The Math Behind Correlation and LSRL **Graded Activity**

Name: _____ Class: _____ Date: _____

PART 1: Correlation

1) Complete the following table and find means, standard devitions and zscores for both X and Y.

		Calculate		
X	**Y**	**Z_X**	**Z_Y**	**$Z_X \bullet Z_Y$**
1	7	-1.250	-1.226	1.532
3	15	-0.555	-0.245	0.136
4	14	-0.208	-0.368	0.077
7	20	0.833	0.368	0.306
8	29	1.180	1.472	1.737
	∑	**0.000**	**0.000**	**3.788**

$Z_X = \frac{X_i - XBAR}{Sx}$

$Z_Y = \frac{Y_i - YBAR}{Sy}$

XBAR= 4.60

S_X = 2.881

n=5

YBAR= 17.00

S_Y = 8.155

What do these summations mean?

The sum of ZScores for each variable X and Y is zero; which is their means. Remember the mean of Zscores is 0 and their standard deviation is 1.

2) Calculate the Correlation Coefficient by hand using the following formula.

$$\text{Calculate the Corr. Coef. (r) } = \frac{\sum Z_X \bullet Z_Y}{n-1} = \frac{3.788}{4} = 0.9471$$

3) Compare your calculated Correlation Coefficient against the "r" for the observed values of X and Y using your calculator. T**hey should be the same. If not, go back and look for your mistake.**

The coer. coef. for the observed (actual) data: r=.94706

The coer. coef. for the zscores is also: r=.9471

4) Rewrite the above Correlation Coefficient calculation using mean, standard deviation, sample size with appropriate subscripts and usage of sigma.

$$r = \frac{\sum (X_i - XBAR)/Sx) \bullet (Y_i - YBAR)/Sx}{n-1} \quad \text{OR}$$

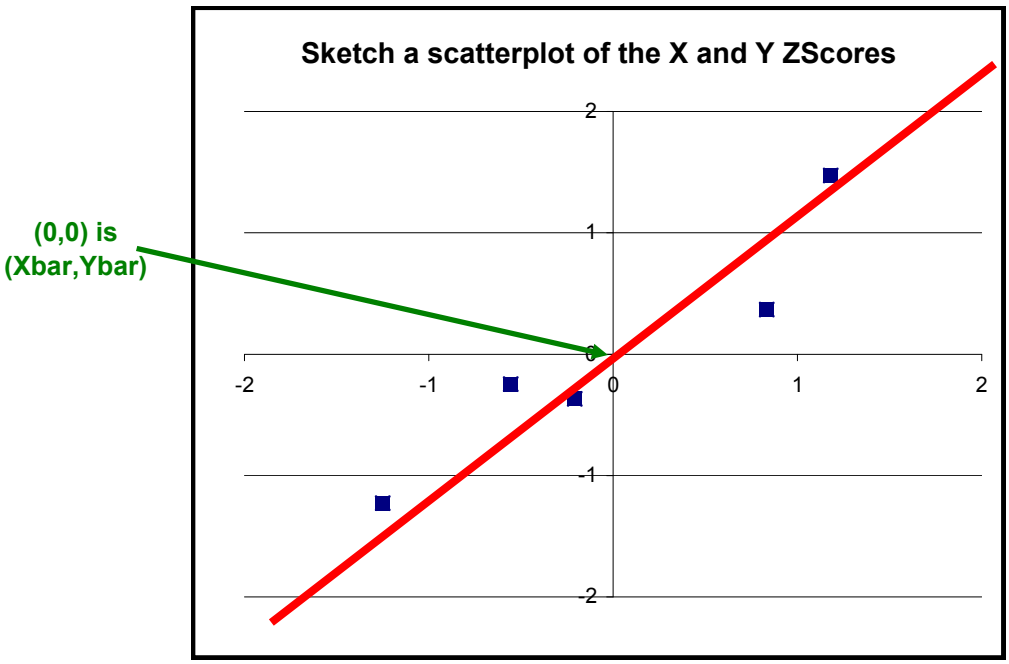
$$r = \frac{1}{n-1} \sum \left[\frac{X_i - XBAR}{Sx} \right] \left[\frac{Y_i - YBAR}{Sy} \right]$$

i=obs #

The Math Behind Correlation and LSRL

Graded Activity

5) Now, create a scatterplot of your Zscores.



