

## Slides from Lines of Fit (Tutorial 26)

# Lines of Fit

Fred Dillon

This handout contains selected slides to use when reviewing this tutorial topic with or without the video. To access all slides, open thumbnail link on the tutorial interface.

## Lines of Fit

- Use a graphing calculator to plot bivariate data
- Determine a visual line of fit
- Calculate the correlation coefficient and the least squares regression line by hand and with technology

The average weight of adult males in the years that have passed since 1960

Years Since 1960	Weight in Pounds
0	166.3
8	170.5
17	174.4
25	178.1
30	182.5
35	185.3
42	191.0

Based on data from National Center for Health Statistics at the Centers for Disease Control and Prevention [CDC]

*Bivariate data* refers to two sets of data that are possibly related.

Years Since 1960	Weight in Pounds
0	166.3
8	170.5
17	174.4
25	178.1
30	182.5
35	185.3
42	191.0

Based on data from National Center for Health Statistics at the Centers for Disease Control and Prevention [CDC]

*What is the relationship between these two data sets?*

Years Since 1960	Weight in Pounds
0	166.3
8	170.5
17	174.4
25	178.1
30	182.5
35	185.3
42	191.0

Based on data from National Center for Health Statistics at the Centers for Disease Control and Prevention [CDC]

## Create a Scatter Plot of the Bivariate Data

# Tutorials for High School Mathematics

## Slides from Lines of Fit (Tutorial 26)

On a graphing calculator

1. Enter data for the years after 1960 in (L1) and data for weight in (L2) with the following keystrokes

STAT ENTER 0 ENTER 8 ENTER

Hit STAT, then go to EDIT and enter the data points in L1 and L2

2. Set up the graph, go to STAT PLOT using the following keystrokes

2nd Y= ENTER ENTER ▾ ENTER ▾

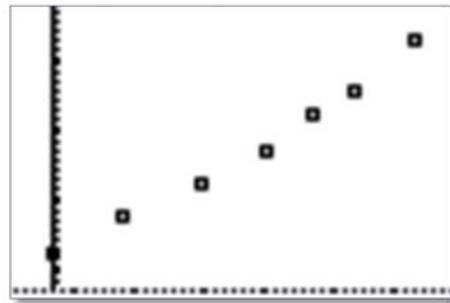
2nd 1 ENTER 2nd 2 ENTER

3. Set the scale for the graph, go ZOOMSTAT using the following keystrokes

ZOOM 9

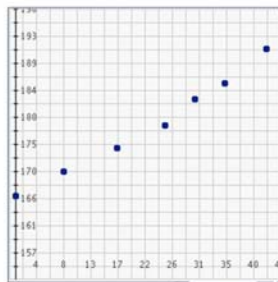
4. Hit ENTER to view graph

The calculator displays this scatter plot showing a strong linear association.



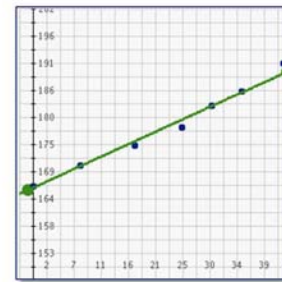
<http://illuminations.nctm.org/ActivityDetail.aspx?ID=146>

This view, from an online applet, has the added quality of a visible scale. The applet can also add a manual line of fit.



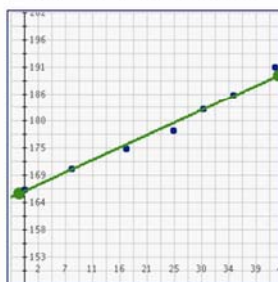
<http://illuminations.nctm.org/ActivityDetail.aspx?ID=146>

A *manual line of fit* is a line that visually seems to fit the pattern of the points. It is also called a *visual line of fit*.



N = 7  
 student guess  
 $Y = 0.55x + 166.01$

In many cases, the manual line of fit is sufficient to allow simple predictions about the data.

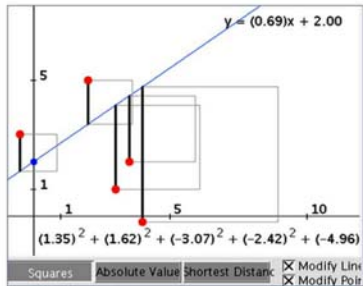


N = 7  
 student guess  
 $Y = 0.55x + 166.01$

## Meaning of the Least Squares Regression Line (LSRL)

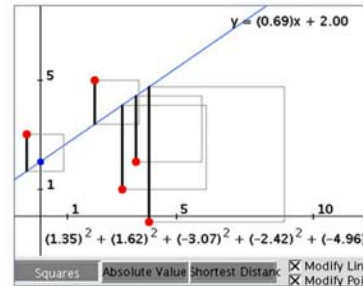
## Slides from Lines of Fit (Tutorial 26)

This view illustrates how the least squares regression line is calculated.



