

## CHAPTER 6: EXPLORING DATA RELATIONSHIPS

How should we look for relationships in our data? How much can we trust the relationships that we find?

Does the grade on Exam 1 predict the grade on Exam 2 in a math class?

Does the temperature influence the rate at which crickets chirp?

Does the weight on a spring change the amount the spring stretches?

A *response variable* measures an outcome or result of a study.

An *explanatory variable* is a variable that we think explains or causes changes in the response variable.

### 6.1 Displaying Relationships: Scatterplots

#### EXAMPLE

The data below is the average monthly temperature ( $^{\circ}\text{F}$ ) and the average daily kWh of electricity used at a particular home each month.

		AUG	SEP									
Temp ( $^{\circ}\text{F}$ )	86	85	83	77	69	62	55	51	52	61	71	80
Electricity (kWh)	76	70	66	38	26	22	25	33	30	26	38	60

Which is the explanatory variable and which is the response variable?

**(A) The temperature is the response variable**

**(B) The electricity is the response variable**

(c) Epstein 2013

Chapter 6

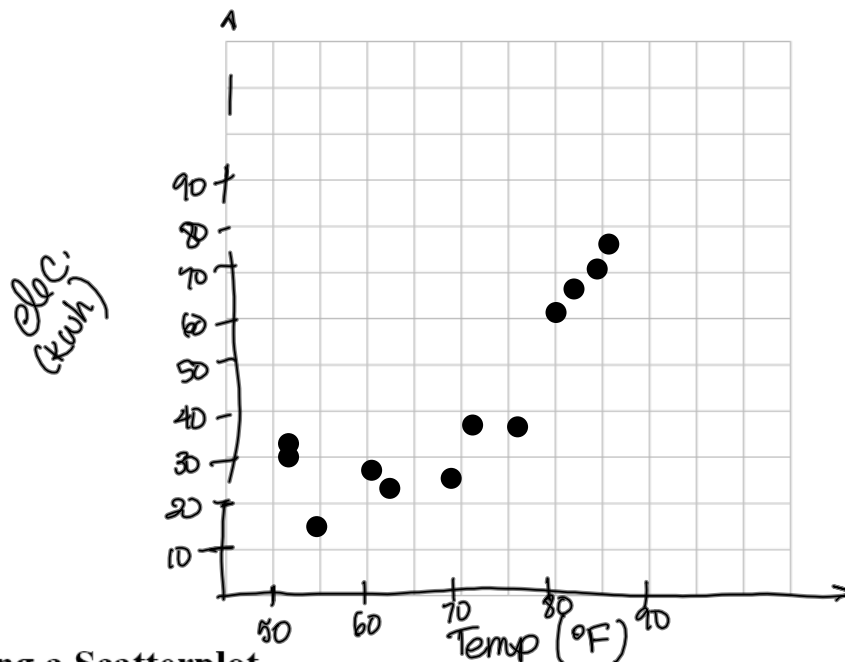
page 2

A **scatterplot** is a graph of two values with explanatory variable on the horizontal axis and the response variable on the vertical axis.

### EXAMPLE

The data below is the average monthly temperature ( $^{\circ}\text{F}$ ) and the average daily kWh of electricity used at a particular home each month. Display this information in a scatterplot.

Temp ( $^{\circ}\text{F}$ )	86	85	83	77	69	62	55	51	52	61	71	80
Electricity (kWh)	76	70	66	38	26	22	25	33	30	26	38	60



### Examining a Scatterplot

- Look at the overall pattern:
  - Is the form linear or not?
  - What is the general direction?
  - How strong is the relationship
- Look for outliers – do any data points fall outside of the general pattern?

(c) Epstein 2013

Chapter 6

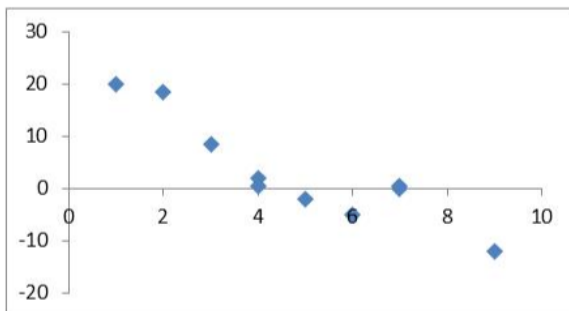
page 3

Two variables are **positively associated** if an *increase* in one variable tends to accompany an *increase* in the other variable.

