Math 2311

Bekki George – bekki@math.uh.edu

Office Hours: MW 11am to 12:45pm in 639 PGH

Online Thursdays 4-5:30pm

And by appointment

Class webpage: http://www.math.uh.edu/~bekki/Math2311.html

5.4 – 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) = $(y - \hat{y})$

predicted y from LSRL formula

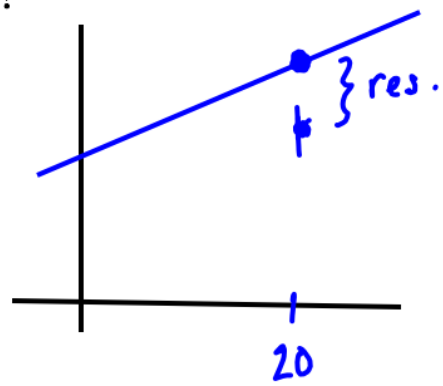
from data

Example:

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?

$y = 25$ lbs

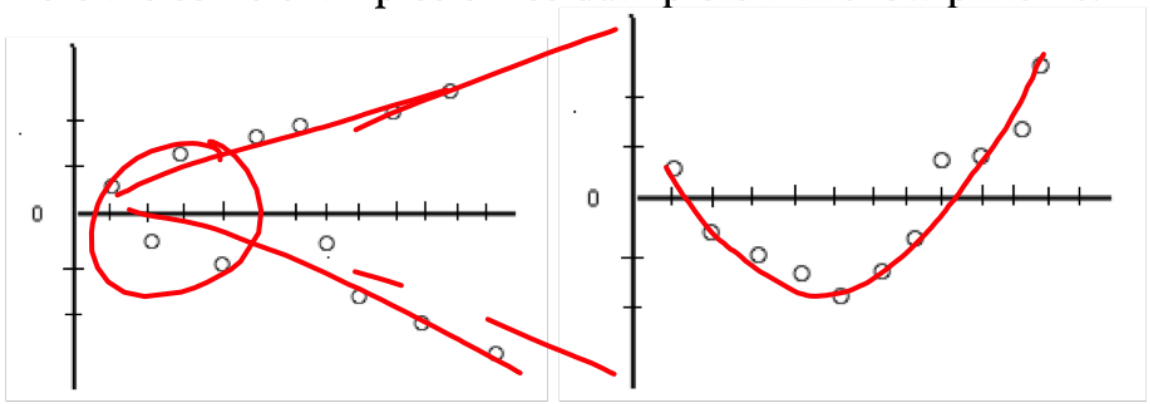
$\hat{y}(20) = 16.6 + .65(20) = 29.6$ lbs



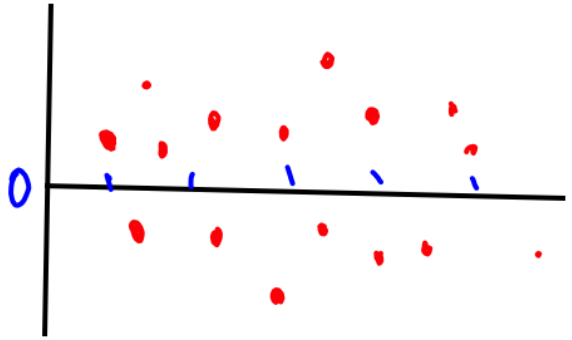
residual = $25 - 29.6 = -4.6$ lbs.

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.

Here are some examples of residual plots that show patterns:



Want:



No pattern

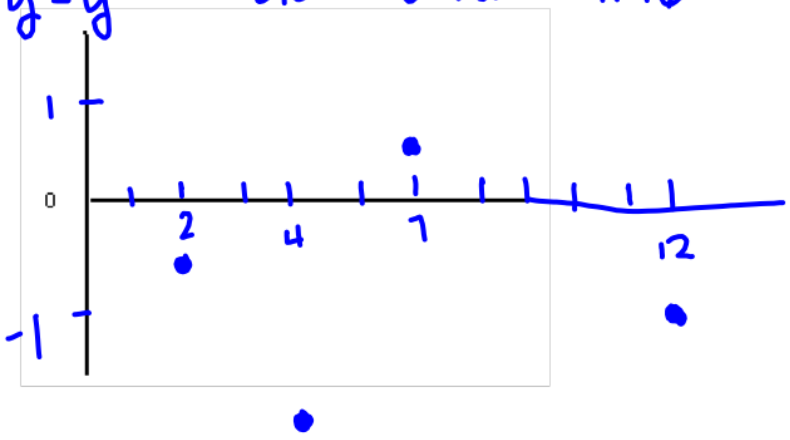
★ sum of all residuals should be 0

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. (this is a subset of entire set)

