5. Associations between and within the time series

In this Chapter we'll explore how to take advantage of measuring correlation between time series and then extend this concept to measuring correlation within the time series. A concept of autocorrelation and partial autocorrelation will be introduced, as well as various methods describing how to calculate them. These two concepts are most crucial for any stochastic model building attempt and have much broader connotations in time series analysis.

5.1 Correlation

A measure that tells us if any two variables are related is called covariance. Let's assume that we have two variables x and y, each with n observations. In this case the sample covariance is calculated using the formula:

$$\sigma_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad (5.1.1)$$

This measure, as well as the individual variable standard deviations, can be used to extract one single coefficient measuring correlation between the two variables. A general formula is:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad (5.1.2)$$

From this general formula a number of specific correlation coefficients can be derived and the one most popular is the so called Pearson's Product Moment correlation coefficient. It is defined as:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \qquad (5.1.3)$$

Where *x* and *y* are the observations belonging to two different time series, $\bar{x}$ and $\bar{y}$ are the respective means. The coefficient measures only linear correlation and can be applied in Excel by using either =PEARSON() or =CORREL() function. We'll use the later one.

To illustrate the use of =CORREL() function, we'll take the closing daily values of Dow Jones Industrial Composite Index (DJI) and NASDAQ between 4 January 2016 and 7 September 2016. As the DJI values are in the region of ten thousand and the NASDAQ values are in the region of tens, we'll show them on the same graph where the left axis is scaled for the DJI data and the right axis is scaled for the NASDAQ data.

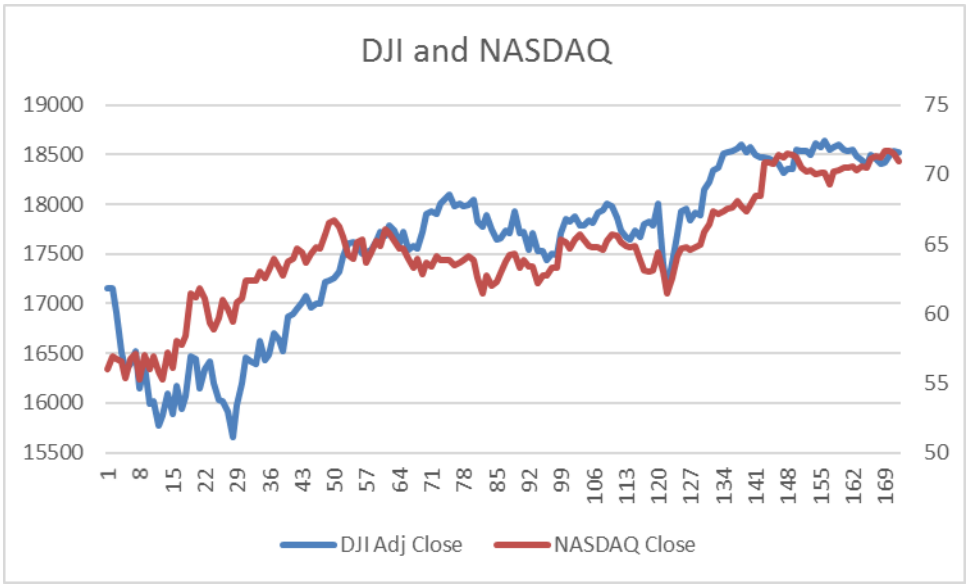


Fig. 5.1.1 Chart showing the DJI and NASDAQ closing values between 4 January 2016 and 7 September 2016

The two time series appear to be highly correlated and the calculated coefficient of correlation confirms it. The value is 0.914 (cell F2 in Fig 5.1.2.).

	A	B	C	D	E	F	G	H
1	Date	DJI Adj Close	NASDAQ Close		r=			
2	04-01-16	17148.93945	56.046013		0.844	=CORREL(B2:B52,C2:C52)		
3	05-01-16	17158.66016	56.8985					
4	06-01-16	16906.50977	56.729983		r-squared=			
5	07-01-16	16514.09961	56.601118		0.71216	=E2^2		
6	08-01-16	16346.4502	55.322391					
7	11-01-16	16398.57031	56.779545					
8	12-01-16	16516.2207	57.076927					
9	13-01-16	16151.41016	55.243088					
10	14-01-16	16379.04981	56.987713					
11	15-01-16	15988.08008	55.99645					

Fig.5.1.2 DJI and NASDAQ values and the coefficient of correlation (only first ten out of fifty values shown)

We know that the correlations coefficient varies between -1 and +1, with zero implying that there is no linear correlation between the two variables. In our case the value of 0.844 means that these two series are highly correlated.

As the two time series are highly correlated, the coefficient of determination, R-Squared, is also going to be very high. We calculated it using two different methods, the first one by just squaring the value of the correlation coefficient (cell F4 in Fig. 5.1.2) and the second one by using a dedicated Excel function =RSQ() for the coefficient of determination (cell F5 in Fig. 5.1.2).

The coefficient of determination, whose value is 0.71 (cells F4 or F5 in Fig 5.1.2), implies that 71% of the variations in one variable can be explained by the variations in another variable. This is quite high as it leaves the remaining 29% of variations due to other influences not captured by the association between these two variables.

It is important to emphasise that the correlation coefficient and the coefficient of determination are only useful where there is linear relationship between variables.

We'll now use the same principle of correlation, but not between two variables. We'll apply it to a single variable, but the one that is lagged on itself.

5.2 Autocorrelation

Let us take just the DJI time series and see what happens if we shift the values by one observation. What we get is called a lagged variable and Fig. 5.2.1 shows a simple way to achieve this in column C.

	A	B	C	D
1	Date	DJI	Lag 1	
2	31-08-16	18400.88	18419.3	C2=B3
3	01-09-16	18419.3	18491.96	C3=B4
4	02-09-16	18491.96	18538.12	C4=B5
5	06-09-16	18538.12		

Fig. 5.2.1 Lagging the variable by one period

Needless to say, we could lag this variable *k* times, where *k* is any number, usually not bigger than n/3 (n is the total number of observations). In our case we have a series with 50 observations (n=50), which means that *k* should not be more than 17 (k=50/3≈17). Fig. 5.2.2 shows just the first five lagged time series for the DJI series.

