

CHAPTER 6 – EXPLORING DATA RELATIONSHIPS

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

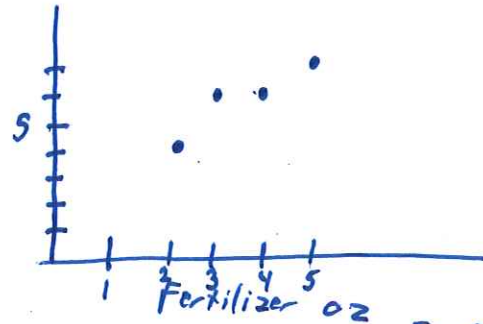
An *explanatory variable* is a variable that we think explains or causes changes in the response variable.

For the studies shown below, decide which is the explanatory variable and which is the response variable. Then display the data in a scatterplot.

(a) A farmer fertilized her tomato plants with different amounts of fertilizer as shown below:

tomatoes (lbs)	4	6	6	7
fertilizer (oz)	2	3	4	5

response var
↓
tomatoes in lb



← explanatory variable

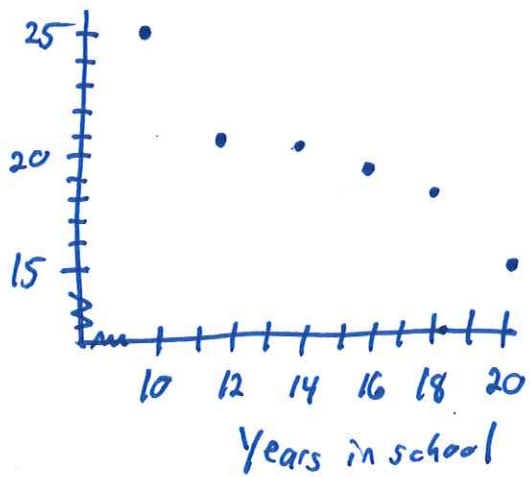
(b) A study was done on the number of hours of television watched by adults and the number of years of school. The results were

← explanation

← response

years in school	10	12	14	16	18	20
hours of television per week	25	21	20.5	19	18	15

hrs of TV/week



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.

How is the amount of fertilizer related to the amount of tomatoes?

positively assoc.

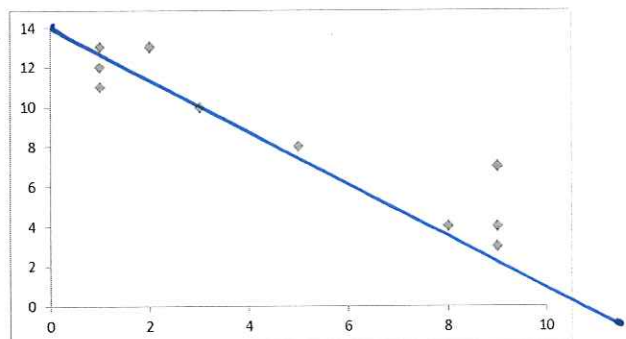
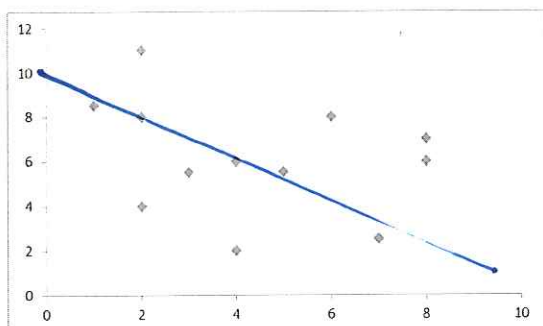
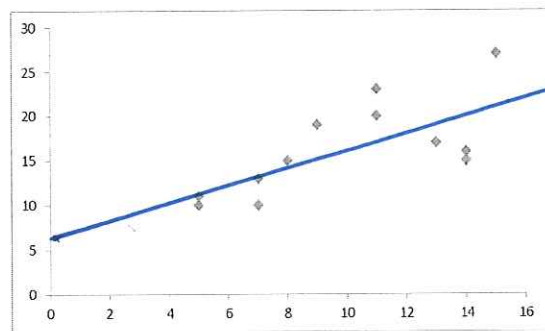
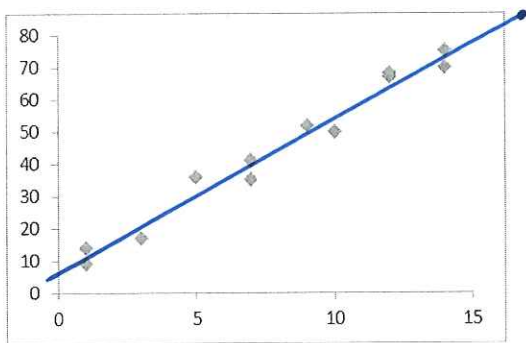
How is the amount of school related to the amount of television watched?

