MATH 3307
Lesson 20

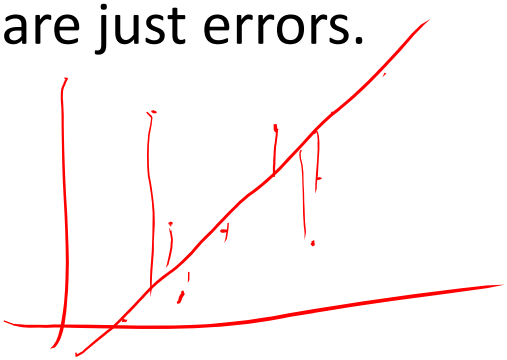
Residuals

Every Data Point has its own residual

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 - $\hat{y}$).

(Actual y) - (LSRL y)



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

1. (after finding LSRL), y - <LSRL equation>

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-(LSRL)

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?

$\hat{y} = 16.6 + 0.65t$

$(20, 25)$

$t \quad y$

$y - \hat{y}$

$25 - (16.6 + 0.65 \cdot 20)$

$y \quad \hat{y} \quad t$

```
> 25-(16.6+0.65*20)
[1] -4.6
```

Meaning this child's weight is 4.6 pounds less than predicted.

Interpreting the Plots of Residuals

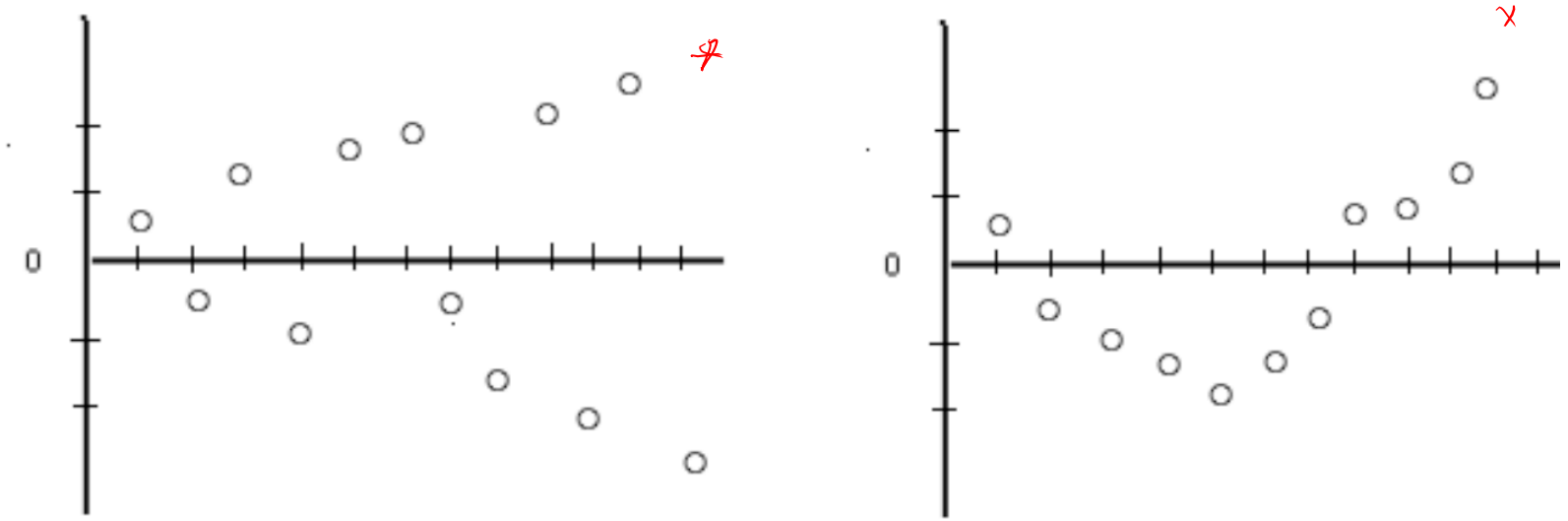
The plot of the residual values against the x values can tell us a lot about our LSRL model. Plots of residuals may display patterns that would give some idea about the appropriateness of the model. If the functional form of the regression model is incorrect, the residual plots constructed by using the 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.

Patterns appearing in the Residual Plot: the model was not the best one to use.

No Pattern in the Residual Plot: the model is an accurate representation.

