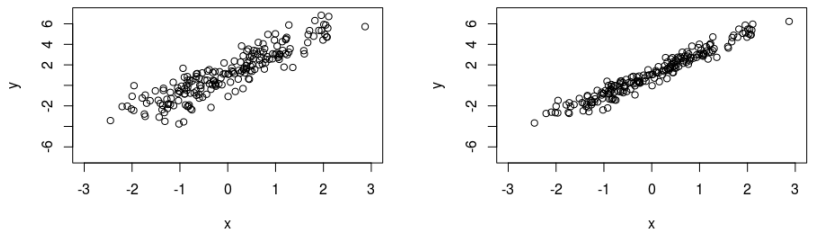
Estimation and testing for $Corr(X, Y)$

- For each of the following two plots, observations for two variables X and Y are plotted.



The data are generated according to the following two models respectively:

$$\text{Model 1: } Y = 1 + 2X + Z,$$

$$\text{Model 2: } Y = 1 + 2X + 0.5Z,$$

where X and Z are independent $N(0, 1)$ random variables. The plot in the right panel is for the data generated from Model 2, which shows a stronger linear relation between X and Y.

- Suppose that for two variables X and Y, we have n pairs of observations for (X, Y): (X_1, Y_1), ..., (X_n, Y_n). The strength of linear relation between two variables X and Y based on their observations is often reflected by the sample correlation.

- The sample correlation (or the sample correlation coefficient) between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is

$$\frac{\text{the sample covariance between } (X_1, \ldots, X_n) \text{ and } (Y_1, \ldots, Y_n)}{S_X S_Y}, \quad (1)$$

where S_X and S_Y are the sample standard deviations for $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively, and the sample covariance between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is defined as

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Here $\bar{X}$ and $\bar{Y}$ are the sample means for $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively. It is common to use r to denote the sample correlation.

- Suppose that (X_1, Y_1), ..., (X_n, Y_n) are IID pairs of observations and $(X_i, Y_i) \sim (X, Y)$. Then for large n,

$$\begin{aligned} &\text{the sample covariance between } (X_1, \ldots, X_n) \text{ and } (Y_1, \ldots, Y_n) \\ &\approx E(X - \mu_X)(Y - \mu_Y), \end{aligned} \tag{2}$$

  where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

  - The expectation $E(X-\mu_X)(Y-\mu_Y)$ is called the covariance between X and Y, which is denoted by $Cov(X, Y)$.
  - From (2), for large n, the sample correlation between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is close to

$$\frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}},$$

    since $S_X \approx \sqrt{Var(X)}$ and $S_Y \approx \sqrt{Var(Y)}$.

- The quantity

$$\frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

  is called the correlation between X and Y and is denoted by $Corr(X, Y)$.

- $Corr(X, Y)$ is a measurement of the strength of linear relation between X and Y. Below are some properties of the correlation.

  - $|Corr(X, Y)| \leq 1$.
  - $|Corr(X, Y)| = 1$ means $X = a + bY$ (or $Y = a + bX$) for some constants a, b.
  - If X and Y are independent, then $Corr(X, Y) = 0$.
  - $Corr(aX + b, cY + d) = Corr(X, Y)$ for constants a, b, c, d if $ac > 0$.

  For details and other properties of $Cov(X, Y)$ and $Corr(X, Y)$, see the handout "Covariance and correlation".

- Suppose that we have IID observations (X_1, Y_1), ..., (X_n, Y_n) such that $(X_i, Y_i) \sim (X, Y)$, then $Corr(X, Y)$ can be well approximated by r: the sample correlation between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ if n is large.

  - $|r| \leq 1$.
  - $|r| = 1$ means $X_i = a + bY_i$ (or $Y_i = a + bX_i$) for $i = 1$, ..., n for some constants a, b.

– The sample correlation coefficient between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is the same as the sample correlation coefficient between $(aX_1 + b, \ldots, aX_n + b)$ and $(cY_1 + d, \ldots, cY_n + d)$ for constants a, b, c, d if $ac > 0$.

- When computing the sample correlation, the formula

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \left(\sum_{i=1}^{n} X_i Y_i\right) - n\bar{X}\bar{Y}$$

is often used for computing the sample covariance.

- Example 1. 下表爲101/04/19 – 101/04/24 期間, 中華電和台灣大的每日股票收盤價. 計算這段期間中華電收盤價和台灣大收盤價的sample correlation coefficient.

	中華電收盤價	台灣大收盤價
101/04/19	88.3	91.1
101/04/20	87.6	90.5
101/04/23	88.3	91.6
101/04/24	89.7	93

Sol. 考慮將中華電收盤價-87, 台灣大收盤價-90, 以轉換後的資料計算 sample covariance and sample correlation coefficient. 轉換後的中華電收盤價爲 1.3, 0.6, 1.3, 2.7, sample mean 爲$(1.3+0.6+1.3+2.7)/4 = 5.9/4$, sample variance 爲

$$\frac{1}{4-1}\left(1.3^2 + 0.6^2 + 1.3^2 + 2.7^2 - 4(5.9/4)^2\right) = 9.31/12.$$

轉換後的台灣大收盤價爲 1.1, 0.5, 1.6, 3, sample mean 爲$(1.1 + 0.5 + 1.6 + 3)/4 = 6.2/4$, sample variance 爲

$$\frac{1}{4-1}\left(1.1^2 + 0.5^2 + 1.6^2 + 3.0^2 - 4(6.2/4)^2\right) = 13.64/12.$$

轉換後的中華電收盤價和轉換後的台灣大收盤價之sample covariance 爲

$$\begin{aligned} & \frac{1}{4-1}\left(1.3 \times 1.1 + 0.6 \times 0.5 + 1.3 \times 1.6 + 2.7 \times 3 - 4(5.9/4)(6.2/4)\right) \\ = \quad & \frac{1}{3}\left(\frac{11.06}{4}\right), \end{aligned}$$

而sample correlation coefficient 爲

$$\frac{(11.06/4)/3}{\sqrt{(9.31/12)(13.64/12)}} = 0.9814611.$$

- R commands. 假設兩組樣本$(X_1, \ldots, X_n)$及$(Y_1, \ldots, Y_n)$已分別儲存於R中的兩向量變數x及 y.

