Chapter 4: Scatterplots and Correlation

- Explanatory and response variables
- Displaying relationships: Scatterplots
- Interpreting Scatterplots
- Adding categorical variables to Scatterplots
- Correlation-measure of linear association

SCATTERPLOTS-An Example

A new fertilizer was tested on several plots of land to see if there is a relationship to the crop yield;

Plot of Land	Pounds of Fertilizer	Bushels of Beans (000's)
1	2	4
2	1	3
3	3	4
4	2	3
5	4	6
6	5	5
7	3	5

How would you graph this?

What is meant by a response and explanatory variable and which is which?

Explanatory Variables

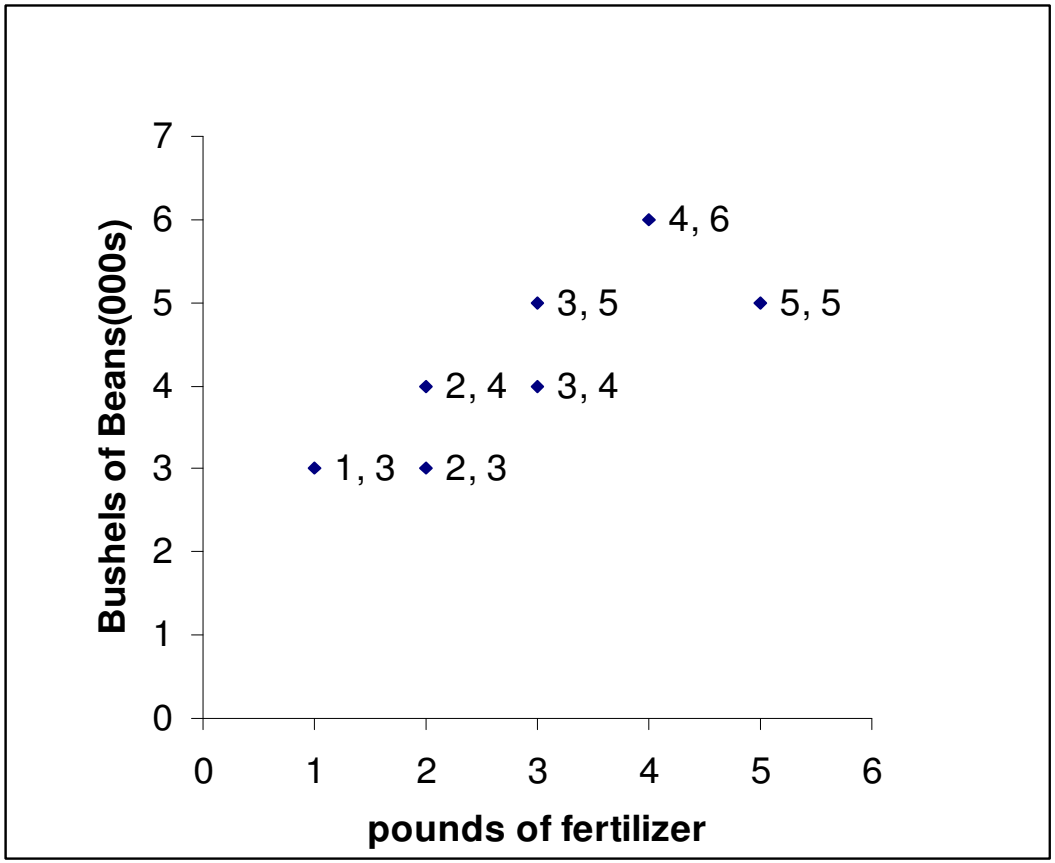
- variable used to predict or explain another(response) variable
- also called independent variables

Response Variables

- variables whose values are predicted from another(explanatory) variable
- also called a dependent variable
- does not imply changes are caused by the explanatory variable

Scatterplot (for class example):

