Introduction to Data Science CS61 June 12 - July 12, 2018

Dr. Ash Pahwa

Lesson 4: Statistics

Lesson 4.1: Covariance & Correlation

Outline

- Covariance
- Properties of Covariance
- Correlation Coefficient
- Properties of Correlation

Univariate and Bivariate data

- We examined a single variable
  - Univariate data
    - Mean, Median, Mode, Standard Deviation, Variance
- Now we will examine 2 variables
  - Bivariate data
    - Covariance, Correlation
    - Regression

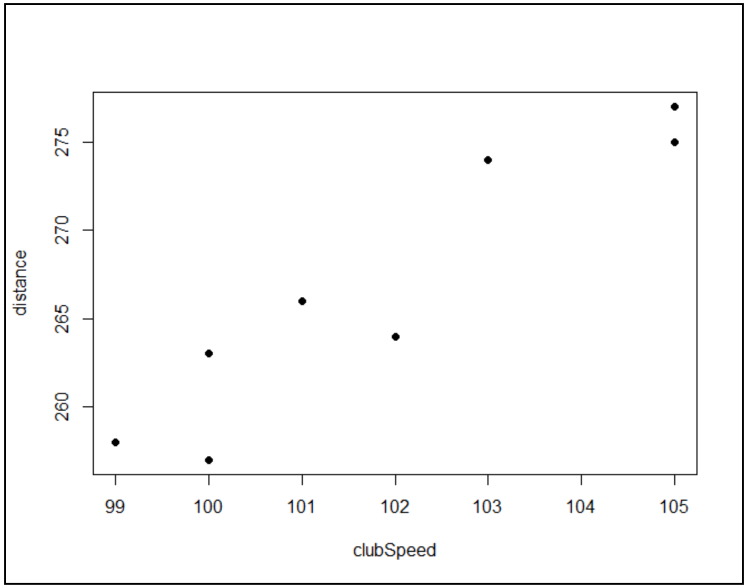
Bivariate Variables

- Variables
  - Response variables
  - Explanatory or predictor variable
- Response variable
  - Whose value can be explained by the explanatory variable or predictor variable

Scatter Plot

Club-head speed (mph)	Distance (yards)
100	257
102	264
103	274
101	266
105	277
100	263
99	258
105	275

- How to graphically represent bivariate data
- A scatter diagram is a graph that shows the relationship between 2 quantitative variables on the same individual



Covariance

Covariance - Definition

- Covariance
  - The covariance measures the direction of the linear relationship between two quantitative variables.
  - If the values of x and y become large or small, the covariance coefficient will also become large or small

