LESSON 16 Visualizing and Analyzing Residuals

LEARNING OBJECTIVES

- Today I am: creating a residual plot.
- So that I can: determine if a line was the best model for my data.
- I'll know I have it when I can: use a residual plot to tell if a set of data is linear or nonlinear.

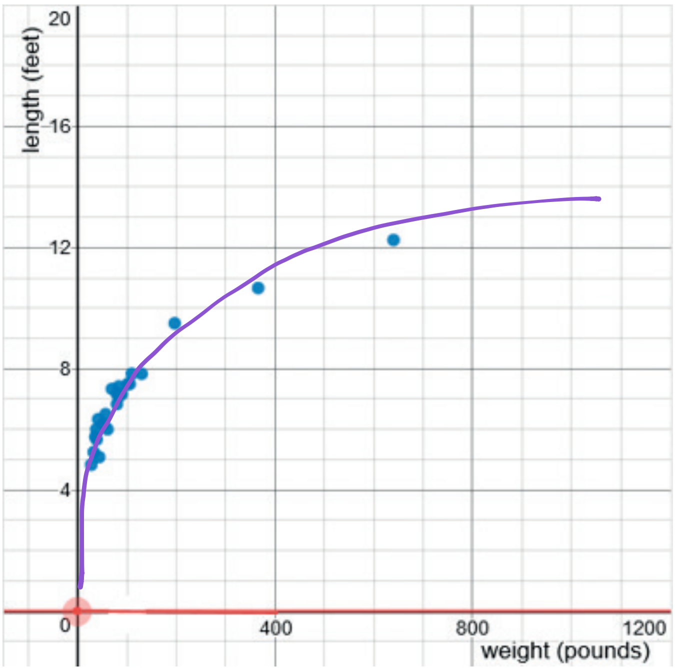
Opening Exploration

You will need: a Chrome book

An enormous alligator lurks in the swamp. Can scatterplots and least-squares regression tell if you have enough animal tranquilizers to stay safe?

1. Go to student.desmos.com and type in your teacher's class code for the Desmos activity *Alligator Swamp*.

2. Gia's equation, $y = \sqrt[3]{1.54x - 28.93} + 2.52$, was pretty complicated. Do you think a piecewise function may have worked for the alligator data? Explain your thinking.



3. Next, we'll look at residual plots with the Desmos activity *Line of Best Fit*. Again, you'll use your teacher's class code.

