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

College algebra
Linear Regression (section 4.2)

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.

$y = 8x + 4$

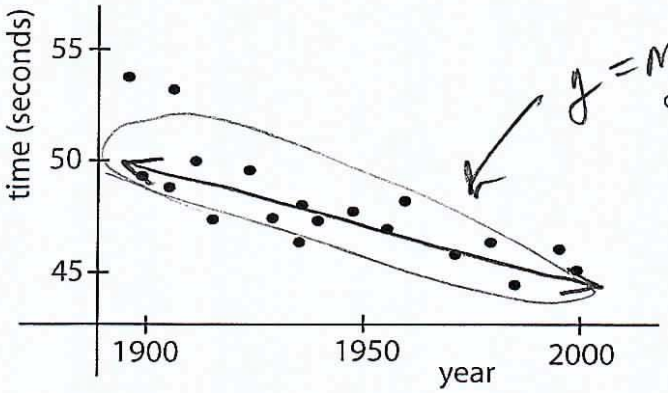
Can you make up an example?

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.

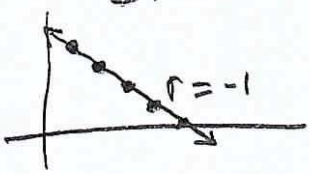
Notice how the scatter plot takes on a linear pattern. If we were to find the equation of the line that best fits this pattern of points, we could use it to predict the time it takes to run the 400 meter dash in any given year. That is the idea of regression.



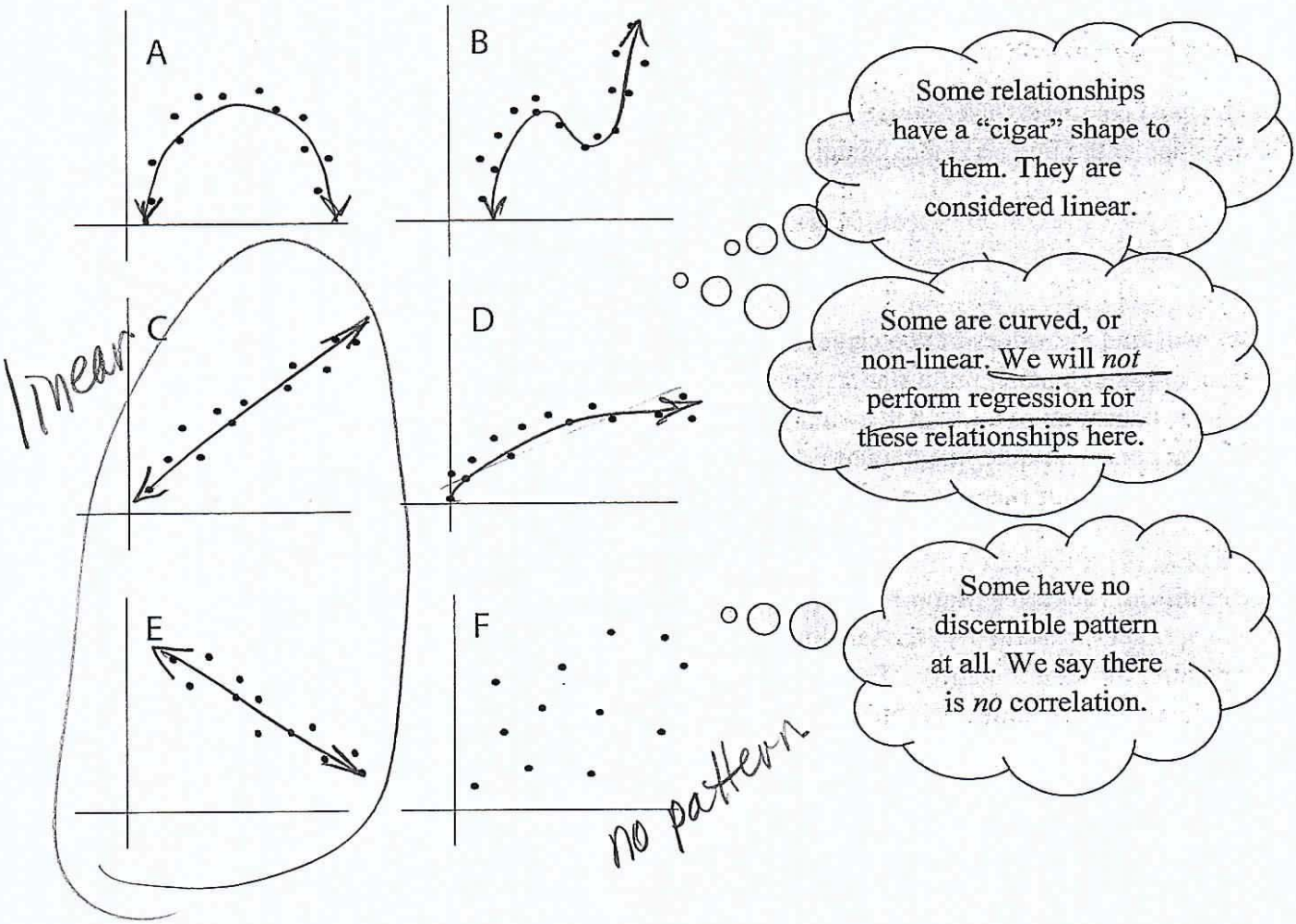
$y = mx + b$

Draw in a single straight line that shows the pattern of the data.

If ...



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.



expl 3: What is the main difference between the graphs in parts *c* and *e* above? What would that imply about how the variables are related?

c is increasing (pos. slope)
e is decreasing (neg. slope)

We could take two of our many points and find the line that goes through them. That would give us a line that shows the trend of the data. However, that is *not* good enough for us. We want to use all of the points somehow. That is where the methods of regression come in.