6) Find the regression equation for the Zscores. Plot the line on your scatter plot

a) Write the regression equation: **Yhat = 0 + .947X**

b) What is the slope **0.947**

c) What is the yintercept **0 or (0,0)**

In your own words, how does the y-intercept illustrates the concept of regression towards the mean?

The least square regression line is going through the mean-mean value for both X and Y.
because for standardized data the mean is zero.
So for standardized data (Xbar,Ybar) is thet point (0,0); which is also the y-intercept.

d) What is the Correlation Coefficien for the ZScores: **0.947**

e) Has the correlation changed from the observed data to the zcsores? **NO**
EXPLAIN what you think this means?
The correlation coef. does not change when we add or subtract constants or multiply it by positive constants.

f) What do you notice about the slope of zscores for X and Y? **Slope(b) and r are the same.**
The slope is defined as follows: b =r *(Sx/Sy)
For standardized data, the standard devitation is 1, hence b=r.

7) Some key questions to answer:

a) What are the units for Zscores? **Zscores have no units**

b) What are the units for Correlation? **Correlation has no units**

c) What are the 3 conditions for Correlation? **Linearity, no outliers, quantitative data**

Graded Activity

PART 2: LSRL (least square regression line)

(8) On a sheet of graph paper plot X and Y and use units of 1
(i.e use the entire page - leave the bottom 8 lines blank).

(9) Use a color pencil and eyeball your best estimate of the regression line.
Draw this line lightly and used a dashed line.

a) Write the regression equation FOR YOUR LINE: **Yhat = 7 + 2X**

10) Now use your calculator to find the LSRL. Plot 2 points and label the ordered pairs. Now use a different color and draw this line.

a) Write the LSRL regression equation: **Yhat = 4.67 + 2.68X**

b) What is the slope **2.68** c) What is the yintercept **4.67**

c) comment on how close your line was to the actual LSRL:

My slope was in the ball park but my y-intercept was too high

11) Now use your calculator to find the LSRL. Plot 2 points and label the ordered pairs. Now use a different color and draw this line. **Possible points:**

1) the yintercept (0,4.67)
2) mean-mean (4.6, 17)
3) pick a large value of X to draw a straight line (10,31.47) ***Yhat=4.67+2.68(10)=31.47***

~~12) Now use a third color pencil and draw the residual lines. Residuals are the vertical lines from the observed point to the point on our predicted LSRL.~~

13) Make the necessary calculations to fill in the following table.

yhat=predicted value for Y

Calculate

X	Y	YHAT	Y-YHAT	$(Y-YHAT)^2$
1	7	7.349	-0.349	0.122
3	15	12.711	2.289	5.240
4	14	15.392	-1.392	1.936
7	20	23.434	-3.434	11.791
8	29	26.114	2.886	8.326
	Σ		**0.000**	**27.416**

aka residuals =
OBSERVED - PREDICTED

aka residuals^2
(also called SSRESID - Residual Sum of Squares)

XBAR= 4.60
S_X = 2.881

YBAR= 17.00
S_Y = 8.155

r= 0.947
r^2= 0.897

What does this summation mean in context?
The sum of the residuals (deviations between actual and predicted) is zero.

Why do we have to square?
Like with standard deviation, we want the average deviation and square to set a positive value (otherwise, the average deviation is zero).

The Math Behind Correlation and LSRL

Graded Activity

14) Calculate the slope of the LSRL using the following formula.

$$b = r \cdot \frac{S_Y}{S_X} \qquad \frac{0.947 \quad \frac{8.155}{2.881}}{} = 2.681$$

15) Graph and label the line Y=Ybar

16) Plot and label the coordinates of the MEAN-MEAN (xbar,ybar) point.

In your own words, explain the concept of regression towards the mean?

There are many lines that could be drawn through a data set.

The least square regression line always goes through the mean-mean value for both X and Y.

