AUA CS108, Statistics, Fall 2020

Lecture 16

Michael Poghosyan

02 Oct 2020

Contents

Reminder

Recall the definitions of the Sample Covariance and Correlation Coefficient between Datasets x and y of the same size (with denominator n):

$$cov(x, y) = s_{xy} = \frac{\displaystyle\sum_{k=1}^{n}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{n}$$

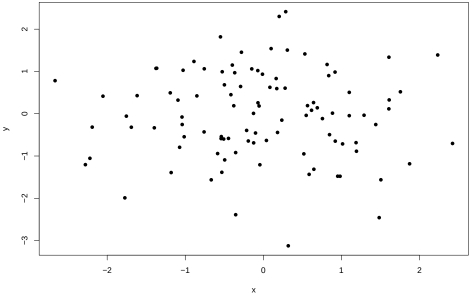
and

$$cor(x, y) = \rho_{xy} = \frac{cov(x, y)}{\sqrt{Var(x) \cdot Var(y)}} = \frac{cov(x, y)}{sd(x) \cdot sd(y)} = \frac{s_{xy}}{s_x \cdot s_y},$$

Examples:

Some simulations:

```
x <- rnorm(100); y <- rnorm(100);
plot(x,y, pch=16)
```



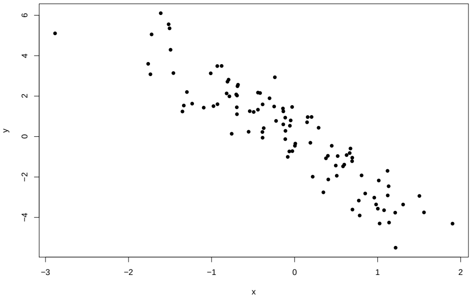

```
c(cor(x,y), cov(x,y))
```

```
## [1] -0.01712577 -0.01878710
```

Examples:

Some simulations:

```
x <- rnorm(100); y <- -2.4*x + rnorm(100);
plot(x,y, pch=16)
```




```
c(cor(x,y), cov(x,y))
```

```
## [1] -0.9106775 -2.0742413
```

Examples:

Let us now use the `state.x77` Dataset from **R**:

```r
head(state.x77)
```

```
##            Population Income Illiteracy Life Exp Murder HS Gr
## Alabama          3615   3624        2.1    69.05   15.1    41
## Alaska            365   6315        1.5    69.31   11.3    66
## Arizona          2212   4530        1.8    70.55    7.8    58
## Arkansas         2110   3378        1.9    70.66   10.1    39
## California      21198   5114        1.1    71.71   10.3    62
## Colorado         2541   4884        0.7    72.06    6.8    63
```

Examples:

Let us now use the `state.x77` Dataset from **R**:

```r
head(state.x77)
```

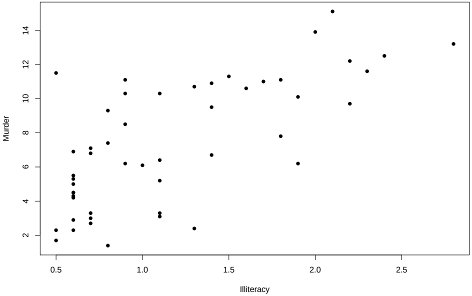
```
##            Population Income Illiteracy Life Exp Murder HS Gr
## Alabama          3615   3624        2.1    69.05   15.1    41
## Alaska            365   6315        1.5    69.31   11.3    66
## Arizona          2212   4530        1.8    70.55    7.8    58
## Arkansas         2110   3378        1.9    70.66   10.1    39
## California      21198   5114        1.1    71.71   10.3    62
## Colorado         2541   4884        0.7    72.06    6.8    63
```

It is not of the DataFrame format, so we change it to DataFrame:

```r
state <- as.data.frame(state.x77)
```

Examples:

```r
plot(Murder~Illiteracy, data = state, pch=16)
```




```r
cor(state$Illiteracy, state$Murder)
```

```
## [1] 0.7029752
```

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Say, we want to have Datasets x, y of size n with $cor(x, y) = \rho \in (-1, 1)$.

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Say, we want to have Datasets x, y of size n with $cor(x, y) = \rho \in (-1, 1)$.