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.

reference only

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.

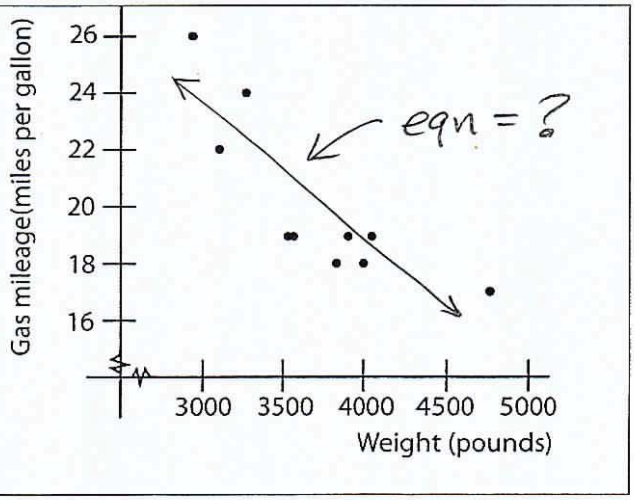
We will analyze r to see how well our line fits the data.

Return to page 1. Can you estimate r for the 400 meter dash scatter plot? Is it positive or negative? What does that mean about the relationship between the time it takes to run 400 meters and the year?

$r \approx -0.8$

expl 4: 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?

Car	Weight, x (pounds)	Gas Mileage, y (mpg)
Buick LaCrosse	4724	17
Cadillac XTS	4006	18
Chevy Cruze	3097	22
Chevy Impala	3555	19
Chrysler 300	4029	19
Dodge Charger	3934	19
Dodge Dart	3242	24
Ford Focus	2960	26
Ford Mustang	3530	19
Lincoln MKZ	3823	18



expl 4 (continued):

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

$f(x) \approx -0.00468x + 37.357$

$r \approx -0.841576$

If your calculator does *not* output r, read the worksheet.

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

The value of r is about −0.84 and so is close to −1. This means the line is a good fit

c.) The slope of this line can be thought of as the **average rate of change**. Interpret the slope of your line with regard to gas mileage and vehicle weight.

$\text{slope} = \frac{-0.00468 \text{ mpg}}{1 \text{ lb}}$

$m = \frac{y_2 - y_1}{x_2 - x_1}$

→ The gas mileage of a car decreases by 0.00468 mpg for every additional pound of car weight, on average.

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

$f(x) = -0.00468x + 37.357$

x = weight of car

$f(x)$ = gas mileage

$f(3000) \approx 23$ mpg

What is the Domain of the Function?

You will be asked what the domain of the function is. There are two answers, one with a statistical bent and one with an algebraic bent. The book will expect the algebraic answer.

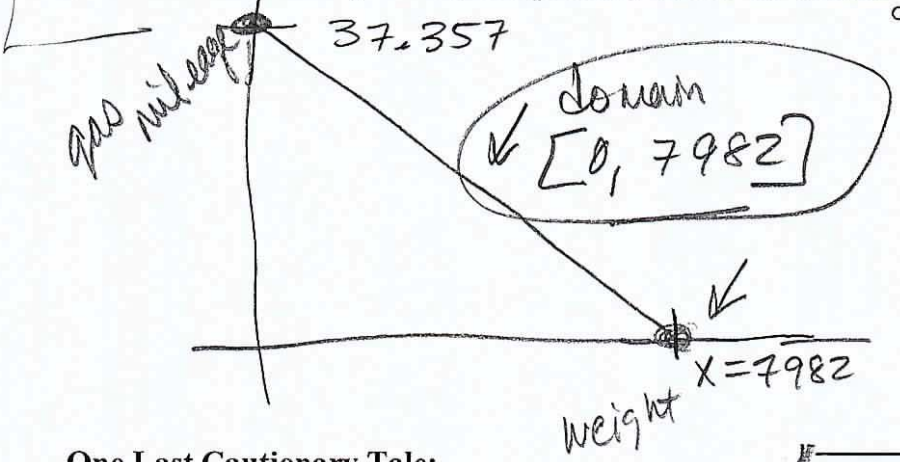
Statistically speaking:

Truly, our regression equation should only be used to predict gas mileages for weights that are close to those values given in the original data. (This is called **interpolation**, as opposed to **extrapolation** which is a no-no.) So, what is the domain (or acceptable values for x) of our regression equation?

→ x should be in range from 2960 to 4724 lbs.

Algebraically speaking:

If you simply consider the function we found and ask what the domain is, we will get a different answer. Draw a graph of the function now, labeling the axes, and mark the x-intercept. Think about what the x and y variables represent; what is the domain of this function?



Again, you are expected to do this when asked.

$0 = -0.00468x + 37.357$

$x \approx 7982$

One Last Cautionary Tale:

Some problems will define the independent variable strangely to newcomers. Here, the x variable is "decade". They use this notation to define the x-values as 1, 2, 3, 4, and 5; they do *not* use the year spans as that would *not* work in the regression formulas.

Decade, x	Major Hurricanes Striking Atlantic Basin, H
1921–1930, 1	17
1931–1940, 2	16
1941–1950, 3	29
1951–1960, 4	33
1961–1970, 5	27

They do a similar thing here by denoting t takes on values 1, 2, 3, 4, 5, and 6 as opposed to the actual years. They do this often so that the calculations involve smaller numbers.

Year, t	Percent below Poverty Level, p
2005, 1	14.5
2006, 2	14.6
2007, 3	15.0
2008, 4	15.7
2009, 5	17.1
2010, 6	18.5

Source: U.S. Census Bureau

Be sure to enter these properly in the calculator.