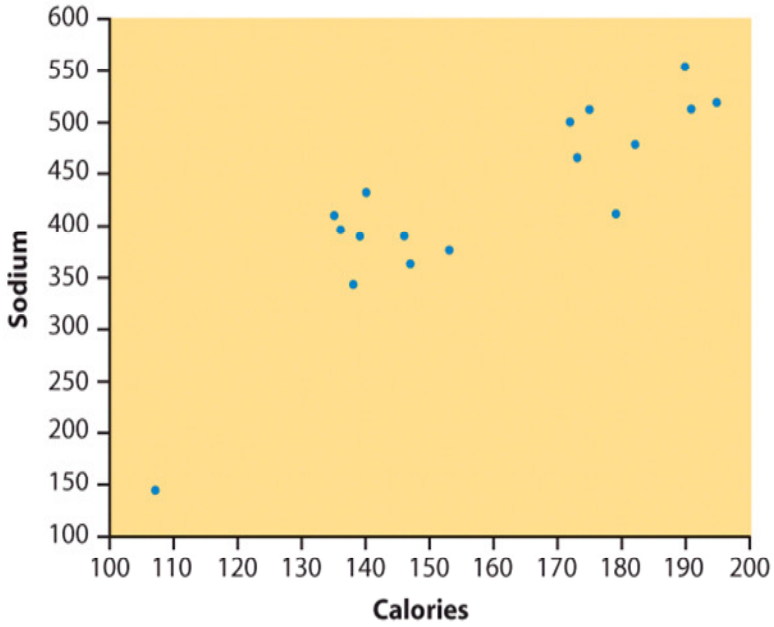
- Labeling of points is optional(wouldn't do with larger samples)



Scatterplots

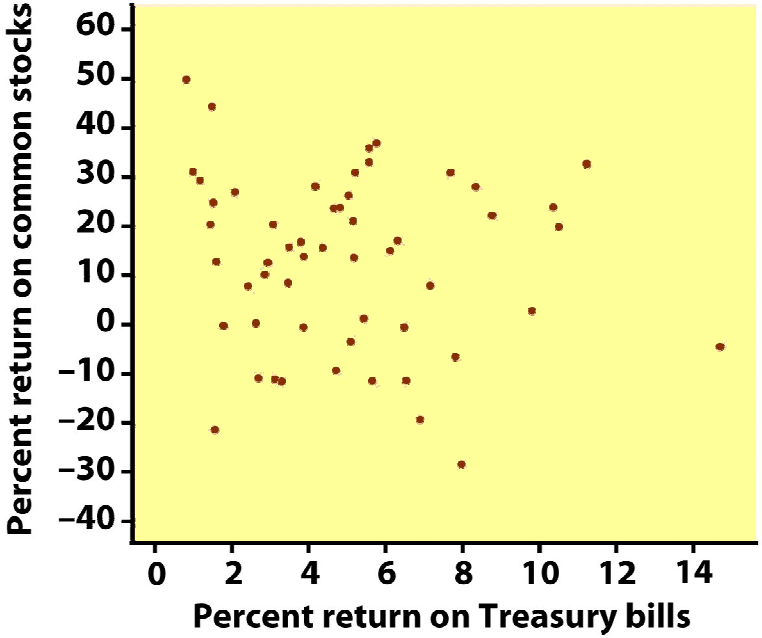
- show the relationship between 2 quantitative variables measured on the same cases
- for n cases, have n points on the plot (x_i,y_i)
- response variable appears on the y axis and explanatory variable appears on the x axis(**when you know the distinction**)
- Doesn't have to join at (0,0). Should be centered at the midpoint for each axis

Some plots don't have clear explanatory and response variables.



Do calories explain sodium amounts?

Does percent return on Treasury bills explain percent return on common stocks?



Interpreting Scatterplots

Look for an overall pattern and deviations(or outliers, i.e. value(s) that fall outside overall relationship pattern)

- **Direction;** positive, negative or none. See comments on **association** below

- **Form or overall shape ;**linear, curved, cyclical, clusters, or none apparent

- **Strength***;* strong (where strong is very little scatter from form identified), moderate or weak

