

MATH 3307

Lesson 20

# Residuals

A **residual** value is the difference between an actual observed  $y$  value and the corresponding predicted  $y$  value,  $\hat{y}$ . Residuals are just errors.

Residual = error = (observed – predicted) =  $(y - \hat{y})$ .

Using Rstudio, there are two ways you can do this:

1. (after finding LSRL),  $y - \langle \text{LSRL equation} \rangle$
2. `residuals(LSRL name or command)` *Note: this will give residuals referenced by their place in line, not by their x-value.*

Using the TI Calculator:

(After finding LSRL, and storing in Y1), STAT, Edit, (in L3), L2-Y1

## Examples:

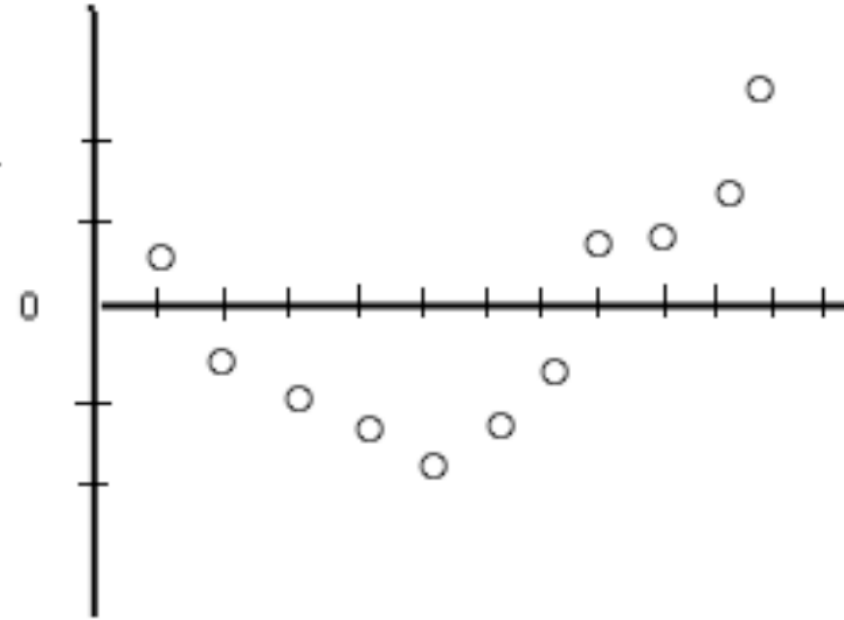
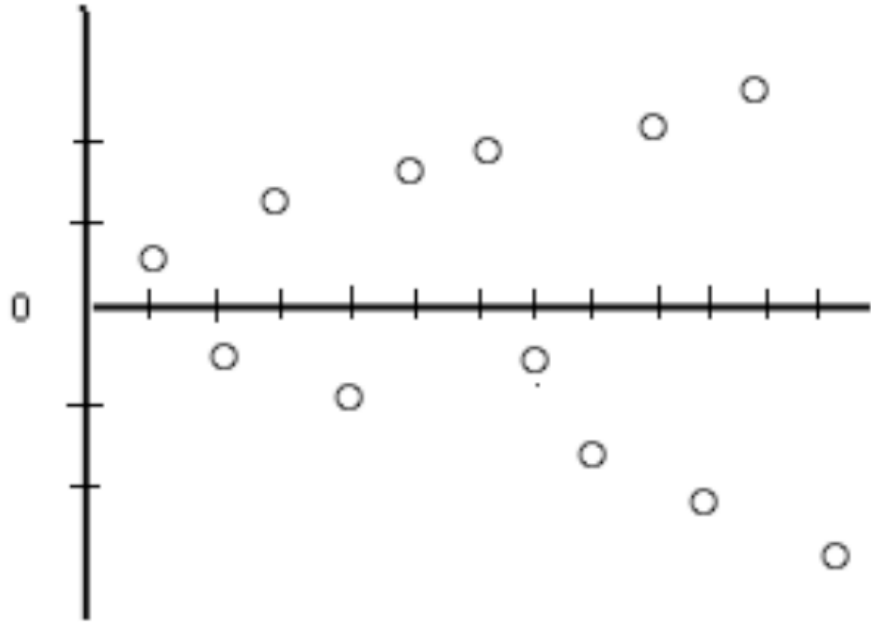
A least-squares regression line was fitted to the weights (in pounds) versus age (in months) of a group of many young children. The equation of the line is  $\hat{y} = 16.6 + 0.65t$ , where  $\hat{y}$  is the predicted weight and  $t$  is the age of the child. A 20-month old child in this group has an actual weight of 25 pounds. What is the residual weight, in pounds, for this child?

# Interpreting the Plots of Residuals

The plot of the residual values against the  $x$  values can tell us a lot about our LSRL model. Plots of residuals may display patterns that would give some idea about the appropriateness of the model. If the functional form of the regression model is incorrect, the residual plots constructed by using the model will often display a pattern. The pattern can then be used to propose a more appropriate model. When a residual plot shows no pattern, it indicates that the proposed model is a reasonable fit to a set of data.

# Interpreting the Plots of Residuals

Here are some examples of residual plots that show patterns:



# Example:

A data set produced the regression equation  $\hat{y} = 17.3 - .96x$ . Below are four of the data points. Draw a residual plot for these data points.

x	2	7	4	12
y	15	11	12	5



## Example:

The following data was collected comparing score on a measure of test anxiety and exam score:

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

Construct a scatterplot.

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

Find the LSRL and fit it to the scatter plot.



Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

Find  $r$  and  $r^2$

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

Does there appear to be a linear relationship between the two variables? Based on what you found, would you characterize the relationship as positive or negative? Strong or weak?

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

Interpret the slope in terms of the problem

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

Find the values of the residuals and plot the residuals.

Measure of test anxiety	23	14	14	0	7	20	20	15	21
Exam score	43	59	48	77	50	52	46	51	51

What does this plot reveal?

Is it reasonable to conclude that test anxiety caused poor exam performance? Explain

# Another example

Year	1790	1800	1810	1820	1830	1840	1850	1860	1870	1880
People per square mile	4.5	6.1	4.3	5.5	7.4	9.8	7.9	10.6	10.09	14.2
Year	1890	1900	1910	1920	1930	1940	1950	1960	1970	1980
People per square mile	17.8	21.5	26	29.9	34.7	37.2	42.6	50.6	57.5	64

Examine the LSRL to determine if it is a good model for this data

# Residual Meanings

Since the residuals show how far the data falls from the LSRL, examining the values of the residuals will help us to gauge how well the LSRL describes the data. The sum of the residuals is always 0 so the plot will always be centered around the x-axis.

An **outlier** is a value that is well separated from the rest of the data set. An outlier will have a large absolute residual value.

An observation that causes the values of the slope and the intercept in the line of best fit to be considerably different from what they would be if the observation were removed from the data set is said to be **influential**.

Age(years)	9	10	11	12	13	14	15	16
Time (sec)	34.8	34.2	32.9	29.1	28.4	22.4	25.2	24.9

This example showed that there is an influential point. Let's investigate

Determine the LSRL

Determine the value for  $r$ .

Determine the value for  $r^2$ .

Find and plot the residuals:

Is the LSRL a good fit to the data?