STAT 234 Lecture 23A
Sample Covariance and Correlation
Section 12.5

Yibi Huang
Department of Statistics
University of Chicago

Sample Covariance

Given n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, *sample covariance* s_{xy} is a measure of the *direction* and *strength* of the linear relationship between X and Y, defined as

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Sample Covariance

Given n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, *sample covariance* s_{xy} is a measure of the *direction* and *strength* of the linear relationship between X and Y, defined as

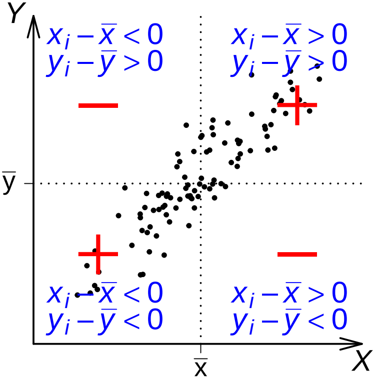
$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

- $s_{xy} > 0$: Positive linear relation;
- $s_{xy} < 0$: Negative linear relation
- The *magnitude* of covariance reflects the *strength* of the relation
- The covariance of a variable X with itself is its *sample variance*

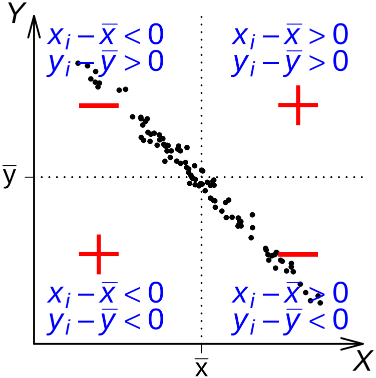
$$s_{xx} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s_x^2$$

Sample Covariance Reflects the Direction of a Linear Relation

What is the sign of $(x_i - \bar{x})(y_i - \bar{y})$?

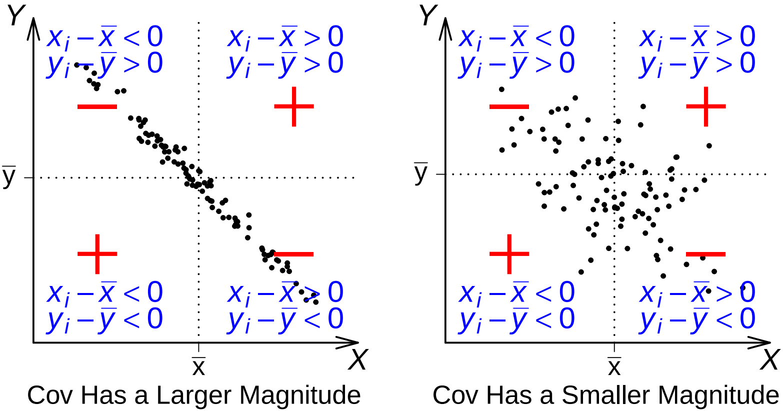


Cov > 0 as most points have $(x_i - \bar{x})(y_i - \bar{y}) > 0$



Cov < 0 as most points have $(x_i - \bar{x})(y_i - \bar{y}) < 0$

Sample Covariance Reflects the Strength of a Linear Relation



Covariance is of a smaller magnitude in the right plot than in the left because the $(x_i - \bar{x})(y_i - \bar{y})$ of most points in the left plot are of the different signs and get cancelled out when adding up.

How Large the Covariance is Large Enough?

It can be shown in the next slide that

$$|s_{xy}| \leq s_x s_y = (\text{SD of } X) \times (\text{SD of } Y)$$

How Large the Covariance is Large Enough?

It can be shown in the next slide that

$$|s_{xy}| \leq s_x s_y = (\text{SD of } X) \times (\text{SD of } Y)$$

Moreover, the sample covariance reaches its maximum possible magnitude if and only if all the points (x_i, y_i) fall on a straight line.

How Large the Covariance is Large Enough?

It can be shown in the next slide that

$$|s_{xy}| \leq s_x s_y = (\text{SD of } X) \times (\text{SD of } Y)$$

Moreover, the sample covariance reaches its maximum possible magnitude if and only if all the points (x_i, y_i) fall on a straight line.