Association

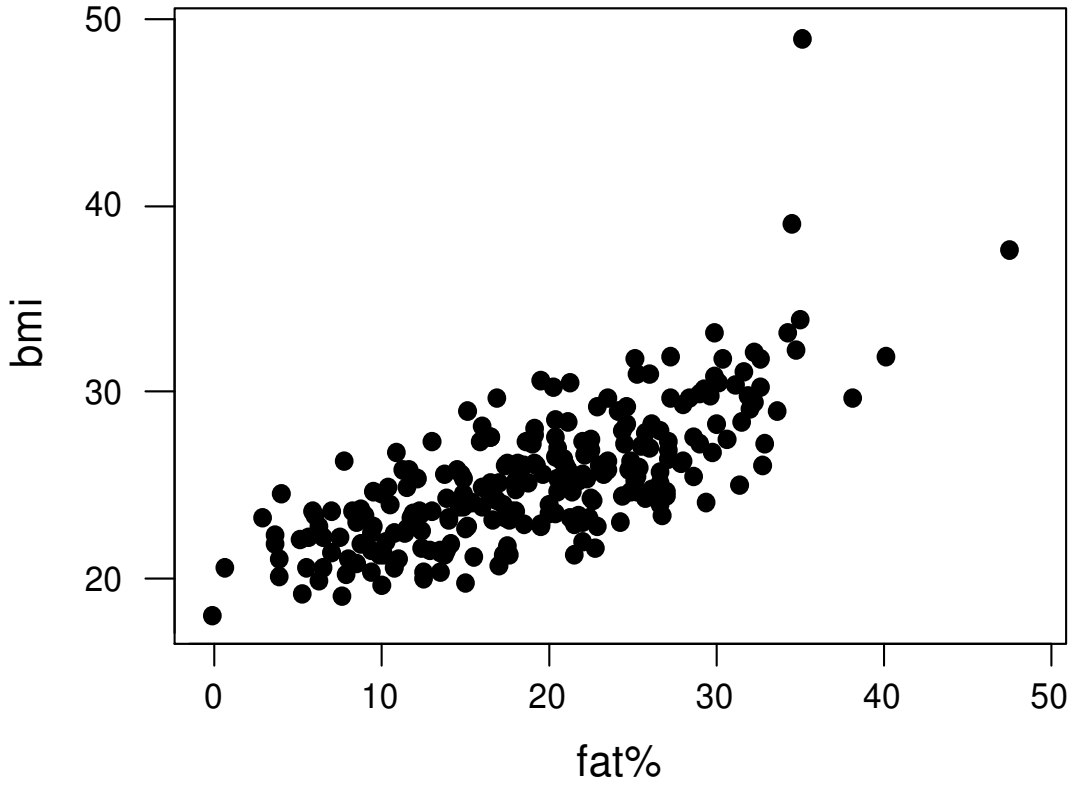
- knowledge of one variable helps you know something of the other

- association does not imply causation

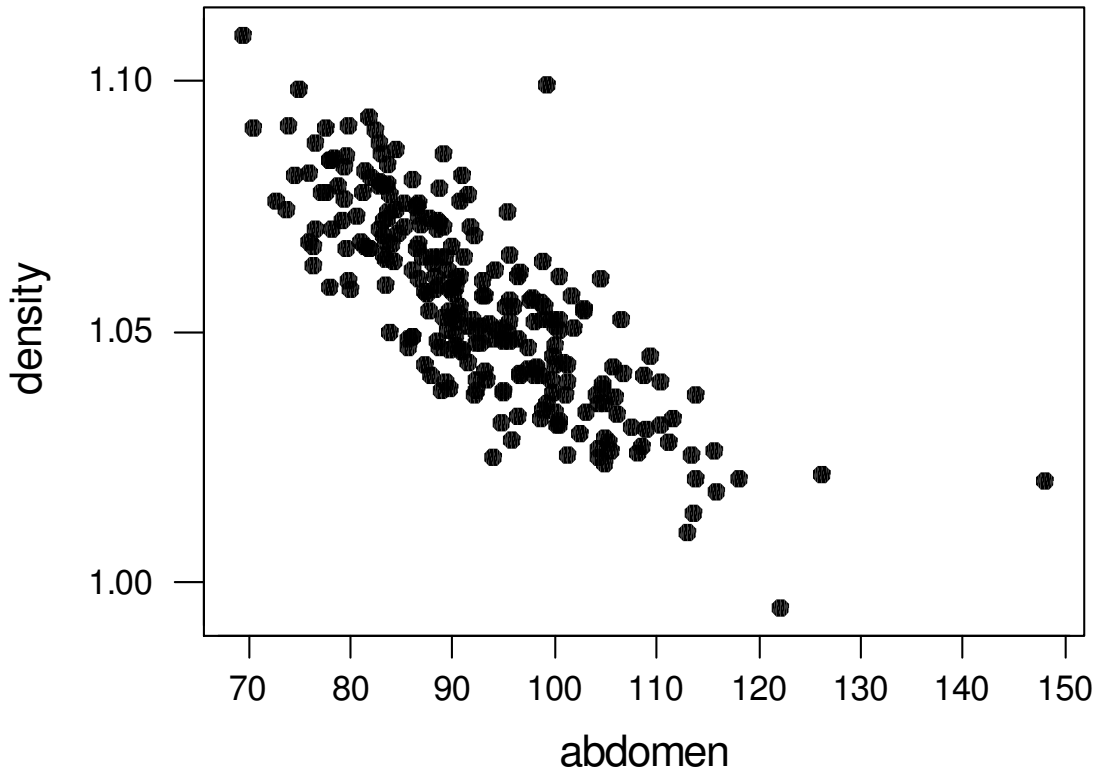
- **Positive Association**-above average values of one variable accompany above average values of the other (similarly for below average)