The mean for X is 4.6; and the mean for Y is 17.

So you can see from the graph the mean-mean point, (4.6, 17), is on the least square regression line.

17) The RESIDUALS are (Y-YHAT). Look at your graph and check the residuals for each point. . Label each residual.

18) Create a residual plot. Use the space on the bottom of your graph paper. Plot X vs Residual and use the same scale as the x-axis as your scatter plot. What is the pattern of the residual plot and what does this mean?

To assess the overall fit of a line, the residual plot gives the "big" picture of how well our line fits the data.

A residual plot that displays a random scatter of points confirms that our linear model is appropriate for our data.

We should also look to see if there are any unusually large residuals that could be outliers.

19) Calculate the standard deviation for the residuals using the following formula.

$$s_e = \sqrt{\frac{\sum \text{residuals}^2}{n-2}} = \frac{27.416}{3} = 3.023$$

take square root of entire expression

In your words, interpret s_e

Roughly speaking, s_e is a typical amount by which an observation deviates from the LSRL. Remember that standard deviation for one variable is a typical distance from the mean. Now for 2 variable data, we measure typical vertical distance from the LSRL.

The Math Behind Correlation and LSRL

Graded Activity

20) What is r^2? a) r^2 = **0.89**

b) Give it's actual name of and write the definition: b) r^2 = **Coefficient of determination**

Definition:

The *coefficient of determination*, r^2, gives the proportion of variability in y that can be explained by the linear association with x.

r^2 represents the percent of the data that is the closest to the line of best fit and $0 \leq r^2 \leq 1$

c) In your own words, explain what r^2 means in the context of this problem.

In this example, 89% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation).

The other 11% of the total variation in y remains unexplained.

21) Write a few paragraphs summarizing the process of LSRL.

1) check scatter plot for linearity
2) find the mean and standard devitions for x and y.
3) check r to find the strenth of the linear association
4) look at residual plot to make sure the pattern is random and look for large residuals
5) check r2 to see the strength of the model
6) find the predicted equation Yhat = a + bx
7) use the LSRL equation to make predictions but remember extrapolation is dangerous.

See AP GREEN formula sheets for equations you must be able to use

$$\hat{y} = b_0 + b_1 x$$

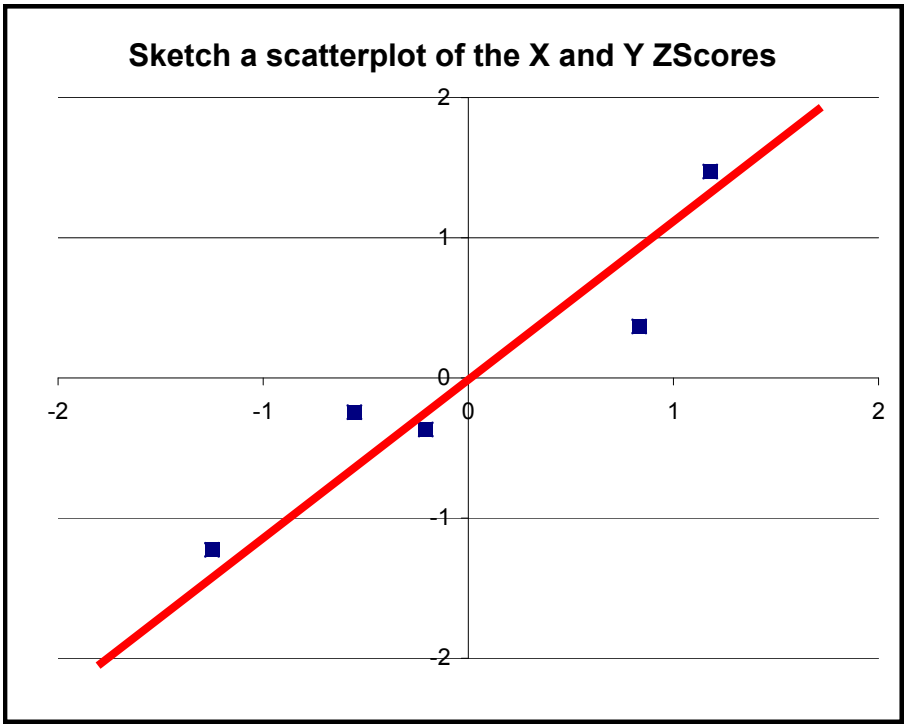
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})}$$

