Econometrics: Covariance and Correlation

Ryan Safner

Fall 2017

1 Variance

Recall the variance of a random variable X, denoted $var(X)$ or σ^2, is the expected value (probability-weighted average) of the squared deviations of X_i from it's mean (or expected value) $\bar{X}$ or $E(X)$.[1]

$$\begin{aligned}\sigma_X^2 &= E(X - E(X)) \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 p_i\end{aligned}$$

Note if all possible values of X_i are equally likely (or we don't know the probabilities), we can write variance as a simple average of squared deviations from the mean:

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Variance has some useful properties:

1. **The variance of a constant is 0**

   $$var(c) = 0 \text{ iff } P(X = c) = 1$$

   If a random variable takes the same value (e.g. 2) with probability 1.00, E(2)=2, so the average squared deviation from the mean is 0, because there are never any values other than 2.

2. **The variance is unchanged for a random variable plus/minus a constant**

   $$var(X \pm c)$$

   Since the variance of a constant is 0.

3. **The variance of a scaled random variable is scaled by the square of the coefficient**

   $$var(aX) = a^2 var(X)$$

4. **The variance of a linear transformation of a random variable is scaled by the square of the coefficient**

   $$var(aX + b) = a^2 var(X)$$

[1] Note there will be a different in notation depending on whether we refer to a population (e.g. μ_X) or to a sample (e.g. $\bar{X}$). As the overwhelming majority of cases we will deal with samples, I will use sample notation for means).

2 Covariance

For two random variables, X and Y, we can measure their **covariance** (denoted $cov(X, Y)$ or $\sigma_{X,Y}$)[2] to quantify how they vary *together*. A good way to think about this is: when X is above its mean, would we expect Y to also be above its mean (and covary positively), or below its mean (and covary negatively). Remember, this is describing the *joint* probability distribution for two random variables.

$$\sigma_{X,Y} = E\big[(X - \bar{X})(Y - \bar{Y})\big]$$

Again, in the case of equally probable values for both X and Y, covariance is sometimes written:

$$\sigma_{X,Y} = \frac{1}{N}\sum_{i=1}^{n}(X - \bar{X})(Y - \bar{Y})$$

Covariance also has a number of useful properties:

1. **The covariance of a random variable X and a constant c is 0**

$$cov(X, c) = 0$$

2. **The covariance of a random variable and itself is the variable's variance**

$$cov(X, X) = var(X)$$
$$\sigma_{X,X} = \sigma_X^2$$

3. **The covariance of a two random variables X and Y each scaled by a constant a and b is the product of the covariance and the constants**

$$cov(aX, bY) = a \times b \times cov(X, Y)$$

4. **If two random variables are independent, their covariance is 0**

$$cov(X, Y) = 0 \text{ iff } X \text{ and } Y \text{ are independent:} E(XY) = E(X) \times E(Y)$$

[2] Again, to be technically correct, $\sigma_{X,Y}$ refers to populations, $s_{X,Y}$ refers to samples, in line with population vs. sample variance and standard deviation. Recall also that sample estimates of variance and standard deviation divide by $n - 1$, rather than n. In large sample sizes, this difference is negligible.

3 Correlation

Covariance, like variance, is often cumbersome, and the numerical value of the covariance of two random variables does not really mean much. It is often convenient to normalize the covariance to a decimal between -1 and 1. We do this by dividing by the product of the standard deviations of X and Y. This is known as the **correlation coefficient** between X and Y, denoted $corr(X,Y)$ or $\rho_{X,Y}$ (for populations) or $r_{X,Y}$ (for samples):

$$\begin{aligned} r_{X,Y} &= \frac{cov(X,Y)}{sd(X)sd(Y)} \\ &= \frac{E\big[(X-\bar{X})(Y-\bar{Y})\big]}{\sqrt{E\big[X-\bar{X}\big]}\sqrt{E\big[Y-\bar{Y}\big]}} \\ &= \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \end{aligned}$$

Note this also means that covariance is the product of the standard deviation of X and Y and their correlation coefficient:

$$\begin{aligned} \sigma_{X,Y} &= r_{X,Y}\sigma_X\sigma_Y \\ cov(X,Y) &= corr(X,Y) \times sd(X) \times sd(Y) \end{aligned}$$

Another way to reach the (sample) correlation coefficient is by finding the average joint Z-score of each pair of (X_i, Y_i):

$$\begin{aligned} r_{X,Y} &= \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}))}{s_X s_Y} && \text{Definition of sample correlation} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right) && \text{Breaking into separate sums} \\ &= \frac{1}{n}\sum_{i=1}^{n}(Z_X)(Z_Y) && \text{Recognize each sum is the z-score for that r.v.} \end{aligned}$$

Correlation has some useful properties that should be familiar to you:

1. Correlation is between -1 and 1
2. A correlation of -1 is a downward sloping straight line
3. A correlation of 1 is an upward sloping straight line
4. A correlation of 0 implies no relationship

Note another way of relationg the