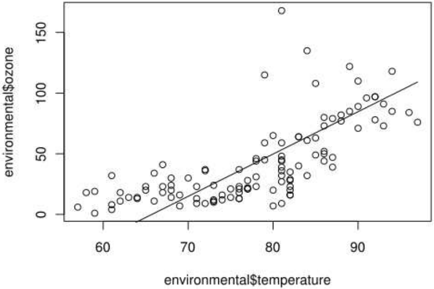
Math 207: Statistics

Chapter 8: Correlation



Dr. Ralph Wojtowicz

SHENANDOAH UNIVERSITY

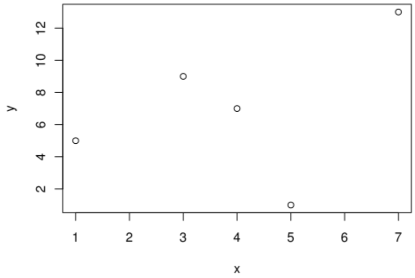
Scatter Diagrams

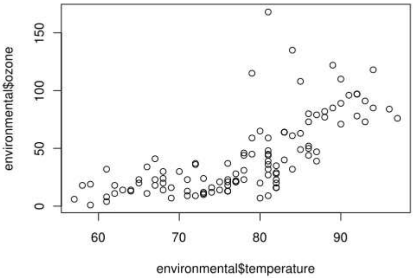
- Example from page 132 of our text

```
> x <- c(1, 3, 4, 5, 7)
> y <- c(5, 9, 7, 1, 13)
> plot(x, y)
```

- Example using an R environmental data set

```
> library(lattice)
> plot(environmental$temperature, environmental$ozone)
```





The Correlation Coefficient

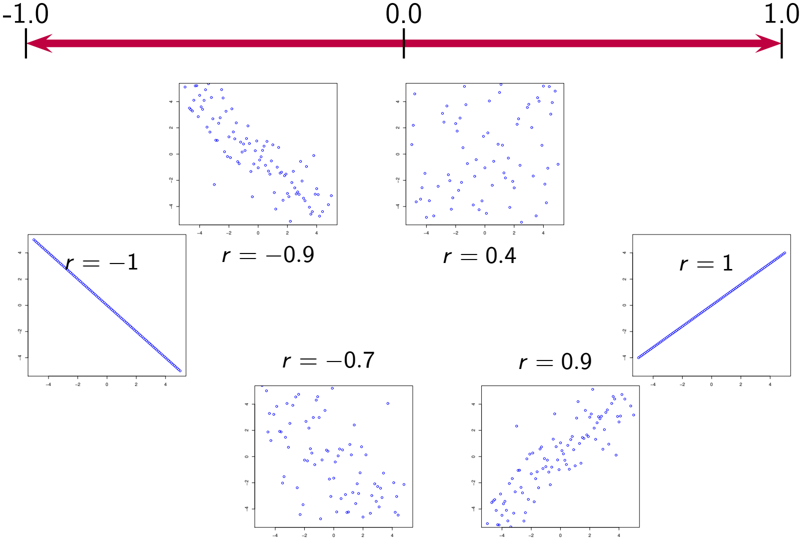
- Given lists $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, the correlation coefficient:
  - Is a measure of linear association between the lists
  - Is a measure of the clustering of the (x_i, y_i) points around a line
  - Is a number between -1 and 1
  - Is defined by:

$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \text{mean}_x}{\text{SD}_x} \right) \left(\frac{y_i - \text{mean}_y}{\text{SD}_y} \right)$$

$$= \text{average of the } x \text{ and } y \text{ values measured in standard units}$$

- A positive correlation means that the cloud of (x_i, y_i) points slopes up
- A negative correlation means that the cloud of (x_i, y_i) points slopes down

Magnitude of the Correlation Coefficient



Computing r

$$r = \frac{1}{n} \sum \left(\frac{x_i - \text{mean}_x}{\text{SD}_x} \right) \left(\frac{y_i - \text{mean}_y}{\text{SD}_y} \right)$$

x	y	z_x	z_y	$z_x\, z_y$
1	5	$-3/2$	$-1/2$	$3/4$
3	9	$-1/2$	$1/2$	$-1/4$
4	7	0	0	0
5	1	$1/2$	$-3/2$	$-3/4$
7	13	$3/2$	$3/2$	$9/4$

Computing r

$$r = \frac{1}{n}\sum\left(\frac{x_i - \text{mean}_x}{\text{SD}_x}\right)\left(\frac{y_i - \text{mean}_y}{\text{SD}_y}\right)$$

x	y	z_x	z_y	$z_x\,z_y$
1	5	$-3/2$	$-1/2$	$3/4$
3	9	$-1/2$	$1/2$	$-1/4$
4	7	0	0	0
5	1	$1/2$	$-3/2$	$-3/4$
7	13	$3/2$	$3/2$	$9/4$

$$\text{mean}(x) = \frac{1}{5}(1+3+4+5+7) = 4 \qquad \text{SD}(x) = \sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}} = 2$$

$$\text{mean}(y) = \frac{1}{5}(5+9+7+1+13) = 7 \qquad \text{SD}(y) = \sqrt{\frac{(5-7)^2+(9-7)^2+(7-7)^2+(1-7)^2+(13-7)^2}{5}} = 4$$

Computing r

$$r = \frac{1}{n}\sum\left(\frac{x_i - \text{mean}_x}{\text{SD}_x}\right)\left(\frac{y_i - \text{mean}_y}{\text{SD}_y}\right)$$

x	y	z_x	z_y	$z_x\, z_y$
1	5	$-3/2$	$-1/2$	$3/4$
3	9	$-1/2$	$1/2$	$-1/4$
4	7	0	0	0
5	1	$1/2$	$-3/2$	$-3/4$
7	13	$3/2$	$3/2$	$9/4$

$$\text{mean}(x) = \frac{1}{5}(1+3+4+5+7) = 4 \qquad \text{SD}(x) = \sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}} = 2$$

$$\text{mean}(y) = \frac{1}{5}(5+9+7+1+13) = 7 \qquad \text{SD}(y) = \sqrt{\frac{(5-7)^2+(9-7)^2+(7-7)^2+(1-7)^2+(13-7)^2}{5}} = 4$$

- Convert the x values to standard units. For example,

$$x = 1 \quad \text{becomes} \quad z_x = \frac{1-4}{2} = -3/2 \qquad \text{and} \qquad x = 3 \quad \text{becomes} \quad z_x = \frac{3-4}{2} = -1/2$$

Computing r

$$r = \frac{1}{n} \sum \left(\frac{x_i - \text{mean}_x}{\text{SD}_x} \right) \left(\frac{y_i - \text{mean}_y}{\text{SD}_y} \right)$$

x	y	z_x	z_y	$z_x\, z_y$
1	5	$-3/2$	$-1/2$	$3/4$
3	9	$-1/2$	$1/2$	$-1/4$
4	7	0	0	0
5	1	$1/2$	$-3/2$	$-3/4$
7	13	$3/2$	$3/2$	$9/4$

$$\text{mean}(x) = \frac{1}{5}(1+3+4+5+7) = 4 \qquad \text{SD}(x) = \sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}} = 2$$

$$\text{mean}(y) = \frac{1}{5}(5+9+7+1+13) = 7 \qquad \text{SD}(y) = \sqrt{\frac{(5-7)^2+(9-7)^2+(7-7)^2+(1-7)^2+(13-7)^2}{5}} = 4$$

- Convert the x values to standard units. For example,

$$x = 1 \quad \text{becomes} \quad z_x = \frac{1-4}{2} = -3/2 \qquad \text{and} \qquad x = 3 \quad \text{becomes} \quad z_x = \frac{3-4}{2} = -1/2$$

- Convert the y values to standard units: For example,

$$y = 5 \quad \text{becomes} \quad z_y = \frac{5-7}{4} = -1/2 \qquad \text{and} \qquad y = 9 \quad \text{becomes} \quad z_y = \frac{9-7}{4} = 1/2$$

Computing r

$$r = \frac{1}{n} \sum \left(\frac{x_i - \text{mean}_x}{\text{SD}_x} \right) \left(\frac{y_i - \text{mean}_y}{\text{SD}_y} \right)$$

x	y	z_x	z_y	$z_x\, z_y$
1	5	$-3/2$	$-1/2$	$3/4$
3	9	$-1/2$	$1/2$	$-1/4$
4	7	0	0	0
5	1	$1/2$	$-3/2$	$-3/4$
7	13	$3/2$	$3/2$	$9/4$

$$\text{mean}(x) = \frac{1}{5}(1+3+4+5+7) = 4 \qquad \text{SD}(x) = \sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}} = 2$$

$$\text{mean}(y) = \frac{1}{5}(5+9+7+1+13) = 7 \qquad \text{SD}(y) = \sqrt{\frac{(5-7)^2+(9-7)^2+(7-7)^2+(1-7)^2+(13-7)^2}{5}} = 4$$

- Convert the x values to standard units. For example,

  $$x = 1 \quad \text{becomes} \quad z_x = \frac{1-4}{2} = -3/2 \qquad \text{and} \qquad x = 3 \quad \text{becomes} \quad z_x = \frac{3-4}{2} = -1/2$$

- Convert the y values to standard units: For example,

  $$y = 5 \quad \text{becomes} \quad z_y = \frac{5-7}{4} = -1/2 \qquad \text{and} \qquad y = 9 \quad \text{becomes} \quad z_y = \frac{9-7}{4} = 1/2$$

- Compute the products $z_x\, z_y$.

Computing r

$$r = \frac{1}{n}\sum\left(\frac{x_i - \text{mean}_x}{\text{SD}_x}\right)\left(\frac{y_i - \text{mean}_y}{\text{SD}_y}\right)$$

x	y	z_x	z_y	$z_x\, z_y$
1	5	$-3/2$	$-1/2$	$3/4$
3	9	$-1/2$	$1/2$	$-1/4$
4	7	0	0	0
5	1	$1/2$	$-3/2$	$-3/4$
7	13	$3/2$	$3/2$	$9/4$

$$\text{mean}(x) = \frac{1}{5}(1+3+4+5+7) = 4 \qquad \text{SD}(x) = \sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}} = 2$$

$$\text{mean}(y) = \frac{1}{5}(5+9+7+1+13) = 7 \qquad \text{SD}(y) = \sqrt{\frac{(5-7)^2+(9-7)^2+(7-7)^2+(1-7)^2+(13-7)^2}{5}} = 4$$

- Convert the x values to standard units. For example,

$$x = 1 \quad \text{becomes} \quad z_x = \frac{1-4}{2} = -3/2 \qquad \text{and} \qquad x = 3 \quad \text{becomes} \quad z_x = \frac{3-4}{2} = -1/2$$

- Convert the y values to standard units: For example,

$$y = 5 \quad \text{becomes} \quad z_y = \frac{5-7}{4} = -1/2 \qquad \text{and} \qquad y = 9 \quad \text{becomes} \quad z_y = \frac{9-7}{4} = 1/2$$

- Compute the products $z_x\, z_y$.
- Compute the correlation coefficient: $r = \frac{1}{5}\left(\frac{3}{4} - \frac{1}{4} + 0 - \frac{3}{4} + \frac{9}{4}\right) = \frac{1}{5}\,\frac{8}{4} = \frac{2}{5} = 0.4$

Computing r

$$r = \frac{1}{n}\sum\left(\frac{x_i - \text{mean}_x}{\text{SD}_x}\right)\left(\frac{y_i - \text{mean}_y}{\text{SD}_y}\right)$$

x	y	z_x	z_y	$z_x\,z_y$
1	5	$-3/2$	$-1/2$	$3/4$
3	9	$-1/2$	$1/2$	$-1/4$
4	7	0	0	0
5	1	$1/2$	$-3/2$	$-3/4$
7	13	$3/2$	$3/2$	$9/4$

$$\text{mean}(x) = \frac{1}{5}(1+3+4+5+7) = 4 \qquad \text{SD}(x) = \sqrt{\frac{(1-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(7-4)^2}{5}} = 2$$

$$\text{mean}(y) = \frac{1}{5}(5+9+7+1+13) = 7 \qquad \text{SD}(y) = \sqrt{\frac{(5-7)^2+(9-7)^2+(7-7)^2+(1-7)^2+(13-7)^2}{5}} = 4$$

- Convert the x values to standard units. For example,

  $$x = 1 \quad \text{becomes} \quad z_x = \frac{1-4}{2} = -3/2 \qquad \text{and} \qquad x = 3 \quad \text{becomes} \quad z_x = \frac{3-4}{2} = -1/2$$

- Convert the y values to standard units: For example,

  $$y = 5 \quad \text{becomes} \quad z_y = \frac{5-7}{4} = -1/2 \qquad \text{and} \qquad y = 9 \quad \text{becomes} \quad z_y = \frac{9-7}{4} = 1/2$$

- Compute the products $z_x\,z_y$.
- Compute the correlation coefficient: $r = \frac{1}{5}\left(\frac{3}{4} - \frac{1}{4} + 0 - \frac{3}{4} + \frac{9}{4}\right) = \frac{1}{5}\,\frac{8}{4} = \frac{2}{5} = 0.4$
- In R:

```
x <- c(1, 3, 4, 5, 7)
y <- c(5, 9, 7, 1, 13)
cor(x, y)
```

The SD Line

- Given lists $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, the SD line
  - Is a linear approximation to the cloud of (x_i, y_i) points
  - Is defined by

$$(y - \text{mean}_y) = (\text{sign } r)\left(\frac{\text{SD}_y}{\text{SD}_x}\right)(x - \text{mean}_x)$$

    where r is the correlation coefficient.
  - It goes through the point of averages: $(\text{mean}_x,\ \text{mean}_y)$.
  - It's slope is $\pm\frac{\text{SD}_y}{\text{SD}_x}$.
- For every increase of 1 SD_x in the x-direction, there is an increase of 1 SD_y in the y-direction.
- If $r > 0$, the slope of the SD line is $\frac{\text{SD}_y}{\text{SD}_x}$.
- If $r < 0$, the slope of the SD line is $-\frac{\text{SD}_y}{\text{SD}_x}$.

SD Line Calculation

x	y	z_x	z_y	$z_x\, z_y$	y values predicted by SD line	SD error
1	5	$-3/2$	$-1/2$	$3/4$	1	4
3	9	$-1/2$	$1/2$	$-1/4$	5	4
4	7	0	0	0	7	0
5	1	$1/2$	$-3/2$	$-3/4$	9	-8
7	13	$3/2$	$3/2$	$9/4$	13	0

- SD line equation:

$$(y - \text{mean}_y) = (\text{sign } r)\left(\frac{\text{SD}_y}{\text{SD}_x}\right)(x - \text{mean}_x)$$

- Subsitute the values from Slide 5:

$$(y - 7) = (+1)\left(\frac{4}{2}\right)(x - 4)$$

which simplifies to

$$(y - 7) = 2\,(x - 4)$$