Negatively assoc.

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.

Sketch the regression line for the scatterplots below:



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

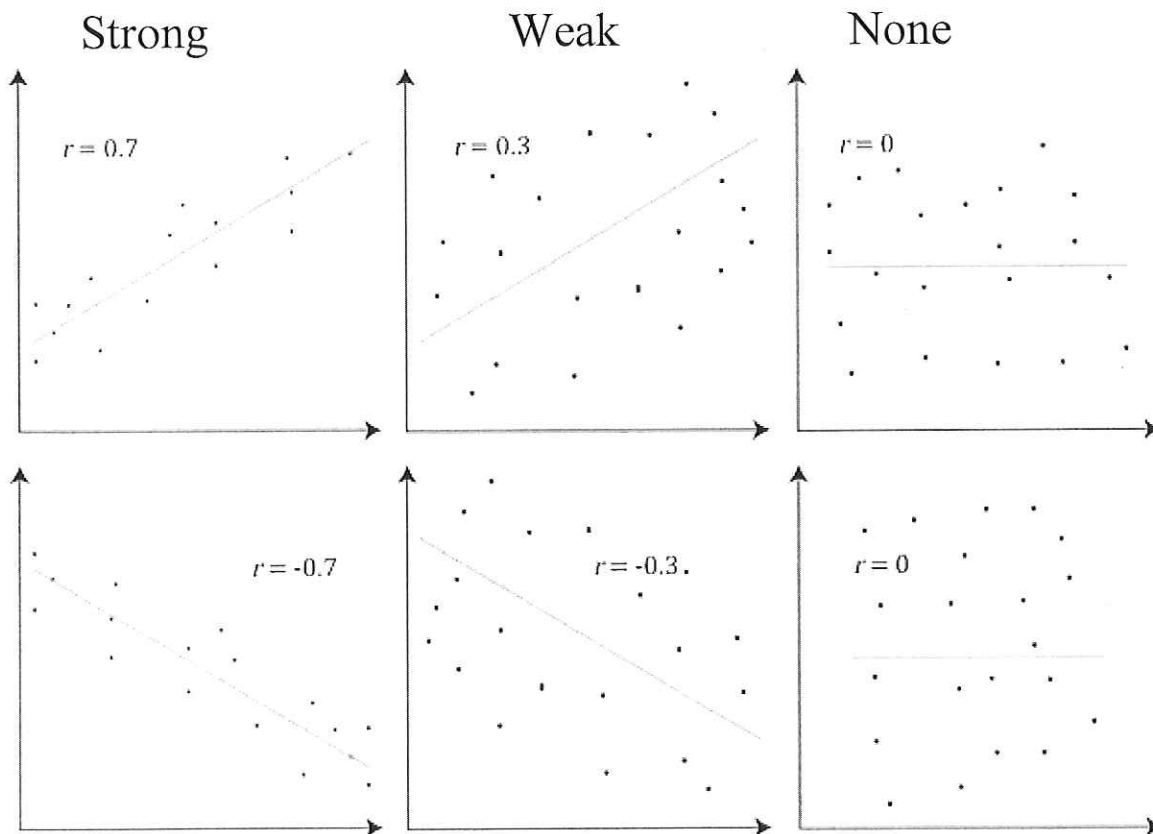
The value of the correlation is a number r that is between -1 and 1 , inclusive. That is, $-1 \leq r \leq 1$.

