Set 3: Paired Data, Scatter Plots, Correlation

Stat 260 A01: May 15, 2024

We have previously seen **univariate data**, which consists of observations of a single characteristic/attribute, and is typically denoted as $x_1, x_2, \ldots, x_n$.

Data that arises in pairs is referred to as **bivariate data**, often denoted as:

(x_1, y_1), (x_2, y_2), (x_n, y_n) . . .

Example 1: A biologist captures 5 loggerhead sea turtles (*Caretta caretta*) and records their head-to-tail lengths (in cm) and their blood plasma arsenic levels (in ppb).

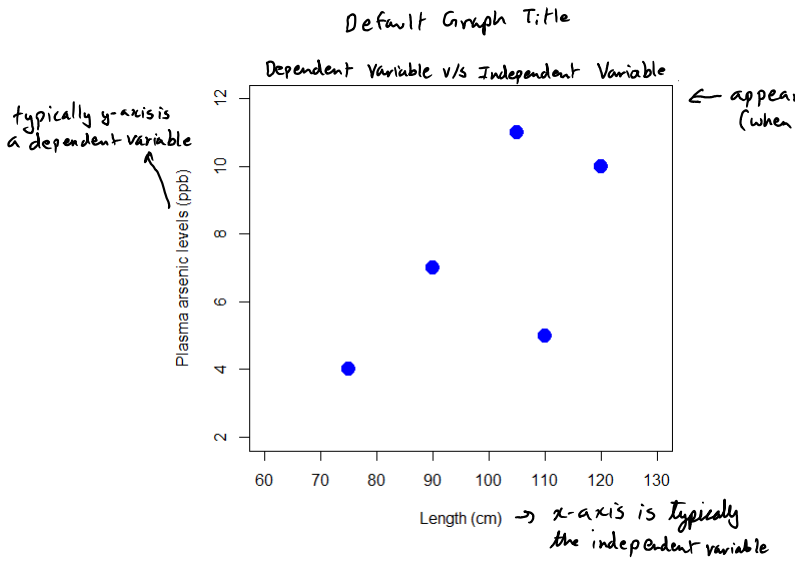
	Length (cm)	Arsenic (ppb)	
Turtle 1	75	4	75, 4
Turtle 2	110	5	110, 5
Turtle 3	90	7	90, 7
Turtle 4	105	11	105, 11
Turtle 5	120	10	120, 10

Standard deviation of x $s_x = 17.67$

Standard deviation of y $s_y = 3.05$

$\bar{x} = 100$

$\bar{y} = 7.4$



typically y-axis is a dependent variable

← appears to be some positive relationship (when one val increases so does the other)

x-axis is typically the independent variable

Covariance

Covariance is a measure of the association between two random variables.

Recall: sample variance $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = S_{xx}$ ← always ≥ 0

For bivariate data, we can calculate the **sample covariance**: (written as cov)

$$S_{xy} = \text{cov}(x, y) = \text{cov}_{x,y} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

← can be positive, negative or zero

Example 1 Continued... Determine the covariance for loggerhead turtle data.

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i-\bar{x})(y_i-\bar{y})$
75	4	−25	−3.4	85
110	5	10	−2.4	−24
90	7	−10	−0.4	4
105	11	5	3.6	18
120	10	20	12.6	52
				Total = 135

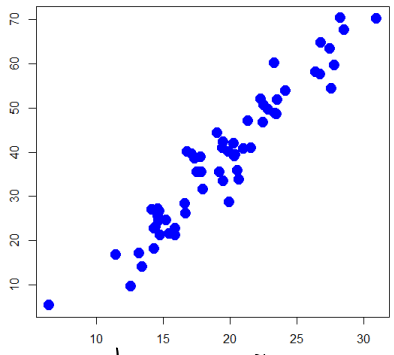
$$\text{cov}(x_i, y_i) = \frac{1}{5-1}(135) = 33.75$$

positive covariance indicates a positive relationship between the variables (likewise, a negative covariance indicates a negative relationship)

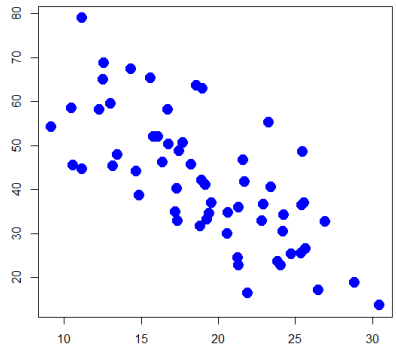
Like variance, a single covariance does not tell us as much about the strength of that relationship

---

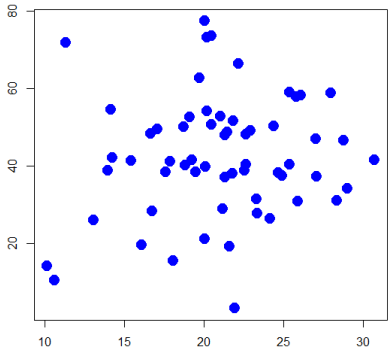
Coefficient of Correlation



Shows positive linear relationship
$r = 0.96$



moderate / weak negative linear relationship
$r = -0.7$



No linear relationship $r = 0.12$

The **correlation coefficient**, r, measures the strength of the linear relationship between x and y.

- r close to $+1$ indicates a strong positive linear relationship.
- r close to -1 indicates a strong negative linear relationship.
- r close to 0 indicates no linear relationship.

always $-1 \le r \le 1$

The **correlation coefficient**, r, measures the strength of linear relationship between x and y;

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{S_{xy}}{S_x \cdot S_y}$$

$$= \frac{\text{covariance of } (x,y)}{\text{Std deviation of } x \cdot \text{standard deviation of } y}$$

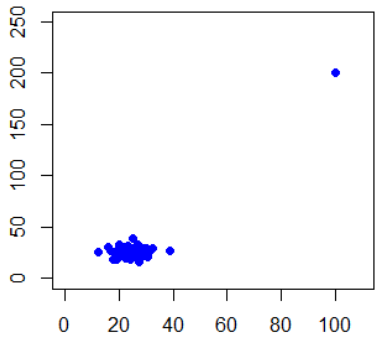
Example 1 Continued... Determine the correlation coefficent of the loggerhead turtle data.

Length (cm)	Arsenic (ppb)
75	4
110	5
90	7
105	11
120	10

$S_{xy} = 33.75$, $S_x = 17.67$, $S_y = 3.05$

$\therefore r = \frac{S_{xy}}{S_x \cdot S_y} = \frac{33.75}{17.67 \cdot 3.05} = 0.629$

→ some indication of a positive linear relationship
→ since 0.629 isn't very close to 1 it is likely a fairly weak linear relationship



Warnings about r:

→ r is very strongly influenced by outliers
→ r is good for values close to -1, 0, +1 but everything inbetween is more difficult to c

Textbook Readings: Swartz 2.5. EPS 7.8
Practice problems: Swartz: Produce scatterplots and compute the sample correlation using your calculator for the data given in 7.1, 7.3, 7.5. Comment on your findings. [Answers r=.312, .986, .707]