**A9Wf Autocorrelation**

Autocorrelations and the autocorrelation function are two elementary tools used in several more advanced statistical techniques. If you ever decide to model the strength of two variables or try to find the best line that will fit the variable, in order to predict its future values, you will need to use autocorrelations. This section will just introduce the concept and provide the necessary foundation to be able to understand some of these more advanced techniques. We will cover these techniques in subsequent chapters.

### Lagged variables

Before we explain the concept of autocorrelation and how it is calculated, we need to understand the concept of lagging, or lagged variables. For ordinary correlation coefficient, we said that we measured how closely any two variables are related. What if we try to measure the relationship of one variable, but with itself? How do we do this?

Imagine a variable, for example several UK visits abroad, which we already included in all the examples in this chapter. However, for the simplicity sake, we will only look at 15 observations between April 2017 and June 2018. Let’s now “lag” this variable by dropping the first observation and creating a new “variable”, then by dropping the first two observations and creating the second artificial variable, etc. Figure 1 shows the example.

**Example 1**



Figure 1 An example of lagging the variable

**Excel solution**

Date = Cell B3:B17 Value

Variable = Cell C3:C17 Value

Lag 1 = Cell D3 Formula:=C4

 Cell D4 Formula:=C5

 Copy to D5:D17

Lag 2 = Cell E3 Formula:=C5

 Cell E4 Formula:=C6

 Copy to E5:E17

Lag 3 = Cell F3 Formula:=C6

 Cell F4 Formula:=C7

 Copy to F5:F17, etc.

How many lagged variables does it make sense to produce will depend on application. We will see shortly that for the purposes of this application it does not make sense to produce more than n/3 lagged variables, where n is the number of observations in the data set.

Now that we have lagged the original variable, we can ask ourselves if we could calculate the correlation between this variable and its lagged values? Yes, we can, and it is called the autocorrelation. In fact, we can calculate autocorrelations between the original time series and the lag one, lag two, etc. The series of these autocorrelation coefficients will form the autocorrelation function (ACF).

### The autocorrelation function (ACF)

The equation for the autocorrelation function is given by equation (1):

$r\_{k}=\frac{\sum\_{t=1}^{n-k}(y\_{t}-\overbar{y})(y\_{t+k}-\overbar{y})}{\sum\_{t=1}^{n}(y\_{t}-\overbar{y})^{2}}$ (1)

Where rk are the autocorrelation coefficients, n is the number of observations in the variable, k is the number of lags, yt are individual observations at the point t and $\overbar{y}$ is the mean value. Note that the series of autocorrelation coefficients rk constitute the autocorrelation function (ACF).

In Example 1, we have only 14 observations. This means that the maximum that makes sense to calculate is 4 (14/3 ≈ 4). However, just to illustrate the point how the autocorrelation coefficients are calculated, we will calculate 10 of them.

To calculate the autocorrelation coefficient for the first lag, equation (1) is implemented as follows:

$r\_{1}=\frac{\sum\_{t=1}^{14-1}(y\_{t}-\overbar{y})(y\_{t+1}-\overbar{y})}{\sum\_{t=1}^{14}(y\_{t}-\overbar{y})^{2}} $= $\frac{(y\_{1}-\overbar{y})(y\_{2}-\overbar{y})+(y\_{2}-\overbar{y})(y\_{3}-\overbar{y})+…+(y\_{13}-\overbar{y})(y\_{14}-\overbar{y})}{(y\_{1}-\overbar{y})^{2}+(y\_{2}-\overbar{y})^{2}+…+(y\_{14}-\overbar{y})^{2}}$

For the lag 2, the calculation is:

$$r\_{2}=\frac{\sum\_{t=1}^{14-2}(y\_{t}-\overbar{y})(y\_{t+2}-\overbar{y})}{\sum\_{t=1}^{14}(y\_{t}-\overbar{y})^{2}}=\frac{(y\_{1}-\overbar{y})(y\_{3}-\overbar{y})+(y\_{2}-\overbar{y})(y\_{4}-\overbar{y})+…+(y\_{12}-\overbar{y})(y\_{14}-\overbar{y})}{(y\_{1}-\overbar{y})^{2}+(y\_{2}-\overbar{y})^{2}+…+(y\_{14}-\overbar{y})^{2}}$$

For the lag 3, the calculation is:

$$r\_{3}=\frac{\sum\_{t=1}^{14-3}(y\_{t}-\overbar{y})(y\_{t+3}-\overbar{y})}{\sum\_{t=1}^{14}(y\_{t}-\overbar{y})^{2}}=\frac{(y\_{1}-\overbar{y})(y\_{4}-\overbar{y})+(y\_{2}-\overbar{y})(y\_{5}-\overbar{y})+…+(y\_{11}-\overbar{y})(y\_{14}-\overbar{y})}{(y\_{1}-\overbar{y})^{2}+(y\_{2}-\overbar{y})^{2}+…+(y\_{14}-\overbar{y})^{2}}$$

etc

For the lag 10, the calculation is:

$$r\_{10}=\frac{\sum\_{t=1}^{14-10}(y\_{t}-\overbar{y})(y\_{t+10}-\overbar{y})}{\sum\_{t=1}^{14}(y\_{t}-\overbar{y})^{2}}=\frac{(y\_{1}-\overbar{y})(y\_{11}-\overbar{y})+(y\_{2}-\overbar{y})(y\_{12}-\overbar{y})+(y\_{3}-\overbar{y})(y\_{13}-\overbar{y})+(y\_{4}-\overbar{y})(y\_{14}-\overbar{y})}{(y\_{1}-\overbar{y})^{2}+(y\_{2}-\overbar{y})^{2}+…+(y\_{14}-\overbar{y})^{2}}$$

**Example 3**

We will use the same variable as in Example 1 above. We will show how to calculate autocorrelation function in Excel. We are using two different methods, each using different Excel functions, to show how to calculate autocorrelations as illustrated in Figures 2 and 3. Although, as per the convention, the number of autocorrelations should not be more than n/3 (which is 4 in our case), we will calculate 10 of them, just to better illustrate how the formulae work.



