

Ever wonder how they  
come up with those  
equations given in algebra?

My Algebra book tells me that a female's life expectancy is related to her birth year by the equation  $y = 0.12x + 81.35$  where  $y$  is the life expectancy of a female born in the US  $x$  years after 2017. Another problem tells me that the number of bacteria in a Petri dish,  $y$ , at time  $t$  (hours) obeys the equation  $y = 980e^{0.02t}$ .

How do they get these equations? Most often, the answer is regression. The idea is that they obtain many paired data values, like (birth year, life expectancy) or (hour, number of bacteria). They plot these points and perform statistics called **regression analysis**. We will explore some problems, letting the calculators do the heavy lifting.

Let's cover the basics first.

Given a scatter plot of data, we will find the  
line or curve that best fits the pattern of points.

**Recall: Definition: Linear relationship:** A **linear relationship** is a relationship between two variables, often denoted by  $x$  and  $y$ , where the graph is a **straight line**.

The most commonly used equation that describes a linear relationship is  $y = mx + b$ . Here  $m$  is the slope of the line,  $b$  is the  $y$ -intercept, and  $(x, y)$  is a generic point on the line.

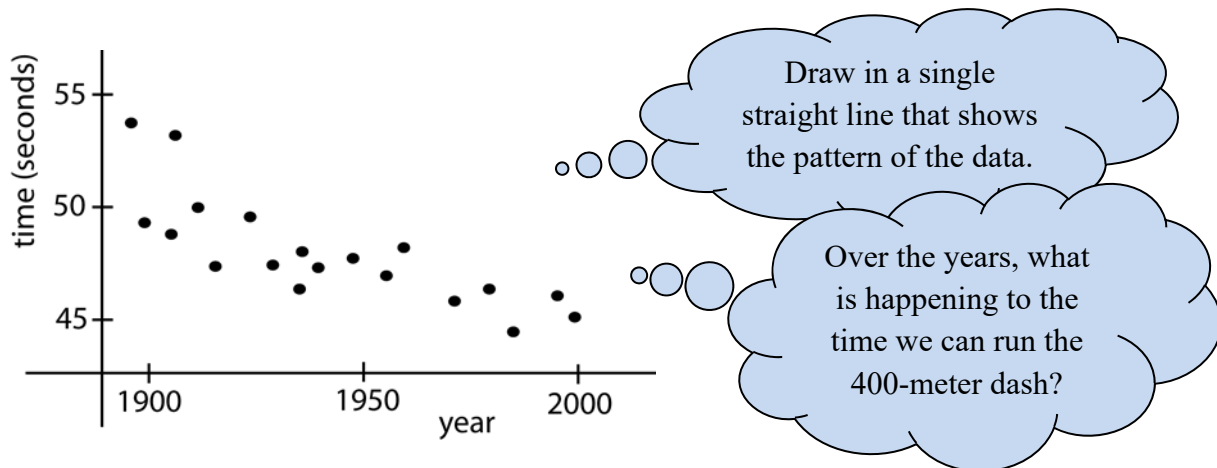
Can you make  
up an example?

We will find the values of two characteristics for many individuals and organize that data in the form of ordered pairs. For instance, we might ask many adults for their income and years of college education or look up the winning times for running a particular race along with the year. We then make a scatter diagram of these points and look for a consistent trend among the points. This is the idea of **regression**.

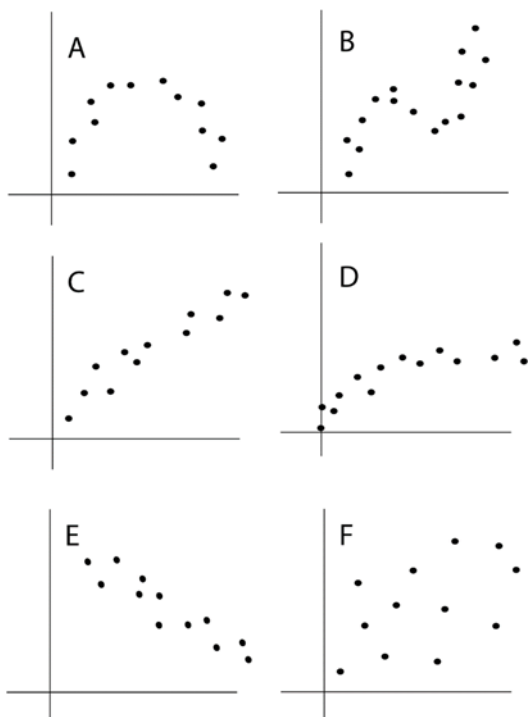
**Definition:** A **scatter plot** (or **scatter diagram**) is a graph that shows the relationship between two variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The **independent variable** is plotted on the horizontal axis, and the **dependent variable** is plotted on the vertical axis.