4. What did you notice in the last screen if you dragged the black line onto the residual axis? Why does this happen?

Residual plot → Scatterplot

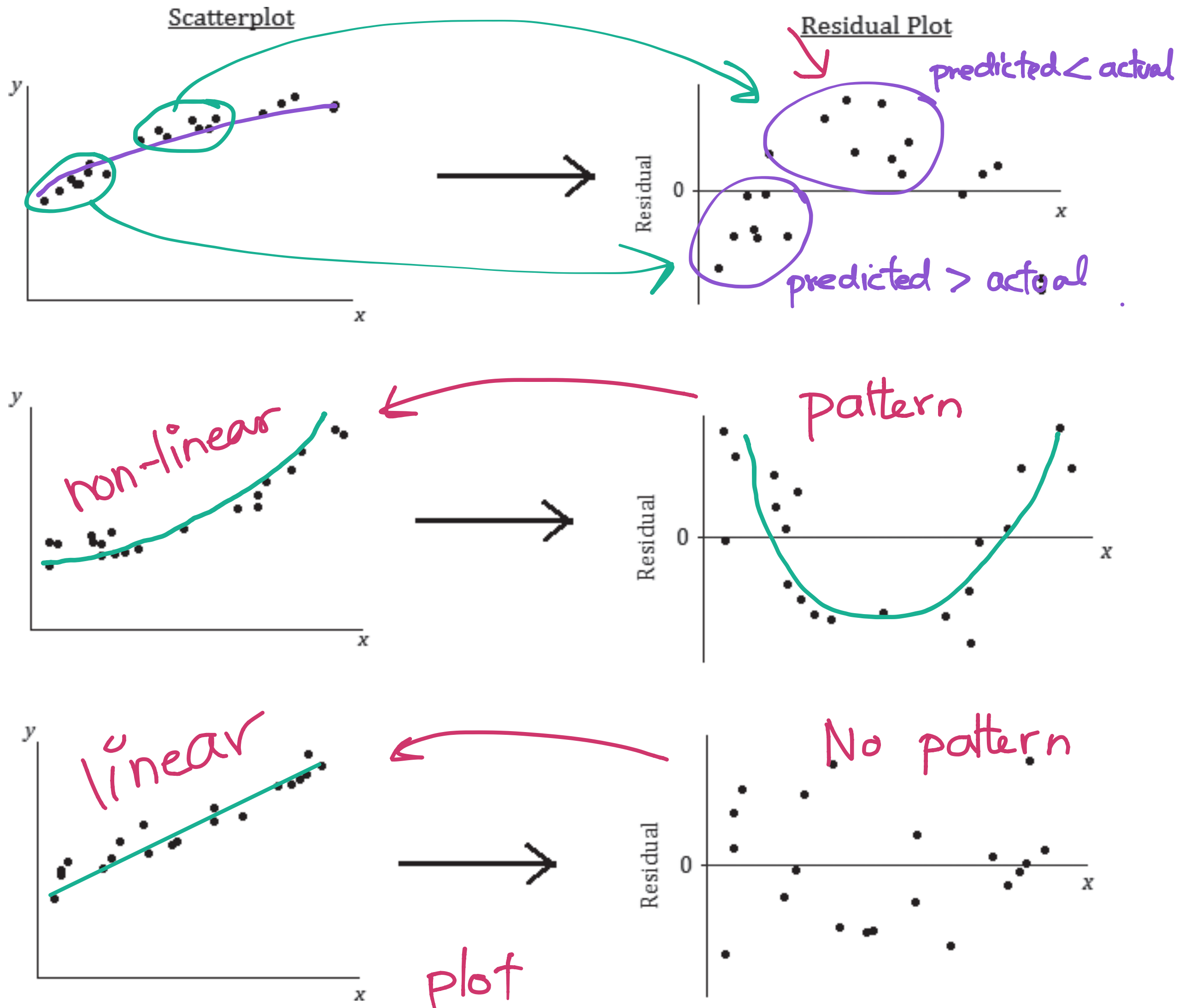
- No pattern → linear

- pattern → non-linear

Exploratory Activity—How Good Is Our Line?

In general, if the points in a residual plot are randomly scattered, then a linear model is the best fit. If the points in a residual plot have a pattern (exponential or quadratic) then a linear model is NOT the best fit.

5. Let's look at more examples of scatterplots and their corresponding residual plot. For each one, describe the shape of the original scatterplot and the distribution in the residual plot. What conclusions can you make?



- If the residual plot has a pattern, then the scatterplot has a nonlinear model.
- If the residual plot has no pattern, then the scatterplot has a linear

model.

6. A. What does it mean when there is a curved pattern in the residual plot?

The relationship might not be linear

B. What does it mean when the points in the residual plot appear to be scattered at random with no visible pattern?

The relationship might be linear.

