**Linear Regression**

The purpose of linear regression is to capture the relationship between two variables that appear to have a **linear** relationship.

**Examples:** SUNLIGHT + PLANT GROWTH
HEIGHT + WEIGHT
STUDY TIME + GRADE

---

The first step in performing regression procedures is to create a **scatter plot** of the data and determine if there appears to be a linear relationship.

This is performed in the calculator by first entering the data in L1 and L2. To do this, press the **[STAT]** key and choose **1:Edit**. Then enter the x-values of your data set into **L1** and the y-values of your data set into **L2**.

Now we are ready to create a scatter plot of the data.

First check to make sure there are no functions entered in **[Y=]**. If there are, clear them.

Now press **[2nd] [Y=]** and choose **1:Plot 1 [ENTER]**. Choose **On** and the **Type:** (scatter plot).

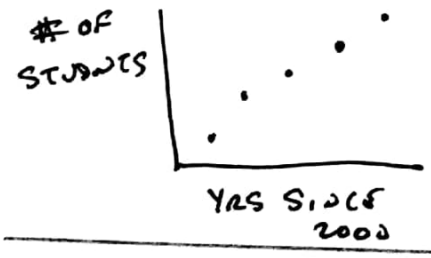
In **Xlist:** type **[2ND] [1]** for L1 and in **Ylist:** type **[2ND] [2]** for L2. (If L1 and L2 are already in Xlist and Ylist, you do not need to type them.

To see your scatter plot, press **[ZOOM] 9:ZoomStat**. Now examine the graph to see if there appears to be a linear relationship. If it is linear, we can find a line of best fit. If the data does not appear linear, a different type of model might be more appropriate.

**Example.** The table below shows the number of students in Arlington High School along with the years since 2000.

| Years since 2000 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Number of stiudents | 2892 | 3042 | 3087 | 3136 | 3296 |

Enter the data into your calculator and create a scatter plot. Sketch the plot below and be sure to label the axes. Does there appear to be a linear relationship?

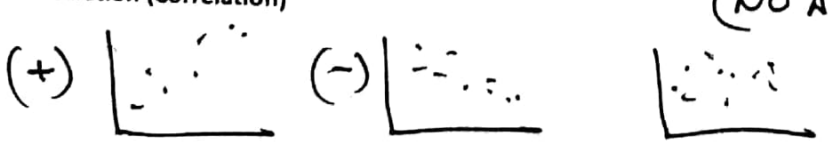



ROUGH LINEARITY
POSITIVE

AS YRS ↑, # OF STUDNTS ↑

---

**Types of Association (Correlation)**



The next step in linear regression is determining a linear equation that best fits the data. This line is often referred to as the **"line of best fit"** or the **"least squares regression line."**

To use the calculator to find the equation, press **[STAT] <CALC> 4:LinReg(ax+b)**. This should produce the LinReg Screen.

Xlist: L1

Ylist: L2

FreqList: (leave blank)

Store RegEQ: (use **[VARS] <Y_VARS> 1:Function 1:Y1** to store your equation in Y1=)

Calculate **[ENTER]**

This should produce the regression results screen that shows y=ax+b along with the slope and y-intercept of the regression line. Additionally, the equation has been stored in Y1. If you graph the scatter plot the line will show up on the graph.

**Example (cont).**

1. Compute a least squares regression line for the previous data set. Write your equation below and clearly label your variables.

$\hat{y} = 90.2x + 2729.8$

$\hat{y}$ = PREDICTED # OF STUDENTS

$x$ = YRS SINCE 2000

2. Interpret the meaning of the slope of the regression line in real life terms.

$\frac{90.2}{1}$ $\frac{\text{STUDENTS}}{\text{YR}}$ FOR EACH ADD'L YR SINCE 2000, # OF STUDENTS IS PREDICTED TO INCREASE BY 90.2 STUDENTS.

3. Interpret the meaning of the y-intercept of the regression line in real life terms.

IN 2000, MODEL PREDICTS # OF STUDENTS TO BE ABOUT 2730.

4. Use your regression line to predict the population of Arlington High School in 2012. Round your answer to the nearest whole number. What does this assume about the population of Arlington High School?

$2012 - 2000 = 12$ $\quad \hat{y}(12) = 3812.2 \Rightarrow 3812$ STUDENTS.

ASSUMES POP. CONTINUES TO GROW @ SAME RATE.