x	2	7	4	12
y	15	11	12	5
$\hat{y}$	15.38	10.58	13.46	5.78
res. $y - \hat{y}$	-0.38	0.42	-1.46	-.78

← plug x into $\hat{y}$



Popper 11

1. What is the residual for the point $(1, 2)$ given $\hat{y} = 2.75 - 1.06x$?

x y

a. -0.31
b. 0.31
c. 1.69
d. -1.69
e. none of these

$\hat{y}(1) =$ _____

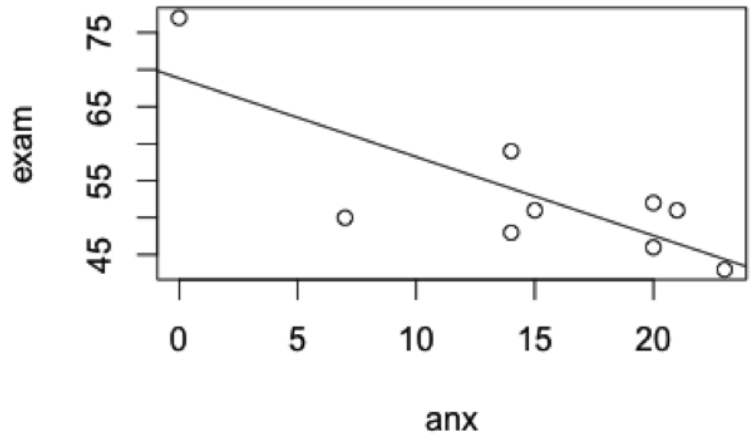
$res = y - \hat{y}$

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.

```
> anx=c(23,14,14,0,7,20,20,15,21)
> exam=c(43,59,48,77,50,52,46,51,51)
> plot(anx,exam)
```



Find the LSRL and fit it to the scatter plot.

$\hat{y} = 68.838 - 1.064x$

Find r and r^2.

```
> cor(anx,exam)
[1] -0.7877352
> (-0.7877352)^2
[1] 0.6205267
```

```
> examline=lm(exam~anx)
> examline

Call:
lm(formula = exam ~ anx)

Coefficients:
(Intercept)          anx
     68.838       -1.064
```

slope

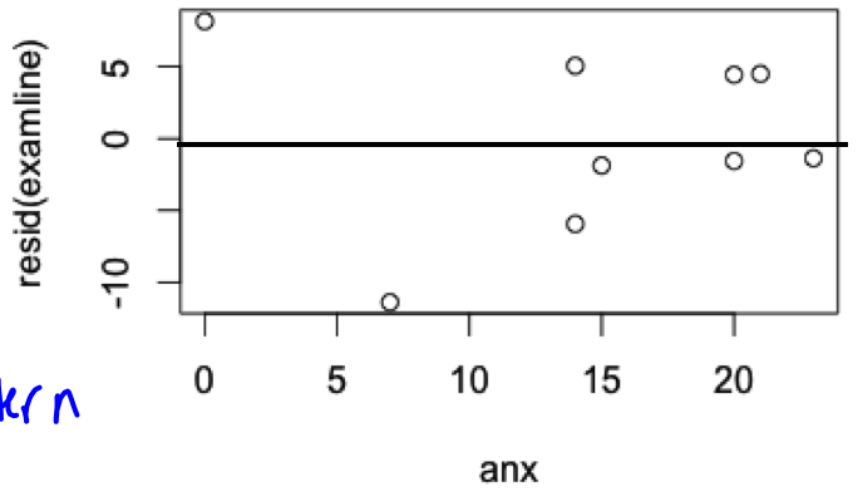
```
> abline(examline)
```

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?

moderate

Interpret the slope in terms of the problem. (-1.064) for every 1 increase in test anxiety score, there is a decrease of 1.064 in the exam score.