One of the possible methods: take a Matrix

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

which is **Positive Definite**, take any 2D vector, say $\mu = [0, 0]^T$, and generate a Sample of size n from the Bivariate Normal Distribution $\mathcal{N}(\mu, \Sigma)$.

Examples:

Question: How to generate samples x, y with some given Correlation Coefficient?

Answer: Say, we want to have Datasets x, y of size n with $cor(x, y) = \rho \in (-1, 1)$.

One of the possible methods: take a Matrix

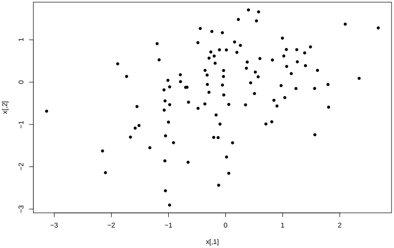
$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

which is **Positive Definite**, take any 2D vector, say $\mu = [0, 0]^T$, and generate a Sample of size n from the Bivariate Normal Distribution $\mathcal{N}(\mu, \Sigma)$.

Then, the $cor(x, y)$ will be approximately ρ (and it will approach ρ as $n \to +\infty$).

Example

```
rho <- 0.35
covmatrix <- matrix(c(1,rho, rho, 1), nrow = 2)
mu <- c(0,0)
x <- mvtnorm::rmvnorm(100, mean = mu, sigma = covmatrix)
plot(x, pch = 16)
```




```
cor(x)
```

```
##              [,1]      [,2]
## [1,] 1.0000000 0.3965342
## [2,] 0.3965342 1.0000000
```

Properties of the Sample Covariance

- $cov(x, y) = cov(y, x)$;

Properties of the Sample Covariance

- $cov(x, y) = cov(y, x)$;
- For any Datasets x, y, z and real numbers α, β,

$$cov(\alpha \cdot x + \beta \cdot y, z) = \alpha \cdot cov(x, z) + \beta \cdot cov(y, z);$$

Properties of the Sample Covariance

- $cov(x, y) = cov(y, x)$;
- For any Datasets x, y, z and real numbers α, β,

$$cov(\alpha \cdot x + \beta \cdot y, z) = \alpha \cdot cov(x, z) + \beta \cdot cov(y, z);$$

- For any Dataset x,

$$cov(x, x) = var(x)$$

Properties of the Sample Correlation Coefficient

- $cor(x, y) = cor(y, x)$;

[1]Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2]Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Properties of the Sample Correlation Coefficient

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Properties of the Sample Correlation Coefficient

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Properties of the Sample Correlation Coefficient

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$
- For any Datasets x, y,

$$-1 \leq \rho_{xy} \leq 1;$$

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Properties of the Sample Correlation Coefficient

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$
- For any Datasets x, y,

$$-1 \leq \rho_{xy} \leq 1;$$

- $\rho_{xy} = 1$ iff there exists a constant $a > 0$ and $b \in \mathbb{R}$ such that[1] $y_i = a \cdot x_i + b$ for any $i = 1, ..., n$.

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Properties of the Sample Correlation Coefficient

- $cor(x, y) = cor(y, x)$;
- If $\alpha > 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = cor(x, y)$
- If $\alpha < 0$ and $\beta \in \mathbb{R}$, then $cor(\alpha \cdot x + \beta, y) = -cor(x, y)$
- For any Datasets x, y,

$$-1 \leq \rho_{xy} \leq 1;$$

- $\rho_{xy} = 1$ iff there exists a constant $a > 0$ and $b \in \mathbb{R}$ such that[1] $y_i = a \cdot x_i + b$ for any $i = 1, ..., n$.
- $\rho_{xy} = -1$ iff there exists a constant $a < 0$ and $b \in \mathbb{R}$ such that[2] $y_i = a \cdot x_i + b$ for any $i = 1, ..., n$.

---

[1] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).
[2] Or $x_i = a \cdot y_i + b$ for any $i = 1, ..., n$ (maybe for another a and b).

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

- If $|cov(x, y)| > |cov(z, t)|$, we cannot state that the relationship between x and y is stronger than the relationship between z and t.