	A	B	C	D	E	F	G
1	Date	DJI	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5
2	03-08-16	18355	18352.05	18543.53	18529.29	18533.05	18495.66
3	04-08-16	18352.05	18543.53	18529.29	18533.05	18495.66	18613.52
4	05-08-16	18543.53	18529.29	18533.05	18495.66	18613.52	18576.47
5	08-08-16	18529.29	18533.05	18495.66	18613.52	18576.47	18636.05
6	09-08-16	18533.05	18495.66	18613.52	18576.47	18636.05	18552.02
7	10-08-16	18495.66	18613.52	18576.47	18636.05	18552.02	18573.94
8	11-08-16	18613.52	18576.47	18636.05	18552.02	18573.94	18597.7
9	12-08-16	18576.47	18636.05	18552.02	18573.94	18597.7	18552.57
10	15-08-16	18636.05	18552.02	18573.94	18597.7	18552.57	18529.42
11	16-08-16	18552.02	18573.94	18597.7	18552.57	18529.42	18547.3

Fig. 5.2.2 A variable (DJI Index) lagged by five periods

In the same fashion as we calculated the correlation coefficient between the DJI and NASDAQ variable, we can calculate the correlation coefficient between any variable and its lagged values. In so doing, we obtain a series of autocorrelation coefficients, and when we put them together, they form what is known as the autocorrelation function.

The formula for the autocorrelation function is:

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \qquad (5.2.1)$$

In the above formula *k* is the number of autocorrelations in the function (we said that k_{max}=n/3), *n* is the number of observations in the series and $\bar{y}$ is the mean value.

Note that the series of autocorrelation coefficients r_k constitute the autocorrelation function.

We will now show several simple ways to calculate the autocorrelation function using standard Excel functions.

To translate unfriendly looking formula (5.2.1) into the spreadsheet syntax and in order to demonstrate how the autocorrelation function is calculated, let us use the time series for DJI closing values between 28 June 2016 and 7 September 2016.

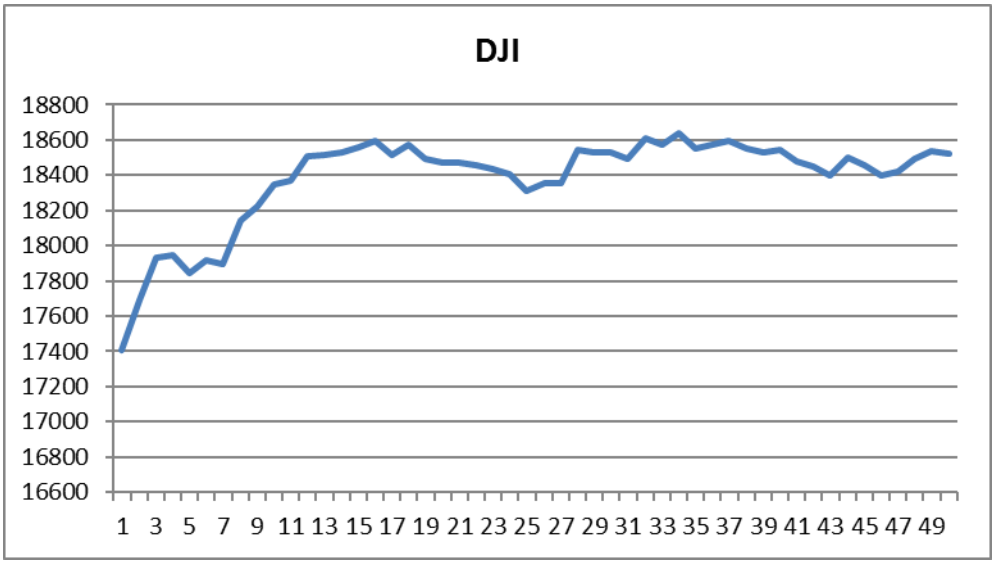


Fig. 5.2.3 DJI daily closing values 28 June 2016 and 7 September 2016

To translate formula (5.2.1) into Excel "speak", we just need to use two standard functions, i.e. =SUMPRODUCT() and =DEVSQ().

Below in Fig. 5.2.4 in column D we can see that we first calculated deviations of every observation from the mean, i.e. $(y_t - \bar{y})$. In cell D3 we entered a formula =C4-AVERAGE($C$3:$C$52) and copied it down.

We'll first look to the denominator part of formula (5.2.1). The formulae in column E contain expression =…./ DEVSQ($D$3:$D$52), which is an Excel version of the denominator from formula (5.2.1). Column D contain deviations from the mean, but according to the formula (5.2.1) they need to be squared, i.e. $(y_t - \bar{y})^2$. Rather than squaring every cell and then summing them up, we have used Excel function =DEVSQ(). A very simple and elegant solution.

Let's take a closer look at the cells in column E. The first cell E3 contains formula:

=(SUMPRODUCT($D$3:D51,$D4:D$52)/DEVSQ($D$3:$D$52)). We'll now explain the first part of this formula that corresponds to the numerator in formula (5.2.1).

The numerator in equation (5.2.1) states that we should calculate the sum of product of two expressions. They are $(y_t - \bar{y})$ and $(y_{t+k} - \bar{y})$. This means:

For r_1: $(y_1 - \bar{y})(y_2 - \bar{y})+(y_2 - \bar{y})(y_3 - \bar{y})+(y_3 - \bar{y})(y_4 - \bar{y})+\ldots+(y_{49} - \bar{y})(y_{50} - \bar{y})$
For r_2: $(y_1 - \bar{y})(y_3 - \bar{y})+(y_2 - \bar{y})(y_4 - \bar{y})+(y_3 - \bar{y})(y_5 - \bar{y})+\ldots+(y_{48} - \bar{y})(y_{50} - \bar{y})$
.
.
For r_{18}: $(y_1 - \bar{y})(y_{19} - \bar{y})+(y_2 - \bar{y})(y_{20} - \bar{y})+(y_3 - \bar{y})(y_{21} - \bar{y})+\ldots+(y_{42} - \bar{y})(y_{50} - \bar{y})$

	A	B	C	D	E
1					M1
2	PERIOD	DATE	DJI	Deviation	ACF
3	1	28-06-16	17409.72	-974.74	0.7995
4	2	29-06-16	17694.68	-689.78	0.6631
5	3	30-06-16	17929.99	-454.47	0.5752
6	4	01-07-16	17949.37	-435.09	0.5114
7	5	05-07-16	17840.62	-543.84	0.3876
8	6	06-07-16	17918.62	-465.84	0.2759
9	7	07-07-16	17895.88	-488.58	0.1184
10	8	08-07-16	18146.74	-237.72	0.0404
11	9	11-07-16	18226.93	-157.53	-0.0367
12	10	12-07-16	18347.67	-36.79	-0.0695
13	11	13-07-16	18372.12	-12.34	-0.1154
14	12	14-07-16	18506.41	121.95	-0.1052
15	13	15-07-16	18516.55	132.09	-0.1046
16	14	18-07-16	18533.05	148.59	-0.0888
17	15	19-07-16	18559.01	174.55	-0.0649
18	16	20-07-16	18595.03	210.57	-0.0148
19	17	21-07-16	18517.23	132.77	-0.0083
20	18	22-07-16	18570.85	186.39	0.0328
21	19	25-07-16	18493.06	108.60	0.0389
22	20	26-07-16	18473.75	89.29	0.0456
23	21	27-07-16	18472.17	87.71	
24	22	28-07-16	18456.35	71.89	