$$\text{Data}:\{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$$

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

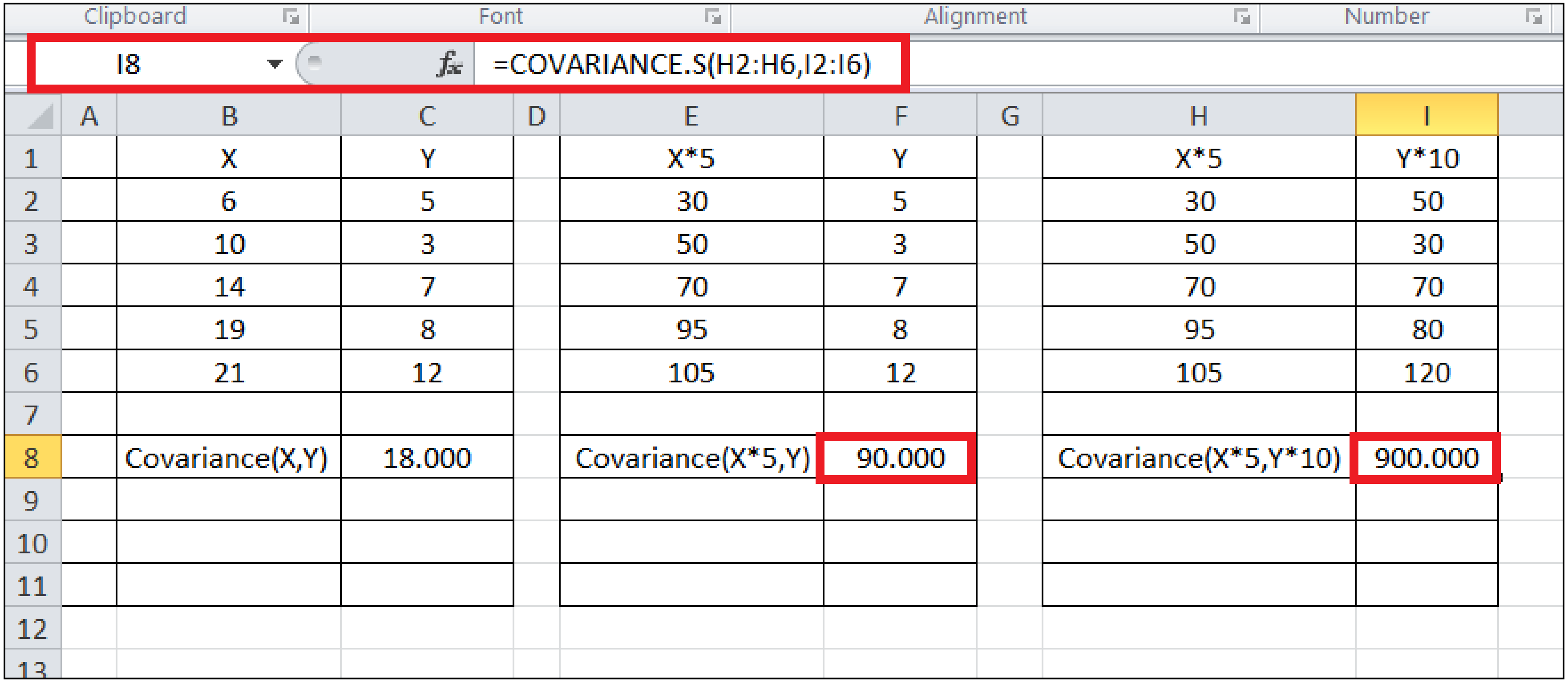
Covariance Example 1

$$\mathrm{Cov}(X,Y) = S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$

	X	Y		Deviation in X	Deviation in Y	Product	
1	6	5		6 - 14 = -8	5 - 7 = -2	-8 * -2 = 16	
2	10	3		10 - 14 = -4	3 - 7 = -4	-4 * -4 = 16	
3	14	7		14 - 14 = 0	7 - 7 = 0	0 * 0 = 0	
4	19	8		19 - 14 = 5	8 - 7 = 1	5 * 1 = 5	
5	21	12		21 - 14 = 7	12 - 7 = 5	7 * 5 = 35	
Mean	14	7		Add up the products = 16 + 16 + 0 + 5 + 35 = 72			
				Divide by (n-1) = (5 - 1) = 4			
	Covariance	18		Cov=72/4=18			

Property of Covariance

$$\mathrm{Cov}(a+bX,c+dY)=bd\,\mathrm{Cov}(X,Y)$$



I8 | f_x | =COVARIANCE.S(H2:H6,I2:I6)

	A	B	C	D	E	F	G	H	I
1		X	Y		X*5	Y		X*5	Y*10
2		6	5		30	5		30	50
3		10	3		50	3		50	30
4		14	7		70	7		70	70
5		19	8		95	8		95	80
6		21	12		105	12		105	120
7									
8		Covariance(X,Y)	18.000		Covariance(X*5,Y)	90.000		Covariance(X*5,Y*10)	900.000
9									
10									
11									
12									
13									

