Section 5.8 Covariance and correlation

Tuesday, November 15, 2016 12:31 PM

Section 5.8 Covarianc...

5.8 Covariance and correlation

5.8.1 Covariance

TERMINOLOGY: Suppose that Y_1 and Y_2 are random variables (discrete or continuous) with means $E(Y_1) = \mu_1$ and $E(Y_2) = \mu_2$, respectively. The **covariance** between Y_1 and Y_2 is given by

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &\equiv E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \\ &= E(Y_1Y_2) - E(Y_1)E(Y_2). \end{aligned}$$

$$= E[Y_1Y_2 - \mu_1Y_2 - \mu_2Y_1 + \mu_1\mu_2]$$

$$= E(Y_1Y_2) - \mu_1E(Y_2) - \cancel{\mu_2E(Y_1)} + \cancel{\mu_1\mu_2}$$

The latter expression is often easier to work with and is called the **covariance computing formula**. The covariance is a numerical measure that describes how two variables are linearly related.

- If $\text{Cov}(Y_1, Y_2) > 0$, then Y_1 and Y_2 are positively linearly related.
- If $\text{Cov}(Y_1, Y_2) < 0$, then Y_1 and Y_2 are negatively linearly related.
- If $\text{Cov}(Y_1, Y_2) = 0$, then Y_1 and Y_2 are not linearly related.

RESULT: If Y_1 and Y_2 are independent, then $\text{Cov}(Y_1, Y_2) = 0$.

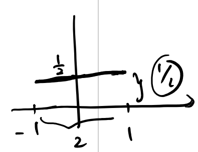
Proof. Suppose that Y_1 and Y_2 are independent. Using the covariance computing formula,

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= E(Y_1Y_2) - E(Y_1)E(Y_2) \\ &= E(Y_1)E(Y_2) - E(Y_1)E(Y_2) = 0. \ \square \end{aligned}$$

IMPORTANT: If two random variables are independent, then they have zero covariance. However, zero covariance does not necessarily imply independence, as we see now.

Example 5.13. *An example of two dependent variables with zero covariance.* Suppose that $Y_1 \sim \mathcal{U}(-1, 1)$, and let $Y_2 = Y_1^2$. It is straightforward to show that

$$\begin{aligned} E(Y_1) &= 0 \\ E(Y_1Y_2) = E(Y_1^3) &= 0 \\ E(Y_2) = E(Y_1^2) = V(Y_1) &= 1/3. \end{aligned}$$



$$E(Y_1^3) = \int_{-1}^{1} y_1^3 \times \frac{1}{2}\, dy_1$$

$$= \frac{1}{2} \times \frac{1}{4} y^4 \Big|_{-1}^{1}$$

$$= \frac{1}{8} - \frac{1}{8} = 0$$

Thus,

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = 0 - 0(1/3) = 0.$$

However, clearly Y_1 and Y_2 are not independent; in fact, they are perfectly related! It is just that the relationship is not linear (it is quadratic). The covariance only measures linear relationships. $\square$

Example 5.14. Gasoline is stocked in a tank once at the beginning of each week and then sold to customers. Let Y_1 denote the proportion of the capacity of the tank that

Thus,

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = 0 - 0(1/3) = 0.$$

However, clearly Y_1 and Y_2 are not independent; in fact, they are perfectly related! It is just that the relationship is not linear (it is quadratic). The covariance only measures linear relationships. □

Example 5.14. Gasoline is stocked in a tank once at the beginning of each week and then sold to customers. Let Y_1 denote the proportion of the capacity of the tank that is available after it is stocked. Let Y_2 denote the proportion of the capacity of the bulk tank that is sold during the week. Suppose that the random vector (Y_1, Y_2) has joint pdf

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Compute $\text{Cov}(Y_1, Y_2)$.

Solution. It is perhaps easiest to use the covariance computing formula

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2).$$

The marginal distribution of Y_1 is beta(3, 1). The marginal distribution of Y_2 is

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, the marginal first moments are

$$\begin{aligned} E(Y_1) &= \frac{3}{3+1} = 0.75 \\ E(Y_2) &= \int_0^1 y_2 \times \frac{3}{2}(1 - y_2^2)dy = 0.375. \end{aligned}$$

Now, we need to compute $E(Y_1Y_2)$. This is given by

$$E(Y_1Y_2) = \int_{y_1=0}^1 \int_{y_2=0}^{y_1} y_1y_2 \times 3y_1 dy_2 dy_1 = 0.30.$$