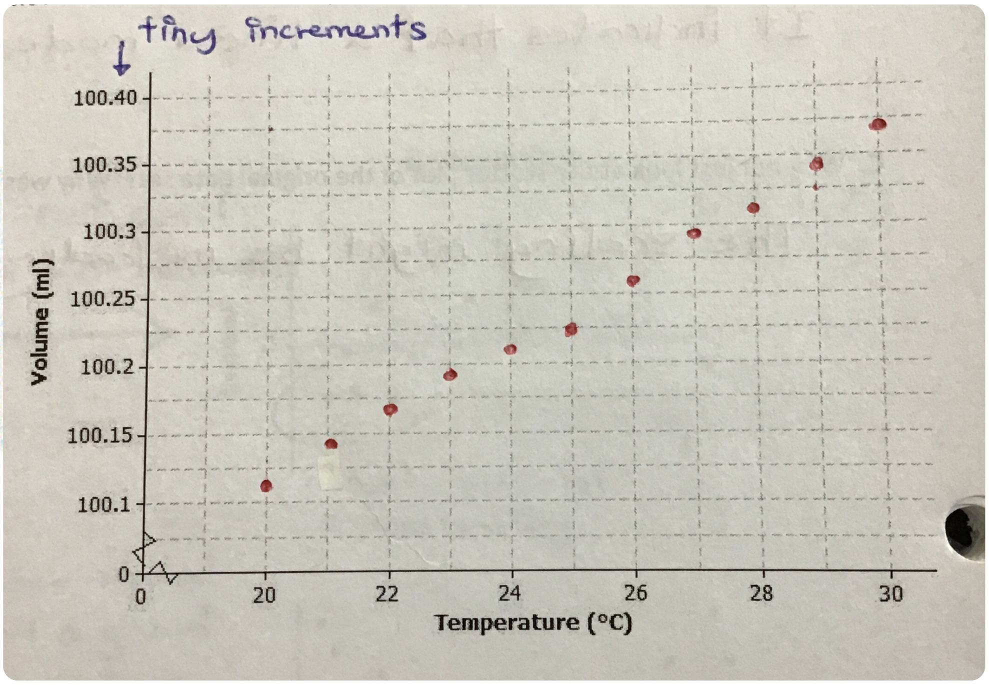
C. Why not just look at the scatter plot of the original data set? Why was the residual plot necessary?

Why Do You Need the Residual Plot?

Sometimes graphs can appear to be linear but they aren't. The scale on the vertical or horizontal axis can skew the look of the data. We'll look at an example to see how this happens.

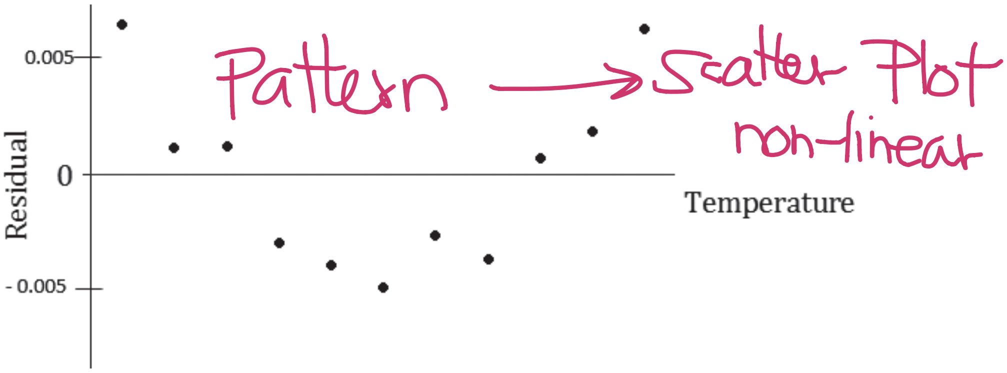
Water expands as it heats. Researchers measured the volume (in milliliters) of water at various temperatures. The results are shown below along with the scatter plot of the data set.

Temperature (°C)	Volume (ml)
20	100.125
21	100.145
22	100.170
23	100.191
24	100.215
25	100.239
26	100.266
27	100.290
28	100.319
29	100.345
30	100.374



Just by looking at the graph, we would suspect that the data is linear. But once we look at the residuals we see a different picture.

7. Below is the residuals graph using the linear equation, $y = 0.024918x + 99.621$. Do you see a clear curve in the residual plot? What does this say about the original data set?



The residual plot indicates a non-linear relationship.

Residuals—Calculating Prediction Errors

Let's look at a different example with data on animal pregnancies. The gestation time for an animal is the typical duration between conception and birth. The longevity of an animal is the typical lifespan for that animal. The gestation times (in days) and longevities (in years) for 13 types of animals are shown in the table below and a scatterplot has been constructed of that data.