Find the values of the residuals and plot the residuals.



What does this plot reveal?

Good - no pattern

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

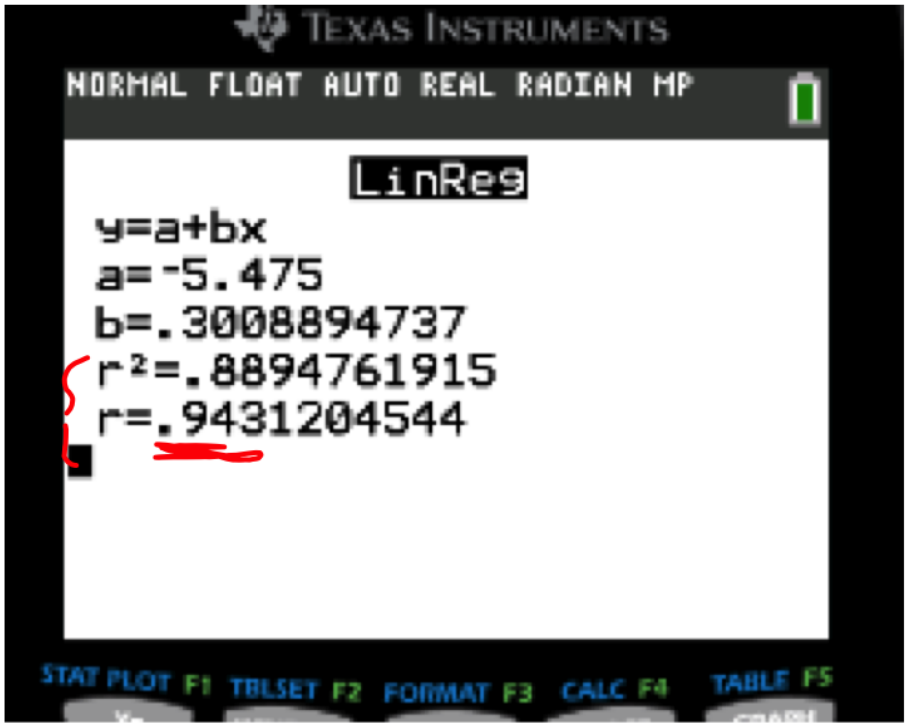
No

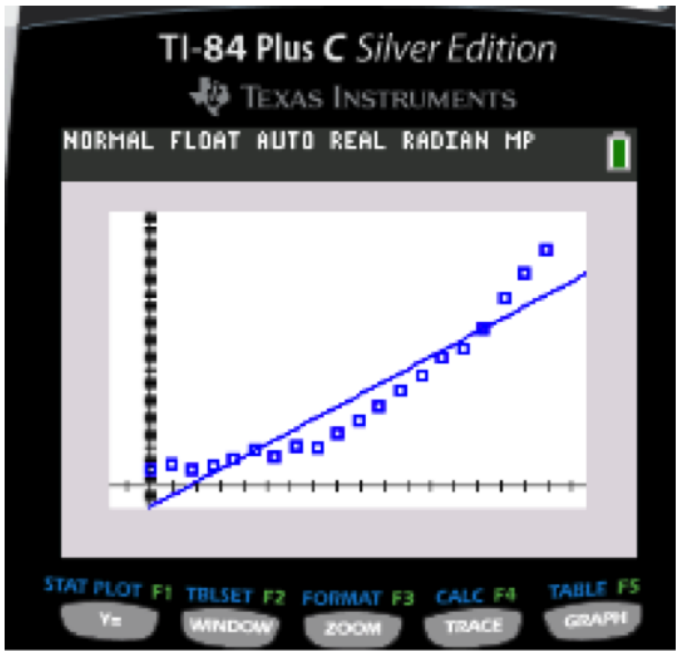
Another example:

		0	10	20	30	40	50	60	70	80	90
X	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
		100	110	120	130	140	150	160	170	180	190