Thus, the covariance is

$$\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = 0.30 - (0.75)(0.375) = 0.01875. \quad \square$$

$$\begin{aligned} f_{Y_1}(y_1) &= \int f_{Y_1,Y_2}(y_1,y_2)\,dy_2 \\ &= \int_0^{y_1} 3y_1\,dy_2 \\ &= 3y_1 \int_0^{y_1} dy_2 \\ &= 3y_1 \left(y_2 \Big|_0^{y_1}\right) \\ &= 3y_1^2 \quad 0<y_1<1 \\ &= 3y_1^{3-1}(1-y_1)^{1-1} \end{aligned}$$

$$\begin{aligned} f_{Y_2}(y_2) &= \int_{y_2}^1 3y_1\,dy_1 \\ &= \frac{3}{2}y_1^2 \Big|_{y_2}^1 \\ &= \frac{3}{2} - \frac{3}{2}y_2^2, \quad 0<y_2<1 \end{aligned}$$

IMPORTANT: Suppose that Y_1 and Y_2 are random variables (discrete or continuous).

$$\begin{aligned} V(Y_1+Y_2) &= V(Y_1)+V(Y_2)+2\text{Cov}(Y_1,Y_2) \\ V(Y_1-Y_2) &= V(Y_1)+V(Y_2)-2\text{Cov}(Y_1,Y_2). \end{aligned}$$

Proof. Suppose that Y_1 and Y_2 are random variables with means $E(Y_1)=\mu_1$ and $E(Y_2)=\mu_2$, respectively. Let $Z=Y_1+Y_2$. From the definition of variance, we have

$$\begin{aligned} V(Z) &= E[(Z-\mu_Z)^2] \\ &= E\{[(Y_1+Y_2)-E(Y_1+Y_2)]^2\} \\ &= E[(Y_1+Y_2-\mu_1-\mu_2)^2] \\ &= E\{[(Y_1-\mu_1)+(Y_2-\mu_2)]^2\} \\ &= E[(Y_1-\mu_1)^2+(Y_2-\mu_2)^2+2\underbrace{(Y_1-\mu_1)(Y_2-\mu_2)}_{\text{cross product}}] \\ &= E[(Y_1-\mu_1)^2]+E[(Y_2-\mu_2)^2]+2E[(Y_1-\mu_1)(Y_2-\mu_2)] \\ &= V(Y_1)+V(Y_2)+2\text{Cov}(Y_1,Y_2). \end{aligned}$$

That $V(Y_1-Y_2)=V(Y_1)+V(Y_2)-2\text{Cov}(Y_1,Y_2)$ is shown similarly. □

$$\begin{aligned} V(Z) &= E(Z^2)-E(Z)^2 \\ &= E((Y_1+Y_2)^2) - [E(Y_1+Y_2)]^2 \\ &= E(Y_1^2+2Y_1Y_2+Y_2^2) - (\mu_1+\mu_2)^2 \\ &= E(Y_1^2)+2E(Y_1Y_2)+E(Y_2^2) - \mu_1^2 - 2\mu_1\mu_2 - \mu_2^2 \\ &= V(Y_1)+V(Y_2)+2\text{Cov}(Y_1,Y_2) \end{aligned}$$

RESULT: Suppose that Y_1 and Y_2 are **independent** random variables (discrete or continuous).

$$\begin{aligned} V(Y_1+Y_2) &= V(Y_1)+V(Y_2) \\ V(Y_1-Y_2) &= V(Y_1)+V(Y_2). \end{aligned}$$

Proof. In the light of the last result, this is obvious. □

Example 5.15. A small health-food store stocks two different brands of grain. Let Y_1 denote the amount of brand 1 in stock and let Y_2 denote the amount of brand 2 in stock (both Y_1 and Y_2 are measured in 100s of lbs). The joint distribution of Y_1 and Y_2 is

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} 24y_1y_2, & y_1>0,\ y_2>0,\ 0<y_1+y_2<1 \\ 0, & \text{otherwise.} \end{cases}$$

What is the variance for the total amount of grain in stock? That is, find $V(Y_1+Y_2)$.