Covariance

- Covariance, unlike correlation, doesn’t have to be between -1 and 1.
- Covariance doesn’t give us a real sense of how negatively they are
  - If X and Y have large values, the covariance will be large as well
  - If suppose covariance is -100, it doesn’t give us a real sense of how negatively related they are

Correlation

Correlation - Definition

- The correlation measures the strength and direction of the linear relationship between two quantitative variables.
- Correlation is usually written as “r”
  - ‘r’ can vary between -1 and 1

$$\text{Total data points} = \mathrm{n}$$

$$\text{x values}: \bar{\mathrm{x}} \text{ is the mean, } \sigma_{\mathrm{x}} \text{ is the standard deviation}$$

$$\text{y values}: \bar{\mathrm{y}} \text{ is the mean, } \sigma_{\mathrm{y}} \text{ is the standard deviation}$$

$$\mathrm{r} = \frac{\sum (\mathrm{x} - \bar{x})(\mathrm{y} - \bar{y})}{\sqrt{\sum (\mathrm{x} - \bar{x})^2 \sum (\mathrm{y} - \bar{y})^2}} = \frac{\sum (\mathrm{x} - \bar{x})(\mathrm{y} - \bar{y})}{(\mathrm{n} - 1)\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}}$$

Covariance & Correlation

- Covariance

$$\text{Data}: \{(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)\}$$

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