Animal	Gestation Time (days)	Longevity (years)
Baboon	187	20
Black Bear	219	18
Beaver	105	5
Bison	285	15
Cat	63	12
Chimpanzee	230	20
Cow	284	15
Dog	61	12
Fox (Red)	52	7
Goat	151	8
Lion	100	15
Sheep	154	12
Wolf	63	5

Data Source: *Core Math Tools,* http://nctm.org

© Scenic Shutterbug/Shutterstock.com

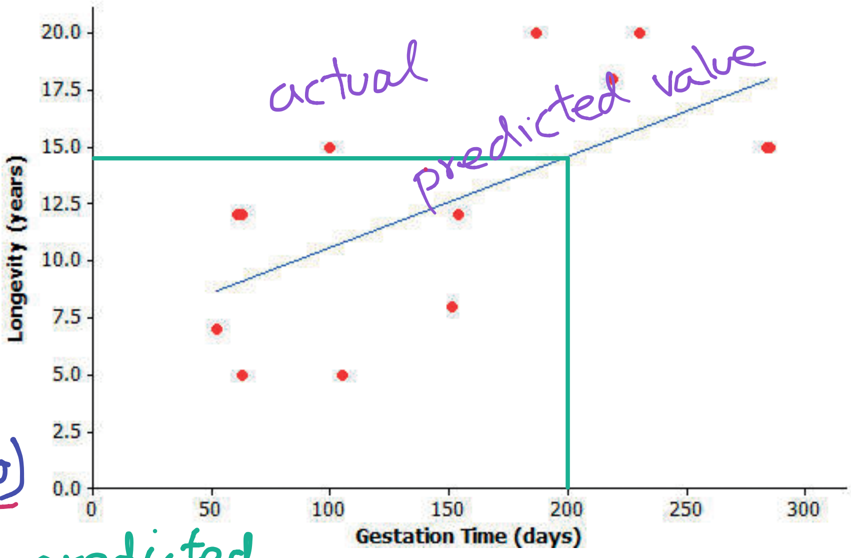
© vahamrick/Shutterstock.com

© TTphoto/Shutterstock.com

$y = b + mx$

The equation of the least squares line is $y = 6.642 + 0.03974x$, where x represents the gestation time (in days), and y represents longevity (in years).

The least squares line has been added to the scatter plot at the right.



X

8. Suppose a particular type of animal has a gestation time of 200 days. Approximately what value does the line predict for the longevity of that type of animal?

$y = 6.642 + 0.03974(200)$

$y \approx 14.6$ years ← predicted

9. Would the value you predicted in Exercise 8 necessarily be the exact value for the longevity of that type of animal? Could the actual longevity of that type of animal be longer than predicted? Could it be shorter?

No, the actual longevity could be longer or shorter. This is a prediction

10. You can investigate further by looking at the types of animals included in the original data set. Take the lion, for example. Its gestation time is 100 days. You also know that its longevity is 15 years, but what does the least squares line *predict* for the lion's longevity?

Substituting $x = 100$ days into the equation, you get $y = 6.642 + 0.03974(\underline{100}) \approx$ or $\underline{10.6}$ ← predicted value $6.642 + 3.974$

The least squares line predicts the lion's longevity to be approximately $\underline{10.6}$ years.

11. How close is this to being correct? More precisely, how much do you have to add to get the lion's true longevity of 15?

$15 - 10.6 = 4.4$

residual

Actual − Predicted = Residual

Residuals as Prediction Errors

In previous exercises, you found out how much needs to be added to the predicted value to find the actual value. In order to find this, you have been calculating the residual. It is summarized as

* residual = actual y-value − predicted y-value.

The residuals for all of the points in our animal longevity example are shown in the table below.

Animal	Gestation Time (days)	Longevity (years)	Residual (years)
Baboon	187	20	5.9
Black Bear	219	18	2.7
Beaver	105	5	−5.8
Bison	285	15	−3.0
Cat	63	12	2.9
Chimpanzee	230	20	4.2
Cow	284	15	−2.9
Dog	61	12	2.9
Fox (Red)	52	7	−1.7
Goat	151	8	−4.6
Lion	100	15	4.4
Sheep	154	12	−0.8
Wolf	63	5	−4.1

© GUDKOV ANDREY/ Shutterstock.com

© Leena Robinson/Shutterstock.com

© schubbel/Shutterstock.com

© Betty Shelton/Shutterstock.com

12. These residuals show that the actual longevity of an animal should be within six years of the longevity predicted by the least squares line. Where is the "within six years" coming from?