SOLUTION: We know that

$$V(Y_1 + Y_2) = V(Y_1) + V(Y_2) + 2\mathrm{Cov}(Y_1, Y_2).$$

Marginally, Y_1 and Y_2 are both beta$(2, 3)$; see Example 5.6. Thus,

$$E(Y_1) = E(Y_2) = \frac{2}{2+3} = \frac{2}{5}$$

and

$$V(Y_1) = V(Y_2) = \frac{2(3)}{(2+3+1)(2+3)^2} = \frac{1}{25}.$$

We need to compute $\mathrm{Cov}(Y_1, Y_2)$. Note that

$$E(Y_1Y_2) = \int_{y_1=0}^{1} \int_{y_2=0}^{1-y_1} y_1y_2 \times 24y_1y_2 dy_2 dy_1 = \frac{2}{15}.$$

Thus,

$$\begin{aligned} \mathrm{Cov}(Y_1, Y_2) &= E(Y_1Y_2) - E(Y_1)E(Y_2) \\ &= \frac{2}{15} - \left(\frac{2}{5}\right)\left(\frac{2}{5}\right) \approx -0.027. \end{aligned}$$

Finally,

$$\begin{aligned} V(Y_1 + Y_2) &= V(Y_1) + V(Y_2) + 2\mathrm{Cov}(Y_1, Y_2) \\ &= \frac{1}{25} + \frac{1}{25} + 2(-0.027) \approx 0.027. \ \square \end{aligned}$$

$= 0.08 - 0.054 = 0.026$

RESULTS: Suppose that Y_1 and Y_2 are random variables (discrete or continuous). The covariance function satisfies the following:

(a) $\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(Y_2, Y_1)$

(b) $\mathrm{Cov}(Y_1, Y_1) = V(Y_1)$.

(c) $\mathrm{Cov}(a + bY_1, c + dY_2) = bd\mathrm{Cov}(Y_1, Y_2)$, for any constants a, b, c, and d.

Proof. Exercise. □

$E(Y_1 Y_2)$

$$= \iint_K y_1 y_2 f_{Y_1,Y_2}(y_1,y_2)\, dy_2\, dy_1$$

$$= \int_0^1 \int_0^{1-y_1} 24 y_1^2 y_2^2 \, dy_2\, dy_1$$

$$= \int_0^1 24 y_1^2 \left[\int_0^{1-y_1} y_2^2\, dy_2\right] dy_1$$

$$= \int_0^1 8 y_1^2 \left(y_2^3 \Big|_0^{1-y_1}\right) dy_1$$

$$= \int_0^1 8 y_1^2 (1-y_1)^3\, dy_1$$

$$= 8 \int_0^1 y_1^2 (1-y_1)^3\, dy_1$$

Beta(3,4) Kernel

$$= 8 \times \frac{\Gamma(3)\Gamma(4)}{\Gamma(3+4)}$$

$$= 8 \times \frac{2!\,3!}{6!} = \frac{8 \times 2}{6 \times 5 \times 4}$$

$$= \frac{2}{15}$$

5.8.2 Correlation

GENERAL PROBLEM: Suppose that X and Y are random variables and that we want to predict Y as a linear function of X. That is, we want to consider functions of the form $Y = \beta_0 + \beta_1 X$, for fixed constants β_0 and β_1. In this situation, the "error in prediction" is given by

$$Y - (\beta_0 + \beta_1 X).$$

This error can be positive or negative, so in developing a measure of prediction error, we want one that maintains the magnitude of error but ignores the sign. Thus, we define the **mean squared error of prediction**, given by

$$Q(\beta_0, \beta_1) \equiv E\{[Y - (\beta_0 + \beta_1 X)]^2\}.$$

A two-variable calculus argument shows that the mean squared error of prediction $Q(\beta_0, \beta_1)$ is minimized when

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{}$$

A two-variable calculus argument shows that the mean squared error of prediction $Q(\beta_0, \beta_1)$ is minimized when

$$\beta_1 = \frac{\text{Cov}(X,Y)}{V(X)}$$

and

$$\beta_0 = E(Y) - \left[\frac{\text{Cov}(X,Y)}{V(X)}\right] E(X) = E(Y) - \beta_1 E(X).$$

Note that the value of β_1, algebraically, is equal to

$$\begin{aligned}\beta_1 &= \frac{\text{Cov}(X,Y)}{V(X)} \\ &= \left[\frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}\right] \frac{\sigma_Y}{\sigma_X} \\ &= \rho\left(\frac{\sigma_Y}{\sigma_X}\right),\end{aligned}$$

where

$$\rho = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

The quantity ρ is called the **correlation coefficient** between X and Y.