Two variables are **negatively associated** if an *increase* in one variable tends to accompany a *decrease* in the other variable.

EXAMPLE

Examine the scatterplots below and determine if there is a positive or negative association between the explanatory and response variables. Is it a linear relationship? Is it a strong relationship? Are there outliers?

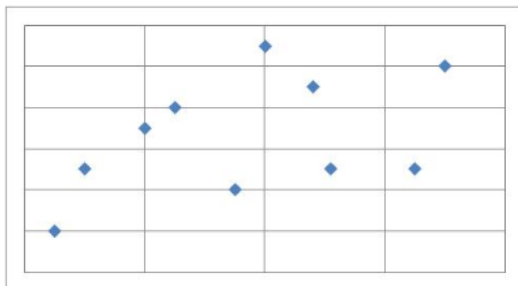


How are the variables in the graph to the right associated?

(A) positively

**(B) negatively**

(C) not associated

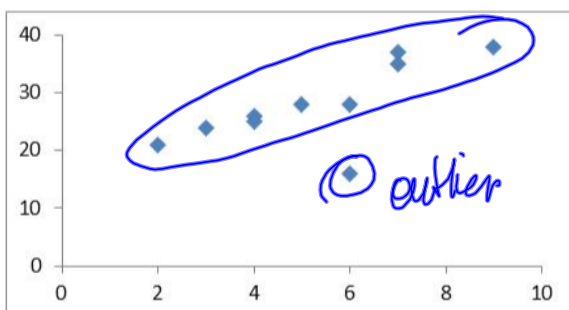


How are the variables in the graph to the right associated?

(A) positively

(B) negatively

**(C) not associated**



How are the variables in the graph to the right associated?

(A) positively

(B) negatively

(C) not associated

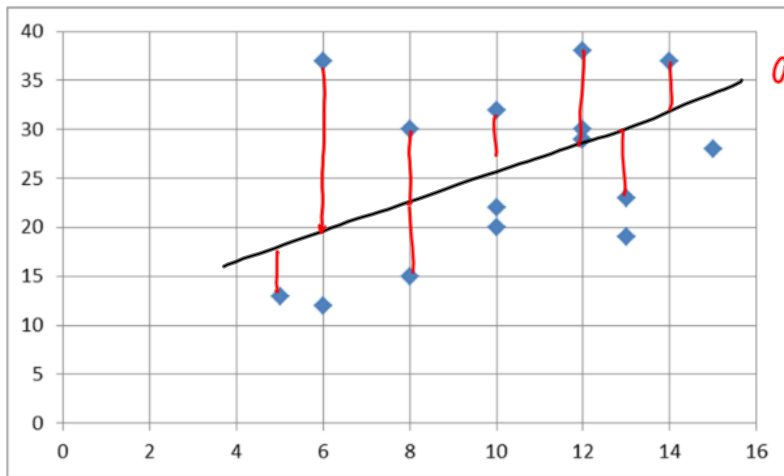
**6.2 Making Predictions: Regression Lines**

A **regression line** is a straight line that describes how the response variable changes as the explanatory variable changes.

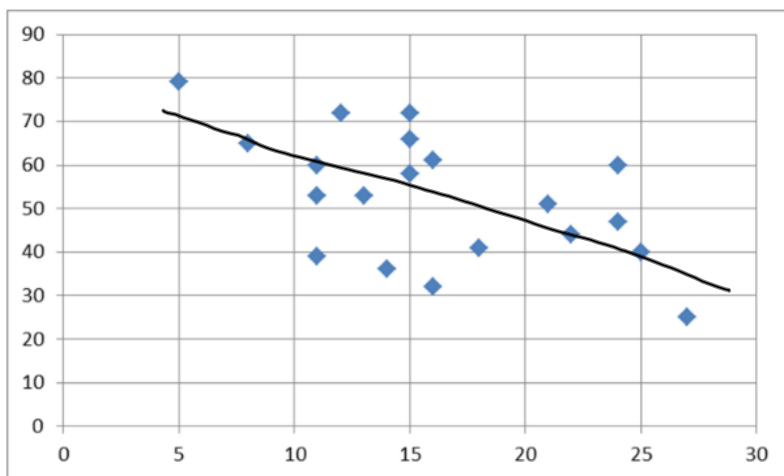
The regression line is a line that is as close as possible to all the points.

EXAMPLE

Draw the estimated regression line for the scatterplots in the previous example and the scatterplots below.



$ax+b=y$



TRY This

Clicker Question

Did you try this?