R 指令	用途
`var(x)`	計算 x 的 sample variance
`cov(x,y)`	計算 x 和 y 的 sample covariance
`length(x)`	計算 x 的長度, 即n.

執行以下R指令可用原始資料計算Example 1 中的sample correlation coefficient.

```
x=c(88.3, 87.6, 88.3, 89.7)
y=c(91.1, 90.5, 91.6, 93)
cov(x,y)/sqrt(var(x)*var(y))
```

- Problem set-up. Suppose that (X_1, Y_1), ..., (X_n, Y_n) form a random sample of n pairs and $(X_i, Y_i) \sim (X, Y)$. Let $\rho = Corr(X, Y)$. We would like to
  - estimate ρ and
  - test
$$H_0 : \rho = 0 \text{ versus } H_1 : \rho \neq 0$$

  based on (X_1, Y_1), ..., (X_n, Y_n). Since for large n, ρ can be approximated by r: the sample correlation between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, the test will be based on r.

- Suppose that (X_1, Y_1), ..., (X_n, Y_n) are IID pairs of observations and $(X_i, Y_i) \sim (X, Y)$. Let $\rho = Corr(X, Y)$. Suppose that the distribution for (X, Y) is bivariate normal, then under H_0: $\rho = 0$,
$$\sqrt{n-2}\left(\frac{r}{\sqrt{1-r^2}}\right) \sim t(n-2)$$
  for $n \geq 3$. Let
$$T = \sqrt{n-2}\left(\frac{r}{\sqrt{1-r^2}}\right), \tag{3}$$
  then for testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, one can reject H_0 at level a if
$$|T| > t_{a/2, n-2}.$$

- Note. When the distribution for (X, Y) is bivariate normal, $Corr(X, Y) = 0$ implies that X and Y are independent.

- Example 2. Suppose that $n = 12$ and (X_1, Y_1), ..., (X_n, Y_n) form a random sample of n pairs and the (observed) sample correlation coefficient is 0.32. Let $\rho = Corr(X_1, Y_1)$. Can we conclude that $\rho \neq 0$ at the 0.05 significant level?

  Sol. $t_{0.05/2,12-2} = t_{0.025,10} = 2.228$ and the observed T statistic

$$T = \sqrt{10}\left(\frac{0.32}{\sqrt{1-(0.32)^2}}\right) = 1.068092$$

  Since the observed $|T| < 2.228$, we can not conclude that $\rho \neq 0$ at the 0.05 significant level.

- Example 3. Suppose that two samples $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ are stored in two vectors `x` and `y` in R respectively. Suppose that we have the following R outputs:

```
> var(x)
[1] 0.7758333
> var(y)
[1] 1.136667
> cov(x,y)
[1] 0.9216667
> length(x)
[1] 4
```

  Find the sample correlation coefficient between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$. Can we conclude that $Corr(X_1, Y_1) \neq 0$ at level 0.05?

  Sol. The sample correlation coefficient between $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ is $0.9216667/\sqrt{(0.7758333 \times 1.136667)} = 0.981461$. The observed T statistic for testing $H_0 : Corr(X_1, Y_1) = 0$ is

$$\sqrt{4-2}\left(\frac{0.981461}{\sqrt{1-(0.981461)^2}}\right) = 7.241893.$$

  Since $t_{0.05/2,4-2} = t_{0.025,2} = 4.303 < |7.241893|$, we can conclude that $Corr(X_1, Y_1) \neq 0$ at level 0.05 (assuming (X_1, Y_1), ..., (X_n, Y_n) are IID bivariate normal).

- One-side tests. Suppose that (X_1, Y_1), ..., (X_n, Y_n) are IID pairs of observations and $(X_i, Y_i) \sim (X, Y)$. Let $\rho = Corr(X, Y)$. Suppose that the distribution for (X, Y) is bivariate normal and T is still defined according to (3).

- For testing $H_0 : \rho \leq 0$ versus $H_1 : \rho > 0$, we reject H_0 at level a if $T > t_{a,n-2}$.
- For testing $H_0 : \rho \geq 0$ versus $H_1 : \rho < 0$, we reject H_0 at level a if $T < -t_{a,n-2}$.

- Example 4. Suppose that (X_1, Y_1), ..., (X_n, Y_n) form a random sample of n pairs and the sample correlation coefficient is 0.32. Let ρ be the correlation between X_1 and Y_1.

  (a) Suppose that $n = 12$. Can we conclude that $\rho > 0$ at the 0.05 significant level?

  (b) Suppose that $n = 212$. Can we conclude that $\rho > 0$ at the 0.05 significant level?

  Sol.

  (a) No, the observed $T = \sqrt{12-2}(0.32)/\sqrt{1-(0.32)^2)} = 1.068092 < t_{0.05,10} = 1.812$.

  (b) Yes, the observed $T = \sqrt{212-2}(0.32)/\sqrt{1-(0.32)^2)} = 4.894611 > t_{0.05,200} = 1.653$.