Fig. 5.2.4 Calculating autocorrelations via SUMPRODUCT() and DEVSQ() function

As we can see that the subscripts for y_t in the first expression in the brackets of the product, for k=1, go from 1 to 49, and for the second expression from 2 to 50. The first expression in the brackets for k=2 goes from 1 to 48, and the second from 3 to 50. For k=3, the ranges go from 1 to 47 and from 4 to 50, etc., until finally for k=18 the ranges go from 1 to 42 and 18 to 50. We translated this into Excel, by using the =SUMPRODUCT() function. So the first part of the formula between E3:E20 is:

E3 = SUMPRODUCT($D$3:D51,D4:$D$52)
E4 = SUMPRODUCT($D$3:D50,D5:$D$52)
E5 = SUMPRODUCT($D$3:D49,D6:$D$52)
.
E20 = SUMPRODUCT($D$3:D34,D21:$D$52)

No doubt the =SUMPRODUCT() function is perfectly suited to calculate the numerator of function (5.2.1), except that one of the ranges is going in the opposite direction(C3:C51, C3:C50, C3:C49, etc.). The problem is that if we copied this formula down from E3, we would manually have to change the numbers, as they would be increasing rather than decreasing.

	A	B	C	D	E
1					**M1**
2	**PERIOD**	**DATE**	**DJI**	**Deviation**	**ACF**
3	1	28-06-16	17409.721	=C3-AVERAGE($C$3:$C$52)	=(SUMPRODUCT($D$3:D51,$D4:D$52)/DEVSQ($D$3:$D$52))
4	2	29-06-16	17694.68	=C4-AVERAGE($C$3:$C$52)	=(SUMPRODUCT($D$3:D50,$D5:D$52)/DEVSQ($D$3:$D$52))
5	3	30-06-16	17929.99	=C5-AVERAGE($C$3:$C$52)	=(SUMPRODUCT($D$3:D49,$D6:D$52)/DEVSQ($D$3:$D$52))
6	4	01-07-16	17949.369	=C6-AVERAGE($C$3:$C$52)	=(SUMPRODUCT($D$3:D48,$D7:D$52)/DEVSQ($D$3:$D$52))
7	5	05-07-16	17840.619	=C7-AVERAGE($C$3:$C$52)	=(SUMPRODUCT($D$3:D47,$D8:D$52)/DEVSQ($D$3:$D$52))

Fig. 5.2.5 Formulae for calculating autocorrelations

One option is to use the function =OFFSET. To illustrate how to do this, in Fig. 5.2.6 we created a new column F that calculates the same autocorrelations, but using the =OFFSET() function. In this case the formula in cell F3 is written as:

=SUMPRODUCT(OFFSET($D$3:$D$52,0,0,COUNT($D$3:$D$52)-A3),_
_OFFSET($D$3:$D$52,A3,0,COUNT($D$3:$D$52)-A3))/DEVSQ($D$3:$D$52)

	A	B	C	D	E	F	G
1					M1	M1a	M2
2	PERIOD	DATE	DJI	Deviation	ACF	ACF	ACF
3	1	28-06-16	17409.72	-974.74	0.7995	0.7995	0.7995
4	2	29-06-16	17694.68	-689.78	0.6631	0.6631	0.6631
5	3	30-06-16	17929.99	-454.47	0.5752	0.5752	0.5752
6	4	01-07-16	17949.37	-435.09	0.5114	0.5114	0.5114
7	5	05-07-16	17840.62	-543.84	0.3876	0.3876	0.3876
8	6	06-07-16	17918.62	-465.84	0.2759	0.2759	0.2759
9	7	07-07-16	17895.88	-488.58	0.1184	0.1184	0.1184
10	8	08-07-16	18146.74	-237.72	0.0404	0.0404	0.0404
11	9	11-07-16	18226.93	-157.53	-0.0367	-0.0367	-0.0367
12	10	12-07-16	18347.67	-36.79	-0.0695	-0.0695	-0.0695

Fig 5.2.6 Three different versions of calculating the autocorrelations.

The same effect could have been achieved using the function =INDEX(), as in column G:

=SUMPRODUCT($D$3:INDEX($D$3:$D$52,ROWS(D4:D$52)),$D4:D$52)/_
_DEVSQ($D$3:$D$52).

	E
1	M1
2	ACF
3	=(SUMPRODUCT($D$3:D51,$D4:D$52)/DEVSQ($D$3:$D$52))
4	=(SUMPRODUCT($D$3:D50,$D5:D$52)/DEVSQ($D$3:$D$52))
5	=(SUMPRODUCT($D$3:D49,$D6:D$52)/DEVSQ($D$3:$D$52))
6	=(SUMPRODUCT($D$3:D48,$D7:D$52)/DEVSQ($D$3:$D$52))
7	=(SUMPRODUCT($D$3:D47,$D8:D$52)/DEVSQ($D$3:$D$52))

	F
1	M1a
2	ACF
3	=SUMPRODUCT(OFFSET($D$3:$D$52,0,0,COUNT($D$3:$D$52)-A3),OFFSET($D$3:$D$52,A3,0,COUNT($D$3:$D$52)-A3))/DEVSQ($D$3:$D$52)
4	=SUMPRODUCT(OFFSET($D$3:$D$52,0,0,COUNT($D$3:$D$52)-A4),OFFSET($D$3:$D$52,A4,0,COUNT($D$3:$D$52)-A4))/DEVSQ($D$3:$D$52)
5	=SUMPRODUCT(OFFSET($D$3:$D$52,0,0,COUNT($D$3:$D$52)-A5),OFFSET($D$3:$D$52,A5,0,COUNT($D$3:$D$52)-A5))/DEVSQ($D$3:$D$52)
6	=SUMPRODUCT(OFFSET($D$3:$D$52,0,0,COUNT($D$3:$D$52)-A6),OFFSET($D$3:$D$52,A6,0,COUNT($D$3:$D$52)-A6))/DEVSQ($D$3:$D$52)
7	=SUMPRODUCT(OFFSET($D$3:$D$52,0,0,COUNT($D$3:$D$52)-A7),OFFSET($D$3:$D$52,A7,0,COUNT($D$3:$D$52)-A7))/DEVSQ($D$3:$D$52)

	G
1	M2
2	ACF
3	=SUMPRODUCT($D$3:INDEX($D$3:$D$52,ROWS(D4:D$52)),$D4:D$52)/DEVSQ($D$3:$D$52)
4	=SUMPRODUCT($D$3:INDEX($D$3:$D$52,ROWS(D5:D$52)),$D5:D$52)/DEVSQ($D$3:$D$52)
5	=SUMPRODUCT($D$3:INDEX($D$3:$D$52,ROWS(D6:D$52)),$D6:D$52)/DEVSQ($D$3:$D$52)
6	=SUMPRODUCT($D$3:INDEX($D$3:$D$52,ROWS(D7:D$52)),$D7:D$52)/DEVSQ($D$3:$D$52)
7	=SUMPRODUCT($D$3:INDEX($D$3:$D$52,ROWS(D8:D$52)),$D8:D$52)/DEVSQ($D$3:$D$52)

Fig 5.2.7 Three different versions of calculating the autocorrelations.