For positive association, $r > 0$.

For negative association, $r < 0$.

For no ^{linear} association, $r = 0$.

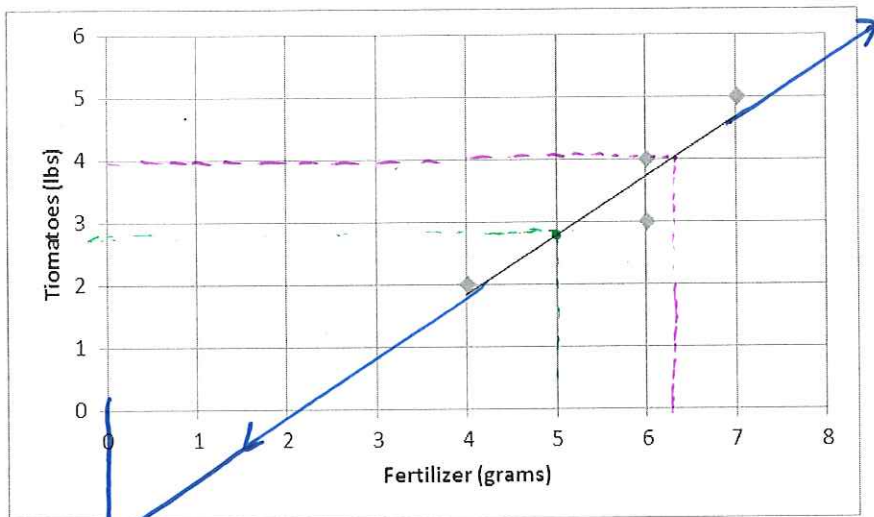
The closer $|r|$ is to 1 , the stronger the association.



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

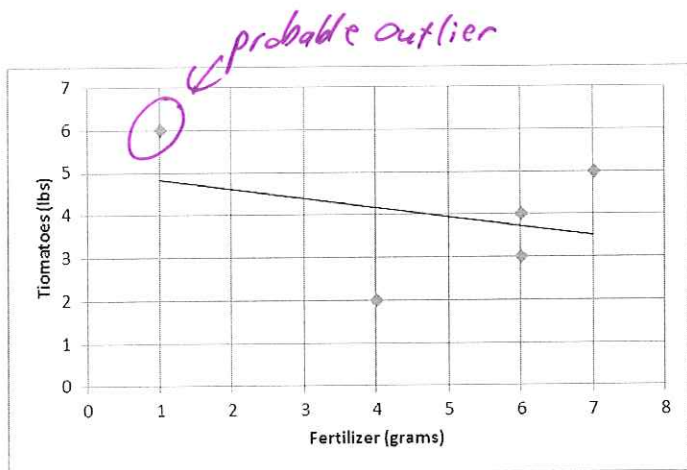
Extrapolation is using the regression line to find values that are *outside* the minimum and maximum values.