expl 1: Take a look at the scatter plot below that shows the relationship between the time it takes the world's fastest men to run the 400-meter dash and the year.

Notice how the scatter plot takes on a linear pattern. If we were to find the equation of the line that best fits this pattern of points, we could use it to predict the time it takes to run the 400-meter dash in any given year. That is the idea of regression.



expl 2: Consider the following scatter plots. Which do you think show a linear relationship? On each graph, draw in the line or curve that mimics the pattern of points.



Some relationships have a “cigar” shape to them. They are considered linear.

Some are curved, or non-linear. We will play with these relationships too.

Some have *no* discernible pattern at all. We say there is *no* correlation.

### Linear Regression:

There is a rather complicated formula to find the line that best fits the data. The method is called the **Least Squares Regression Line** (because of how it is derived) or, simply, the **line of best fit**. Luckily, we are *not* required to do this calculation by hand; we will use the calculator.

### Worksheet: Linear regression on your calculator:

We will explore a couple of examples with step-by-step instructions on how to enter the data, make a scatter plot, and find and graph the regression equation using the calculator.

**Definition: Coefficient of correlation or correlation coefficient**, denoted by  $r$ : This number tells us how well the line fits the pattern of points and if the slope of the line is positive or negative.

The coefficient of correlation ranges from  $-1$  to  $1$ .

If  $r$  is negative, the line has a negative slope.

If  $r$  is positive, the line has a positive slope.

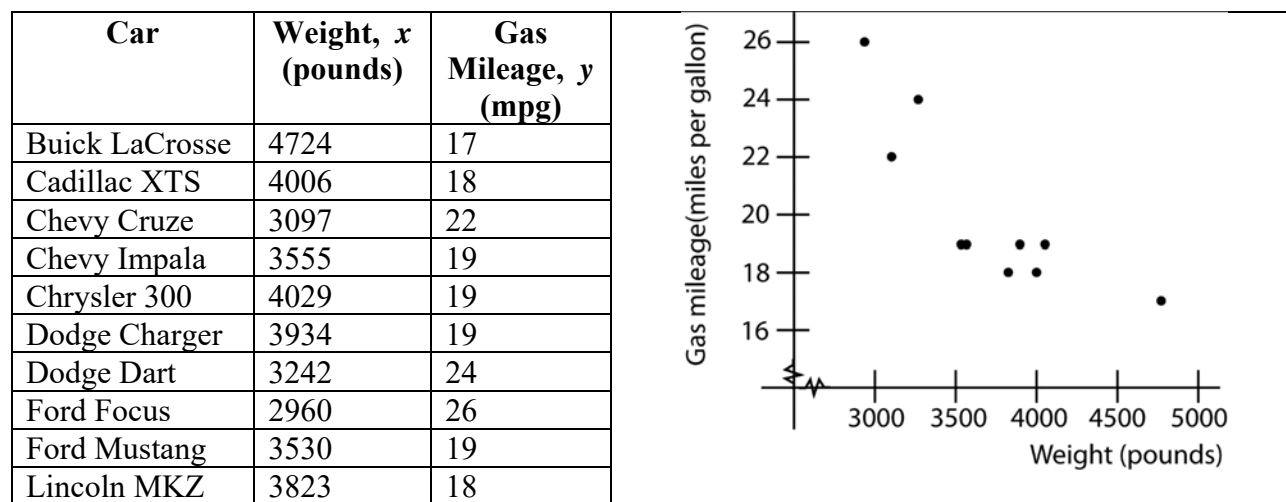
**The closer  $r$  is to  $-1$  or  $1$ , the better the fit.**

**If  $r$  is close to  $0$ , then there is *no linear* pattern.**

We will analyze  $r$  to see how well our line fits the data.