- (A) Yes, before class
- (B) Yes, just now
- (C) No

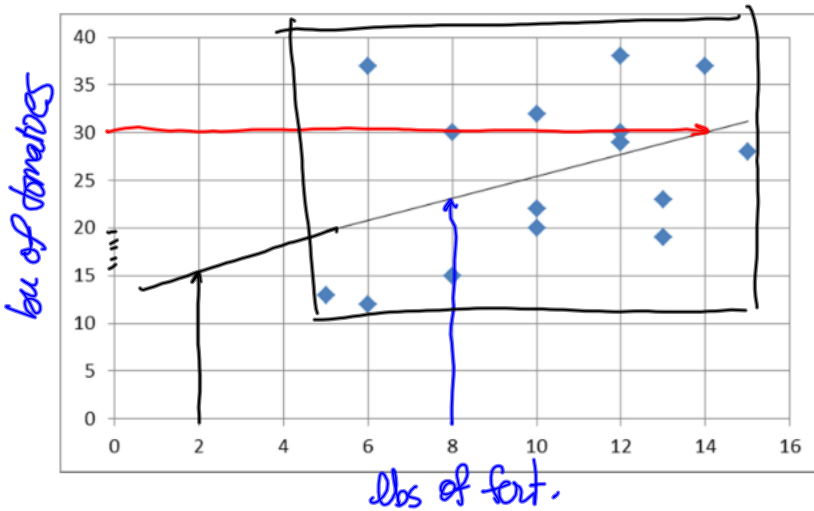
(c) Epstein 2013

Chapter 6

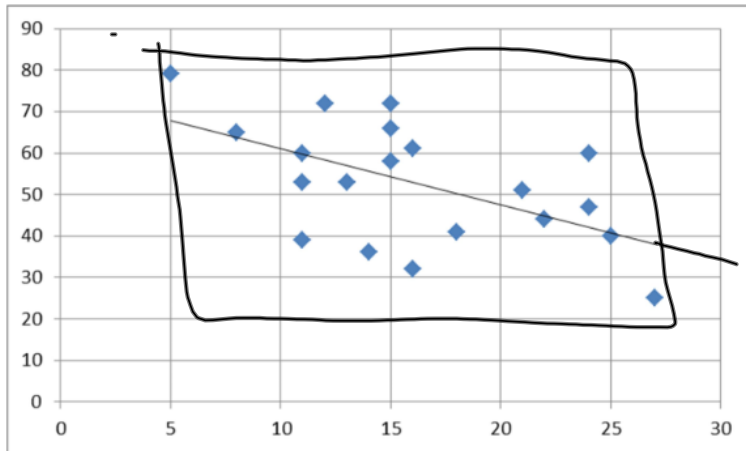
page 5

EXAMPLE

Use the given regression line to make predictions about the variables below.



- (a) If  $x = 8$ , what is  $y$ ?  
23 bu of tom
- (b) If  $y = 30$ , what is  $x$ ?  
14 lbs of fert.
- (c) If  $x=2$ , what is  $y$ ?  
15 bu of tom.



- (a) If  $x = 15$ , what is  $y$ ?  
55
- (b) If  $y = 50$ , what is  $x$ ?  
17 (18)
- (c) If  $x = 30$ , what is  $y$ ?  
32 ✗

**Interpolation** is using the regression line to find values *between* the minimum and maximum data values.

**Extrapolation** is using the regression line to find values that are *outside* the minimum and maximum values.

**Was our answer to part (c) above interpolation or extrapolation?**

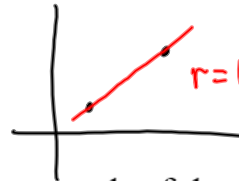
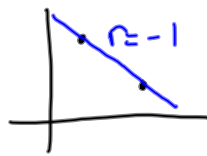
- (A) interpolation
- (B) extrapolation

(c) Epstein 2013

Chapter 6

page 6

**6.3 Correlation**



The **correlation** measures the direction and strength of the straight line relationship between two numerical variables.

The value of the correlation is a number  $r$  that is between -1 and 1, inclusive. That is  $-1 \leq r \leq 1$ .

*inclusive*

For positive association,  $r > 0$ . For negative association,  $r < 0$ . For no association,  $r = 0$ .

The closer  $|r|$  is to 1, the stronger the association.

