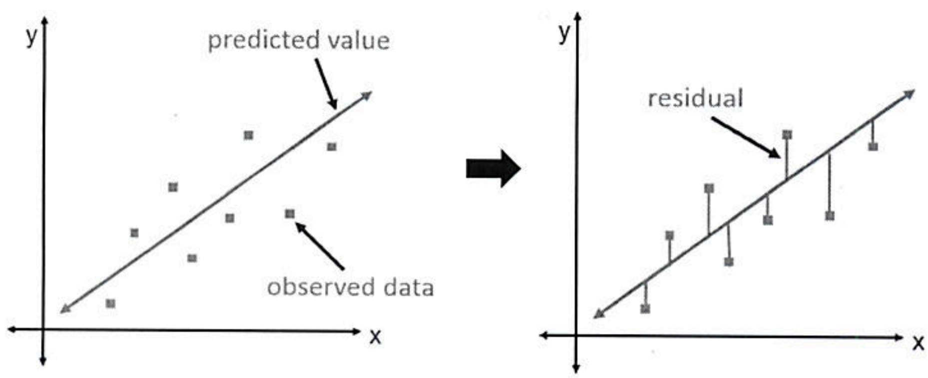
Key Concept 5: Comparing Functions – Modeling and Transformations

Sections

5A – Regression

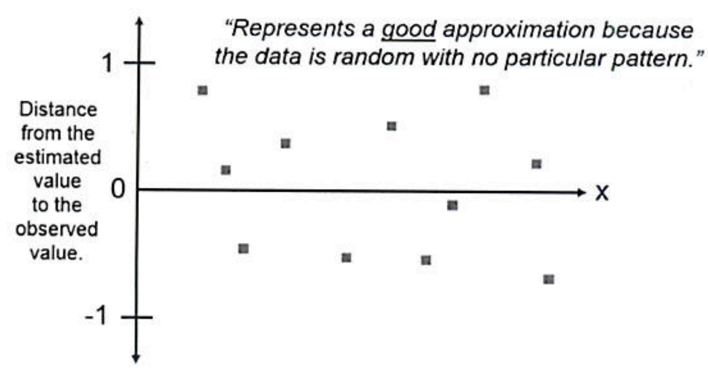
❖ Vocabulary, Formulas, Theories:

- **Scatter Plot**: a plot of two variable data on a coordinate plane. It's known as the "observed data."
- **Regression**: the process of creating a "line of best fit" that shows the relationship between data points.
- **Line of Best Fit**: a line, or curve, plotted on a scatter plot that attempts to model the relationship between data. Equations for these lines, or curves, can be created which are often either linear, quadratic, or exponential. The line of best fit is known as the "estimated values."
- **Residuals**: the vertical distance between the observed data values and the estimated values on the line of best fit.

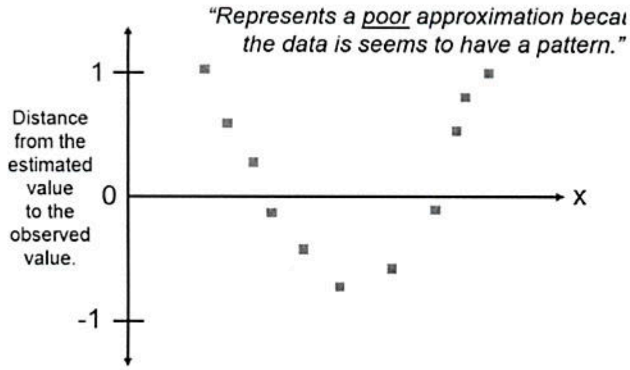


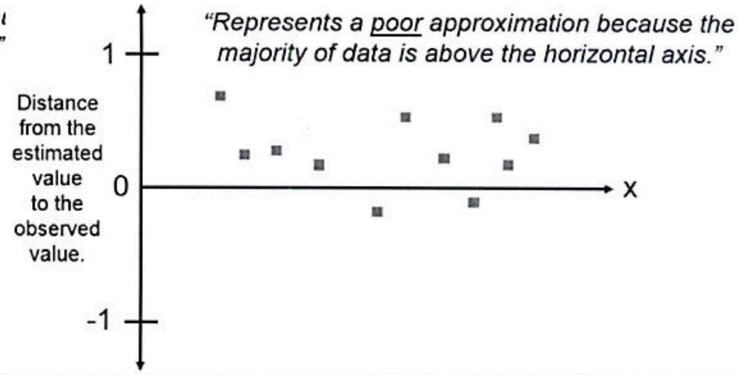
The red vertical lines represent the residual, or the distance from the observed data to the expected value.