5. Use your model to determine between which two years the population is expected to reach 4500.

$90.2x + 2729.8 = 4500$

$-2729.8 \quad -2729.8$

$90.2x = 1770.2$

$x = \frac{1770.2}{90.2} \approx 19.6$ YRS

DURING 2020

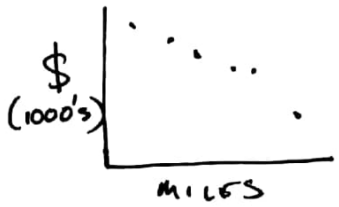
**Interpolation:** PREDICTING VALUES WITHIN DATA SET

**Extrapolation:** PREDICTING VALUES OUTSIDE OF RANGE OF DATA (BAD).

**Application**. A real estate agent is trying to determine the relationship between the distance a 3-bedroom home is from New York City and its average selling price. He records data for the 6 homes shown below. USE 1000's

| Miles from NYC | 10 | 35 | 50 | 65 | 75 | 120 |
|---|---|---|---|---|---|---|
| Price | 755,000 755 | 650,000 650 | 580,000 580 | 505,000 505 | 475,000 475 | 285,000 285 |

a. Create a scatter plot of the data. Sketch your graph below. Label the axes. Does there appear to be a relationship between home price and distance from New York City?




b. Determine a regression model for the data. Write the equation of your model below. Be sure to clearly label your variables.

$\hat{y} = -4.29x + 799.6$ $\hat{y}$ = PREDICTED PRICE (IN 1000S)
$x$ = MILES FROM NYC

c. What are the real life meanings of the y-intercept and the slope of your equation?

$-4.29 \frac{\$}{mi}$ FOR EVERY MILE FROM NYC, PRICE IS PREDICTED TO DROP $4.29 \times 1000$ \$.

PRICE OF HOUSE IN NYC IS PREDICTED TO BE \$799,600.

d. Use your model to determine the predicted price of a home in Woodstock, NY which is located 95 miles from NYC. Would this prediction be *interpolation* or *extrapolation*? Why?

$\hat{y}(95) = \$387.9 \times 1000$ INTERPOLATION BECAUSE 95 MILES IS WITHIN RANGE OF x-VALUES.

e. Use your model to determine, to the nearest mile, the distance from NYC a home would be if the selling price was exactly \$500,000.

$500 = -4.29x + 799.6$
$-799.6 \quad -799.6$
$-299.6 = -4.29x$
$x = 69.8 \Rightarrow$ 70 MILES

(PRACTICE #1-3)
APPLICATIONS →

**Determining How Good a Linear Model Is**

**Residuals** – Obviously when constructing a linear regression model, the actual data points will not be all on the regression line. The distance a point is from the line can be thought of as an error. In other words, how much the model is off from the actual data. This error is called a **residual**.

Residual = (ACTUAL Y-VALUE) − (PREDICTED Y-VALUE)

$= y - \hat{y}$

↑ FROM DATA ← FROM PREDICTION LINE

RESID = A − P

The line of best fit is chosen to minimize the sum of all the residuals. In order to see how good a model fits the data set, a **residual plot** is constructed.

After conducting a linear regression, the residuals will automatically be stored in the calculator. To view the graph of the residuals, the following steps should be taken.

Select **1:Plot 1** and turn it **OFF** ↙ ALPHA KEYS

Select **2:Plot 2** and turn it **ON**. Select the **Scatter Plot** icon. In **Xlist:** enter **L1** and in **Ylist:** enter **RESID**.

Choose **[ZOOM] 9:ZoomStat**

A model that fits a data set well will have a residual plot that is **RANDOMLY SCATTERED** which shows no obvious patterns.

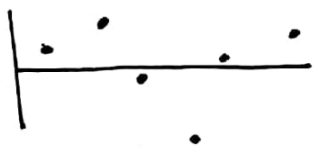
**Example**. Let's return to the New York real estate example.

| Miles from NYC | 10 | 35 | 50 | 65 | 75 | 120 |
|---|---|---|---|---|---|---|
| Price | 755,000 | 650,000 | 580,000 | 505,000 | 475,000 | 285,000 |

1. Enter the data into L1 and L2 and perform a regression. Write the regression equation below being sure to label the variables. $\hat{y} = -4.29x + 799.6$

$\hat{y}$ = PREDICTED HOUSE PRICE (IN 1000's) x = DIST FROM NYC.

2. Create a residual plot and sketch the graph below.



3. What does this residual plot say about the model in general?

GENERALLY GOOD. SCATTERED + RANDOM.

4. Compute the residual value for the house 35 miles from New York City.

$\hat{y}(35) = 645.3$ } $4.63 \times 1000$ OVERPREDICT.
$y(35) = 650$ }

USE GRAPH "TRACE" 4.631703

(5)