- **Negative Association**-above average values of one variable accompany below average values of the other variable(and vice versa)

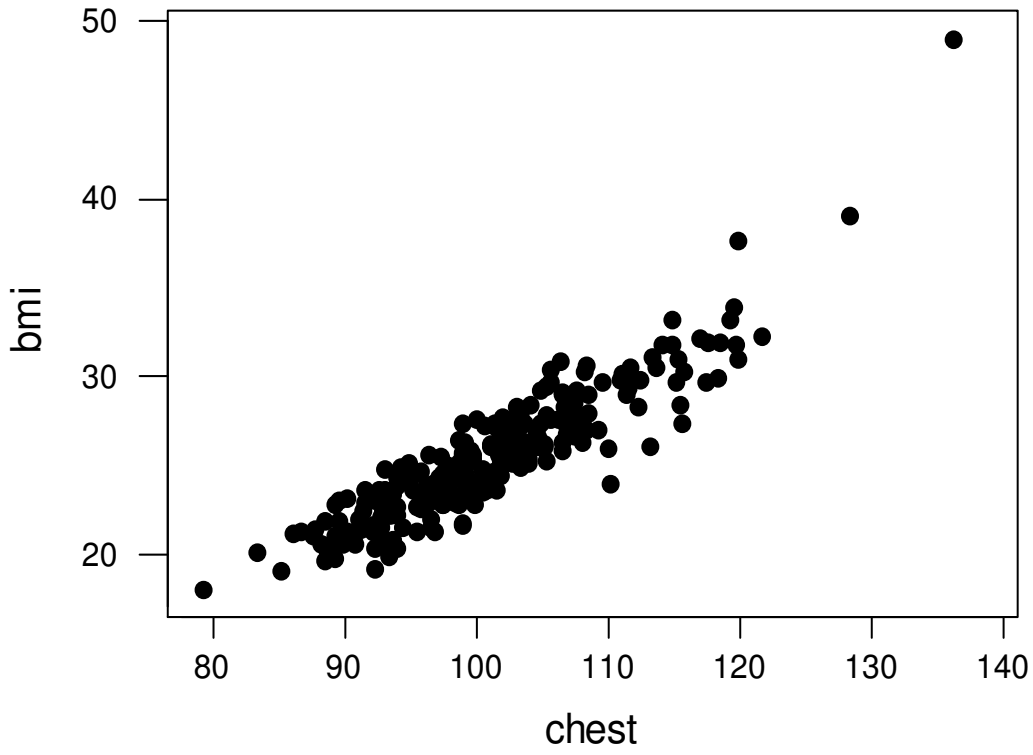
Positive Association



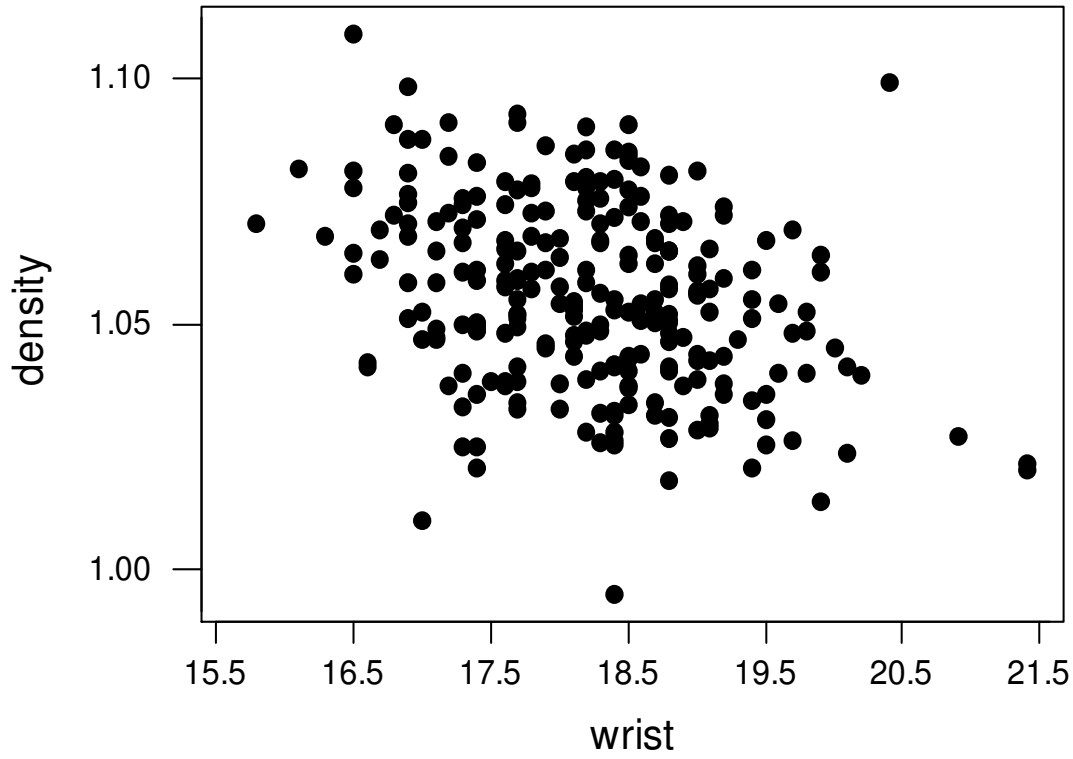
Negative Association

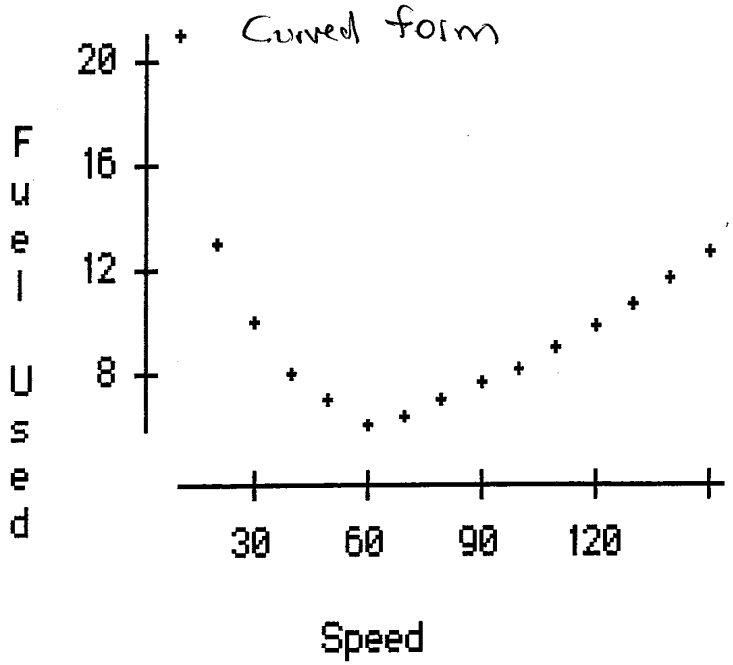