- **Residuals Plot**: a plot that represents all the residuals by highlighting the x-value and the vertical distance between the observed values and estimated values. A residual plot will represent the residuals using $(x,\ residual\ value)$, or $(x, observed - estimated)$. Note that the residual may be negative.
    - If the residual plot shows points that are random with no pattern ☺, the line of best fit is a good estimate of data.



- If the residual plots shows points that don't seem random and tend to have a pattern, the line of best fit is not a good estimation of data. :(





Scatter plots are made up of a bunch of data points which represent the relationship between two categories, or variables. To help describe the relationship between the two variables, we try to match it with a function. This process is called regression. For example, if the data on the scatter plot seems linear (looks like a line), a "line of best fit" can be run through the scatter plot to help describe the relationship. The equation of that "line of best fit" could be used to make predictions and draw conclusions about the data. Watch the next video to understand the vocabulary that is involved in these ideas.

Video - "Modeling Data, Regression, and Residuals - Vocabulary " - MathontheWeb (2:47)

There are plenty of different types of regressions to covers, but one of the most common types is linear regression. Watch the next video to learn how to create a scatter plot, determine a line of best fit, and draw conclusions based on the given data.

Video - "Linear Regression - Example" - MathontheWeb (12:22) CHECK OUT VIDEO TO SEE THE STEPS for Ex1) a) again :)

EX1) The given table has a list of x and y values. Use this data to answer the questions.

a. Use a calculator to determine a linear regression. Nspire gives us: $y = 5.75x - 3.5$
b. Fill in the rest of the chart. ✓
c. Why do the residuals have to be squared? To judge how good the line of best fit is
d. What is the sum of residuals squared? Does a small or large sum represent a more accurate line of best fit? Nspire gives us: 115.25. A small sum represents a more accurate line of best fit.
e. What value can you predict y to be when the value of x is 12? $y = 5.75(12) - 3.5 = 65.5$

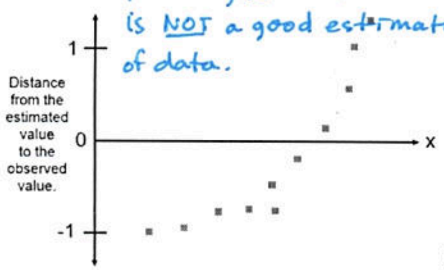
X	Y	Predicted Y-Values	Residual	Residual Squared
1	2	$y = 5.75(1) - 3.5 = 2.25$	$2 - 2.25 = -0.25$	$(-0.25)^2 = 0.0625$
2	13	$y = 5.75(2) - 3.5 = 8$	$13 - 8 = 5$	$(5)^2 = 25$
3	13	$y = 5.75(3) - 3.5 = 13.75$	$13 - 13.75 = -0.75$	$(-0.75)^2 = 0.5625$
4	15	$y = 5.75(4) - 3.5 = 19.5$	$15 - 19.5 = -4.5$	$(-4.5)^2 = 20.25$
5	19	$y = 5.75(5) - 3.5 = 25.25$	$19 - 25.25 = -6.25$	$(-6.25)^2 = 39.0625$
6	36	$y = 5.75(6) - 3.5 = 31$	$36 - 31 = 5$	$(5)^2 = 25$
7	39	$y = 5.75(7) - 3.5 = 36.75$	$39 - 36.75 = 2.25$	$(2.25)^2 = 5.0625$
8	42	$y = 5.75(8) - 3.5 = 42.5$	$42 - 42.5 = -0.5$	$(-0.5)^2 = 0.25$

The example uses a calculator to determine a line of best fit for a linear regression because we knew the data had a linear relationship. That isn't always the case. We might not know if the data is linear. It can have other relationships like quadratic or exponential. To decipher the relationship, we use a residual plot. Watch the next video to realize what type of residual plot to look out for when trying to determine the relationship (or correct regression) of the data.