Interpreting the Plots of Residuals

Here are some examples of residual plots that show patterns:

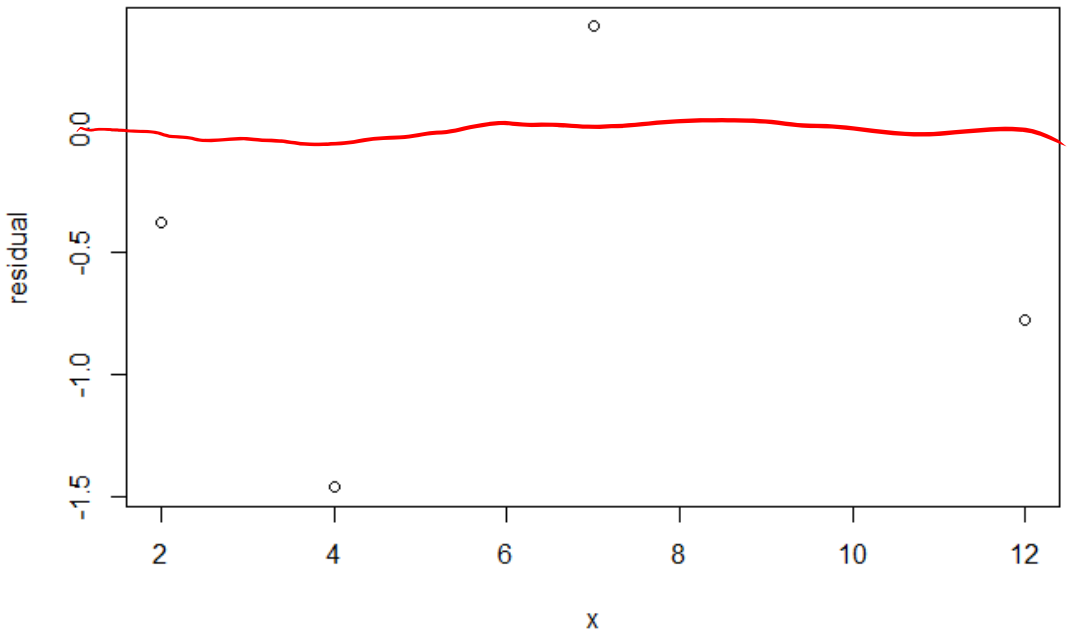


If you can, with some accuracy, predict where one additional point should go, there is a patter in the plot.

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




```
> assign("x",c(2,7,4,12))
> assign("y",c(15,11,12,5))
> residual=y-(17.3-0.96*x)
> plot(x,residual)
```

Based on these four points, there is no pattern evident in the residual plot.

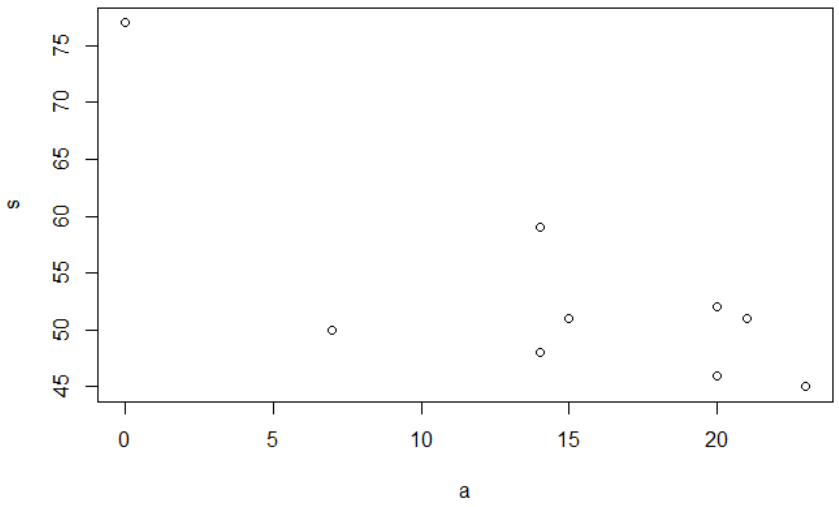
The LSRL is (likely) an accurate model.

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.




```
plot(a,s)
```

To Copy:

assign(“a”,c(23,14,14,0,7,20,20,15,21))

assign(“s”,c(45,59,48,77,50,52,46,51,51))

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.