Thus, one can determine whether a linear relation is strong by comparing the Cov with the product of the SDs of the two variables.

Proof of $|s_{xy}| \le s_x s_y$

For any two sequences $(a_1, a_2, \dots, a_n)$ and $(b_1, b_2, \dots, b_n)$, the Cauchy Schwartz Inequality below is always true

$$\left(\sum_{i=1}^n a_i b_i\right)^2 \le \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right)$$

Moreover, the inequality becomes an equality if and only if

$$\alpha a_i + \beta b_i = 0 \quad \text{for all } i \text{ for some non-zero constants } \alpha \text{ and } \beta.$$

Applying Cauchy Schwartz Inequality with $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$, we get

$$\underbrace{\left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right)^2}_{[(n-1)s_{xy}]^2} \le \underbrace{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)}_{(n-1)s_x^2}\underbrace{\left(\sum_{i=1}^n (y_i - \bar{y})^2\right)}_{(n-1)s_y^2}.$$

Dividing both sides by $(n-1)^2$, and taking square-root, we get

$$|s_{xy}| \le s_x s_y.$$

Proof of $|s_{xy}| \leq s_x s_y$ (Cont'd)

Moreover, recall the the inequality becomes an equality if and only if

$$\alpha a_i + \beta b_i = 0 \quad \text{for all } i \text{ for some nonzero constants } \alpha \text{ and } \beta.$$

Now with $a_i = x_i - \bar{x}$ and $b_i = y_i - \bar{y}$, we get that $|s_{xy}|$ reach its max $s_x s_y$ if and only if

$$\alpha(x_i - \bar{x}) + \beta(y_i - \bar{y}) = 0 \quad \text{for all } i \text{ for some nonzero constants } \alpha \text{ and } \beta,$$

or equivalently all the points (x_i, y_i) fall on the straight line

$$\alpha x_i + \beta y_i = \alpha \bar{x} + \beta \bar{y}$$

Shortcut Formula for the Sample Covariance

There are various formula for computing the sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \frac{\left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}}{n-1}$$

The last one is the *shortcut formula* for calculating the *sample covariance*, similar to the shortcut formula for the sample variance

$$s_x^2 = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2}{n-1}$$

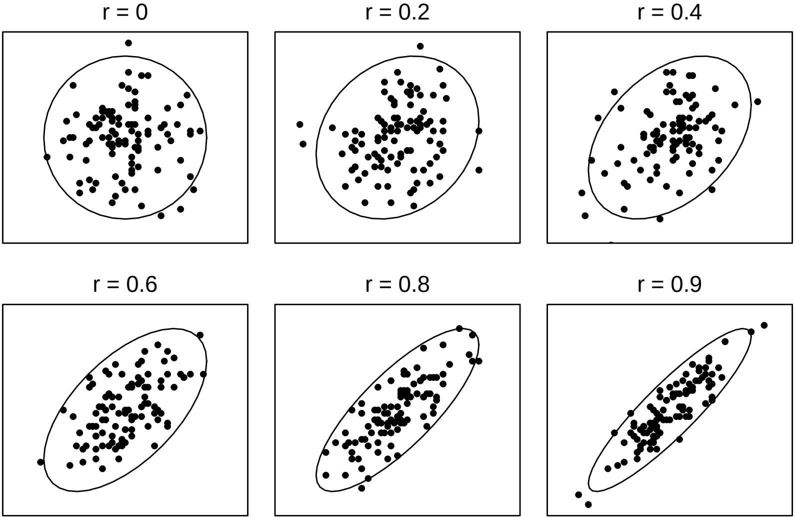
Sample Correlation = Correlation Coefficient r

Given n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, the (sample) **corelation** is defined to be

