14

Notes:

Residual Plot

Created for you by MS. Nhotsoubanh

The **correlation coefficient, r** is a measure of how well the data set is fit by a model. In other words, how well it fits the line of best fit.

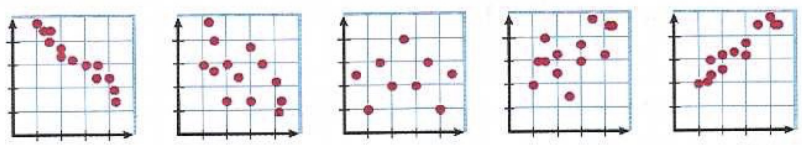
Properties of the Correlation Coefficient *r*

r is a value in the range $-1 \le r \le 1$.

If $r = 1$, the data set forms a straight line with a positive slope.

If $r = 0$, the data set has no correlation.

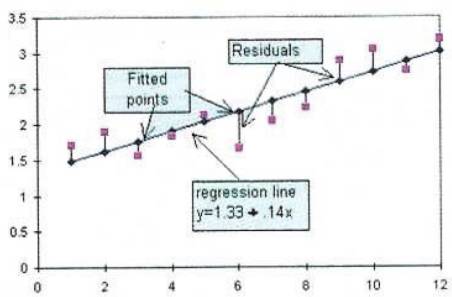
If $r = -1$, the data set forms a straight line with a negative slope.



$r \approx -0.95$ $r \approx -0.6$ $r \approx 0$ $r \approx 0.6$ $r \approx 0.95$

A scatter plot can be modeled as a linear function using the line of best fit, or a linear regression. This line then needs to be tested for a goodness of fit. If a graph is to be a good predictor of data, the line must be accurate in predicting data. This can be found by analyzing the residuals.

Residuals are **error distances**. They measure the goodness of a fit. It is the measurement of how far the data fall from the line of best fit.



To find the residuals of a regression model you find the difference between the observed value (from a table) and the models predicted value (from the equation of regression).

Residual = [Observed Value] – [Predicted Value]

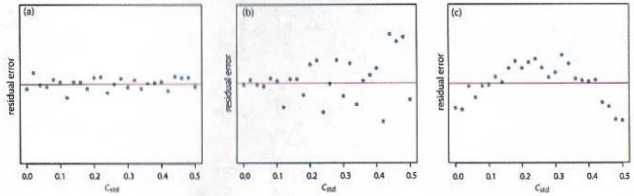
Name: _____

Alg. 1 H – Date: April 15

Glue on page 16

Residual Plots

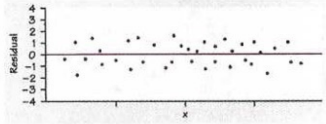
- A Residual Plot is a scatterplot of all of the residual values. They help us assess the fit of a regression line.
- If the regression line captures the overall relationship between x and y, the residuals should have **no systematic pattern**.



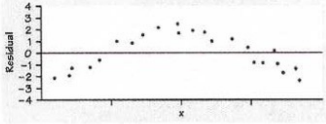
The residuals should add up to 0. If they do add up to zero then the regression line was properly calculated. The goal is to minimize all of the residuals. We want the residuals to be "short" or close to the x-axis. The "longer" or more dispersed the residuals, the further they are from the x-axis.

If the residuals are clustered towards the x-axis, then the regression line is a good predictor of the data.

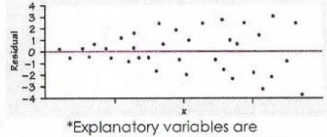
Analyzing Residuals Graphically



The uniform (random) scatter of points indicates that the regression line fits the data well, so the line is a good model.



A curved pattern shows that the relationship is not linear. The points are non-random.



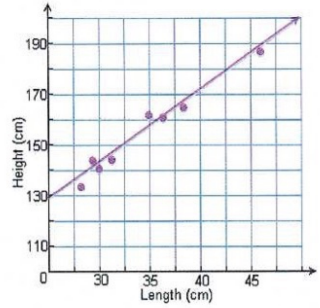
Increasing or decreasing spread about the line. The response variable y has more spread for larger values of the explanatory variable x, so the prediction will be less accurate when x is large.

*Explanatory variables are variables that are tested to try and show causes in statistics.

Ex 1) Anthropologists can use the femur, or thighbone, to estimate the height of a human being. The table shows the results of a randomly selected sample.

a. Make a scatter plot of the data with the graphing calculator. Then compare with the graph below.

Femur Length and Height (cm)			
Length	Height	Predicted Value	Residual Value
36	160	158.8	1.2
32	143	147.1	-4.1
46	187	187.9	-0.9
29	142	138.4	3.6
35	161	155.9	5.1
38	164	164.6	-0.6
30	140	141.3	-1.3
27	131	132.6	-1.6



b. Find the correlation coefficient *r* and the line of best fit.