Figure 2 Method 1 to calculate the autocorrelation function in Excel



Figure 3 Method 2 to calculate the autocorrelation function in Excel

**Excel solution**

Date = Cells B3:B17 Values

Variable = Cells C3:C17 Values

Deviations from

the mean = Cells D3:D17 Formula:=C3-AVERAGE($C$3:$C$17)

Copy down to D4:D17

ACF, Method 1 =

Cells E3 Formula:= =SUMPRODUCT($D$3:D16,$D4:D$17)/DEVSQ($D$3:$D$17))

Cells E4 Formula:= (SUMPRODUCT($D$3:D15,$D5:D$17)/DEVSQ($D$3:$D$17))

Cells E5 Formula:= (SUMPRODUCT($D$3:D14,$D6:D$17)/DEVSQ($D$3:$D$17))

Cells E12 Formula:= (SUMPRODUCT($D$3:D7,$D13:D$17)/DEVSQ($D$3:$D$17))

ACF, Method 2 =

Cell G3 Formula:

=SUMPRODUCT($D$3:INDEX($D$3:$D$17,ROWS(D4:D$17)),$D4:D$17)/DEVSQ($D$3:$D$1

Copy down G4:G12

Equation (1) sums the product of two expressions: ($y\_{t}-\overbar{y})(y\_{t+k}-\overbar{y})$, which is why we used Excel =SUMPRODUCT() function. The denominator in equation (1) is the sum of squared deviations from the mean: $(y\_{t}-\overbar{y})^{2}$, which is the reason we used Excel function =DEVSQ(). This simplifies our calculations in the spreadsheet.

The values in column E are calculated in this manner, but there is one inconvenient operation. The difficulty with the formulae in column E is that the value of one of the references goes backward, i.e. in cell E3 we have in the formula D15, in cell E4 we have D14, in E5 we have D13, etc. This makes Copy/Paste function invalid as it is impossible to copy the top cell and paste the values in descending order (i.e. if you copy E3 to E4, then D15 becomes D14, in E5 it is D13, etc.). In order to “fix” this, we had to manually change all the ascending references to column D into the descending references. Can be time consuming if you have a long time series.

For this reason, we are showing another way to calculate the autocorrelations, and column G contains the same set of values, but calculated using the additional two Excel functions, called =INDEX() and =ROWS(). We will not provide detailed explanation how these functions work. You can find it on the web. It suffices to say that they are easier to implement, and they return identical values as the first method. It is a convention to present the autocorrelation function (ACF) not as a line, but as the series of bars. Figure 4 shows the results of our calculations from Example 2.



Figure 4 Autocorrelations for the UK visits abroad from Jan 2016 and Feb 2017

How do we interpret the autocorrelation coefficients? Just like the ordinary correlation, individual coefficients show us which historical lag is highly correlated with the current value of the variable. This means that in the case of periodicity, or seasonality, we will see the spikes in autocorrelation coefficient for the periods when the periodicity or seasonality takes place. The Example 8.7 shows artificially short data set that we used just to explain the principles.

However, Figure 5 shows the autocorrelation coefficients for the full data set from Example 1, i.e. just the UK visits abroad time series. This gradual drop of the autocorrelation coefficients indicates that the data set has a “memory”. In other words, the correlation between the current data set and the data set lagged by one time period is high, but the correlation between the current data set and the data set lagged by two time periods is smaller. The further we go in the past, the autocorrelations get smaller and eventually drop towards zero, although the drop is gradual. This indicates the “memory”, i.e. today’s data is influenced by yesterday’s, the yesterday’s by the day before yesterday, etc. In later chapters we will use a different expression (non-stationarity).



Figure 5 ACF for UK visits abroad data set

Figure 6 shows a typical autocorrelation function for a seasonal data set. The seasonality is clearly indicated in this artificially created time series. This will be the topic for Chapter 12.



Figure 6 ACF for a periodic data set

The ACF will be instrumental in deciding what kind of data set (time series) are we handling and how.

Because the autocorrelations show the correlation of a single variable with its own historical values (lagged values), you can also think about the autocorrelations as a method that will show you the “memory” of the system. In other words, if we have a spike in the autocorrelation value at the lag 6, for example, this means that the current behaviour of the variable is highly correlated to the behaviour that took place six periods ago. In a way, the variable has memorized what happened before and it is repeating the pattern.

Autocorrelation function is a tool that many more advanced statistical methods rely upon, either as a source of inference, or as a building block for calculating some more complex indicators. The calculation logic and the content fit well with other topics from this chapter, which is the reason we included it here. A more meaningful interpretation of this concept will become obvious as we move into regression analysis and time series analysis topics.