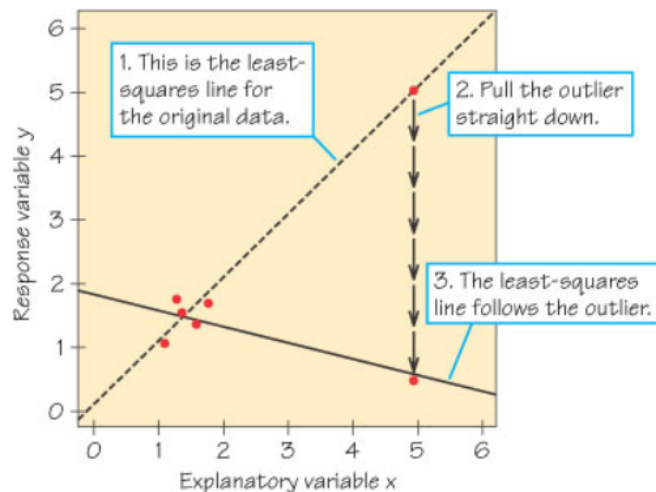
~~The  $r$  value for the Prius mileage is -0.989 and the  $r$  value for the spring distance is 0.998.~~

**6.4 Least Squares Regression**

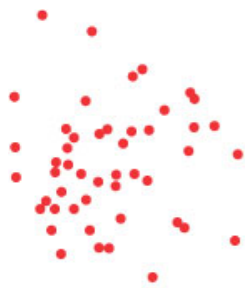
Skip this section.

**6.5 Interpreting Correlation and Regression**

Outliers will have a very strong effect on both the regression line and the correlation!



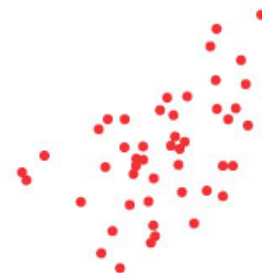
p192



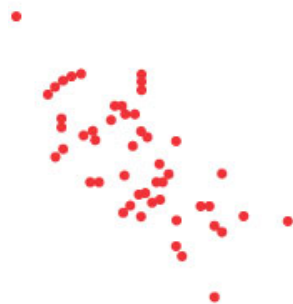
Correlation  $r=0$



Correlation  $r=-0.3$



Correlation  $r=0.5$



Correlation  $r=-0.7$



Correlation  $r=0.9$



Correlation  $r=-0.99$

Page 192 of your textbook

(c) Epstein 2013

Chapter 6

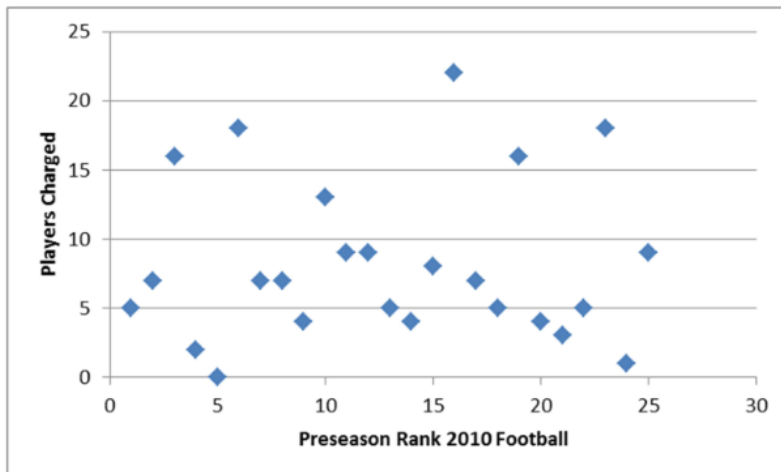
page 7

**CORRELATION DOES NOT MEAN CAUSATION!****EXAMPLE**

Before the introduction of the polio vaccine, investigators noticed a strong positive correlation between the amount of soda drinks and ice cream sold in a week and the incidence of polio. Does that mean that polio is caused by soda and ice cream?

**EXAMPLE**

[http://sportsillustrated.cnn.com/2011/writers/the\\_bonus/02/27/cfb.crime](http://sportsillustrated.cnn.com/2011/writers/the_bonus/02/27/cfb.crime)

**EXAMPLE**

Eating breakfast helps students be successful.

[http://amcbt.indstate.edu/volume\\_30/v30-4p15-19.pdf](http://amcbt.indstate.edu/volume_30/v30-4p15-19.pdf)

**EXAMPLE**

Coffee is bad for you.

<http://www.mayoclinic.com/health/blood-pressure/AN00792>

Coffee is good for you

<http://www.webmd.com/food-recipes/features/coffee-new-health-food>