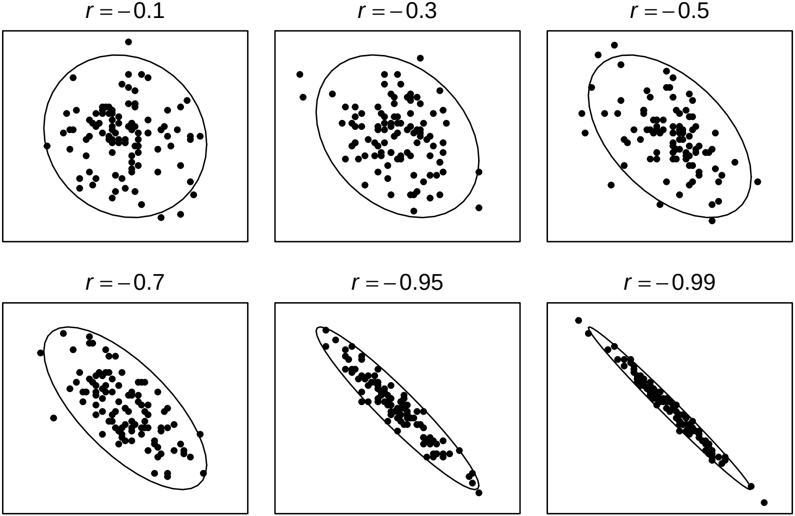
$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- $-1 \leq r \leq 1$ since $|s_{xy}| \leq s_x s_y$
- The closer r is to 1 or -1, the stronger the linear relation
- $r = 1$ or -1 if and only if all the points (x_i, y_i) fall on a straight line

Positive Correlations



Negative Correlations



Sample Correlation r v.s. Population Correlation ρ

Recall in Lecture 11 we introduced the *correlation* between two random variables X, Y,

$$\rho = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} = \frac{\mathrm{E}[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\mathrm{E}[(X - \mu_X)^2]\,\mathrm{E}[(Y - \mu_Y)^2]}}.$$

The sample correlation r

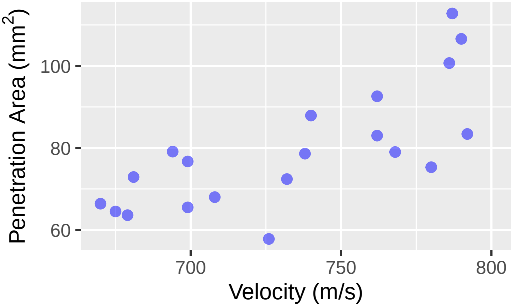
$$r_{xy} = r = \widehat{\rho} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}} = \frac{s_{xy}}{s_x s_y},$$

is an estimate for the population correlation ρ if $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are i.i.d. pairs of observations from the joint distribution of (X, Y).

Example: Armor Strength

Soldiers depend on their body armor for protection. Specimens of UHMWPE body armor were shot with a 7.62 mm round at various firing velocities. The penetration areas were recorded[a].

[a] "Testing of Body Armor Materials-Phase III", 2012, by the US Army and the National Research Council



Velocity (m/s)	Penetration Area (mm^2)
670	66.4
675	64.5
679	63.6
681	72.9
694	79.1
699	76.7
699	65.5
708	68.0
726	57.8
732	72.4
738	78.6
740	87.9
762	92.6
762	83.0
768	79.0
780	75.3
792	83.4
786	100.7
790	106.6
787	112.8

Finding Covariance & Correlation in R

Armor Strength Data and the variables:

```
armor = read.table(
  "http://www.stat.uchicago.edu/~yibi/s234/ArmorStrength.txt",
  header=TRUE)
str(armor)
'data.frame':   20 obs. of  2 variables:
 $ velocity        : int  670 675 679 681 694 699 699 708 726 732 ...
 $ penetration.area: num  66.4 64.5 63.6 72.9 79.1 76.7 65.5 68 57.8 72
```

The R commands `cov()` and `cor()` can calculate the sample covariance and sample correlation between two variables

```
cov(armor$velocity, armor$penetration.area)
[1] 471.0042
cor(armor$velocity, armor$penetration.area)
[1] 0.743148
```

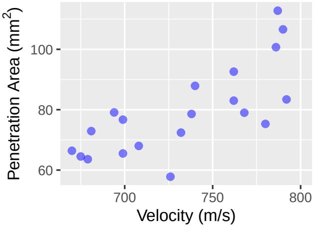
Covariance & Correlation Do Not Distinguish Between X & Y

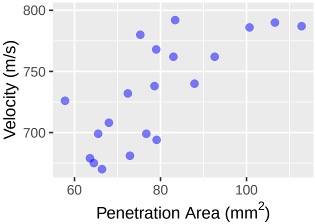
When one uses X to predict Y, X is called the *explanatory variable*, and Y the *response*. Covariance and correlation do not distinguish between X & Y. They treat X and Y symmetrically.