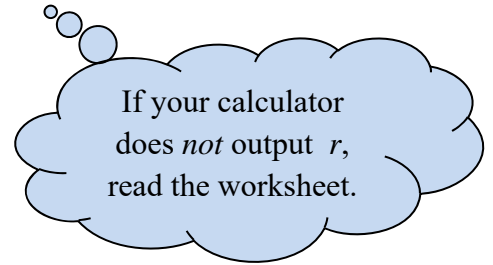
An  $r$  value close to  $0$  does *not* mean there is *no* pattern, just *not a linear* one.

expl 3: Consider the data in the table below. It gives the weights of various cars along with their gas mileages. Look at the scatter plot; do you think the two variables are linearly related?



expl 3 (continued):

a.) Find the least-squares regression line using the calculator. Record the equation of the line and the value of the correlation coefficient  $r$ .



b.) Does your value of  $r$  indicate that the line is a good fit?

c.) My VW Jetta weighs in at 3,000 pounds. Estimate its gas mileage.

**Recall: Definition: Quadratic relationship:** A **quadratic relationship** is a relationship between two variables, often denoted by  $x$  and  $y$ , where the graph is a **parabola**.

The most commonly used equation that describes a quadratic relationship is  $y = ax^2 + bx + c$  where  $a$  is *not* zero. Here,  $(x, y)$  is a generic point on the line.

Can you make up an example?

### Quadratic Regression:

We saw how the pattern of a scatter plot of points could be represented by a single linear equation. But *not* all scatter plots show a linear pattern. Look back at the plots on page 2. Which reminds you of a parabola?

expl 4a: The number of foreign adoptions in the U.S. has declined in recent years, as shown in the table to the right.

i.) Use your calculator to draw a scatter plot and then fit a **quadratic** function to this data. Let  $x$  represent the number of years since 2000. Round your equations' coefficients to three decimal places.

Take note of how they define  $x$ .

Year, $x$	Number of U.S. Foreign Adoptions from Top 15 Countries, $y$
2000, 0	18,120
2001, 1	19,087
2002, 2	20,100
2003, 3	21,320
2004, 4	22,911
2005, 5	22,710
2006, 6	20,705
2007, 7	19,741
2008, 8	17,229
2009, 9	12,782

Notice how the table values first increase and then decrease.

ii.) Use the function from part  $a$  to estimate the number of U.S. foreign adoptions in 2010.

expl 4b: Use your regression equation to estimate the number of adoptions in the year 2050. Why does this value *not* make sense? (Using your regression equation to predict values well outside your original data set is called **extrapolation** and should *not* be done.)

### Interpolation versus Extrapolation:

Truly, our regression equation should only be used to predict foreign adoption numbers for years that are close to those values given in the original data. (This is called **interpolation**, as opposed to **extrapolation** which is a no-no.)

### Worksheet: Exploring Regression:

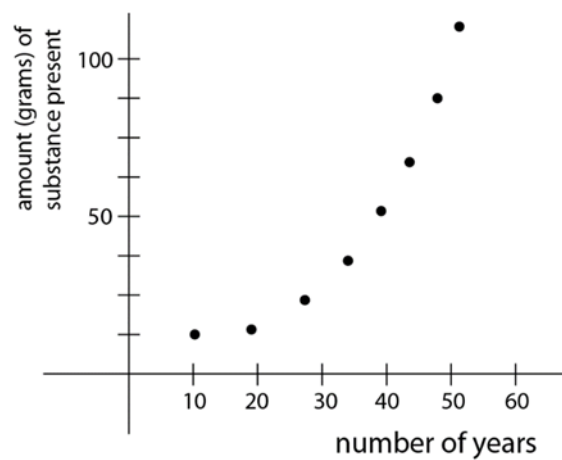
For this worksheet, we do *not* find an equation but rather just draw in a line that best fits the points given (in your personal opinion). You will then use it to *graphically* find some  $y$ -values given  $x$ -values. We will also play with the ideas of interpolation and extrapolation.

