13.3.2 Autocorrelation test

Suppose, we generate an array of random numbers from a distribution. The sequence of numbers can be thought of as a *time series*. The auto-correlation function (ACF) of a time series is the correlation of the variable with itself at differing time lags.

Let us consider a set of N numbers: $x_1,\ x_2,\ x_3,\ x_4,\ \ldots\ x_N$. The *variance* of the numbers:

$$\sigma_x{}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2, \tag{1}$$

where the mean, $\bar{x} = \frac{1}{N}\sum x_i$. We can rewrite the formula (1) in the following form:

$$\sigma_{xx} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})\,(x_i - \bar{x}) \tag{2}$$

For two sets of numbers (x, y), we calculate a similar quantity as follows.

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})\,(y_i - \bar{y}). \tag{3}$$

Formula (3) is called *covariance*. With these we can calculate the *correlation coefficient* between two sets of observations: $x_1,\ x_2,\ x_3,\ \dots\ x_N$ and $y_1,\ y_2,\ y_3,\ \dots\ y_N$:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{4}$$

Now we can calculate a similar quantity between a pair of observations taken from the same set. For example, let us make two sets: $(x_1,\ x_2,\ x_3,\ \dots\ x_{N-1})$ and $(x_2,\ x_3,\ x_4,\ \dots\ x_N)$, each of $(N-1)$ data points. Define a quantity,

$$r_1 = \frac{\sum_{i=1}^{N-1} (x_i - \bar{x}^{(1)})(x_{i+1} - \bar{x}^{(2)})}{\sqrt{\sum_{i=1}^{N-1} (x_i - \bar{x}^{(1)})^2 \sum_{i=1}^{N-1} (x_{i+1} - \bar{x}^{(2)})^2}}. \tag{5}$$

This is called *auto-correlation function*. In this case, the symbol 'r_1' is used to denote that the two sets of data are separated by interval or lag 1. Similarly, we can find out r_2, r_3 and so on. Thus, the auto-correlation is the correlation between successive observations with some gap.

For a sufficiently large set of numbers, we can write, $\bar{x}^{(1)} = \bar{x}^{(2)} = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ [Note: $\bar{x}^{(1)} = \frac{1}{N-1}\sum_{i=1}^{N-1} x_i$ *etc.*]. The formula (5) is usually approximated as

$$r_1 = \frac{\sum_{i=1}^{N-1} (x_i - \bar{x})(x_{i+1} - \bar{x})}{\frac{(N-1)}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{6}$$

This is again simplified for large N,

$$r_1 = \frac{\sum_{i=1}^{N-1} (x_i - \bar{x})(x_{i+1} - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{7}$$

In general, the *auto-correlation coefficient* at lag or gap k,

$$r_k = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \tag{8}$$

Usually, we write

$$c_k = \frac{1}{N}\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x}) \tag{9}$$

as *auto-covariance* at lag k. When $k = 0$ (zero lag), the above function becomes the *covariance* of the time series. Some people, write the formula (9) as $c_k = \frac{1}{N-k}\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x})$. The *auto-correlation function* (ACF) is thus defined as,

$$r_k = \frac{c_k}{c_0} \tag{10}$$

Formula (9) and (10) are generally used to compute *autocorrelation* of a given time series.

$$c_k = \frac{1}{N}\sum_{i=1}^{N-k}(x_i - \bar{x})(x_{i+k} - \bar{x})$$

$$r_k = \frac{c_k}{c_0}$$