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

LinReg(a+bx) L1, L2, Y1

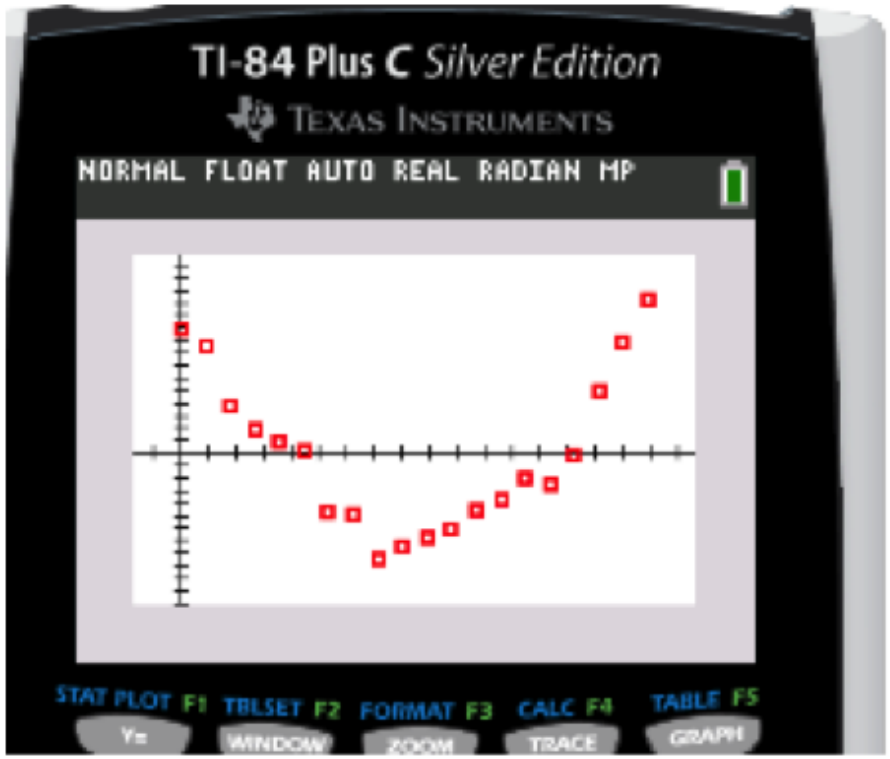




$L_1 \rightarrow x$

$L_2 \rightarrow y$

$L_3 = L_2 \div Y_1(L_1)$



⟸ pattern

may not be a good model

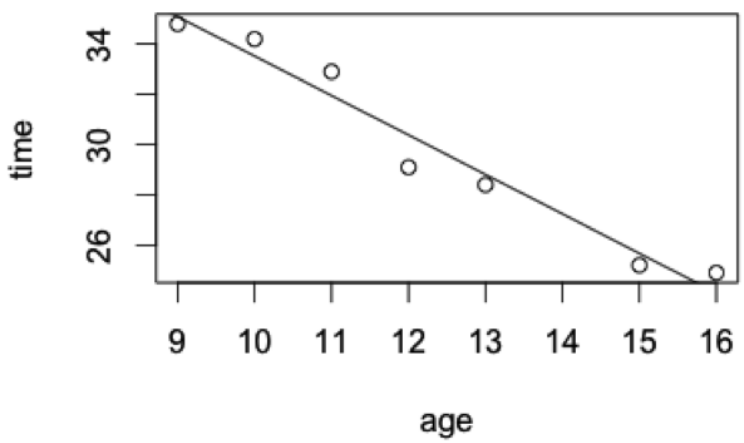
Since the residuals show how far the data falls from the LSRL, examining the values of the residuals will help us to gauge how well the LSRL describes the data. The sum of the residuals is always 0 so the plot will always be centered around the x-axis.

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.**

Example 4 (from text): Johnny keeps track of his best swimming times for the 50 meter freestyle from each summer swim team season. Here is his data:

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



Popper 11

2. Association implies causation

a. True
b. False

3. If a correlation coefficient has a value of 0.9679, that means the data has a linear relationship and we do not have to look at the residual plot.

a. True
b. False

4. A residual plot with a pattern means a strong linear relationship

a. True
b. False

5. Choose E