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

- If $|cov(x, y)| > |cov(z, t)|$, we cannot state that the relationship between x and y is stronger than the relationship between z and t. But if $|cor(x, y)| > |cor(z, t)|$, we can.

Pros/Cons of Sample Covariance and Correlation Coefficient

- Covariance is *linear*, correlation is not
- Correlation is scale-invariant: if we will change the scale of one or both Datasets, then the Correlation Coefficient will not be changed (but the Covariance will be).

Say, if x is a Dataset of heights of some persons, in centimeters, y their weights in grams, and if x' will be the same heights Dataset using meters as units, and y' will be the weights in Kg-s, then $cov(x, y)$ and $cov(x', y')$ will not be the same, but $cor(x, y) = cor(x', y')$.

- If $|cov(x, y)| > |cov(z, t)|$, we cannot state that the relationship between x and y is stronger than the relationship between z and t. But if $|cor(x, y)| > |cor(z, t)|$, we can.

So it is not easy to interpret the magnitude of the covariance, but the magnitude of the correlation coefficient is the strength of the linear relationship.

Pros/Cons of Sample Covariance and Correlation Coefficient

- An important drawback of the Sample Correlation Coefficient is that it is sensitive to outliers.

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

- The sign of Covariance and Corelation Coefficient shows the direction of the relationship: if

$$cov(x, y) > 0, \quad \text{equivalently, if} \quad cor(x, y) > 0,$$

then if x is increasing, then y also tends to be larger.

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

- The sign of Covariance and Corelation Coefficient shows the direction of the relationship: if

  $$cov(x, y) > 0, \quad \text{equivalently, if} \quad cor(x, y) > 0,$$

  then if x is increasing, then y also tends to be larger. And if

  $$cov(x, y) < 0, \quad \text{equivalently, if} \quad cor(x, y) < 0,$$

  then if x is increasing, then y tends to be smaller.

Covariance and Correlation Coefficient, again

So what are showing Covariance and Correlation Coefficient:

- The sign of Covariance and Corelation Coefficient shows the direction of the relationship: if

$$cov(x,y) > 0, \quad \text{equivalently, if} \quad cor(x,y) > 0,$$

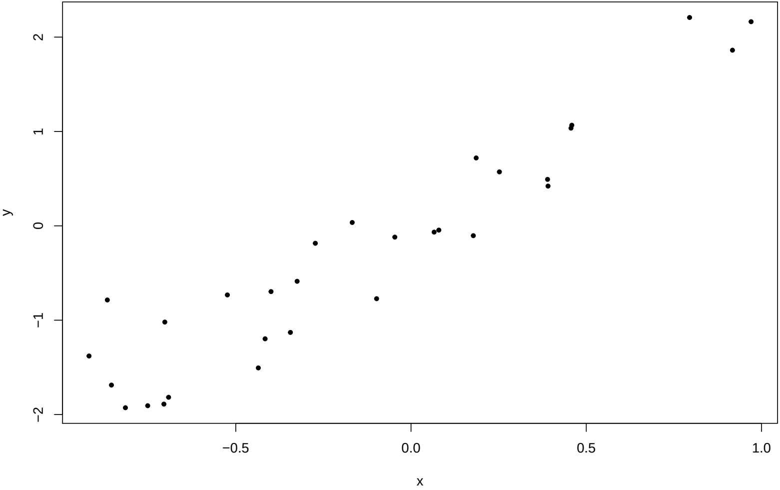
  then if x is increasing, then y also tends to be larger. And if

$$cov(x,y) < 0, \quad \text{equivalently, if} \quad cor(x,y) < 0,$$

  then if x is increasing, then y tends to be smaller.
- The magnitude of the Correlation Coefficient shows the strength of the Linear Relationship.