- Correlation

$$r = \frac{S_{xy}}{\sigma_x \sigma_y}$$

$$\sigma_x = \text{Standard Deviation for x}$$

$$\sigma_y = \text{Standard Deviation for y}$$

$\rho(\text{rho}) = \text{population correlation}$

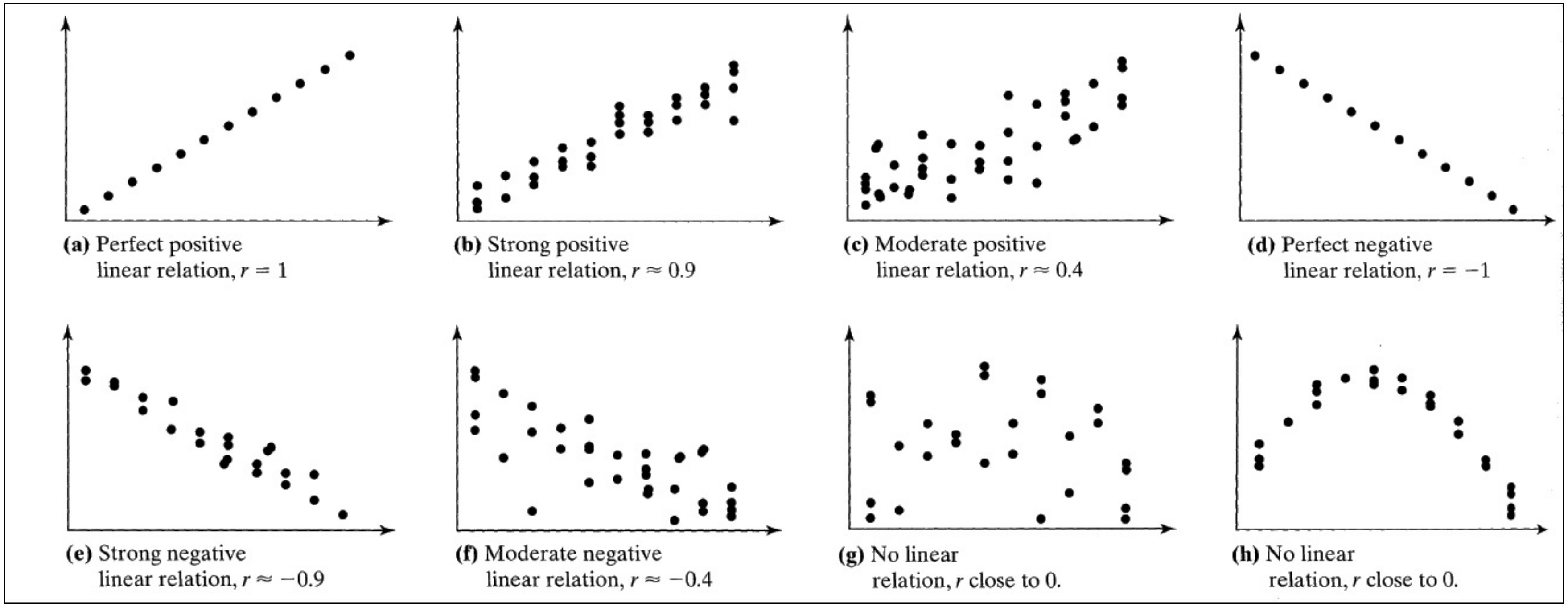
$r = \text{sample correlation}$

Correlation Example 1

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)\sigma_x\sigma_y}$$

	X	**Y**		**Deviation in X**	**Deviation in Y**	**Product**	
1	6	5		6 - 14 = -8	5 - 7 = -2	-8 * -2 = 16	
2	10	3		10 - 14 = -4	3 - 7 = -4	-4 * -4 = 16	
3	14	7		14 - 14 = 0	7 - 7 = 0	0 * 0 = 0	
4	19	8		19 - 14 = 5	8 - 7 = 1	5 * 1 = 5	
5	21	12		21 - 14 = 7	12 - 7 = 5	7 * 5 = 35	
Mean	14	7		Add up the products = 16 + 16 + 0 + 5 + 35 = 72			
Standard Dev	6.20	3.39		Divide by (n-1) * σ_x * σ_y = (5 - 1)* 6.20 * 3.39 = 84.07			
	Correlation	0.855		r = 72/84.07 = 0.855			

Correlation - Values



(a) Perfect positive linear relation, $r = 1$

(b) Strong positive linear relation, $r \approx 0.9$

(c) Moderate positive linear relation, $r \approx 0.4$

(d) Perfect negative linear relation, $r = -1$

(e) Strong negative linear relation, $r \approx -0.9$

(f) Moderate negative linear relation, $r \approx -0.4$

(g) No linear relation, r close to 0.