We'll briefly digress here just to explain the mechanics of this third approach.

ROWS() function just counts the rows in a range. Let's assume we have a small array of numbers A1:A5. If in cells B1:B5 we insert the function ROWS(), we'll get the following results:

B1=ROWS(A1:A$A5) gives the value of 5
B2=ROWS(A2:A$A5) gives the value of 4
B3=ROWS(A3:A$A5) gives the value of 3
.
B5=ROWS(A5:A$A5) gives the value of 1

As we can see the cells in column B will give us the decreasing number of rows in column A. Unfortunately we cannot insert this function directly into the =SUMPRODUCT() function, but we can combine it with the =INDEX() function.

Following the above example we said that the function =SUMPRODUCT() should be used in the following way:

=SUMPRODUCT($A$1:$A5,A1:$A$5)
=SUMPRODUCT($A$1:$A4,A2:$A$5)
.
=SUMPRODUCT($A$1:$A1,A5:$A$5)

However, if we copy this formula down, the first range will increase in value and we would have to change the cell references manually. Fortunately, by combining functions INDEX() and ROWS(), we can eliminate this problem:

=SUMPRODUCT($A$1:INDEX($A$1:$A$5,ROWS(A1:$A$5)),A1:$A$5)
=SUMPRODUCT($A$1:INDEX($A$1:$A$5,ROWS(A2:$A$5)),A2:$A$5)

=SUMPRODUCT($A$1:INDEX($A$1:$A$5,ROWS(A5:$A$5)),A5:$A$5)

Now the function can be copied down and the two ranges are multiplied and summed up automatically for every value of *k*.

You can experiment with other Excel functions, such as =COVARIANCE.S(), =STDEV.S() and/or =CORREL() and build a different formulae for calculating the autocorrelations. However, the numbers might just slightly differ from the ones we obtained here. The reason is that these functions will not take into account the proper number of degrees of freedom for the reducing data set.

5.3 Interpreting autocorrelations

The autocorrelation factors, or coefficients, when charted on the graph will represent the autocorrelation function. Fig 5.3.1 shows a typical autocorrelation function. You can use the line graph, but it is customary to use the bar graph. We'll explain this later.

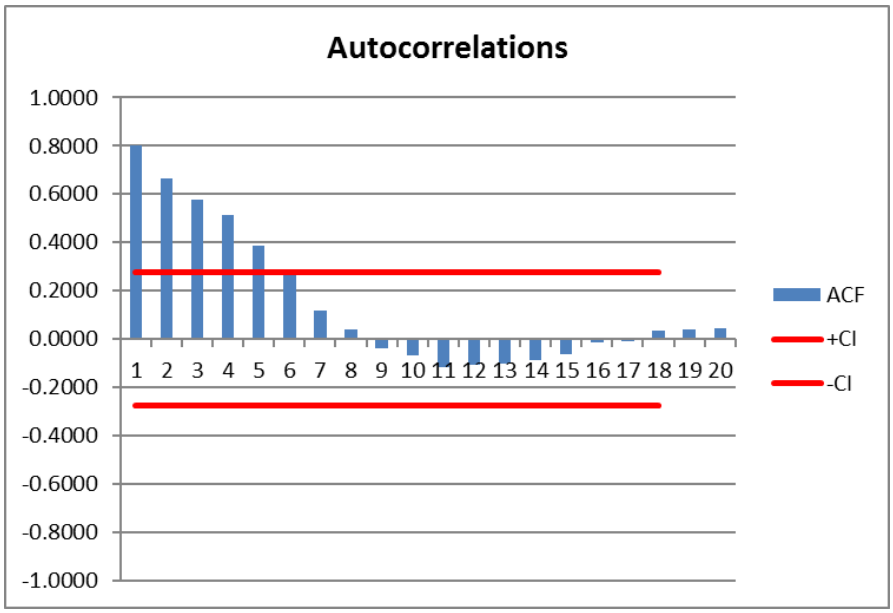


Fig 5.3.1 Autocorrelation function for DJI data

A general rule is, if the autocorrelation coefficients are persistently large, i.e. they are very slowly dropping towards zero, this indicates that the time series is probably non-stationary. The chart in Fig. 5.3.1 indicates that in this time window the DJI values are probably non-stationary. If the dataset was stationary, we would expect only a few autocorrelations to be significantly different from zero.

What do we mean when we say "significantly different from zero"? This means that in order to be called "different from zero", the autocorrelation coefficients need to stay outside some predefined confidence interval. Effectively, we need to calculate the standard error, and the formula is:

$$SE_{ACF} = \frac{1}{\sqrt{n}} \qquad (5.3.1)$$

Where SE_{ACF} is the standard error and *n* is the number of observations in the time series.

It is customary to apply a 95% confidence interval. As we know, a 95% confidence interval is calculated as follows (for details, see the refresher section in Appendix):

$$95\%CI = \pm 1.96\, SE_{ACF} = \pm 1.96 \frac{1}{\sqrt{n}} \qquad (5.3.2)$$

The formula for the standard error $1/\sqrt{n}$ is applicable under the assumption that a time series is completely random. For truly random time series, such as white noise for example, all the autocorrelation coefficients will be zero. The value of 1.96 in equation (5.3.2) represents the area under the normal curve, or the z-score for a 95% confidence interval. In other words, we can assume with 95% certainty that all sample autocorrelations that are within ±1.96 standard errors are **not** different from zero.

In the case of DJI closing values, we had n=50, which gives us the standard error of 0.1414 (a square root of one over *n*). The critical value of ±0.277 (±1.96 × 0.1414) gives us a 95% "corridor". All the autocorrelation values inside this "corridor" are considered to be virtually zero. The two red lines in Fig 5.3.1 represent the 95% confidence interval that all the autocorrelations inside this interval are virtually zero. We can see that from the eighth autocorrelation onwards, they are all inside this corridor.

Most of the software packages use somewhat different method to measure the standard errors. In fact if you look at the most the autocorrelation graphs in numerous books, you'll see that the standard error lines are not completely horizontal. They are narrower for lower lags and then grow gradually wider as the lags increase. Let's show how to calculate standard errors using a different method, so that your autocorrelations calculated in Excel match those calculated by some of these software packages.

5.4 Corrections to the autocorrelations standard error

A modification to equation (5.3.2) that is used by many software packages and textbooks is called the Bartlett's formula. It was published before the era of personal computers when precision was often sacrificed for the sake of computational speed. It is in fact a bit approximate as it gives us a two standard error interval, rather than 1.96, which is exactly 95% confidence interval.