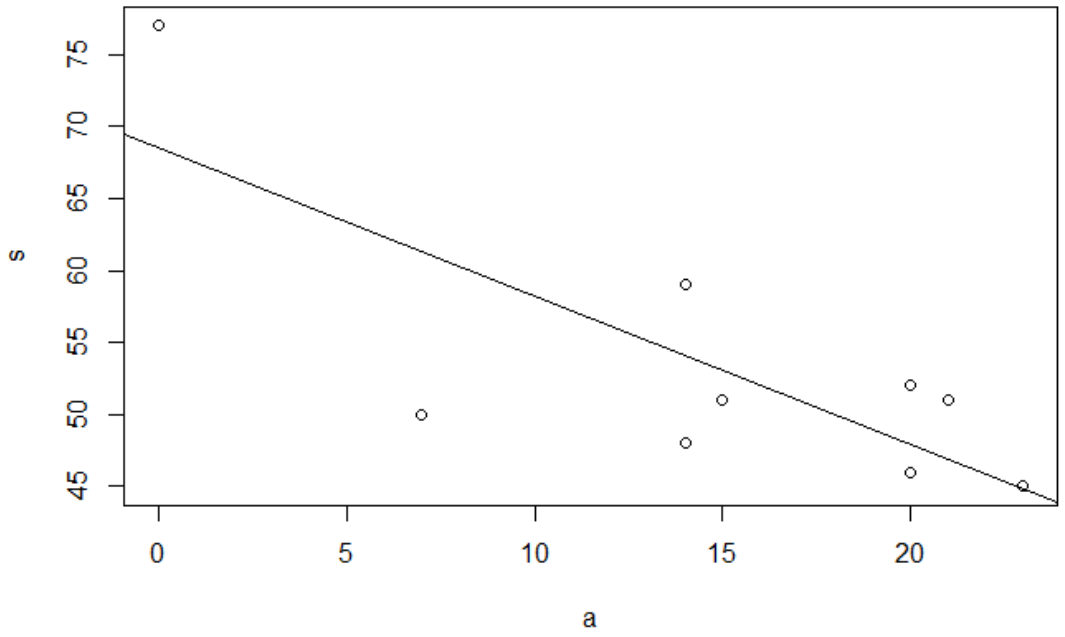
```
> reg=lm(s~a)
> reg

Call:
lm(formula = s ~ a)

Coefficients:
(Intercept)            a
     68.513       -1.027

> abline(reg)
```

Lsloxe



$\hat{y} = -1.027x + 68.513$

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

r: correlation coefficient
r^2: coefficient of determination

```
> cor(a,s)
[1] -0.7783333
> cor(a,s)^2
[1] 0.6058027
```

60% of the variation in exam score can be predicted by the LSRL

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?

There does not appear to be a linear relationship between the variables.

The relationship between them is negative (based on the r-value) and moderate.

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

There is [slope value] [increase/decrease] in [y variable] for every unit of [x variable] increase.

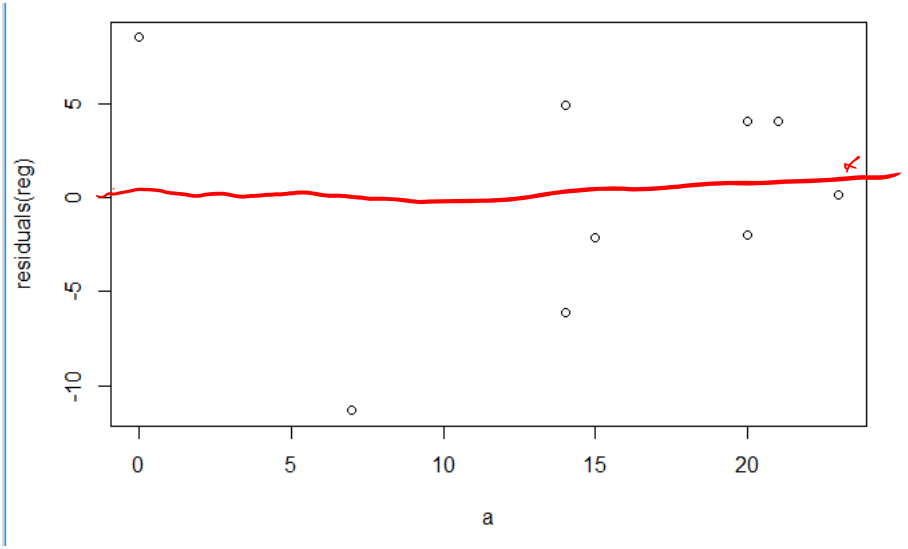
There is a 1.027 decrease in exam score for one unit increase of test anxiety.