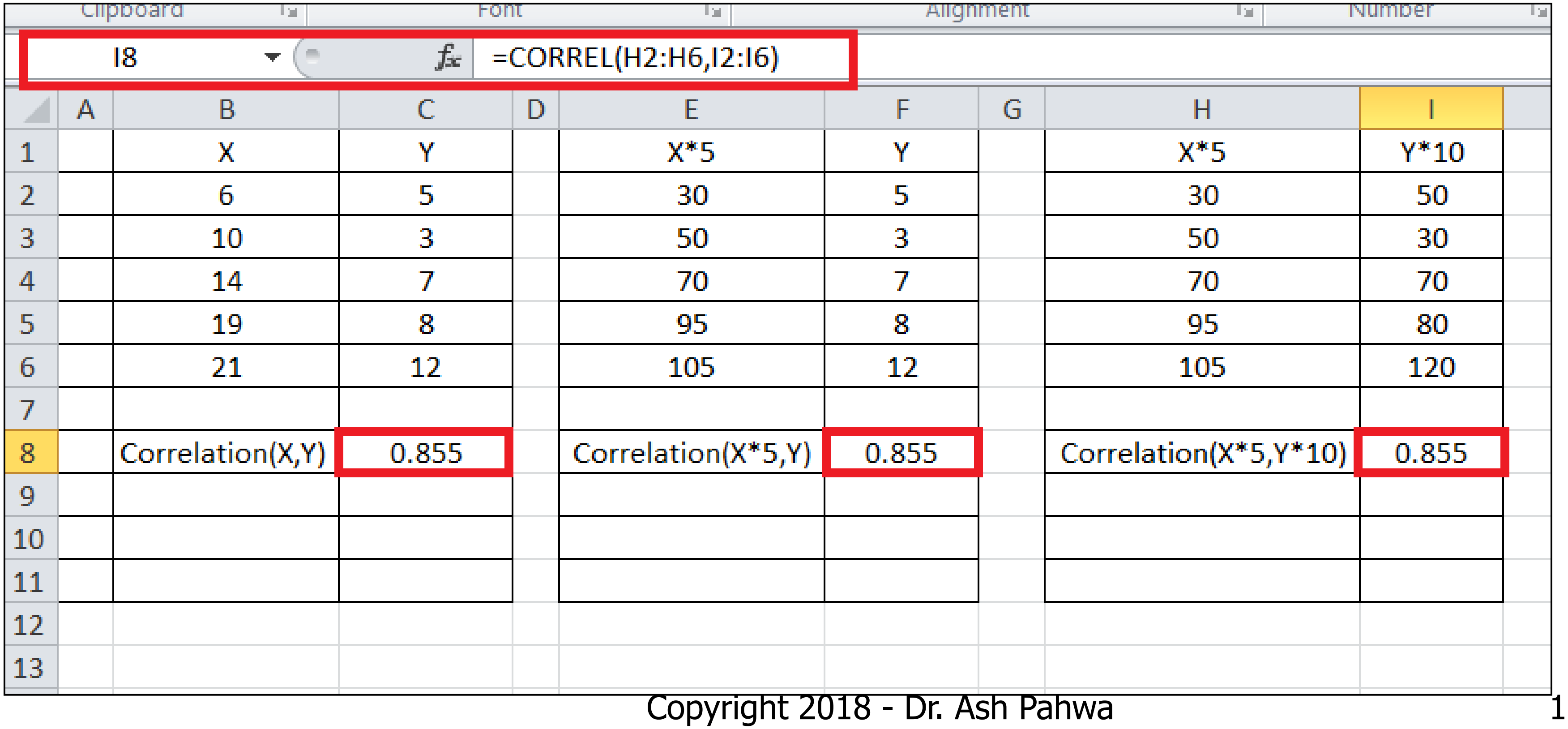
(h) No linear relation, r close to 0.

Properties of the Correlation Coefficient

- Unlike covariance, correlation varies between -1 and 1
  - $-1 <= r <= 1$
  - $r > 0$
    - If y increases as x increases
  - $r < 0$
    - If y decreases as x increases
  - The more x and y are linearly related
    - The closer r will be to -1 or 1