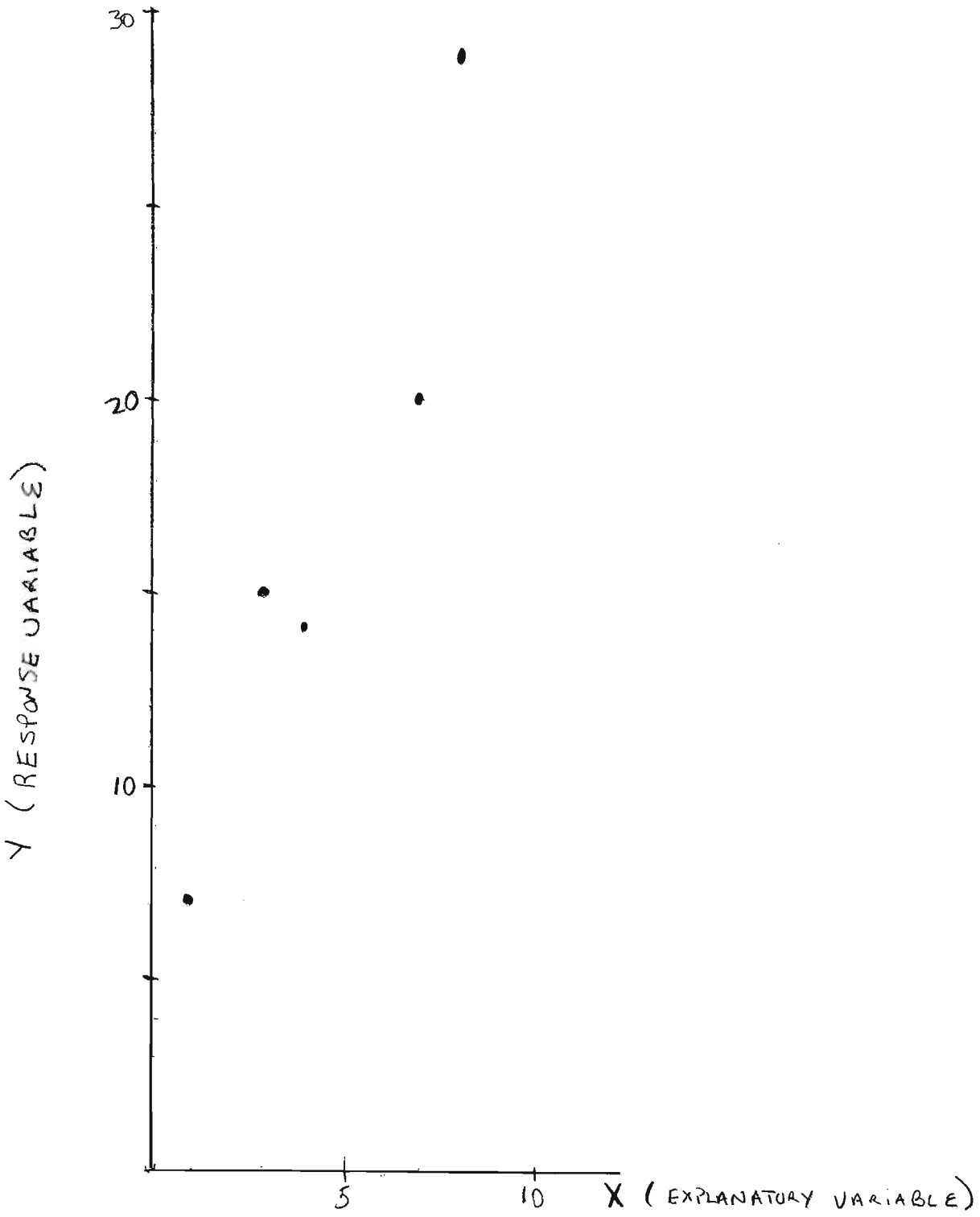
$$r = \frac{1}{n-1} \sum \left(\frac{(x_i - \bar{x})}{s_x} \right) \left(\frac{(y_i - \bar{x})}{s_y} \right)$$