The Bartlett's formula for the standard error for the autocorrelation function is defined by the following equation:

$$SE_{ACF} = \sqrt{\frac{1}{n}\left(1+2\sum_{q=1}^{k-1} r_q^2\right)} \qquad (5.4.1)$$

Where, r_q are the autocorrelation factors and *n* is the number of observations in the time series as before.

Fig. 5.4.1 shows the values of the standard errors and the confidence interval for the autocorrelation function of the DJI daily closing values in the period of 17 September and 25 November 2009.

	A	B	C	D	E	F	G
1	**METHOD 2**						
2	**PERIOD**	**DJI**	**Dev from mean**	**ACF**	**2SE**	**-CI**	**+CI**
3	1	17409.72	-974.74	0.7995	0.141	-0.141	0.141
4	2	17694.68	-689.78	0.6631	0.213	-0.213	0.213
5	3	17929.99	-454.47	0.5752	0.251	-0.251	0.251
6	4	17949.37	-435.09	0.5114	0.276	-0.276	0.276
7	5	17840.62	-543.84	0.3876	0.295	-0.295	0.295
8	6	17918.62	-465.84	0.2759	0.305	-0.305	0.305
9	7	17895.88	-488.58	0.1184	0.310	-0.310	0.310
10	8	18146.74	-237.72	0.0404	0.311	-0.311	0.311
11	9	18226.93	-157.53	-0.0367	0.311	-0.311	0.311
12	10	18347.67	-36.79	-0.0695	0.311	-0.311	0.311
13	11	18372.12	-12.34	-0.1154	0.311	-0.311	0.311
14	12	18506.41	121.95	-0.1052	0.312	-0.312	0.312
15	13	18516.55	132.09	-0.1046	0.313	-0.313	0.313
16	14	18533.05	148.59	-0.0888	0.313	-0.313	0.313
17	15	18559.01	174.55	-0.0649	0.314	-0.314	0.314
18	16	18595.03	210.57	-0.0148	0.314	-0.314	0.314
19	17	18517.23	132.77	-0.0083	0.314	-0.314	0.314
20	18	18570.85	186.39	0.0328	0.314	-0.314	0.314

Fig. 5.4.1 Calculation of the standard error and the confidence interval for the autocorrelation function

The cell E3 is calculated as =1/SQRT(COUNT(B3:B52)), followed by E4 onwards as:

=SQRT((1/COUNT($B$3:$B$52))*(1+2*SUMSQ($D$3)))
=SQRT((1/COUNT($B$3:$B$52))*(1+2*SUMSQ($D$3:D4)))
=SQRT((1/COUNT($B$3:$B$52))*(1+2*SUMSQ($D$3:D5))), etc.

Cell E3 contains a simple equation for the standard error: $1/\sqrt{n}$. Cell E4 uses the equation (5.4.1), as does the range E5:E20. However, note the subscript of r_q^2. The values of q has to be q<k, or as in our case the max value of q is k-1. In other words if we are calculating the standard error for lag 3, we can only use the squared values of the autocorrelation coefficients for lag 2 and 1. If we are calculating the standard error for lag 4, we can only use the squared values of the autocorrelation coefficients for lag 3, 2 and 1, etc.

Cells in column F and G are just positive and negative values from column E and we created separate columns just to make the creation of the graph easier. No other reason.

The same autocorrelation function as in Fig 5.3.1 is given below in Fig. 5.4.2. The two red lines are no longer parallel.

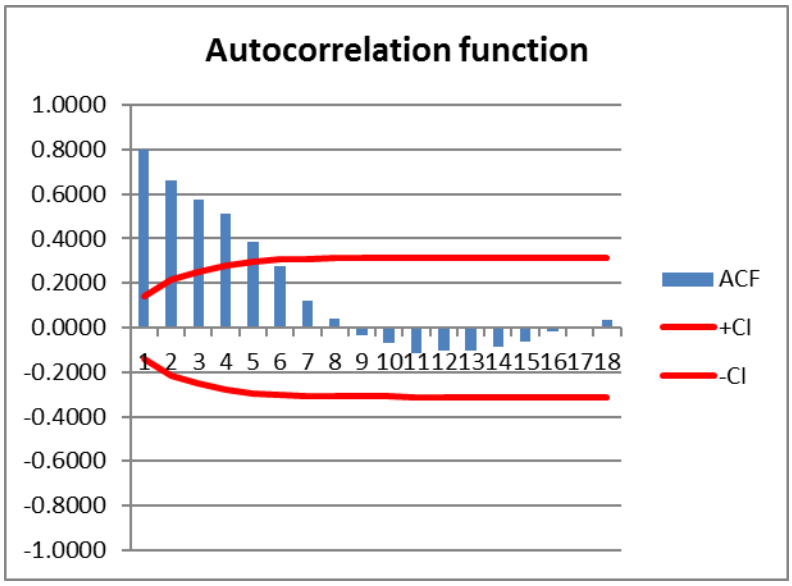


Fig. 5.4.2 The autocorrelation function showing three coefficients to be non-zero

We can see that starting with the sixth autocorrelation coefficient in Fig. 5.4.2 the remaining autocorrelations are all virtually zero.

One very important property of stationary time series (i.e. series that oscillate around some fixed, or stationary, mean value) is that, with the exception of a few initial values, all their autocorrelation coefficients are virtually zero. Later on we introduce a typical example of stationary time series, a white noise process, where in fact all the autocorrelations are equal to zero. Non-stationary time series (those with upward or downward trend) will have a certain number of persistent non-zero autocorrelations (i.e. larger than the critical value for a particular level of significance). This is all covered in Chapter 11.

Sometimes the autocorrelation graphs reveal other properties too. In case we had any doubt about the seasonality of the series, autocorrelations will reveal the periodicity, which can be utilised to select the correct forecasting model. Autocorrelations are clearly one of the most important indicators in the forecasting process and we'll rely on them heavily as we progress towards more complex forecasting methods.

There is a visual difference between SE_{ACF} calculated as $\pm 1.96\frac{1}{\sqrt{n}}$ (let's call it SE1) and SE_{ACF} calculated as $SE_{ACF} = \sqrt{\frac{1}{n}\left(1+2\sum_{q=1}^{k-1} r_q^2\right)}$ (let's call it SE2). These two versions of the standard error do not just create two slightly different intervals around the autocorrelations, but they are intended to measure two different things.

The first standard error (SE1) is designed to test for the randomness of the series of autocorrelation coefficients. As we already said, if the autocorrelation coefficients "stick" outside this boundary, we are 95% certain that they do not belong to a random time series. The

null hypothesis in this case that we are effectively testing is that the series of autocorrelations is random (or equal to zero).

The second standard error (SE2) is designed to show us how every individual autocorrelation coefficient is affected by the ones that precede it. It effectively tests the null hypothesis that the theoretical autocorrelation function has "died out" by that particular lag. If it "sticks" outside the corridor, this means the theoretical autocorrelations that precede it have not been reduced to zero. Sounds complicated, but hopefully Chapter 11 will provide some illumination.