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) = s_{yx};$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{yx}}{s_x s_y} = r_{yx}$$

Swapping the x-, y-axes doesn't change r (both $r \approx 0.74$.)





Scaling Property of Sample Covariance

$$\begin{array}{ccc}
(X,\ Y) & \longrightarrow & (aX+b,\ cY+d) \\
\hline
(x_1,\ y_1) & & (ax_1+b,\ cy_1+d) \\
(x_2,\ y_2) & & (ax_2+b,\ cy_2+d) \\
(x_3,\ y_3) & \Rightarrow & (ax_3+b,\ cy_3+d) \\
\vdots & & \vdots \\
(x_n,\ y_n) & & (ax_n+b,\ cy_n+d)
\end{array}$$

The sample covariance has the scaling property:

$$\begin{aligned}
S_{aX+b,cY+d} &= \frac{1}{n-1}\sum\nolimits_{i=1}^{n}[ax_i+b-(a\bar{x}+b)][cy_i+d-(c\bar{y}+d)] \\
&= \frac{1}{n-1}\sum\nolimits_{i=1}^{n} ac(x_i-\bar{x})(y_i-\bar{y}) \\
&= ac\,S_{XY}.
\end{aligned}$$

Scaling Property of Sample Covariance — Example

Example. When $X =$ `velocity` is measured in feet/sec rather than meter/sec,

- the value of X becomes ≈ 3.28 times as large since

$$1 \text{ meter} \approx 3.28 \text{ feet.}$$

- the covariance between `velocity` and `penetration.area` would become about 3.28 times as large

```
x = armor$velocity
y = armor$penetration.area
cov(x, y)
[1] 471.0042
cov(3.28 * x, y)
[1] 1544.894
cov(x, y) * 3.28
[1] 1544.894
```

Correlation is Scale Invariant

The sample correlation is *scaling invariant* and *has no units*!

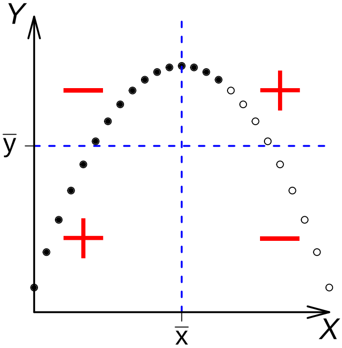
$$r_{aX+b,cY+d} = \frac{S_{aX+b,cY+d}}{S_{aX+b}S_{cY+d}} = \frac{ac\,S_{XY}}{|a|S_X\,|c|S_Y} = (\text{sign of } ac) \times \frac{s_{XY}}{s_X s_Y}$$
$$= (\text{sign of } ac) \times r_{XY}.$$

Example. When `velocity` is measured in ft/s rather than m/s, the value of `velocity` becomes ≈ 3.28 times as large, the correlation between `velocity` and `penetration.area` remain unchanged to be $r \approx 0.74$.

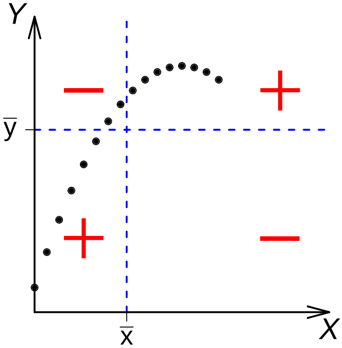
```
cor(x, y)
[1] 0.743148
cor(3.28 * x, y)
[1] 0.743148
```

Correlation Doesn't Reflect Strength of Nonlinear Relations

Both scatter plots below show perfect nonlinear relations. All points fall on the quadratic curve $y = 2 - x^2/2$.



r = 0 (why?)
(black + white dots)



r = 0.91
(black dots only)