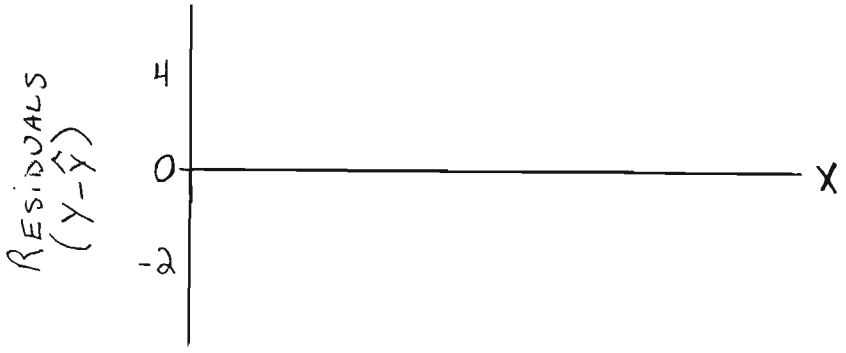
$$b_1 = r \frac{s_y}{s_x}$$

5) Now, create a scatterplot of your Zscores.

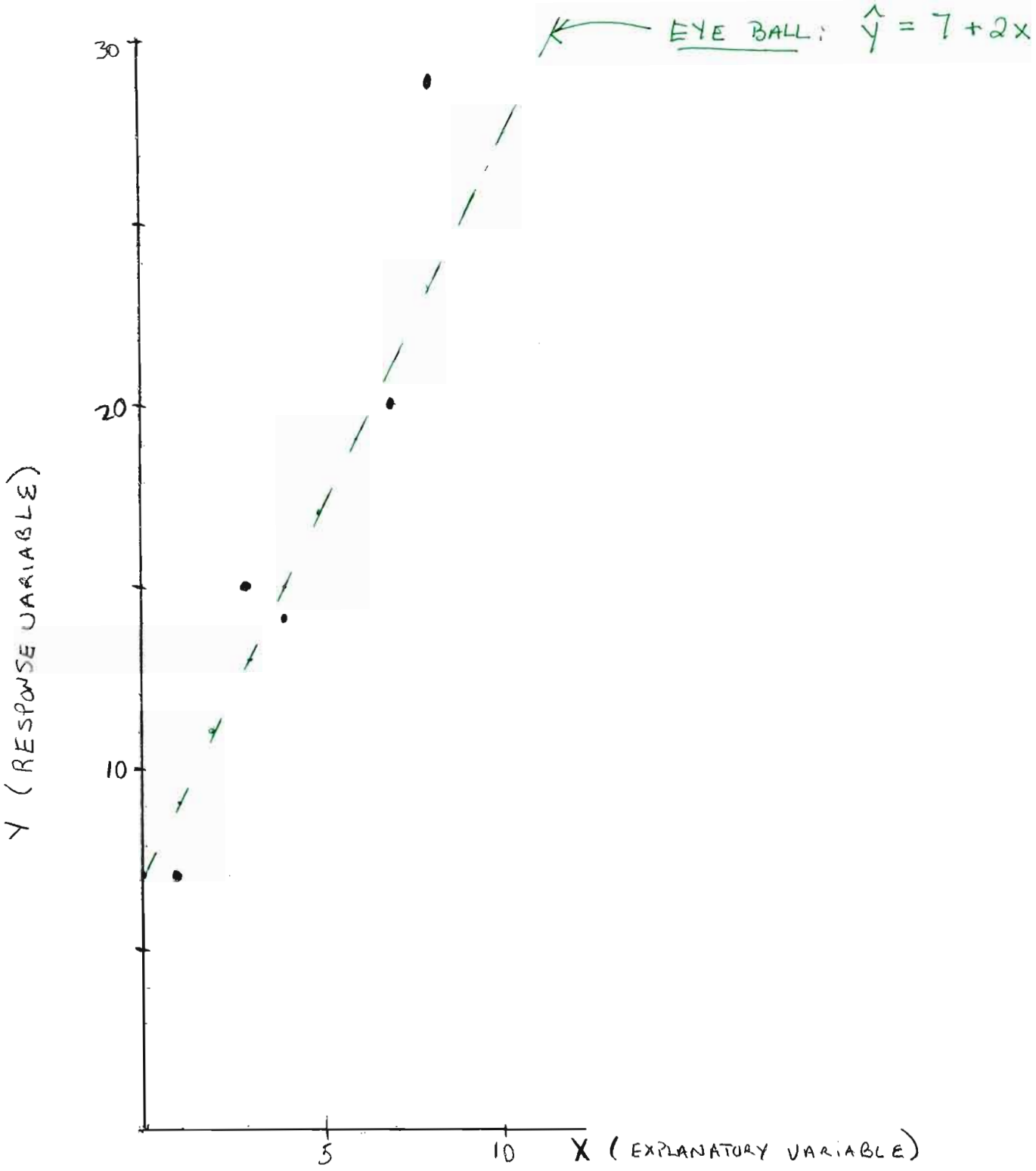