The highest and lowest residual are within six years.

13. Suppose you selected a type of animal that is not included in the original data set, and the gestation time for this type of animal is 270 days. Substituting $x = 270$ into the equation of the least squares line you get

$y = 6.642 + 0.03974(\underline{270})$
$= \underline{17.4}$.

The predicted longevity of this animal is 17.4 years.

17.4 + 6
17.4 − 6

14. Think about what the *actual* longevity of this type of animal might be. Could it be 30 years? How about 5 years?

It's unlikely

15. Judging by the size of the residuals in our table, what kind of values do you think would be reasonable for the longevity of this type of animal?

17 + 6 = 23
17 − 6 = 11

Bisons, Black bear.

16. The gestation time for humans is 270 days. Do the predictions you came up with in Exercises 13–15 make sense? Is this a good model for all animal gestation periods? Why might the predicted value be so far off for humans?

Lesson Summary

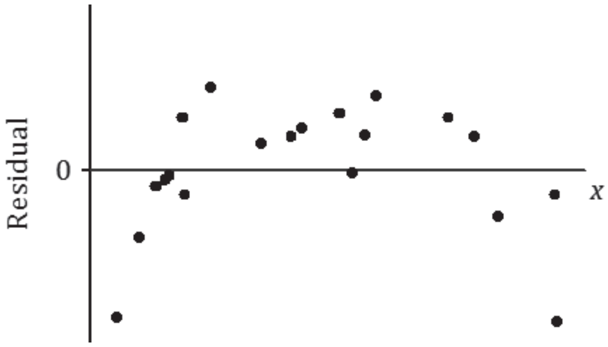
- After fitting a line, the residual plot can be constructed using a graphing utility.
- A curve or pattern in the residual plot indicates a nonlinear relationship in the original data set.
- A random scatter of points in the residual plot indicates a linear relationship in the original data set.

NAME: ____________________ PERIOD: ____________________ DATE: ________________

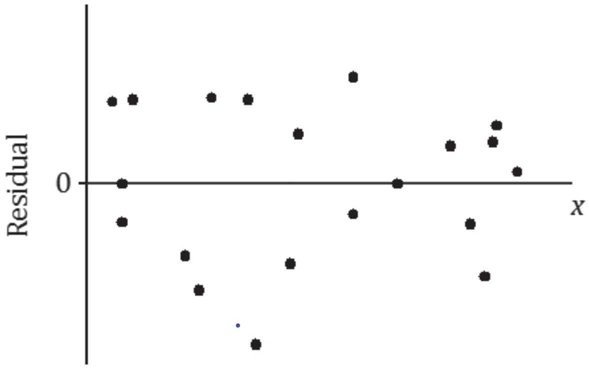
Homework Problem Set

1. For each of the following residual plots, what conclusion could you reach about the relationship between the variables in the original data set? Indicate whether the values would be better represented by a linear or a nonlinear relationship.

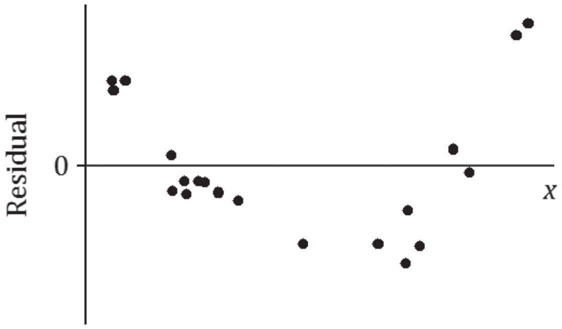
A.



B.

