STAT 516

Covariance and Correlation

Prof. Michael Levine

April 19, 2020

Motivation

- For non-independent variables X and Y $Var(X+Y) \neq Var(X) + Var(Y)$
- Instead, we have

$$
\begin{aligned}
&Var(X+Y) = E(X+Y)^2 - [E(X+Y)]^2 = \cdots \\
&= Var(X) + Var(Y) + 2[E(XY) - E(X)E(Y)]
\end{aligned}
$$

- Let X and Y be two random variables defined on the common sample space Ω
- It is assumed that $E(X)$, $E(Y)$ and $E(XY)$ all exist
- The covariance of X and Y is defined as

$$Cov(X,Y) = E(XY) - E(X)E(Y) = E[(X - E(X))(Y - E(Y))]$$

Interpretation

- Covariance is a measure of whether two random variables X and Y tend to increase or decrease together
- For example, taller people tend to weigh more than shorter people; thus, height and weight usually have a positive covariance
- Covariance is scale dependent and can take arbitrary positive and negative values
- Renormalization is necessary to make it easier to interpret

Correlation

- Let X and Y be two random variables defined on a common sample space Ω such that $Var(X)$ and $Var(Y)$ are finite
- The correlation between X and Y is defined to be

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

Properties of Covariance and Correlation

1. $Cov(X, c) = 0$ for any X and any constant c
2. $Cov(X, X) = Var(X)$ for any X
3. $$Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j)$$

   In particular,
   $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$ and
   $Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + 2\sum_{i<j} Cov(X_i, X_j)$

Properties of Covariance and Correlation

1. For any two independent random variables X and Y, $Cov(X, Y) = \rho_{X,Y} = 0$
2. $\rho_{a+bX,c+dY} = sgn(bd)\rho_{X,Y}$
3. Whenever $\rho_{X,Y}$ is defined, $-1 \leq \rho_{X,Y} \leq 1$
4. $\rho_{X,Y} = 1$ if and only if for some a and $b > 0$ $P(Y = a + bX) = 1$
5. $\rho_{X,Y} = -1$ if and only if for some a and $b < 0$ $P(Y = a + bX) = 1$

Some proofs

- E.g. $Cov(X, c) = E(cX) - E(c)E(X) = c(EX) - c(EX) = 0$
- $Cov(X, X) = E(X^2) - [E(X)]^2 = Var(X)$
- E.g. the property (1) follows since $E(XY) = E(X)E(Y)$ if X and Y are independent

Example I: correlation between min and max in dice rolls

- A fair die is rolled twice; X is the min and Y is the max of two rolls. What is $\rho_{X,Y}$?
- From the joint distribution of X and Y obtained earlier,

$$E(XY) = 1/36 + 2/18 + 4/36 + \cdots = 49/4$$

- From the marginal pmfs of X and Y, $E(X) = 161/36$, $E(Y) = 91/36$, $Var(X) = Var(Y) = 2555/1296$
- Thus, $\rho_{X,Y} = \frac{E(XY)-E(X)E(Y)}{\sqrt{Var(X)Var(Y)}} = 0.48$
- One can show mathematically that this correlation is positive for any number of rolls of a die
- However, it will converge to zero as the number of rolls tends to infinity

Example II: Correlation does not mean Independence

- If $Cov(X, Y) = 0$ X and Y are not necessarily independent
- Take X such that $P(X = \pm 1) = p$, $P(X = 0) = 1 - 2p$ for some $0 < p < \frac{1}{2}$ and define $Y = X^2$
- Note that $E(XY) = E(X^3) = 0$; also, since $E(X) = 0$ we have $E(X)E(Y) = 0$
- Therefore, $Cov(X, Y) = 0$; however, X and Y are not independent
- Indeed, note that $P(Y = 0|X = 0) = 1$ but $P(Y = 0) = 1 - 2p \neq 0$
- More generally, if X has a distribution symmetric around zero and has three finite moments, then X and X^2 always have a zero correlation while not being independent