Unfortunately we have to say that in most of the textbooks these two types of standard errors (SE1 and SE2) are virtually used interchangeably and although not correct, most of the people assume that they just test randomness.

Besides measuring relationships between the variables (correlation) and relationship between variables and their past values lagged by some amount of time (autocorrelations), we can also measure the so-called partial autocorrelation.

5.5 Partial Autocorrelation

To understand the partial autocorrelations is not easy and can be counter-intuitive, so we'll describe a situation that will assist us with understanding of this concept. Let's say that we want to eliminate the influence of certain factors (or, keep them constant) in order to measure the true correlation between other variables. Here is an example. We are interested in the relationship between the number of visitors to the Lake District and a number of warm drinks that are sold to these visitors. Naturally we would expect to see a high correlation coefficient between these two variables. In other words, the more visitors we have, the more of warm drinks we sell. However, on certain days we might end up with the negative correlation, which implies that the more visitors we have the less warm drinks we sell. How could this be possible?

Clearly something else is influencing consumption of warm drinks. On a sunny day people prefer cold drinks, rather than warm drinks. However, on a cold day, although we have fewer visitors, they drink more warm drinks. This third variable, the outside temperature, clearly affects our correlation measurements and we are getting correlation values that do not truly represent the relationship. In a case like this, a partial correlation coefficient is what we want to use.

Effectively, the partial correlation coefficient keeps the variation of one variable constant, whilst measuring the relationship between the other two. In other words, we can "exclude" the effects of daily temperatures out of the relationship when measuring the relationship between the number of visitors and the number of cold drinks sold.

Returning to the issues of autocorrelations, i.e. measuring relationships between a variable and its lagged values, we can draw an analogy and say that we can also measure the partial autocorrelations. In other words, to prevent all the lagged values from affecting our measurements, we'll keep all the lagged values constant except the one that we are correlating with the actual time series. The values obtained in this way are the partial autocorrelations and

they will form the partial autocorrelation function for the series that is used as a basis for calculations.

To recapitulate, if we shift our original series *k* times and measure the correlation between these shifted series, we have, in fact, measured the correlation of the series and its previous values shifted *k* times. The series of these correlation coefficients is called the autocorrelation function. If, on the other hand, we try to measure the correlation of the series and its previous values shifted *k* times, but by keeping the influence of all other shifted series out of the equation, we get the partial autocorrelation coefficients. How do we calculate the partial autocorrelation coefficients?

For the first lag it is easy to calculate the first partial autocorrelation as it is equal to the first autocorrelation, i.e. for k=1:

$$p_{1.1} = r_1 \qquad (5.5.1)$$

For a number of lags greater than 1:

For k=2,3,.....L

$$p_{k.k} = \frac{r_k - \sum_{j=1}^{k-1} p_{k-1.j} r_{k-j}}{1 - \sum_{j=1}^{k-1} p_{k-1.j} r_j} \qquad (5.5.2)$$

Where:

For j=1,2,...,k-1

$$p_{k.j} = p_{k-1.j} - p_{k.k}\, p_{k-1.k-j} \qquad (5.5.3)$$

In other words, to calculate $p_{2.2}$, the second partial autocorrelation, we have to use the equation (5.5.2):

$$p_{2.2} = \frac{r_2 - p_{1.1} r_1}{1 - p_{1.1} r_1}$$

However, in order to calculate $p_{3.3}$ we first have to calculate $p_{2.1}$, which is done by applying equation (5.5.3):

$$p_{2.1} = p_{1.1} - p_{2.2}\, p_{1.1}$$

To expand this example, to calculate the partial autocorrelations, for example for periods 3 to 6, we use the following equations:

$$p_{3.3} = \frac{r_3 - (p_{2.1} r_2 + p_{2.2} r_1)}{1 - (p_{2.1} r_1 + p_{2.2} r_2)}$$

$$p_{4.4}= \frac{r_4 - (p_{3.1}r_3 + p_{3.2}r_2 + p_{3.3}r_1)}{1 - (p_{3.1}r_1 + p_{3.2}r_2 + p_{3.3}r_3)}$$

$$p_{5.5}= \frac{r_5 - (p_{4.1}r_4 + p_{4.2}r_3 + p_{4.3}r_2 + p_{4.4}r_1)}{1 - (p_{4.1}r_1 + p_{4.2}r_2 + p_{4.3}r_3 + p_{4.4}r_4)}$$

$$p_{6.6}= \frac{r_6 - (p_{5.1}r_5 + p_{5.2}r_4 + p_{5.3}r_3 + p_{5.4}r_2 + p_{5.5}r_1)}{1 - (p_{5.1}r_1 + p_{5.2}r_2 + p_{5.3}r_3 + p_{5.4}r_4 + p_{5.5}r_5)}$$

In order to calculate these partial autocorrelations, in between each of them we have to make the following calculations:

$p_{3.1} = p_{2.1} - p_{3.3}\, p_{2.2}$
$p_{3.2} = p_{2.2} - p_{3.3}\, p_{2.1}$
$p_{4.1} = p_{3.1} - p_{4.4}\, p_{3.3}$
$p_{4.2} = p_{3.2} - p_{4.4}\, p_{3.2}$
$p_{4.3} = p_{3.3} - p_{4.4}\, p_{3.1}$, etc.

And finally, for $p_{6.6}$ the same calculations would include:

$P_{6.1} = p_{5.1} - p_{6.6}\, p_{5.5}$
$P_{6.2} = p_{5.2} - p_{6.6}\, p_{5.4}$
$P_{6.3} = p_{5.3} - p_{6.6}\, p_{5.3}$
$P_{6.4} = p_{5.4} - p_{6.6}\, p_{5.2}$
$P_{6.5} = p_{5.5} - p_{6.6}\, p_{5.1}$

We'll take the example of DJI from Fig. 5.4.1 and calculate the partial correlation function. The partial autocorrelation coefficients are presented graphically in Fig. 5.5.1.

