STA 623 – Fall 2010 – Dr. Charnigo

Section 4.5: Covariance and Correlation

Covariance. Suppose X has mean $\mu_X \in (-\infty, \infty)$ and variance $\sigma_X^2 \in (0, \infty)$ and that Y has mean $\mu_Y \in (-\infty, \infty)$ and variance $\sigma_Y^2 \in (0, \infty)$. We define the covariance of X and Y as

$$Cov[X, Y] := E[(X - \mu_X)(Y - \mu_Y)],$$

which may be written as $E[XY] - \mu_X \mu_Y$ by appealing to linearity of expectation.

The covariance is positive when X and Y tend to be larger than average or smaller than average at the same time, while the covariance is negative when X tends to be larger than average when Y is smaller than average and vice versa.

Explorations of covariance. 1. We may ask whether the covariance in fact exists as a finite number, under the conditions above. Let us prove that it does. Put

$$U := X - \mu_X \quad \text{and} \quad V := Y - \mu_Y,$$

so that $E[U] = E[V] = 0$ and $Cov[X, Y] = E[UV]$. Since $(u - v)^2 \geq 0$ for all real u and v, we have $u^2 + v^2 \geq 2uv$. Also, since $(u + v)^2 \geq 0$, we have $u^2 + v^2 \geq -2uv$. Thus $0 \leq |uv| \leq (u^2 + v^2)/2$, and appealing to monotonicity of expectation we have

$$0 = E[0] \leq E[|UV|] \leq E[(U^2 + V^2)/2] = (\sigma_X^2 + \sigma_Y^2)/2 < \infty.$$

Since $E[|UV|]$ is finite, the existence of $E[UV] = Cov[X, Y]$ is ensured.

2. What is the covariance of X with itself?

3. What is the covariance of X and Y if they are independent?

4. Let X have the standard normal distribution (so $\mu_X = 0$ and $\sigma_X^2 = 1$), and put $Y := X^2$. I claim that Y has the chi-square distribution on 1 degree of freedom (so $\mu_Y = 1$ and $\sigma_Y^2 = 2$). To verify my claim, note that Y has cumulative distribution function

$$P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = 2 \int_0^{\sqrt{y}} (2\pi)^{-1/2} \exp[-x^2/2] \, dx \quad \text{for } y \geq 0,$$

so that we may take

$$\frac{d}{dy}P(Y \leq y) = 2(2\pi)^{-1/2}\exp[-y/2]\frac{d}{dy}\sqrt{y} = \frac{1}{\Gamma[1/2]2^{1/2}}y^{-1/2}\exp[-y/2] \quad \text{for } y > 0$$

as probability density function for Y. Clearly, X and Y are not independent. In particular,

$$P(X > 1, Y > 1) = P(X > 1) \neq P(X > 1)P(Y > 1).$$

(Can you explain the two steps above?) On the other hand, we have

$$Cov[X, Y] =$$

5. As the previous item suggests, when people say that covariance describes the association between X and Y, they speak quite loosely. Really the covariance only describes the linear association between X and Y. Nonetheless, covariance is useful. For instance, covariance enables us to calculate variances for sums of random variables. Indeed, for any real constants a and b we have

$$Var[aX + bY] = a^2Var[X] + b^2Var[Y] + 2abCov[X, Y]$$

by linearity of expectation.

Correlation. Suppose X has mean $\mu_X \in (-\infty, \infty)$ and variance $\sigma_X^2 \in (0, \infty)$ and that Y has mean $\mu_Y \in (-\infty, \infty)$ and variance $\sigma_Y^2 \in (0, \infty)$. We define the correlation of X and Y as

$$Corr[X, Y] := \frac{Cov[X, Y]}{\sigma_X \sigma_Y}.$$

Sometimes we use the symbol $\rho_{X,Y}$ to represent correlation (or, if no confusion is possible, we write ρ without any subscript).

Correlation has the advantage of being constrained to lie between -1 and 1, a fact established by your textbook authors using a calculus argument. (Your textbook authors also show that a correlation of ± 1 implies that $Y = a + bX$

for some real constants a and b with probability one.) Hence a correlation of, say, 0.9 may always be regarded as indicative of a strong linear relationship whereas a covariance of (say) 90 may or may not be indicative of a strong linear relationship.

Bivariate normal distribution. Correlation plays a key role in characterizing a bivariate normal distribution. We say that X and Y have a bivariate normal distribution if their joint probability density function is

$$f_{X,Y}(x, y) = (2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2})^{-1}\times$$

$$\exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\}\right],$$

where $\mu_X, \mu_Y \in (-\infty, \infty)$, $\sigma_X, \sigma_Y \in (0, \infty)$, and $\rho \in (-1, 1)$.

All five parameters have the anticipated interpretations: μ_X and σ_X are the mean and standard deviation of the marginal distribution of X, which is univariate normal; μ_Y and σ_Y are the mean and standard deviation of the marginal distribution of Y; and, ρ is the correlation of X and Y. Thus, when speaking of a bivariate normal distribution, zero correlation and independence are equivalent.

Also, the conditional distribution of Y given that $X = x$ turns out to be normal with mean $\mu_Y + \rho(\sigma_Y/\sigma_X)(x - \mu_X) =: \mu_{Y|X}$ and variance $\sigma_Y^2(1-\rho^2) =: \sigma_{Y|X}^2$. This shows that simple linear regression, with which you are already familiar from STA 602, is like estimating the parameters of a bivariate normal distribution. Noting that the "slope" is $\rho(\sigma_Y/\sigma_X) = Cov[X, Y]/Var[X]$, we can reason out a formula for the slope estimate:

Moreover, if we take the Residual Mean Square as an estimate of $\sigma_{Y|X}^2$ and the sample variance of Y as an estimate of σ_Y^2, an estimate of ρ^2 is suggested: