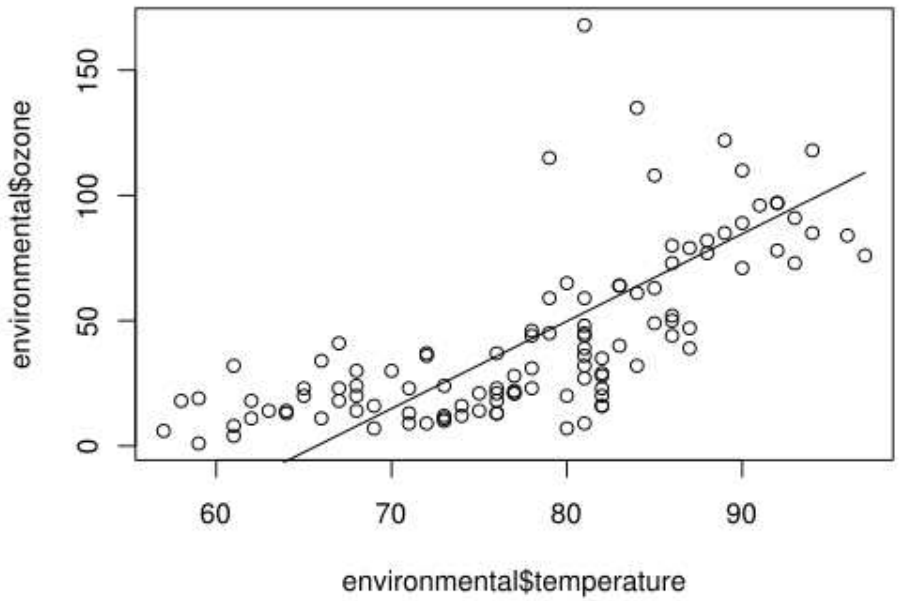
Math 314: Statistics

Chapter 8: Correlation



Dr. Ralph Wojtowicz

CME Department

Shepherd UNIVERSITY

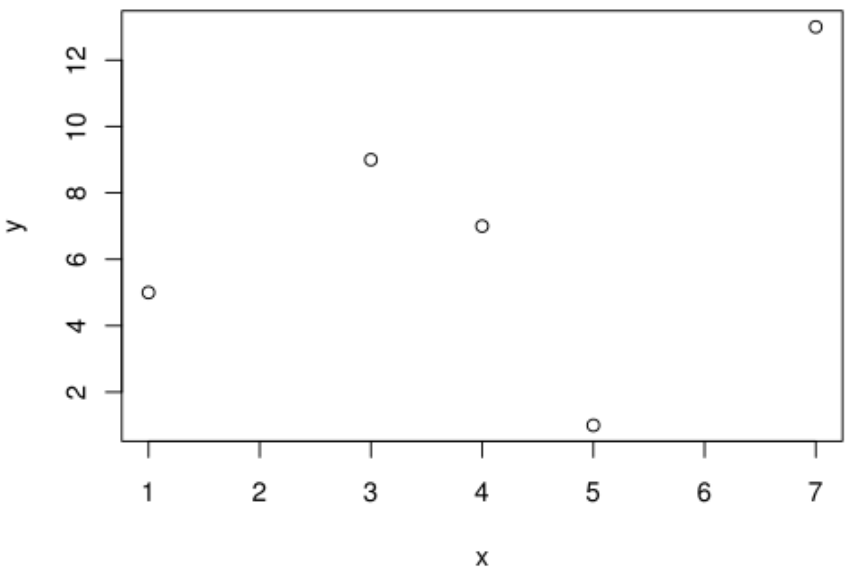
Scatter Diagrams

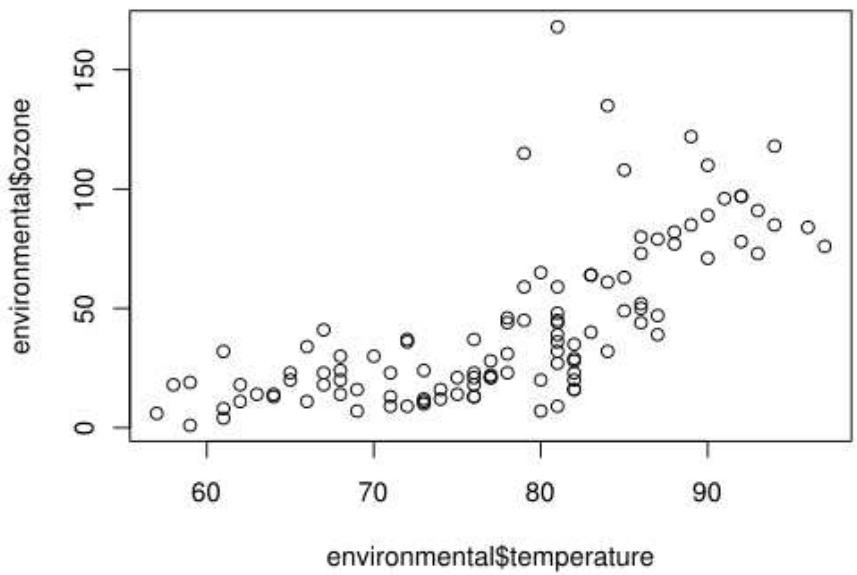
- Example from page 132 of our text

```
> 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

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 below (r is the correlation coefficient):

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

- Here is an example in R (see page 132 of our text)

```
> x <- c(1, 3, 4, 5, 7)
> y <- c(5, 9, 7, 1, 13)
> source("http://www.adjoint-functors.net/SDline.R")
> SDline(x, y)
> $meanX        > $meanY
>[1] 4          >[1] 7
> $slope        > $correlation
>[1] 2          >[1] 0.4
```