c) Does the point (\bar{x}, \bar{y}) lie on the regression line? Why (not)?

Yes. The “center of mass” (0.247, 1023.857) lies on the regression line. If you “plug in” the value 0.247 into the regression equation, $y = 5806x - 411$, you get approximately 1023.857.

8. For any one set of data, how many different “regression lines” are there?


Exactly one.

9. If you have a set of data of (x,y) values and you reverse the two values in each ordered pair to get a new data set of (y,x) values, how do the correlation coefficients compare? Why?

They’re the same. The correlation coefficient describes the relationship between the x and y values. It does not change if we reverse the roles of x and y.

SECTION THREE

1. Find the line of best fit for the following data set:

x_1	 $f(x_1)$
0	4
1	14
2	24
3	34
4	44
5	54

$$y = 10x + 4$$

2. What observations would you make about this data set?

**The data points all lie on one straight line.
correlation coefficient is +1.
Data has a perfect positive correlation.**

3.

A regression line for a set of data is $y = 10x + 4$. If the sum of squared residuals is 0, what is the correlation coefficient?

Either +1 or -1.

SECTION FOUR

1. Julio found out the total sales tax collected in Florida each year from 1960-2000 and the numbers of shark attacks in those years.

He plotted “number of shark attacks vs. sales tax revenues” on a coordinate plane by plotting the data points (“sales tax revenue”, “number of shark attacks”).

He found the correlation coefficient to be 0.9 .

True or False?

True (a) In general, as the number of shark attacks increases, the sales tax revenue also increases.

False (b) In general, as the number of shark attacks increases, the sales tax revenue decreases.

False (c) An increase in sales tax revenue **causes** an increase in shark attacks.

Correlation is not causation!

True (d) Sales tax revenue has a relatively strong positive correlation with number of shark attacks.

SECTION FIVE

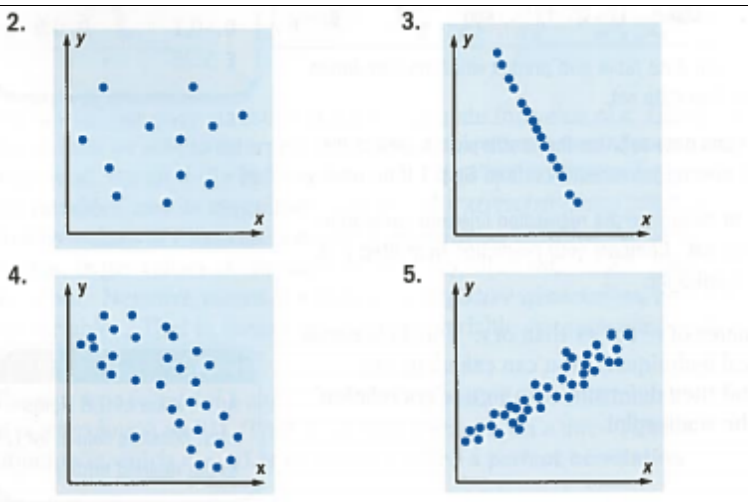
1. Of all the “good fitting lines” for a data set, how do we determine which one is the “best fitting line”?

The line for which the “sum of squared residuals” is the smallest.

For 2-5, match the following:

- (a) r is very close to -1
- (b) r is very close to -0.9
- (c) r is approximately -0.6
- (d) r is very close to 0
- (e) r is very close to +0.6
- (f) r is very close to +0.9
- (g) r is very close to +1

- 2. d
- 3. a
- 4. c
- 5. f



6. For a data set, residual equals ...

observed value minus predicted value.

7. Consider the linear model $f(x) = 7.3x + 4$.

(a) Find the “predicted value” at $x = 1$.

$$f(1) = 7.3(1) + 4 = 11.3$$

(b) Find the “observed value” if the residual at $x = 1$ is -0.2 .

$$\begin{aligned}\text{observed} - \text{predicted} &= -0.2 \\ \text{observed} - 11.3 &= -0.2 \\ \text{observed} &= -0.2 + 11.3 = 11.1\end{aligned}$$

(c) Find the observed value at $x = 3$ if the residual is 1.5 .

$$\text{predicted value} = f(3) = 7.3(3) + 4 = 21.9 + 4 = 25.9$$

$$\begin{aligned}\text{observed} - \text{predicted} &\text{ is } 1.5 \\ \text{observed} - 25.9 &= 1.5 \\ \text{observed} &= 1.5 + 25.9 \\ \text{observed} &= 27.4\end{aligned}$$