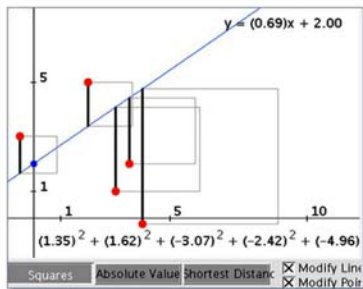
<http://standards.nctm.org/document/eexamples/chap7/7.4/index.htm#applet>

A line of fit is drawn so that the square of the distance of each data point to the line is minimized.



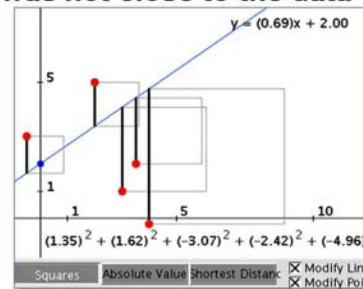
<http://standards.nctm.org/document/eexamples/chap7/7.4/index.htm#applet>

Squaring the distance is done so there are no negative signs on the differences.



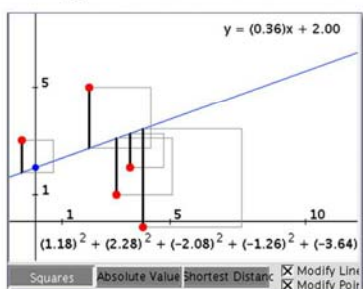
<http://standards.nctm.org/document/eexamples/chap7/7.4/index.htm#applet>

Without squaring to make distances all positive, the differences could have a sum close to zero, even though the "line of fit" was not close to the data values.



<http://standards.nctm.org/document/eexamples/chap7/7.4/index.htm#applet>

In this view the line has been adjusted and the distance squares are further minimized to give an even better line of fit.



## Calculate LSRL

Several statistics are needed:

1. The mean for independent variable  $x$  and dependent variable  $y$  called  $\bar{x}$  and  $\bar{y}$ .
2. The correlation coefficient,  $r$ , used to measure the strength of the linearity of the data.

## Slides from Lines of Fit (Tutorial 26)

### Calculate LSRL

Notes:

- When data is perfectly linear, the  $r$  value is 1 or -1
- When data has no linearity, the  $r$  value is zero
- The closer absolute value of  $r$  is to one, the better the linear fit
- In statistics, a correlation does not imply a causation

### Calculate $r$

- Center by the data to zero. Recall the mean is the balance point of all the data. This is a re-centering of the data from the mean **as a balance point to zero as the balance point**. For each data point, subtract the **mean of the data** from it.

$$X - \bar{X} \quad y - \bar{y}$$

Note: The data is centered at zero to ensure that even when  $x$  and  $y$  values are from widely different sets, the relative sizes of the data will not affect the  $r$  value.

### Calculate $r$

- Divide each value by the standard deviation of its data set.

$$\frac{X - \bar{X}}{S_x} \quad \frac{y - \bar{y}}{S_y}$$

Note: Each  $x$ -value above the mean produces a positive "new" value, while each  $x$ -value below the mean produces a negative "new" value. The same is true for the signs of the  $y$ -values.

### $x$ - and $y$ -values

**Case 1:** If  $x$ - and  $y$ -values are increasing as the data pairs move to larger  $x$ -values, both "new" values are positive.

**Case 2:** If  $x$  is decreasing and  $y$  is decreasing, the values will eventually be below the mean values for each set of data. The "new" values will both be negative.

### $x$ - and $y$ -values

**Case 3:** If  $x$  is decreasing and  $y$  is increasing, the values of  $x$  will eventually be below the mean values for each set of data. The values of  $y$  will grow to be more than the mean value of  $y$ . In this case, the "new" values will have opposite signs.

**Case 4:** If  $x$  increases and  $y$  decreases, the "new" values will also have opposite signs.

### Calculate $r$

- Take the product of each corresponding "new"  $x$ -value and its matching "new"  $y$ -value to produce a set of positive and negative numbers that show a point's position in relation to the mean of  $x$  and the mean of  $y$ .