Line of best fit equation: $y = 2.91x + 54.04$

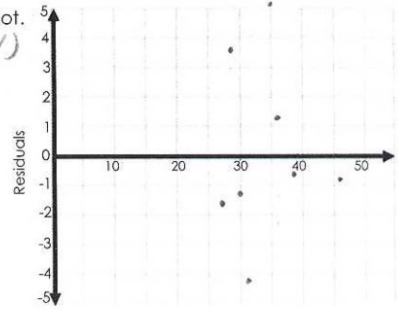
The correlation coefficient is 0.986. What type of correlation does it have? strong positive

c. Interpret the slope of the line of best fit in the context of the problem.

The slope is about 2.91, so for each 1cm increase in femur length, the predicted increase in human's height is 2.91 cm

d. Construct a Residual plot.

(x, residual)
(36, 1.2)
(32, -4.1)
(46, -0.9)
(29, 3.6)
(35, 5.1)
(38, -0.6)
(30, -1.3)
(27, -1.6)



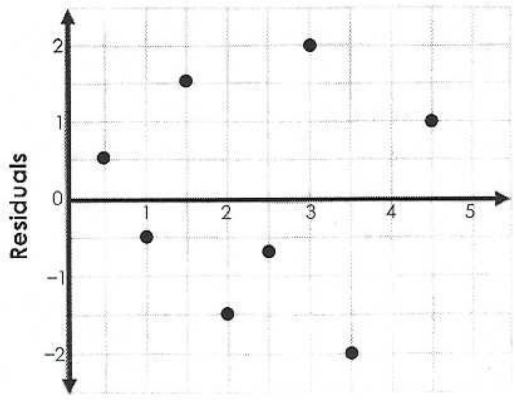
Practice Problems:

Name: Key

Glue on page #17

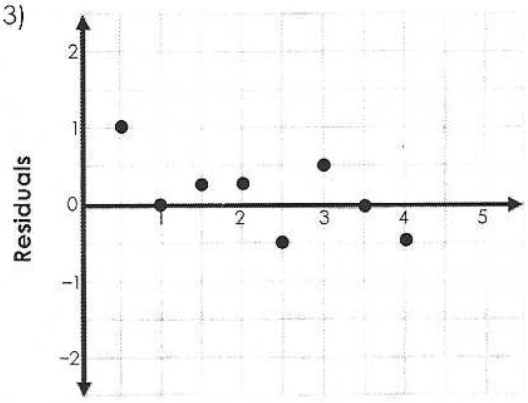
Identify if the linear regression line is a good predictor of the data based on the residuals plot. State why.

Ex 2)



no, this is not a good predictor b/c the residuals are dispersed from the x-axis

Ex 3)

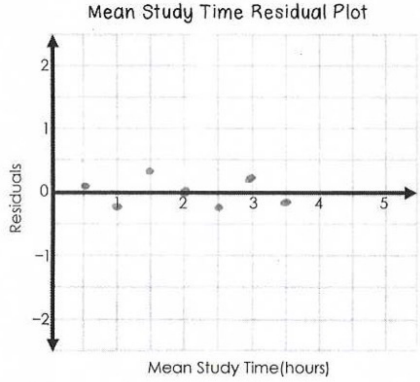


yes, this is a good predictor b/c the residuals are closer to the x-axis. The residuals are random & do not form a pattern

Ex 4) Using the data that Sarah collected from Class 1, let's calculate the residuals.

Class 1 Test Score Analysis			
Mean Study Time (in hours)	Mean Test Score (observed)	Regression Model (Predicted) $y = 8.7x + 58.6$	Residual
0.5	63	62.95	0.05
1	67	67.3	-0.3
1.5	72	71.65	0.35
2	76	76	0
2.5	80	80.35	-0.35
3	85	84.7	0.3
3.5	89	89.05	-0.05

- First calculate the predicted values by using the regression model and the mean study time.
- Subtract the Observed value and the Predicted Value
- Graph the residual



a) What do all the residuals add up to?

$0.05 - 0.3 + 0.35 + 0 - 0.35 + 0.3 - 0.05 = 0$

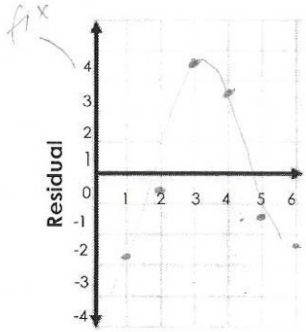
b) Would this regression line be a good predictor of the data?

yes, b/c the residuals are clustered towards the x-axis

Ex 5) Construct the residual plot. Round answers to the nearest tenth.

Line of Best Fit Equation: $y = 4.88x + 3.8$

x	y (Observed Value)	Predicted Value	Residual Value
1	6	8.7	-2.7
2	13	13.6	-0.6
3	22	18.4	3.6
4	26	23.3	2.7
5	27	28.2	-1.2
6	31	33.1	-2.1



Does the residual plot suggest a linear relationship? Explain.

no, b/c the residuals form a curve pattern.