Video - "Residual Plots - Example" - MathontheWeb (2:40)

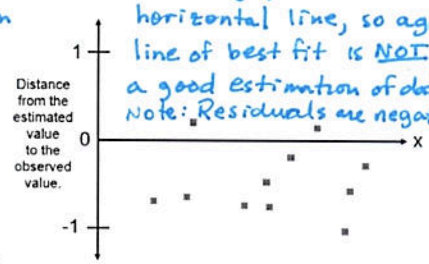
EX2) Determine if the residual plot represents a line of best fit that is a good estimation of data.

Pattern, so line of best fit is NOT a good estimation of data.



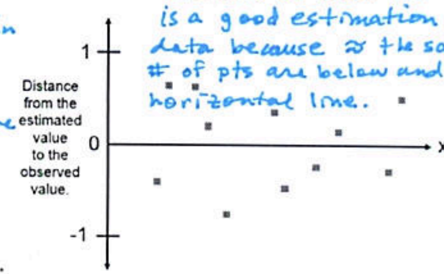
a.

Too many pts. below horizontal line, so again line of best fit is NOT a good estimation of data.
Note: Residuals are negative



b.

The line of best fit is a good estimation of data because ≈ the same # of pts are below and above horizontal line.



c.

Remember that the "line of best fit" isn't always a line. It can be a curve too. Scatter plots that have data that look quadratic, or even exponential, suggest those types of regressions be used. Calculators can handle this rather easily.

Video - "Regression - Example" - MathontheWeb (7:56)

EX3) Determine if the data is linear, quadratic, or exponential.

a.

x	0	1	2	3	4	5	6	7	8	9	10
y	0.38	1.33	1.51	2.85	4.01	6.92	8.96	12.88	16.21	21.50	23.99

Use Nspire to run a regressions and check residual plots.
The data is quadratic :)

WATCH VIDEO

b.

x	0	1	2	3	4	5	6	7	8	9	10
y	0.50	0.98	1.15	1.79	2.58	4.30	6.16	9.54	13.97	22.19	33.41

Regression → Residual Plot
Check if it's linear, quadratic, or exponential. The data is exponential.

❖ Extra Resources:

http://www.mathsisfun.com/data/scatter-xy-plots.html

https://www.youtube.com/watch?v=cGqPVySnDcY

EX4) The height of a child was recorded from the age of 17 months to 30 months.

a. Use a calculator to determine a line of best fit. (determine a linear regression)

Follow the steps on Checkpoint 5A. The line of best fit is: $y = 0.603x + 65.814$

Rounded to 3 decimal places

* b. Fill in the rest of the chart.

c. Why do the residuals have to be squared? Would a small or large sum of residuals squared represent a better line of best fit?

i) The residuals should be squared because they will allow us to judge the how good the line of best fit is.

ii) A small sum represents a better fit.

d. At the age of 35 months, can there be reasonable height predicted? If so, what is it?

$y = 0.603(35) + 65.814 \approx 86.9$ cm

Many factors contribute to a child's growth but we can be fairly sure that the height can possibly be close to 86.9 cm.

Age (months)	Height (cm)	Predicted Height	Residual	Residuals Squared
20	78.0	$y = 0.603(20) + 65.814 \approx 77.9$	$78.0 - 77.9 = 0.1$	$(0.1)^2 = 0.01$
21	78.4	$y = 0.603(21) + 65.814 \approx 78.5$	$78.4 - 78.5 = -0.1$	$(-0.1)^2 = 0.01$
22	78.9	$y = 0.603(22) + 65.814 \approx 79.1$	$78.9 - 79.1 = -0.2$	$(-0.2)^2 = 0.04$
23	79.7	79.7	$79.7 - 79.7 = 0$	$0^2 = 0$
24	80.2	80.3	$80.2 - 80.3 = -0.1$	$(-0.1)^2 = 0.01$
25	81.2	80.9	$81.2 - 80.9 = 0.3$	$(0.3)^2 = 0.09$
26	81.4	81.5	$81.4 - 81.5 = -0.1$	$(-0.1)^2 = 0.01$
27	82.1	82.1	$82.1 - 82.1 = 0$	$(0)^2 = 0$
28	82.5	82.7	$82.5 - 82.7 = -0.2$	$(-0.2)^2 = 0.04$
29	83.4	83.3	$83.4 - 83.3 = 0.1$	$(0.1)^2 = 0.01$
30	83.9	83.9	$83.9 - 83.9 = 0$	$(0)^2 = 0$

* I entered $y = 0.603x + 65.814$ into a graph page. Then I selected menu → Table (7) → Split screen Table (1). This way I get the predicted height values right away.