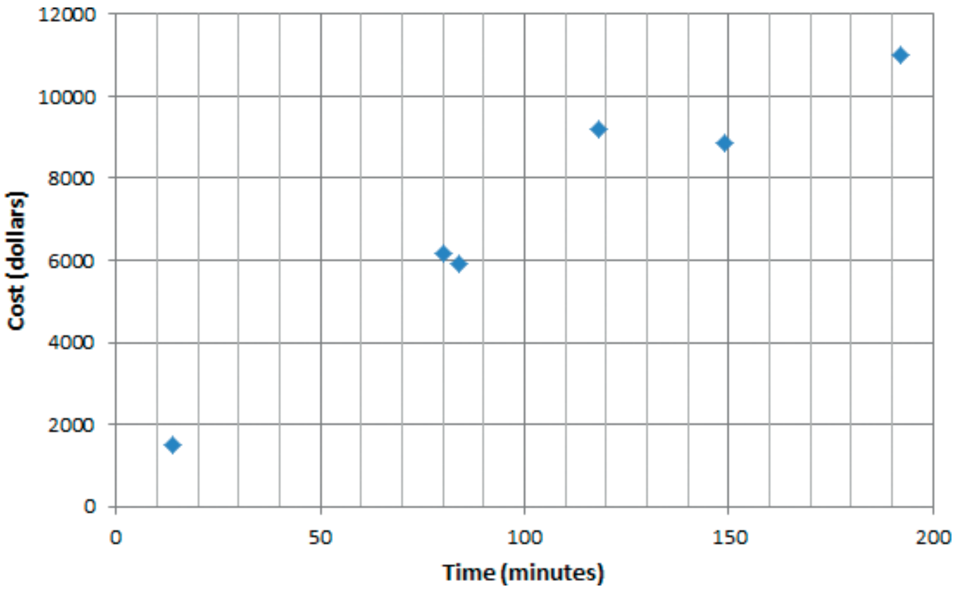


C.



The time spent in surgery and the cost of surgery was recorded for six patients. The results and scatter plot are shown below.

Time (minutes)	**Cost ($)**
14	1,510
80	6,178
84	5,912
118	9,184
149	8,855
192	11,023



2. Calculate the equation of the least squares line relating cost to time. (Indicate slope to the nearest tenth and y-intercept to the nearest whole number.)

$y = 52.71x + 1513.94$

3. Draw the least squares line on the graph above. (Hint: Substitute $x = 50$ into your equation to find the predicted y-value. Plot the point (50, your answer) on the graph. Then substitute $x = 150$ into the equation, and plot the point. Join the two points with a straightedge.)

4. What does the least squares line predict for the cost of a surgery that lasts 118 min? (Calculate the cost to the nearest cent.)

5. How much do you have to add to your answer to Problem 4 to get the actual cost of surgery for a surgery lasting 118 min.? (This is the residual.)

6. Show your answer to Problem 5 as a vertical line between the point for that person in the scatter plot and the least squares line.

7. Remember that the residual is the actual *y*-value minus the predicted *y*-value. Calculate the residual for the surgery that took 149 min. and cost $8,855.

8. Calculate the other residuals, and write all the residuals in the table below. Then graph the residual data on the scatterplot on the previous page.

Time (minutes)	**Cost ($)**	**Predicted Value ($)**	**Residual ($)**
14	1,510		
80	6,178		
84	5,912		
118	9,184		
149	8,855		
192	11,023		

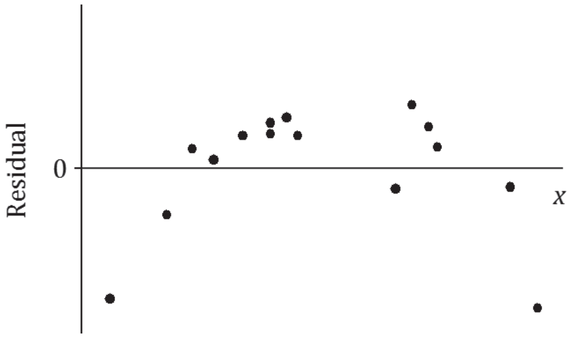
9. Suppose that a surgery took 100 min.

   A. What does the least squares line predict for the cost of this surgery?

   B. Would you be surprised if the actual cost of this surgery were $9,000? Why, or why not?

   C. Interpret the slope of the least squares line.

10. Suppose that after fitting a line, a data set produces the residual plot shown below.



An incomplete scatter plot of the original data set is shown below. The least squares line is shown, but the points in the scatter plot have been erased. Estimate the locations of the original points, and create an approximation of the scatter plot below.