Strong (Positive) Association



Weak (Negative) Association




Curved form
20
16
12
8
Fuel Used
30
60
90
120
Speed

Scatterplots and Categorical Data

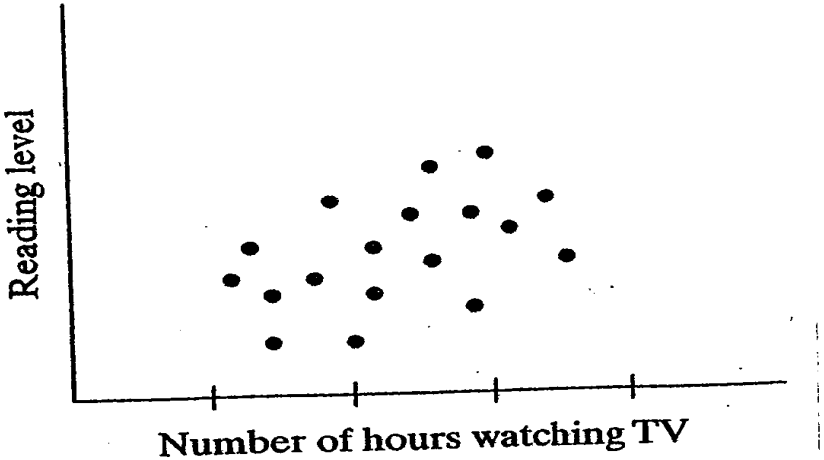
- Use different colors/symbols when you want to add a categorical variable to a scatterplot