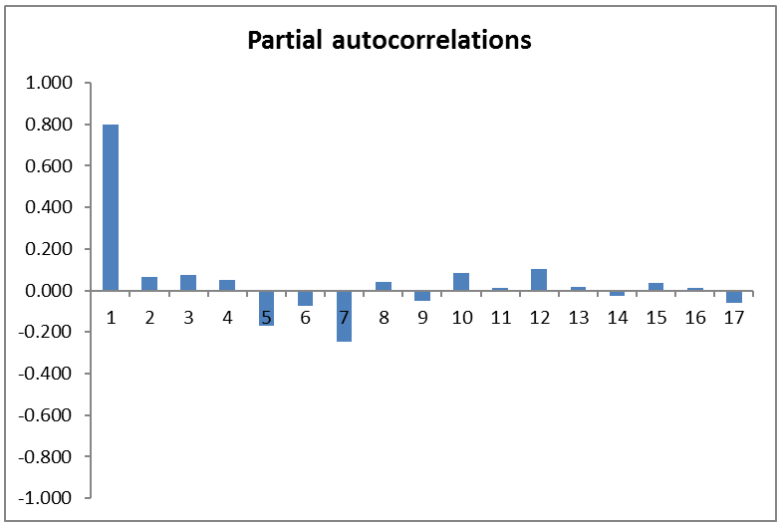


Fig. 5.5.1 The partial autocorrelation coefficients for the values from Fig. 5.4.1

If we literally applied all the above equations in Excel, it is possible to calculate the partial autocorrelations. Fig. 5.5.2 shows this manual approach. However, a "manual" way to calculate partial autocorrelations is very cumbersome.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Lag**	**AC**	**PAC**									
2	k	r_k	$r_{k,j}$	k,1	k,2	k,3	k,4	k,5	k,6	k,7	k,8	k,9
3	1	0.800	0.800	**0.800**								
4	2	0.663	0.066	0.747	**0.066**							
5	3	0.575	0.076	0.742	0.010	**0.076**						
6	4	0.511	0.050	0.738	0.009	0.039	**0.050**					
7	5	0.388	-0.170	0.746	0.016	0.040	0.176	**-0.170**				
8	6	0.276	-0.072	0.734	0.028	0.043	0.177	-0.116	**-0.072**			
9	7	0.118	-0.245	0.716	0.000	0.086	0.187	-0.110	0.108	**-0.245**		
10	8	0.040	0.042	0.727	-0.005	0.091	0.179	-0.113	0.108	-0.275	**0.042**	
11	9	-0.037	-0.049	0.729	-0.018	0.096	0.174	-0.104	0.112	-0.275	0.077	**-0.049**

Fig. 5.5.2 Manually calculated partial autocorrelations

The calculations are executed as follows:

PACF for lag k,1	Cell D3	Formula=B3
PACF for lag k,1	Cells D4:D19	Formula=D3-(D3*E4)
		Formula= D4-(E3*F5)
		Formula= D5-(F5*G6), etc.
PACF for lag k,2	Cell E4	Formula=(B4-(D3*B3))/(1-D3*C3))
PACF for lag k,2	Cells E5:E19	Formula=E4-(D4*F5)
		Formula= E5-(E5*G6)
		Formula= E6-(F6*H7), etc.
PACF for lag k,3	Cell F5	Formula=(B5-(D4*B4+E4*B3))/(1-D4*B3+E4*B4))
PACF for lag k,3	Cells F6:F19	Formula=F5-(D5*G6)
		Formula= F6-(E6*H7), etc.

Excel representation of the same formulae is given in Fig. 5.5.3

	A	B	C	D	E	F
1	**Lag**	**AC**	**PAC**			
2	k	r_k	$r_{k,j}$	k,1	k,2	k,3
3	1	0.800	=D3	**=B3**		
4	2	0.663	=E4	=D3-(D3*E4)	**=(B4-(D3*B3))/(1-(D3*C3))**	
5	3	0.575	=F5	=D4-(E4*F5)	=E4-(D4*F5)	**=(B5-(D4*B4+E4*B3))/(1-(D4*B3+E4*B4))**
6	4	0.511	=G6	=D5-(F5*G6)	=E5-(E5*G6)	=F5-(D5*G6)
7	5	0.388	=H7	=D6-(G6*H7)	=E6-(F6*H7)	=F6-(E6*H7)

	G	H
1		
2	k,4	k,5
3		
4		
5		
6	**=(B6-(D5*B5+E5*B4+F5*B3))/(1-(D5*B3+E5*B4+F5*B5))**	
7	=G6-(D6*H7)	**=(B7-(D6*B6+E6*B5+F6*B4+G6*B3))/(1-(D6*B3+E6*B4+F6*B5+G6*B6))**

Fig. 5.5.3 Excel formulae for manually calculating partial autocorrelations

Cells D3, E4, F5, G6, etc., are $p_{k.k}$ the partial autocorrelation coefficients and the preceding rows (D4, D5:E5, D6:F6, D7:G7, etc.) are $p_{k.j}$ coefficients.

It is quite clear that the above formulae are fairly 'messy'. The simple reason for this is that the formulae get bigger and bigger, as the number of lags *k* increases. However, Excel offers a more elegant way to calculate the same partial correlation function.

5.6 Using Excel functions to calculate partial autocorrelations

In order to use a more elegant way to calculate the partial autocorrelations we need to borrow some equations from stochastic modelling, as well as to dive into the matrix algebra. If this is too difficult for an average reader, skip the explanations and move directly to Excel formulae below. Alternatively, once you understood the chapter on stochastic modelling, return to this section and it will undoubtedly be much clearer.

Autocorrelations and partial autocorrelations are also used to define certain stochastic processes, and one of them is called autoregressive process. The autocorrelation function satisfies the following equation for autoregressive processes:

$$\rho_j = \phi_{k1}\rho_{j-1} + \ldots + \phi_{k(k-1)}\rho_{j-k+1} + \phi_{kk}\rho_{j-k} \qquad (5.6.1)$$

In equation (5.6.1), ρ_j are the autocorrelation coefficients that depend on some lagged value of the coefficients ρ_{j-k}, as well as on ϕ_{kk}, which are the partial autocorrelation coefficients. Using the matrix notation, equation (5.6,1) can be represented as:

$$\begin{bmatrix}\rho_1\\ \rho_2\\ \vdots\\ \rho_k\end{bmatrix} = \begin{bmatrix}1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1}\\ \rho_2 & 1 & \rho_1 & \cdots & \rho_{k-2}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1\end{bmatrix}\begin{bmatrix}\Phi_{k1}\\ \Phi_{k2}\\ \vdots\\ \Phi_{kk}\end{bmatrix} \qquad (5.6.2)$$

The alternative expression for (5.6.2) is:

$$\boldsymbol{\rho}_{\mathbf{k}} = \mathbf{P}_{\mathbf{k}}\boldsymbol{\phi}_{\mathbf{k}} \qquad (5.6.3)$$

This means that:

$$\boldsymbol{\phi}_{\mathbf{k}} = \frac{\boldsymbol{\rho}_{\mathrm{k}}}{\mathbf{P}_{\mathrm{k}}} = \mathbf{P}_{\mathrm{k}}^{-1}\boldsymbol{\rho}_{\mathrm{k}} \qquad (5.6.4)$$

In other words, the matrix that contains the partial autocorrelation coefficients can be calculated from the inverse matrix of all the lagged autocorrelation coefficients multiplied by the vector that contains all the autocorrelations.

