Discussion 1 for PSTAT 174/274

Topic: Review of Covariance and Correlation Functions

Time series analysis relies strongly on use of covariance and correlation functions.

For random vector (X, Y),

• $Cov(X,Y) \stackrel{def}{=} E[(X-EX)(Y-EY)] \equiv E(XY) - E(X)E(Y) \equiv Cov(Y,X)$.

•$FACT$: When X and Y are independent,
$Cov(X,Y) = E[(X-EX)(Y-EY)] = E[(X-EX)] \cdot E[(Y-EY)] = 0$
(independence implies that expectation of a product is product of expectations)

We often work with sums. While expectation of a sum always equals to the sum of expectations, in general, it is not true for variances. On Lecture 1, we showed:

•$FACT$: $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$.

Extending to sum of n variables: $Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n VarX_i + 2\sum_{i<j} Cov(X_i, X_j)$.

We defined

• X and Y are uncorrelated if $Cov(X,Y) = 0$.

Thus, we immediately get:

•$FACT$: All independent variables are uncorrelated. In general, the opposite is not true.

• Example to illustrate that it is possible to have X and Y dependent but uncorrelated:

Let $X \sim U(-1,1)$ be uniform on $(-1,1)$, i.e., $f_X(x) = 1/2$ when $-1 < x < 1$ and zero otherwise. Let $Y = X^2$. Clearly, X and Y are dependent. We show that X and Y are uncorrelated:

$Cov(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - 0 \cdot EY = \int_{-1}^1 (x^3)(1/2)dx = 0.$

Here dependence between X and Y is non-linear, quadratic. Note that the same result will hold for any symmetric bounded r.v. X)

• Discussion of covariance in case of linear relationship between X and Y:

Let $Y = aX + b$, linear dependence, calculate covariance:

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y) = E[X(aX+b)] - E(X) \cdot E(aX+b)] \\ &= aE(X^2) + bE(X) - E(X)[aE(X)+b] = aE(X^2) + bE(X) - a[E(X)]^2 + bE(X) \\ &= aVar(X) \end{aligned}$$

Conclude:
• When X and Y are linearly dependent, i.e., $Y = aX + b$, $Cov(X,Y) \neq 0$.
• When X and Y are linearly dependent, i.e., $Y = aX + b$, and slope of the line $a > 0$,

i.e., X and Y increase together, $Cov(X,Y) = aVar(X) > 0$
• When X and Y are linearly dependent, i.e., $Y = aX + b$, and slope of the line $a < 0$,
i.e., when X increases, Y decreases, $Cov(X,Y) = aVar(X) < 0$.

Thus, covariance is used as a rough guide to mutual dependence, in particular linear dependence. Because covariance can take values from $-\infty$ to $+\infty$, typically, one normalizes it getting a tool called correlation.

Definition The correlation:

$$\rho(X,Y) \equiv \rho_{XY} \equiv Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

• Properties:

- X and Y uncorrelated implies $\rho(X,Y) = 0$.
- $|\rho(X,Y)| \leq 1$ (Cauchy-Schwartz inequality).
- $\rho(aX + b, cY + d) = sign(ac)\rho(X,Y)$, i.e., correlation is essentially unchanged under the change of location and scale which is not true of covariance:
$Cov(aX + b, cY + d) = (ac)Cov(X,Y)$.
- $\rho(X, cX + d) = sign(c)\rho(X,X) = sign(c)$. Thus, if Y is a linear function of X, the absolute value of the $Cor(X,Y)$ achieves its maximum value of 1.

What does the sign of Cor tell us about dependency between variables? Some 5-a type examples:

(a) Let X be height and Y be weight of an individual.
Questions: Is $Cor(X,Y)$ positive or negative? what would you estimate its value to be? If one takes a sample of iid observations $(x_1, y_1), \ldots, (x_n, y_n)$, how would a scatter diagram look like?

Hint: typically taller individuals weigh more, so that weights and heights increase together, i.e., $Cor(X,Y) > 0$, but dependence is not perfectly linear. Expect $\rho(X,Y) \approx 0.6$.

(b) Let X be the age of a used car, Y its price.
Questions: Is $Cor(X,Y)$ positive or negative? what would you estimate its value to be? If one takes a sample of iid observations $(x_1, y_1), \ldots, (x_n, y_n)$, how would a scatter diagram look like?

Hint: slope is negative, $\rho \approx -0.9 < 0$.

(c) Let X be an age of a child and Y number of toys a corresponding child receives for her birthday. Is the correlation coefficient be positive or negative? If Z is the weight of this child, would Cor(X, Z) and Cor(Y, Z) be positive or negative? Drew scatter diagrams.

• Estimation of the correlation coefficient from iid sample $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\hat{\rho}_{X,Y} = \Big(\frac{1}{n-1}\Big)\frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{s_X^2 s_Y^2}},\ s_X^2 \equiv \hat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x})^2,\ \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i.$$

• Notes: in regression problems, observations in a sample is i.i.d. and the order of summation is unimportant.

– If X and Y are independent $\rho = 0$, scatter diagram looks like a ball
– If X and Y are linearly dependent, points lie on a line.
– If $|\rho| \approx 1$, strong linear relationship.

• Questions: when do we use capital letters X, Y and when lower case letters, e.g., x_i as in the formula for sample correlation?

. . . . . .

• Review from PSTAT 120AB:

Question 1: What result from PSTAT 120AB determines that your sample mean is close to the true mean when the sample size is large?

Question 2: What result from PSTAT 120AB allows us to determine the size of the estimation error?

In Time Series: the observations are dependent so 120AB results do not apply directly, one needs ergodic theory.

. . . . . .

Exercise: Calculate covariance for $X_1 = Z_1$ and $X_2 = Z_1 + Z_2$, where Z_i are iid with mean 0 and variance 1.

. . . . . .

Review of Facts about bivariate normal distribution.(Based on §A.3 of [BD])

IMPORTANT FACT: while in general, $Cor(X, Y) = 0$ does NOT imply independence, in case of bivariate Gaussian distribution $Cor(X, Y) = 0$ happens if and only if X and Y are independent. This makes Gaussian r.v.s special.

The fact follows from the structure of bivariate normal pdf as below.

(i) Continuous r.v.'s X and Y have a bivariate normal distribution with the parameters $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and ρ $(-1 \le \rho \le 1, \sigma_i^2 \ge 0)$ if their joint p.d.f. is

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\Big\{-\frac{1}{2(1-\rho^2)}\Big[\Big(\frac{x-\mu_1}{\sigma_1}\Big)^2 - 2\rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big(\frac{y-\mu_2}{\sigma_2}\Big) + \Big(\frac{y-\mu_2}{\sigma_2}\Big)^2\Big]\Big\}$$

(ii) Facts:
(a) the marginal distributions of X and Y are $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ respectively;
(b) the $\mathrm{Cor}(X, Y) = \rho$.
(c) X and Y are independent iff $\rho = 0$.

(d) the conditional distribution of Y given $X = x$, is $N(\mu_2 + \rho(\sigma_2/\sigma_1)(x - \mu_1), (1 - \rho^2)\sigma_2^2)$
(e) the conditional distribution of X given $Y = y$, is $N(\mu_1 + \rho(\sigma_1/\sigma_2)(y - \mu_2), (1 - \rho^2)\sigma_1^2)$

. . . . . .