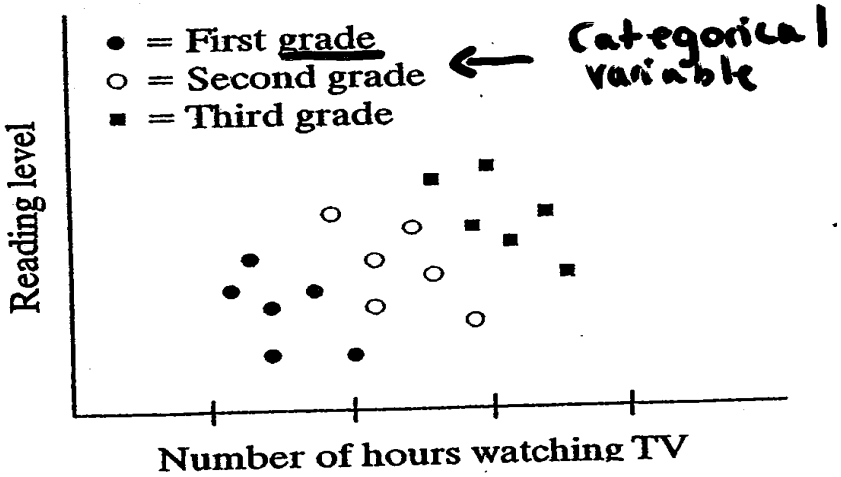
- **See example**
  - scatterplot for 18 primary grade students and reading level scores

  - 2nd scatterplot has grade introduced as a categorical variable

- use different colors/symbols to plot points when you want to add a categorical variable to a scatterplot

SCATTERPLOTS AND CATEGORICAL VARS





CORRELATION(r)

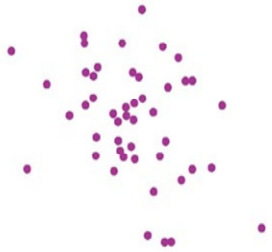
- numerical measure of the strength and direction of the ***linear relationship*** between two *quantitative* variables

- does not distinguish between response(y) and explanatory(x) variable.

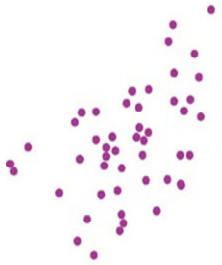
- Values of r fall between -1 and +1

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

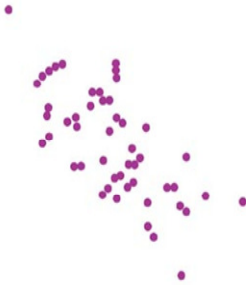
Correlation $r = 0$



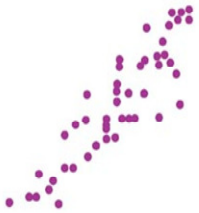
Correlation $r = -0.3$



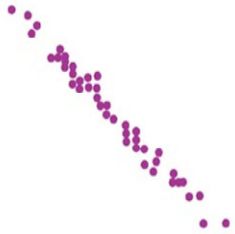
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

Correlation

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- based on standardized values so r has no units
- r is not affected by the measurement units of x and y
- r is a numerical measure that can help overcome problems with visual representations affected by the scales chosen
- r measures only **linear** association
- should do a scatterplot first (see class example-impact of an outlier)
- both x_i and y_i must be *quantitative* variables

Correlation Example- Is Growth linear?

Child	Age(mths)	Height(cms)	$(x_i-\bar{x})/s_x$	$(y_i-\bar{y})/s_y$	$(x_i-\bar{x})(y_i-\bar{y})/s_xs_y$
1	36	86	-1.77	-1.68	2.97
2	48	90	-0.35	-0.46	0.16
3	51	91	0.00	-0.15	0.00
4	54	93	0.35	0.46	0.16
5	57	94	0.71	0.76	0.54
6	60	95	1.06	1.07	1.13
Mean	$\bar{x} = 51$	$\bar{y}=91.5$			
s	$s_x= 8.485$	$s_y= 3.271$			
Total					4.97
r(correlation)					.9944

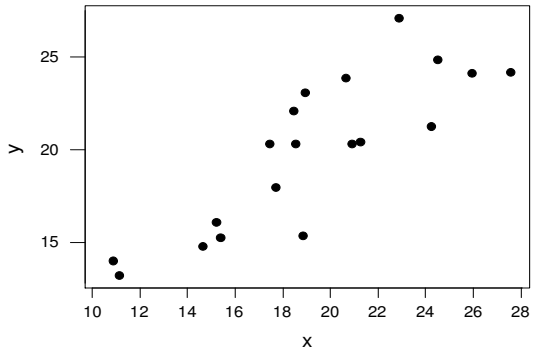
Key steps:

1. calculate $\bar{x}$ and $\bar{y}$ and s_x and s_y
2. standardize x and y values
3. take product of these values for each pair
4. add n values(of 3) and divide by (n-1)

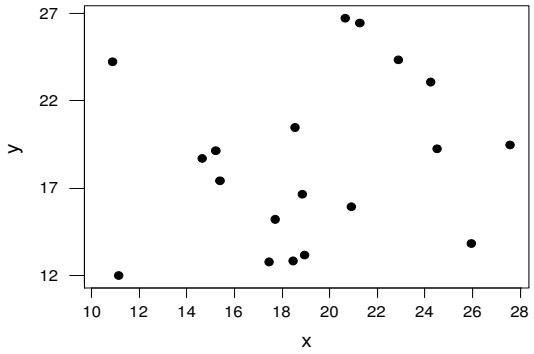
Properties of Correlation

- **r is between -1 and +1**
- **sign indicates the direction**
  - ➢ if r is positive, both variables increase together
  - ➢ if r is negative, one increases while the other decreases
- **magnitude of r indicates the strength**
  - ➢ value near either extremes(+1 or -1) are strong linear relationships
  - ➢ value near 0 indicates a lack of linear relationship
- **r is not resistant (strongly affected by outliers)**
  - ➢ plot data first to check for outliers
  - ➢ value near the center of the plot has less effect on r than values near the sides