SUMMARY: The **best linear predictor** of Y, given X, is $Y = \beta_0 + \beta_1 X$, where

$$\begin{aligned}\beta_1 &= \rho\left(\frac{\sigma_Y}{\sigma_X}\right) \\ \beta_0 &= E(Y) - \beta_1 E(X).\end{aligned}$$

NOTES ON THE CORRELATION COEFFICIENT:

(1) $-1 \leq \rho \leq 1$ (this can be proven using the Cauchy-Schwartz Inequality, from calculus).

(2) If $\rho = 1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 > 0$. That is, X and Y are perfectly positively linearly related; i.e., the bivariate probability distribution of (X, Y) lies entirely on a straight line with positive slope.

(3) If $\rho = -1$, then $Y = \beta_0 + \beta_1 X$, where $\beta_1 < 0$. That is, X and Y are perfectly negatively linearly related; i.e., the bivariate probability distribution of (X, Y) lies entirely on a straight line with negative slope.

(4) If $\rho = 0$, then X and Y are not linearly related.

NOTE: If X and Y are independent random variables, then $\rho = 0$. However, again, the implication does not go the other way; that is, if $\rho = 0$, this does not necessarily mean that X and Y are independent.

NOTE: In assessing the strength of the linear relationship between X and Y, the correlation coefficient is often preferred over the covariance since ρ is measured on a bounded, unitless scale. On the other hand, $\text{Cov}(X, Y)$ can be any real number and its units may not even make practical sense.

Example 5.16. In Example 5.14, we considered the bivariate model

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} 3y_1, & 0 < y_2 < y_1 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

for Y_1, the proportion of the capacity of the tank after being stocked, and Y_2, the proportion of the capacity of the tank that is sold. Compute the correlation ρ between Y_1 and Y_2.

SOLUTION: In Example 5.14, we computed $\text{Cov}(Y_1, Y_2) = 0.01875$, so all we need is σ_{Y_1} and σ_{Y_2}, the marginal standard deviations. In Example 5.14, we also found that

$Y_1 \sim \text{beta}(3, 1)$ and

$$f_{Y_2}(y_2) = \begin{cases} \frac{3}{2}(1 - y_2^2), & 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The variance of Y_1 is

$$V(Y_1) = \frac{3(1)}{(3+1+1)(3+1)^2} = \frac{3}{80} \implies \sigma_{Y_1} = \sqrt{\frac{3}{80}} \approx 0.194.$$

Simple calculations using $f_{Y_2}(y_2)$ show that $E(Y_2^2) = 1/5$ and $E(Y_2) = 3/8$ so that

$$V(Y_2) = \frac{1}{5} - \left(\frac{3}{8}\right)^2 = 0.059 \implies \sigma_{Y_2} = \sqrt{0.059} \approx 0.244.$$

Finally, the correlation is

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_{Y_1}\sigma_{Y_2}} \approx \frac{0.01875}{(0.194)(0.244)} \approx 0.40. \quad \square$$

5.9 Expectations and variances of linear functions of random

$$\text{Cov}(Y_1, Y_2) = \underset{1.}{E(Y_1 Y_2)} - \underset{2.}{\underbrace{E(Y_1)}_{f_{Y_1}(y_1)}}\underset{3.}{\underbrace{E(Y_2)}_{f_{Y_2}(y_2)}}$$

$$\rho = \frac{\text{Cov}(Y_1, Y_2)}{\sigma_{Y_1}\sigma_{Y_2}}$$

$$\sigma_{Y_1} = \sqrt{V(Y_1)} \quad 4. \qquad \sigma_{Y_2} = \sqrt{V(Y_2)} \quad 5.$$

$\sigma_{Y_1}\sigma_{Y_2}$ (0.194)(0.244)

5.9 Expectations and variances of linear functions of random variables

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ are random variables and that $a_1, a_2, ..., a_n$ are constants. The function

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$$

is called a **linear combination** of the random variables $Y_1, Y_2, ..., Y_n$.

EXPECTED VALUE OF A LINEAR COMBINATION:

$$E(U) = E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i)$$

VARIANCE OF A LINEAR COMBINATION:

$$\begin{aligned} V(U) = V\left(\sum_{i=1}^{n} a_i Y_i\right) &= \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum_{i<j} a_i a_j \text{Cov}(Y_i, Y_j) \\ &= \sum_{i=1}^{n} a_i^2 V(Y_i) + \sum_{i\neq j} a_i a_j \text{Cov}(Y_i, Y_j) \end{aligned}$$

$\uparrow$ 4. $\uparrow$ 5

$f_{Y_1}(y_1)$ $f_{Y_2}(y_2)$