Find the sum of the product of each  $x$ -term and each  $y$ -term. This is written as

$$\sum_{i=1}^n \frac{x_i - \bar{x}}{S_x} \cdot \frac{y_i - \bar{y}}{S_y}$$

## Slides from Lines of Fit (Tutorial 26)

### Calculate $r$

4. Find the average value by dividing the sum by the quantity  $n-1$ .

This is the **correlation coefficient** ( $r$ ) which is a measure the strength and direction of the linear fit of the data.

$$r = \frac{1}{n-1} \sum_{i=1}^n \frac{x_i - \bar{x}}{S_x} \cdot \frac{y_i - \bar{y}}{S_y}$$

### Calculate $r$

If  $r$  is **positive**, both  $x$  and  $y$  travel in the same direction so as  $x$  increases,  $y$  increases **OR** as  $x$  decreases,  $y$  decreases.

In both cases, the line has a positive slope and rises as data values move from left to right.

$$r = \frac{1}{n-1} \sum_{i=1}^n \frac{x_i - \bar{x}}{S_x} \cdot \frac{y_i - \bar{y}}{S_y}$$

### Calculate $r$

If  $r$  is **negative**,  $x$  and  $y$  travel in different directions so as  $x$  increases,  $y$  decreases **OR** as  $x$  decreases,  $y$  increases.

In both cases, the line has a negative slope and falls as data values move from left to right.

$$r = \frac{1}{n-1} \sum_{i=1}^n \frac{x_i - \bar{x}}{S_x} \cdot \frac{y_i - \bar{y}}{S_y}$$

Note: Outliers can have a strong effect on  $r$  since it is created from the mean and the standard deviation, both of which are easily influenced by outliers.

One point far from the rest of the data set can make  $r$  seem abnormally high or low.

### Find the Equation for the LSRL

### Find the Equation for the LSRL

Slope is the **correlation coefficient** ( $r$ ) times the quotient of the standard deviation for  $y$  and the standard deviation for  $x$ .

$$m = r \cdot \frac{S_y}{S_x}$$

# Tutorials for High School Mathematics

## Slides from Lines of Fit (Tutorial 26)

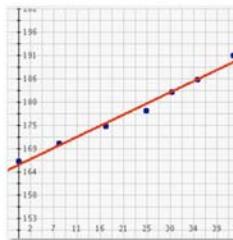
Since the LSRL contains both means, use the point-slope form of a line to find the equation of this line of fit.

Minimizing the squares of distances from the mean forces the line of fit to go through the means,  $(\bar{x}, \bar{y})$  must be a point on the LSRL.

$$y - \bar{y} = r \cdot \frac{S_y}{S_x}(x - \bar{x}) = m(x - \bar{x})$$

$$y = mx - m\bar{x} + \bar{y} = mx + b \text{ or on most calculators, } ax + b$$

### Post 1960/male weight data set:



```
LinReg
y=ax+b
a=.5727070064
b=165.455
r^2=.9851910402
r=.992567902
```

$$\hat{y} \approx .573x + 165.455$$

A  $y$  term with a caret is called “y hat”

$\hat{y}$  is the symbol for the LSRL

### Interpreting the LSQR

Fifty years after 1960, the average weight of an adult American male

$$\hat{y} \approx .573x + 165.455$$

$$\hat{y} \approx .573(50) + 165.455$$

$$\hat{y} \approx 194$$

### Interpreting the LSQR

Fifty years after 1960, the average weight of an adult American male

$$\hat{y} \approx .573x + 165.455$$

$$\hat{y} \approx .573(50) + 165.455$$

$$\hat{y} \approx 194$$

On a graphing calculator

Keystrokes to find the LSRL and to store the equation in the Y= area for graphing

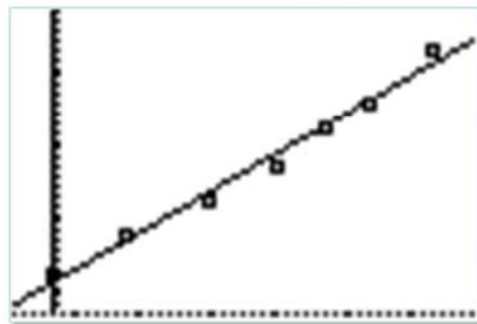
1.

```
LinReg(ax+b) L1,
L2, Y1
```

```
Plot1 Plot2 Plot3
\Y1 .57270700636
941X+165.455
```

2.

3. Hit ENTER to view the graph



## Summary

- Plot data and find a line of fit for linear data to make a prediction
- Calculate the correlation coefficient and the least squares regression line by hand and with technology.