Positive Association

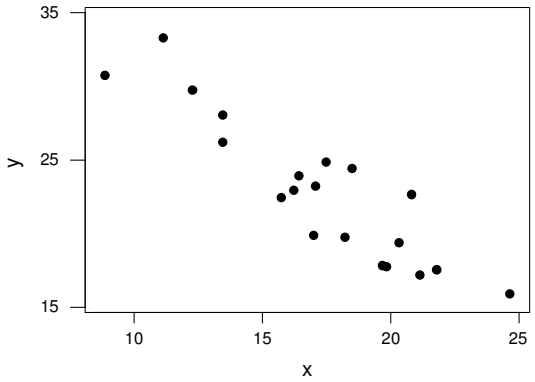


strong $r = 0.843$

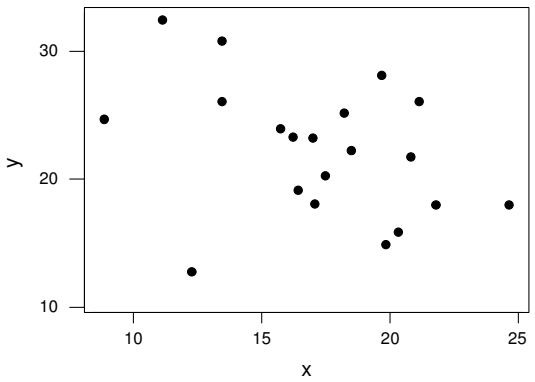


weak $r = 0.189$

Negative Association



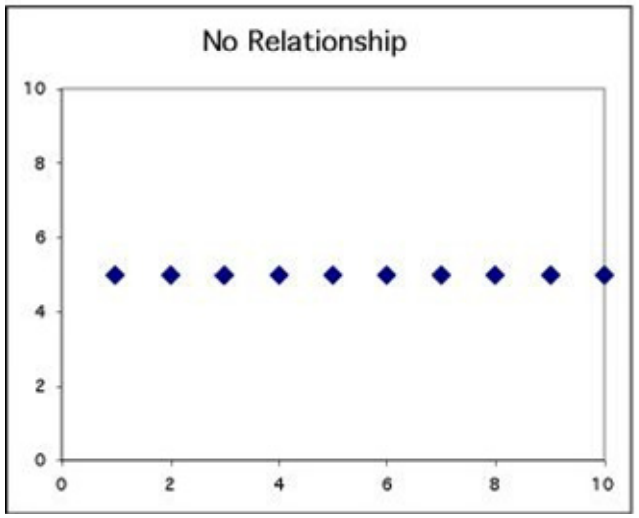
strong $r = -0.906$

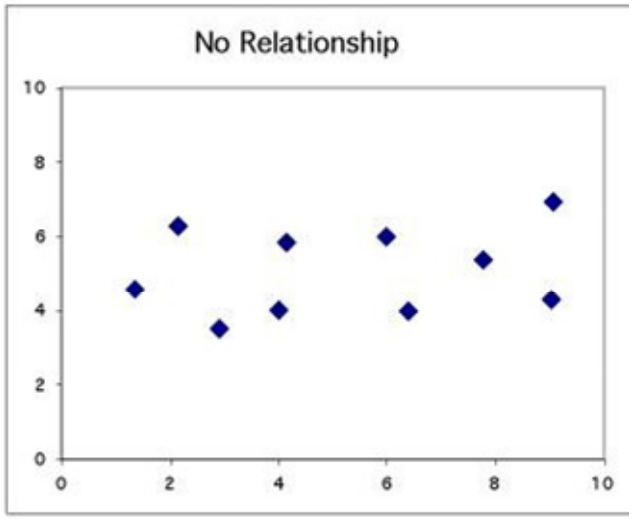


weak $r = -0.368$

No relationship:

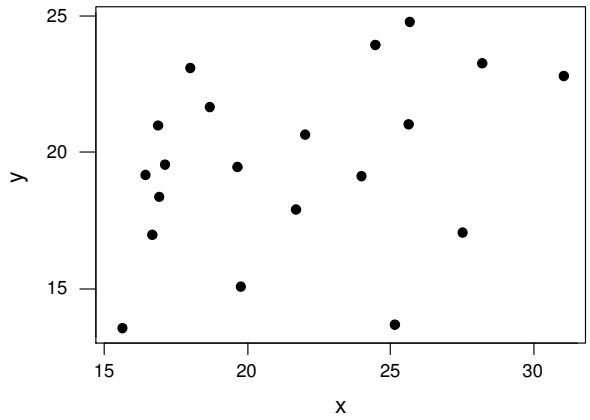
x and *y* vary independently. Knowing *x* tells you nothing about *y*.

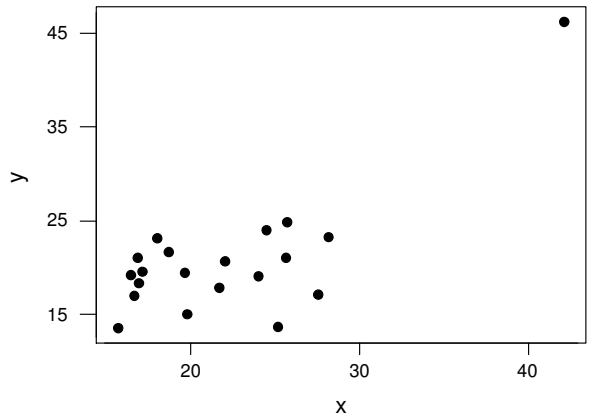




One way to remember this:
The equation for this line is y = 5.
x is not involved.

Effect of Outliers on Correlation

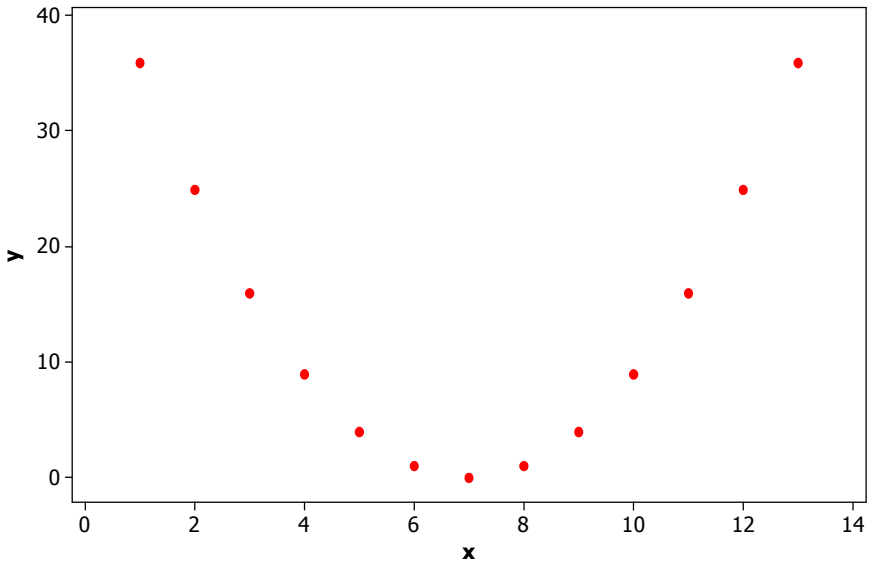




$r = 0.342$ $r = 0.753$

- only one number has been changed between the graphs, the largest x

Correlation Measures Linear Association Only



- the graph shows a perfect quadratic relationship between y and x.
- correlation $r = 0$
- correlation does **not** measure the strength of curved relationships