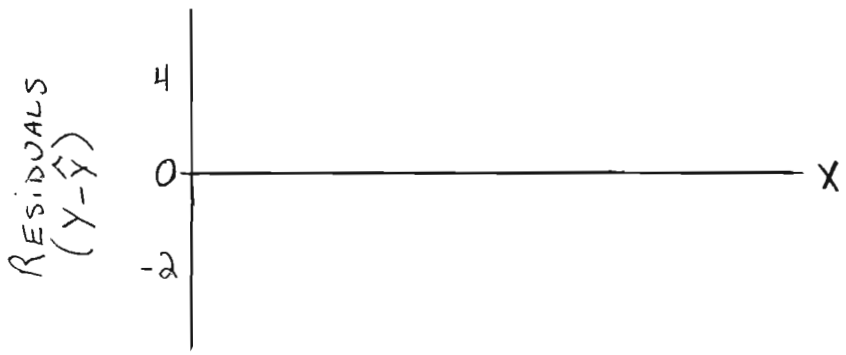
THE MATH BEHIND CORRELATION AND LSRL



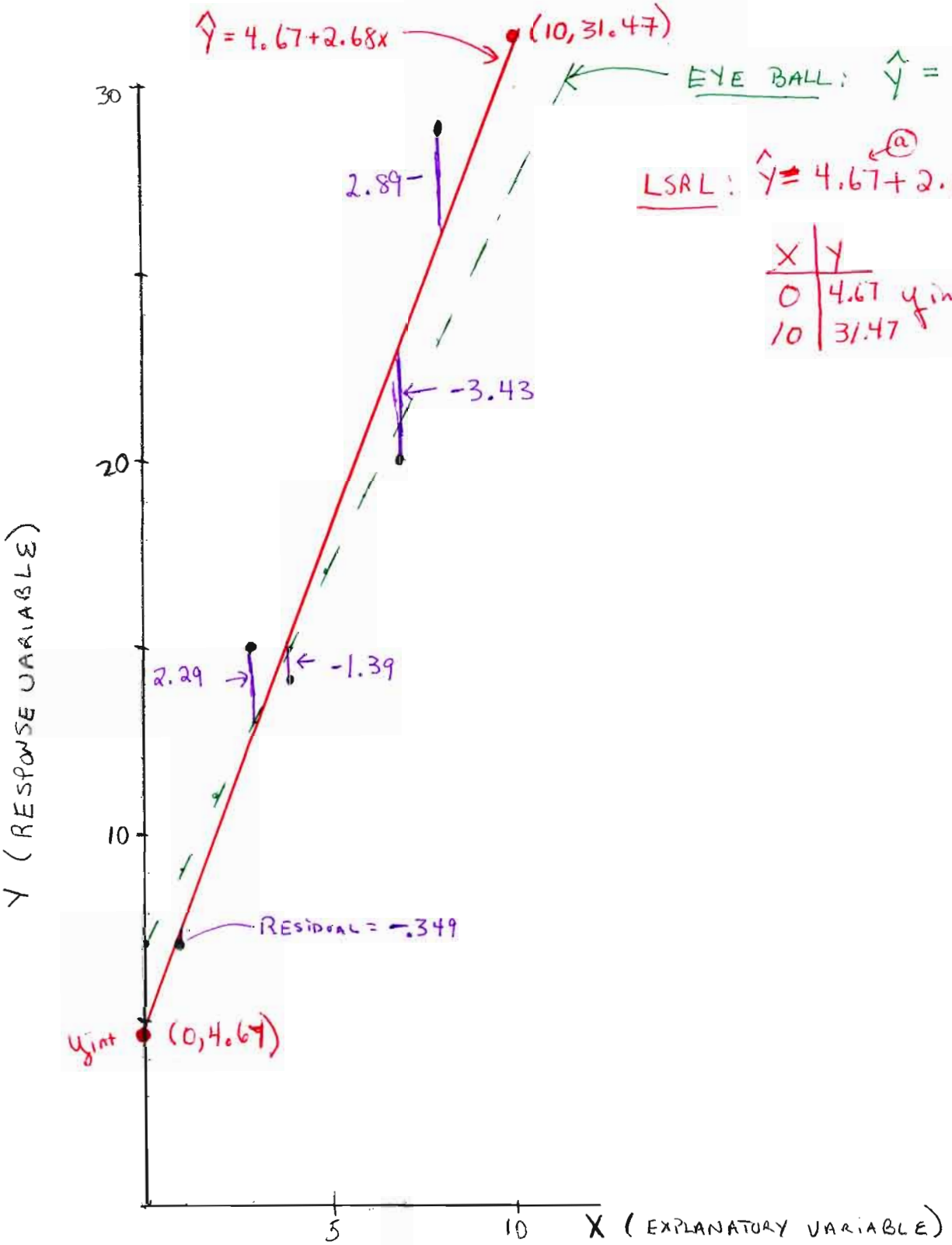