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

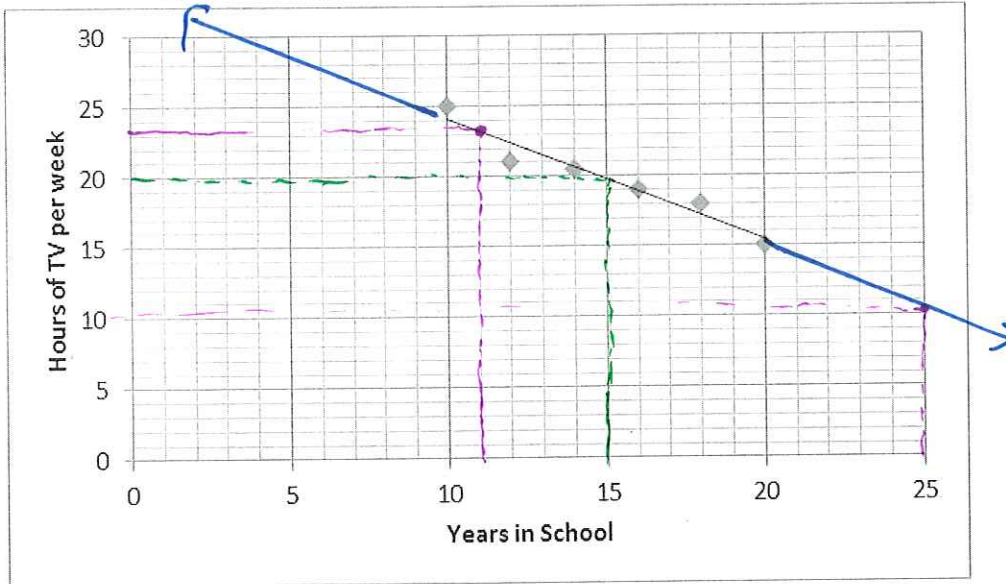


- (a) How many pounds of tomatoes if 5 grams of fertilizer are used?
≈ 2.8 lb of tomatoes
- (b) How much fertilizer was used if 4 pounds of tomatoes were grown?
≈ 6.3 g of fertilizer
- (c) How many tomatoes will be grown if no fertilizer is used?
Negative # of lb of tomatoes ← beware of extrapolation

How would the graph change if we found a measurement of 1 grams of fertilizer was used on a plant that had 6 pounds of tomatoes?



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



(a) If a person had attended school for 11 years, how many hours per week of television is watched? ≈ 23 hr/wk

(b) If a person watches 20 hours of television per week, how many years of school did he attend? ≈ 15 yr of school

(c) How many hours per week of television is watched by a person who attended school for 25 years? ≈ 10 hr/wk

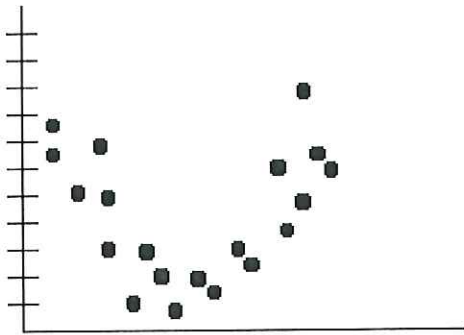
Things to keep in mind when using linear regression:

- Outliers can change things a lot
- Linear regression is only appropriate for linear relationships
- Interpolation is more likely to be accurate than extrapolation

Linear regression not appropriate

CORRELATION DOES NOT IMPLY CAUSATION

SAMPLE EXAM QUESTIONS FROM CHAPTER 6



1. To the left is the scatterplot for data collected on two variables. Would a least squares linear regression equation be useful in describing the relationship between the variables? Why or why not?

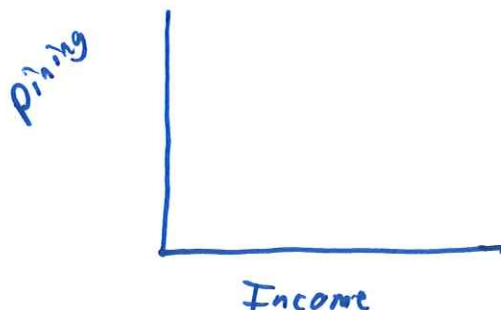
- (A) Yes because the data is approximately linear.
- (B) Yes because the data has a strong relationship.
- (C) No because the data does not have a linear relationship.
- (D) No because the data does not have a strong relationship.

2. Suppose the children of a primary school are surveyed. What type of association would you expect between their ages and their height?

- (A) No association between the ages and heights
- (B) A positive association between the ages and heights
- (C) A negative association between the ages and heights

3. To choose advertising media, a marketing analyst studies the relationship between a consumer's income and the amount spent on restaurant dining. Which variable, consumer's income or dining expenditures, would be the response variable for a least squares regression equation?