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.

```
> residuals(reg)
         1          2          3          4          5
 0.1076109  4.8649194 -6.1350806  8.4873992 -11.3238407
         6          7          8          9
 4.0267137 -1.9732863 -2.1081149  4.0536794
> s-(-1.027*a+68.513)
[1]   0.108   4.865  -6.135   8.487 -11.324   4.027  -1.973
[8]  -2.108   4.054
> plot(a,residuals(reg))
```

Numbers in the top row correspond to position in the table



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?

Due to the pattern appearing in the residual plot, the LSRL is not a good model for this data. There may be a non-linear model that works better.

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

Causality can not be shown by this method. All we can show is that there is a negative relationship between these two variables.

Another example

$\hat{y} = .3009x - 544.0672$

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

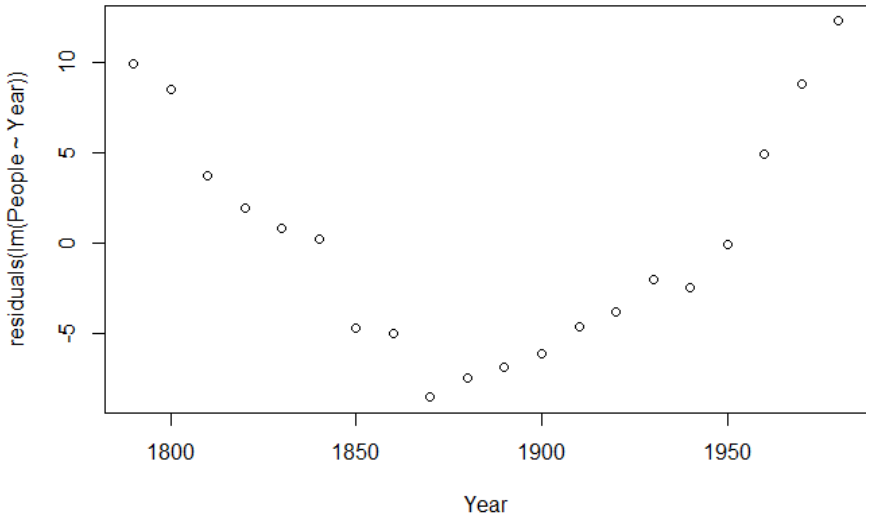
Examine the LSRL to determine if it is a good model for this data

```
> lm(People~Year)

Call:
lm(formula = People ~ Year)

Coefficients:
(Intercept)         Year
  -544.0672       0.3009

> cor(Year,People)^2
[1] 0.8894762
> plot(Year,residuals(lm(People~Year)))
```



Due to the pattern in the residual plot, there is a better, non-linear model.

To copy:

assign(“Year”,c(1790,1800,1810,1820,1830,1840,1850,1860,1870,1880,1890,1900,1910,1920,1930,1940,1950,1960,1970,1980))

assign(“People”,c(4.5,6.1,4.3,5.5,7.4,9.8,7.9,10.6,10.09,14.2,17.8,21.5,26,29.9,34.7,37.2,42.6,50.6,57.5,64))

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.**

Popper 15:

	1	2	3	4	5	6	7	8
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

1. Determine the LSRL

a. y = 50.79 - 1.744x b. y = -1.744 + 50.79x c. y = 50.79 + 1.744x

2. Determine the value for r.

a. -.6536 b. -.8876 c. -.9196

3. Determine the value for r^2.

a. None b. .8457 c. .9934

Find and plot the residuals

4. Residual at x = 9

a. -.292 b. -.760 c. 2.02

5. Residual at x = 12

a. -.292 b. -.760 c. 2.02

6. Residual at x = 16

a. -.292 b. -.760 c. 2.02

7. Is the LSRL a good fit to the data?

a. Yes b. No

There is a pattern in residuals

```
> lm(time~age)

Call:
lm(formula = time ~ age)

Coefficients:
(Intercept)          age
     50.788       -1.744
```

$\hat{y} = -1.744x + 50.788$

```
> cor(age,time)
[1] -0.919593
> cor(age,time)^2
[1] 0.8456513
```

```
> residuals(lm(time~age))
         1          2          3          4          5
-0.2916667  0.8523810  1.2964286 -0.7595238  0.2845238
         6          7          8
-3.9714286  0.5726190  2.0166667
> plot(age,residuals(lm(time~age)))
```