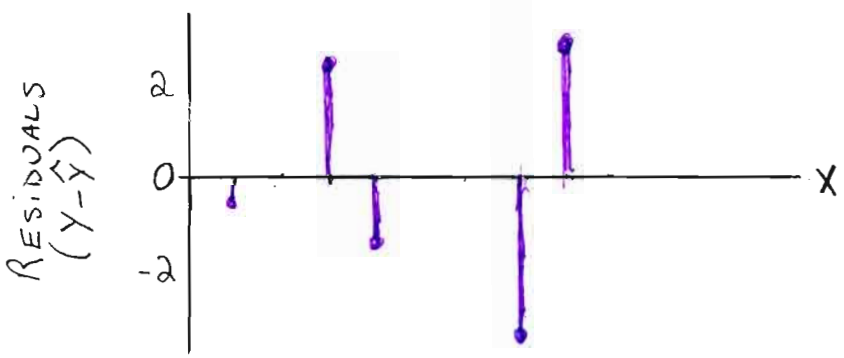
THE MATH BEHIND CORRELATION AND LSRL



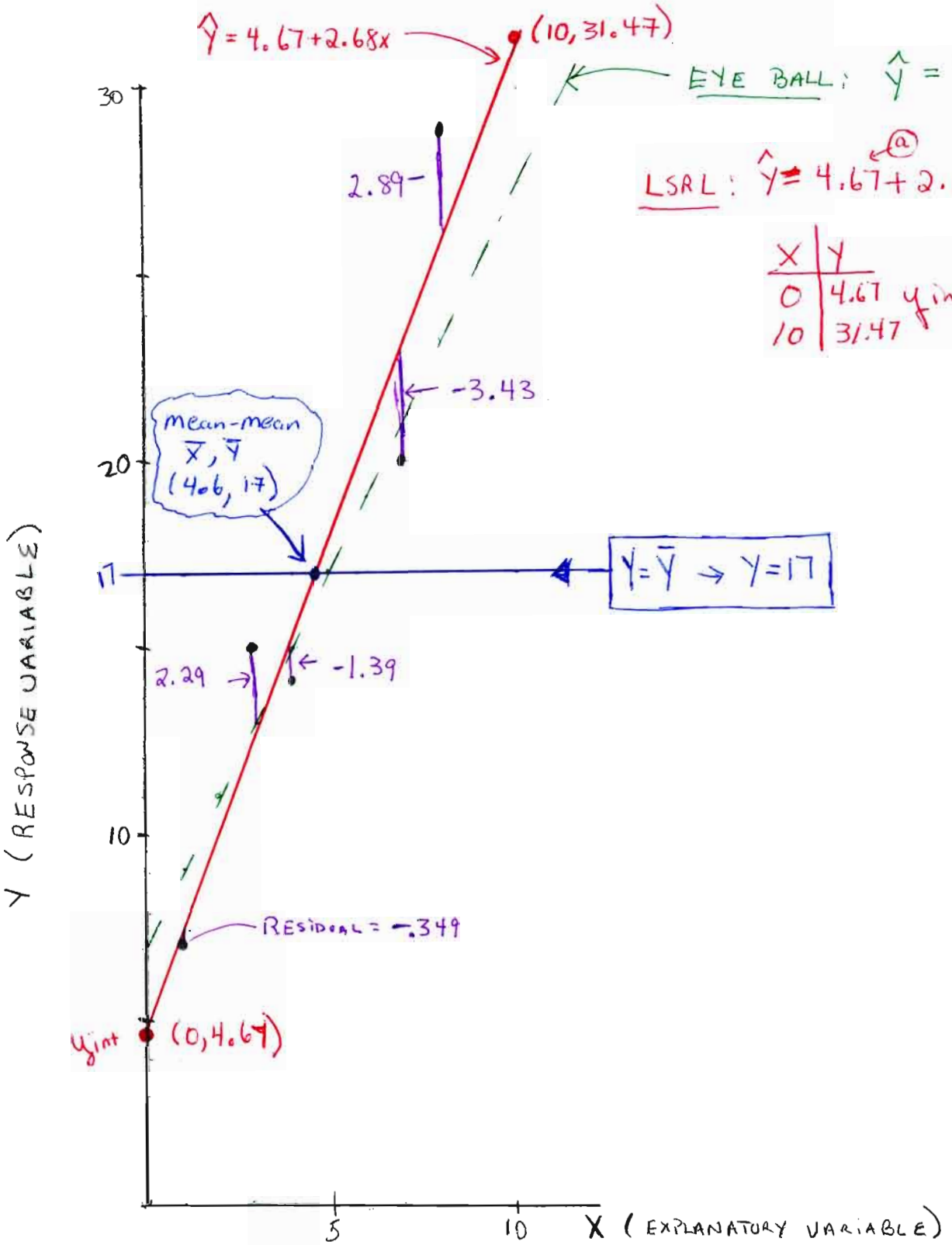


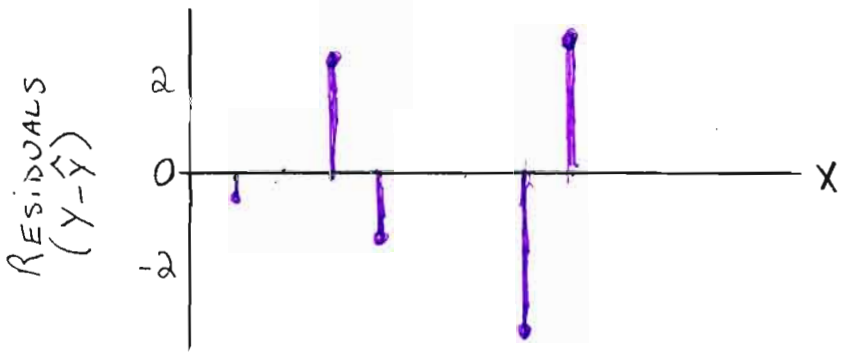
THE MATH BEHIND CORRELATION AND LSRL





THE MATH BEHIND CORRELATION AND LSRL





RESIDUALS IN PURPLE