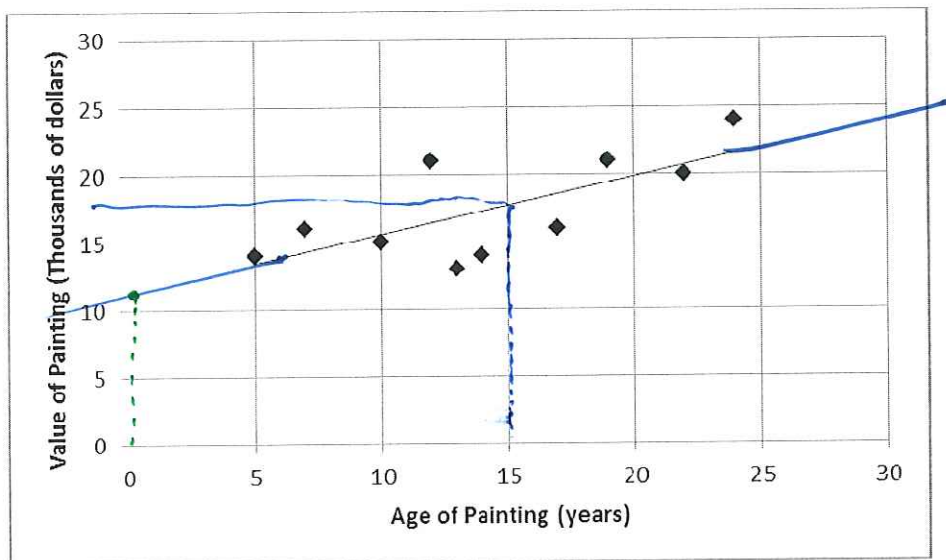
- (A) Consumer's income
- (B) Dining expenditures
- (C) Neither of these
- (D) Need more information



4. When the regression line is used to estimate values between data points this is called

- (A) extrapolation (B) interpolation (C) exfoliation
 (D) interdisciplinary (E) crazy

5. An art collector has a painting regularly appraised to find the value of the item. The results of the appraisal are shown in the scatterplot below along with the regression line. Use the regression line to answer the following questions:



(a) What is the value of the painting when it is 15 years old?

$\approx \$17,000$

(b) What was the initial value of the painting?

$\approx \$11,000$

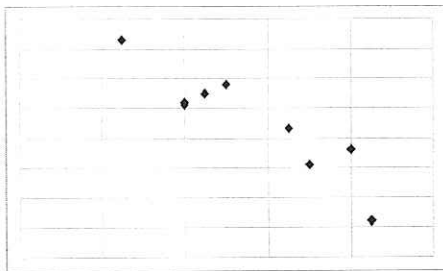
6. Match the scatter plots to the correlations below:

(a) Which scatterplot has a correlation of $r = -0.5$? *iii*

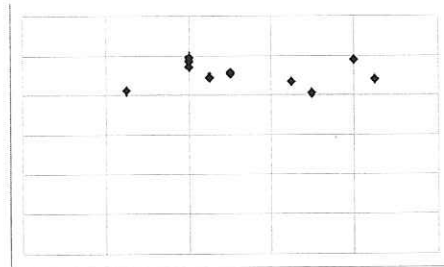
(b) Which scatterplot has a correlation of $r = 0.9$? *✓*

(c) Which scatterplot has a correlation of $r = 0$? *ii*

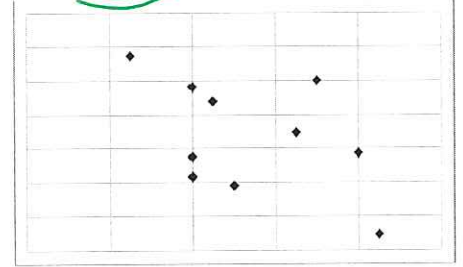
(i) *Strong neg*



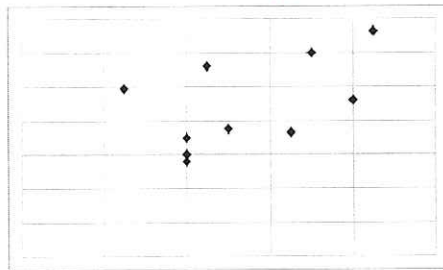
(ii) *zero*



(iii) *Weak neg*



(iv) *Weak pos*



(v) *Strong pos*