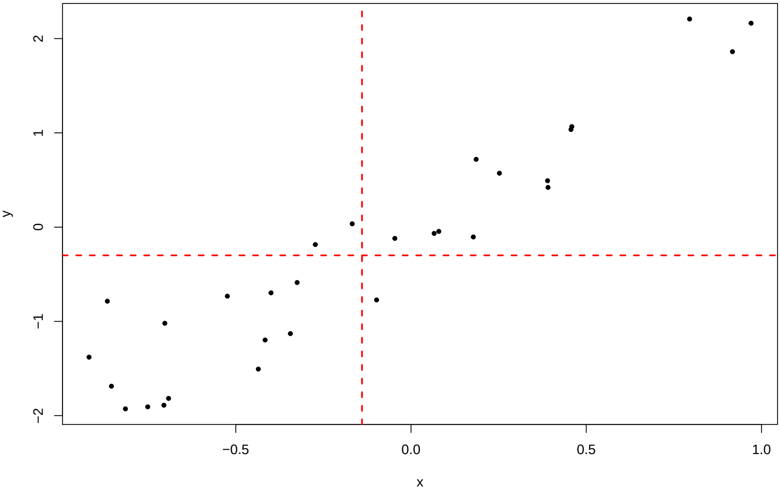
Explanation

Here is a Bivariate Dataset (x, y) with $cov(x, y) > 0$:



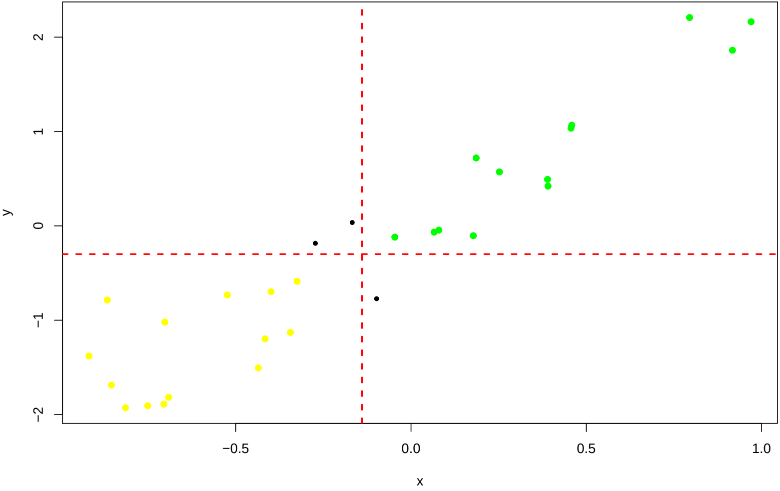
Explanation

Now we add a vertical line through $\bar{x}$ and a horizontal line through $\bar{y}$



Explanation

We color the points in the first and third quadrants:



Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad and \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad and \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Similarly, Points in the 3rd quadrant, yellow points, again contribute positive terms to $cov(x, y)$, since in this case

$$x_k < \bar{x} \quad and \quad y_k < \bar{y}, \quad hence, \quad (x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0.$$

Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad and \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Similarly, Points in the 3rd quadrant, yellow points, again contribute positive terms to $cov(x, y)$, since in this case

$$x_k < \bar{x} \quad and \quad y_k < \bar{y}, \quad hence, \quad (x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0.$$

In the same way, the points in the 2nd and 4th quadrants give negative terms to $cov(x, y)$, as in this case $(x_k - \bar{x}) \cdot (y_k - \bar{y}) < 0$.

Explanation

The points in the 1st quadrant (of the dotted coordinate system, with the center at $(\bar{x}, \bar{y})$), green points, satisfy

$$x_k > \bar{x} \quad \textit{and} \quad y_k > \bar{y},$$

so

$$(x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0,$$

so green points contribute positive terms to

$$cov(x, y) = \frac{1}{n} \cdot \sum_{k=1}^{n} (x_k - \bar{x}) \cdot (y_k - \bar{y}).$$

Similarly, Points in the 3rd quadrant, yellow points, again contribute positive terms to $cov(x, y)$, since in this case

$$x_k < \bar{x} \quad \textit{and} \quad y_k < \bar{y}, \quad \textit{hence}, \quad (x_k - \bar{x}) \cdot (y_k - \bar{y}) > 0.$$

In the same way, the points in the 2nd and 4th quadrants give negative terms to $cov(x, y)$, as in this case $(x_k - \bar{x}) \cdot (y_k - \bar{y}) < 0$. And positive covariance means that the terms for points in the 1st and 3rd quadrants dominate to the ones from 2nd and fourth ones.

Note: Of course, we can have a negative trend and just one strong outlier in the 1st quadrant resulting in a positive covariance.