### Exponential Functions and Regression:

What if the points on our graph do *not* look linear or curved like a parabola? What if they look like this picture to the right?

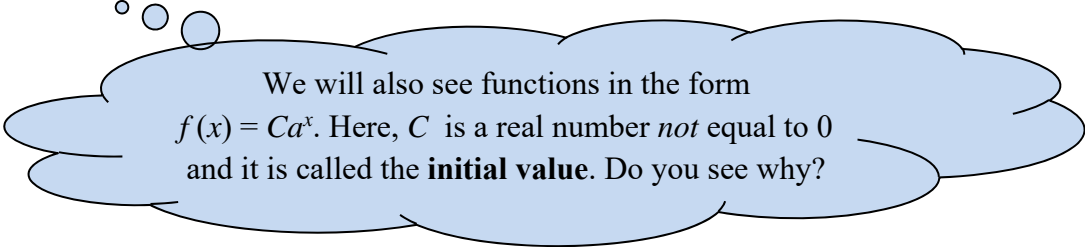
As you go left to right, the points are going up. The rise is slower at first but then takes off rather quickly. Shloop!

This is an exponential relationship.



**Definition: Exponential Function or Relationship:**

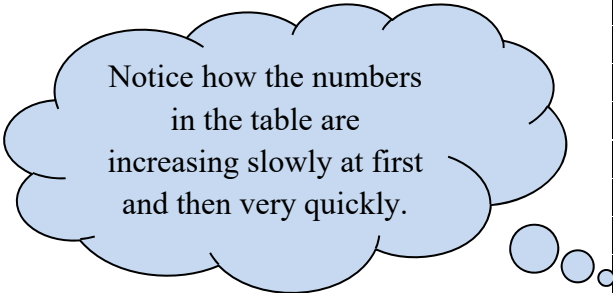
An **exponential function** is of the form  $f(x) = a^x$  where  $a$  is a positive real number *not* equal to 1. The number  $a$  is called the **base** or **growth factor**. Notice the variable  $x$  is in the exponent position. This is what makes it an exponential function.



We will also see functions in the form  $f(x) = Ca^x$ . Here,  $C$  is a real number *not* equal to 0 and it is called the **initial value**. Do you see why?

expl 5: The data in the table below represent the annual revenue of Tesla, Inc. from 2010 to 2018. Answer the following questions.

- a.) Use your calculator to draw a scatter plot using  $x$  to be the number of years since 2010.



Notice how the numbers in the table are increasing slowly at first and then very quickly.

Year	Revenue (\$ Billion)
2010 ( $x = 0$ )	0.12
2011 ( $x = 1$ )	0.20
2012 ( $x = 2$ )	0.41
2013 ( $x = 3$ )	2.01
2014 ( $x = 4$ )	3.20
2015 ( $x = 5$ )	4.05
2016 ( $x = 6$ )	7.00
2017 ( $x = 7$ )	11.76
2018 ( $x = 8$ )	21.46

- b.) Use your calculator to build an **exponential** regression equation. Graph this function on top of the scatter plot to check it visually. Does it fit the data?

- c.) Use your function to predict Tesla's revenue in 2020.

**Instructions for TI Calculators:**

To enter data, you press the **STAT** button and select **Edit** (under **EDIT** menu) to enter the data into the columns **L1** and **L2** (The  $x$ -values go in **L1**; the  $y$ -values go in **L2**). To calculate the regression equations, press **STAT** once again but arrow over to **CALC**. In the list here, you will see **LinReg (ax + b)** for linear regression, **QuadReg** for quadratic regression, and **ExpReg** for exponential regression. There are lots of other types too which we do *not* cover.

Again, we enter the independent ( $x$ ) variable in **L1** and the dependent variable ( $y$ ) in **L2**. Time is usually considered to be the independent variable. Pay attention to instructions given in the homework.

Round coefficients to four decimal places.

Use the **STATPLOT** editor to draw scatter plots as required. Of course, choose window settings that fit the data. Remember, most calculators will allow you to enter a regression equation into the **y-editor** (for **Y1**) by pressing **VARs** > **Statistics**. You then arrow over to **EQ** and select **RegEQ**. This is *not* necessary but looks neat.