Property of Correlation

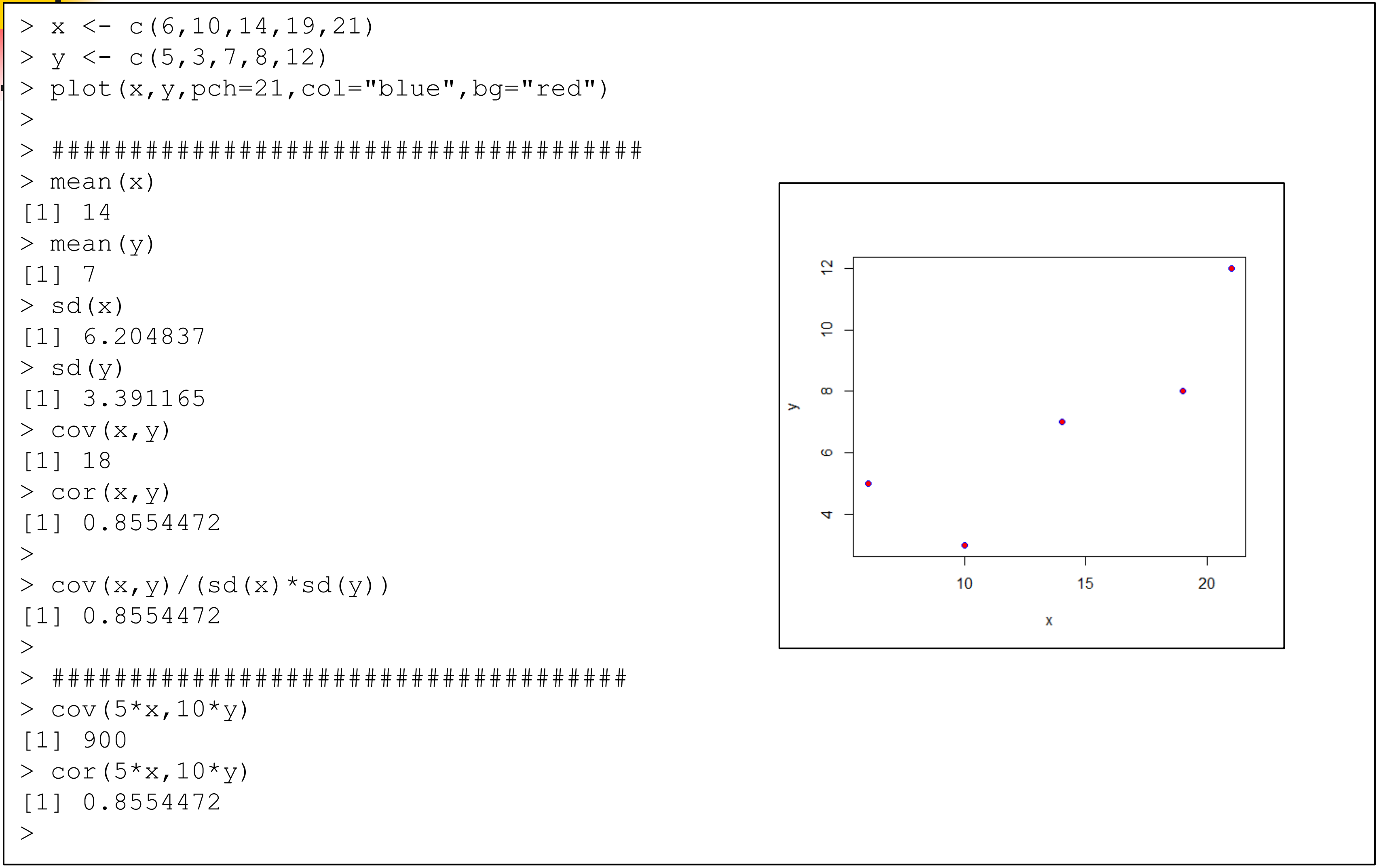
$$\boxed{\text{Corr}(a + bX, c + dY) = \text{sign}(bd)\text{Corr}(X, Y)}$$



I8 =CORREL(H2:H6,I2:I6)

	A	B	C	D	E	F	G	H	I
1		X	Y		X*5	Y		X*5	Y*10
2		6	5		30	5		30	50
3		10	3		50	3		50	30
4		14	7		70	7		70	70
5		19	8		95	8		95	80
6		21	12		105	12		105	120
7									
8		Correlation(X,Y)	0.855		Correlation(X*5,Y)	0.855		Correlation(X*5,Y*10)	0.855
9									
10									
11									
12									
13									

Covariance and Correlation in R



```
> x <- c(6,10,14,19,21)
> y <- c(5,3,7,8,12)
> plot(x,y,pch=21,col="blue",bg="red")
>
> ######################################
> mean(x)
[1] 14
> mean(y)
[1] 7
> sd(x)
[1] 6.204837
> sd(y)
[1] 3.391165
> cov(x,y)
[1] 18
> cor(x,y)
[1] 0.8554472
>
> cov(x,y)/(sd(x)*sd(y))
[1] 0.8554472
>
> #####################################
> cov(5*x,10*y)
[1] 900
> cor(5*x,10*y)
[1] 0.8554472
>
```

Correlation: Python Pandas + Numpy

```
import numpy as np
import pandas as pd
##################################################
# Correlation + Covariance in Pandas
#
df1 = pd.DataFrame({'A':[6,10,14,19,21], 'B':[5,3,7,8,12]})
df1
Out[70]:
    A   B
0   6   5
1  10   3
2  14   7
3  19   8
4  21  12

df1.corr()
Out[71]:
          A         B
A  1.000000  0.855447
B  0.855447  1.000000

df1.cov()
Out[72]:
      A     B
A  38.5  18.0
B  18.0  11.5
```

```
import numpy as np
import pandas as pd
###################################
# Correlation + Covariance in Numpy
#
Alist = [6,10,14,19,21]
Blist =[5,3,7,8,12]
Aarray = np.array(Alist)
Barray = np.array(Blist)

np.corrcoef(Aarray,Barray)
Out[80]:
array([[ 1.        ,  0.85544722],
       [ 0.85544722,  1.        ]])

np.cov(Aarray,Barray)
Out[81]:
array([[ 38.5,  18. ],
       [ 18. ,  11.5]])
```

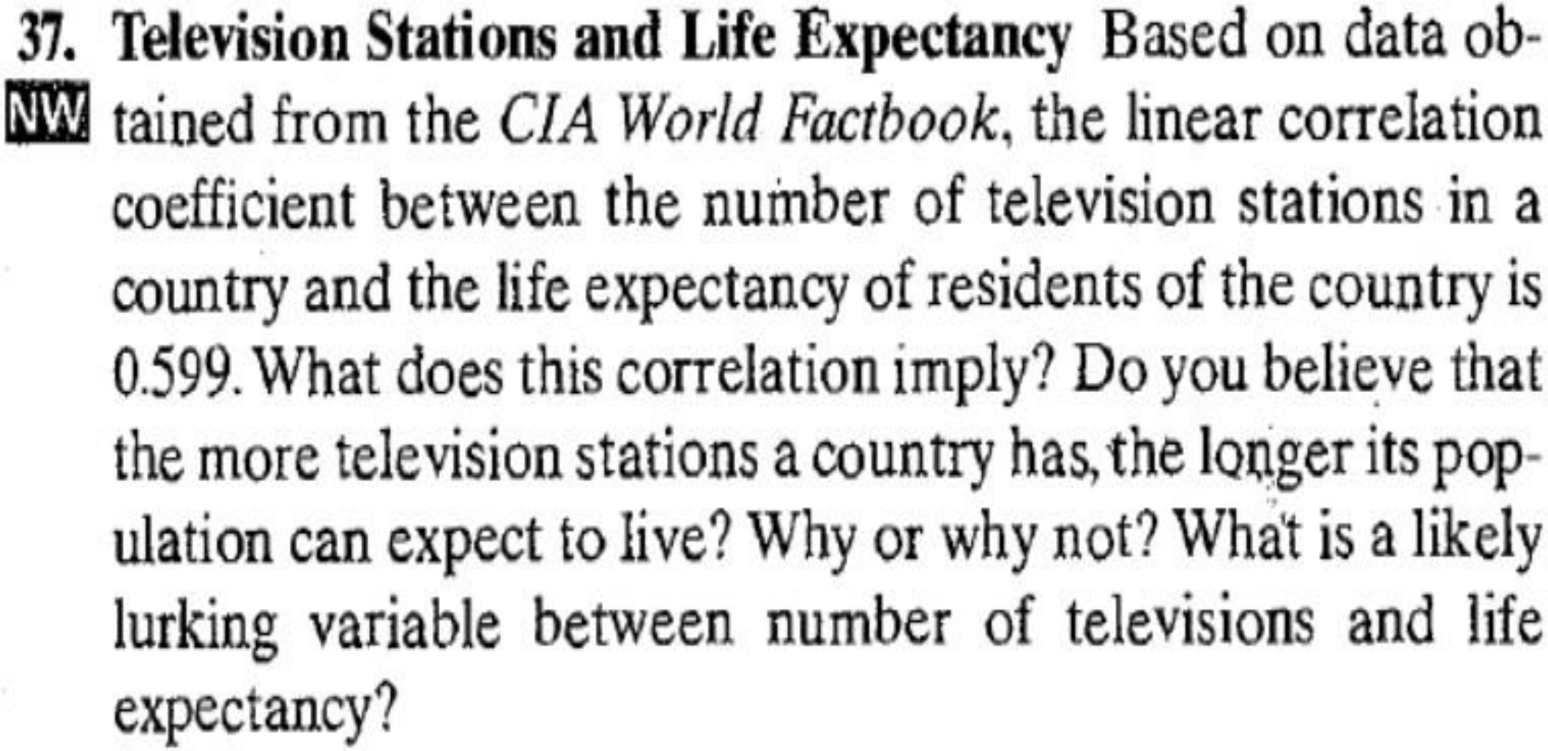
Linear Relationship

- If the correlation between 2 variables is high (close to +1 or -1)
  - We can conclude that there is a linear relationship between 2 variables

Correlation and Causation

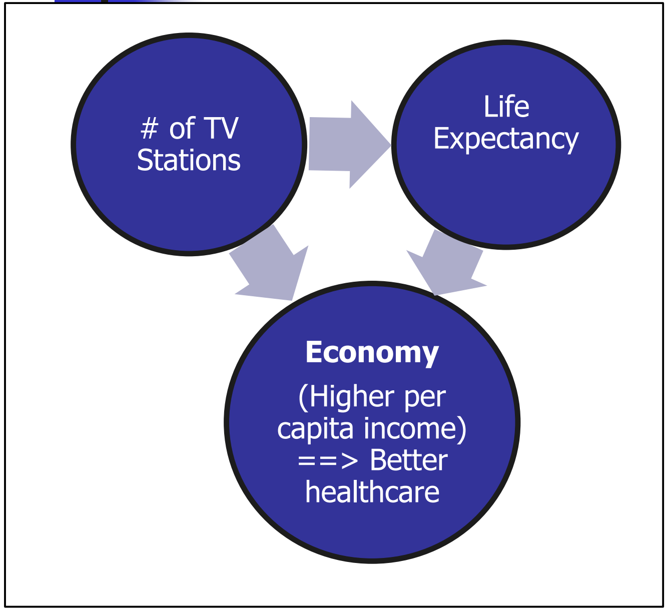
- If 2 variables are correlated
  - We cannot conclude that they have casual relationship
- Lurking variable
  - Third variable that explains the relationship
  - Ac bill goes up, crime rate goes up
    - Ac bill and crime data is highly correlated
  - Lurking variable – temperature
    - Temperature goes up, Ac bill goes up
    - Temperature goes up, crime rate goes up
    - Now it makes sense

Correlation and Causation



37. NW **Television Stations and Life Expectancy** Based on data obtained from the *CIA World Factbook*, the linear correlation coefficient between the number of television stations in a country and the life expectancy of residents of the country is 0.599. What does this correlation imply? Do you believe that the more television stations a country has, the longer its population can expect to live? Why or why not? What is a likely lurking variable between number of televisions and life expectancy?

Correlation and Causation



- The correlation between ‘number of TV stations’ and ‘Life Expectancy’ is 0.599. This means that as the number of TV stations will increase, life expectancy will also increase.
-
- However, correlation does not imply causation. By adding more number of TV stations will not increase life expectancy.
-
- There is a lurking variable here which is ‘Economy’.
  - If economy improves,
    - number of TV stations will also increase.
  - As economy improves,
    - this will lead to higher per capita income,
    - which will lead to better health care,
    - which will lead to higher life expectancy.

Summary

- Covariance
- Properties of Covariance
- Correlation Coefficient
- Properties of Correlation