It sounds very complicated, but in terms of Excel functions this means that in order to calculate the partial autocorrelations, we need only two functions, =MMULT() and =MINVERSE(). The

first function is used to multiply two matrices and the second one returns an inverse of the matrix, which is exactly what we need, as per equation (5.6.4).

We could have used yet another method to calculate the partial autocorrelation coefficients by using the determinants of the matrices via the Excel =MDETERM() function, but this would not be any more elegant than via the =MMULT() and =MINVERSE() function.

The vector ρ_k in the equation (5.6.4) shows that we first need to create a table with all the calculated autocorrelation coefficients. Fig. 5.6.1 shows just the first nine, although the full spreadsheet contains all seventeen.

	A	B	C	D	E	F	G	H	I	J
1	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04
2	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04
3	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12
4	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28
5	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39
6	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51
7	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58
8	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66
9	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80
10	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1

Fig. 5.6.1 Autocorrelation coefficient matrix

Once this matrix has been created, we can start using the matrix functions in Excel that will enable us to calculate the partial autocorrelations. Fig. 5.6.2 shows the final result, where the coloured cells are the full partial autocorrelation coefficients ϕ_{kk}, and other cells vertically above these coloured cells are the partial autocorrelation coefficients ϕ_{kj}, that are necessary for calculation (see example after equation (5.5.3) for details.

	A	B	C	D	E	F	G	H	I	J	K
21		j									
22	k	1	2	3	4	5	6	7	8	9	10
23	1	0.800	0.747	0.742	0.738	0.746	0.734	0.716	0.727	0.729	0.733
24	2		0.066	0.010	0.009	0.016	0.028	0.000	-0.005	-0.018	-0.025
25	3			0.076	0.039	0.040	0.043	0.086	0.091	0.096	0.120
26	4				0.050	0.176	0.177	0.187	0.179	0.174	0.164
27	5					-0.170	-0.116	-0.110	-0.113	-0.104	-0.096
28	6						-0.072	0.108	0.108	0.112	0.097
29	7							-0.245	-0.275	-0.275	-0.283
30	8								0.042	0.077	0.079
31	9									-0.049	-0.111
32	10										0.085

Fig. 5.6.2 Partial autocorrelation coefficients for DJI time series

The cells are calculated as:

PACF for lag 1,1 Cell B23 Value (same as the first autocorrelation coefficient)
PACF for lag 2,1 Cells C23:C24 Formula{=MMULT(MINVERSE($A$1:B2),$A$2:$A3)}
PACF for lag 2,2 Cells D23:D25 Formula{=MMULT(MINVERSE($A$1:C3),$A$2:$A4)}
PACF for lag 3,3 Cells E23:E26 Formula{=MMULT(MINVERSE($A$1:D4),$A$2:$A5)}

Compare the values in coloured cells in Fig. 5.6.2 (B23, C24, D25, E26, etc.) with the ones in column C (C3:C11) in Fig.5.5.2 and you will see that they are identical.

Fig. 5.6.3 shows how the above formula has been implemented.

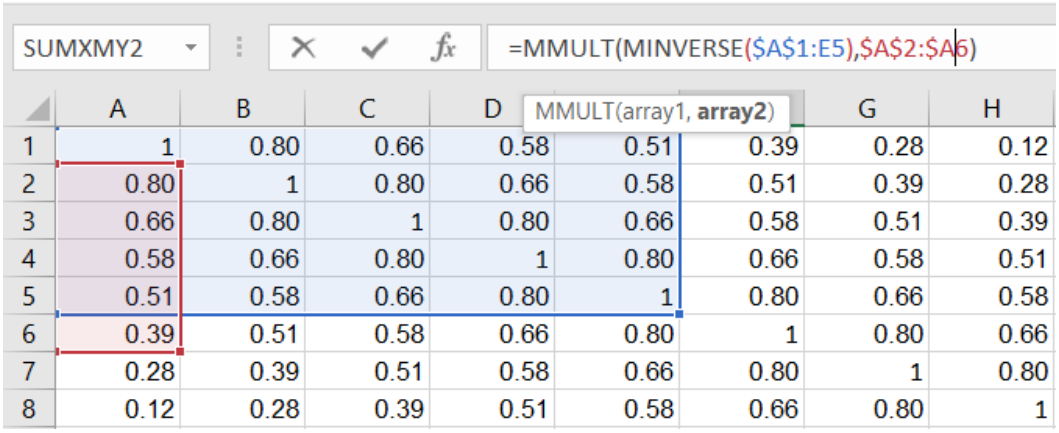
	A	B	C	D	E	F
21		j				
22	k	1	2	3	4	5
23	1	0.8	=MMULT(MINVERSE($A$1:B2),$A$2:$A3)	=MMULT(MINVERSE($A$1:C3),$A$2:$A4)	=MMULT(MINVERSE($A$1:D4),$A$2:$A5)	=MMULT(MINVERSE($A$1:E5),$A$2:$A6)
24	2		=MMULT(MINVERSE($A$1:B2),$A$2:$A3)	=MMULT(MINVERSE($A$1:C3),$A$2:$A4)	=MMULT(MINVERSE($A$1:D4),$A$2:$A5)	=MMULT(MINVERSE($A$1:E5),$A$2:$A6)
25	3			=MMULT(MINVERSE($A$1:C3),$A$2:$A4)	=MMULT(MINVERSE($A$1:D4),$A$2:$A5)	=MMULT(MINVERSE($A$1:E5),$A$2:$A6)
26	4				=MMULT(MINVERSE($A$1:D4),$A$2:$A5)	=MMULT(MINVERSE($A$1:E5),$A$2:$A6)
27	5					=MMULT(MINVERSE($A$1:E5),$A$2:$A6)

Fig. 5.6.3 Calculations for partial autocorrelation coefficients for DJI time series

The curly bracket in the formula {=MMULT(MINVERSE($A$1:B2),$A$2:$A3)} are not entered manually. This is an array formula that is entered by clicking SHIFT, CTRL and ENTER at the same time after the formula has been entered into the formula bar. The curly brackets are added automatically by Excel. To enter this formula follow these steps:

1. We first highlight the cells where the results will be displayed (see in Fig 5.6.2 an example of the range F23:F27)
2. While the cells are highlighted, go to the formula bar and enter the formula: =MMULT(MINVERSE($A$1:E5),$A$2:$A6)
3. Press SHIFT, CTRL and ENTER together
4. The values get populated in the range F23:F27

In our formula above, the first part represents the inverse of the cells A1:E5 (which is equivalent to P_k^{-1} in equation 5.6.4) and A2:A6 is the vector (equivalent to ρ_k in equation 5.6.4). Fig 5.6.4 illustrates the ranges that the formula covers.



	A	B	C	D	E	F	G	H
1	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12
2	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28
3	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39
4	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51
5	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58
6	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66
7	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80
8	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1

Fig. 5.6.4 How matrix functions are used to calculate the partial autocorrelation coefficients

It might appear complicated to start with, but this is a reasonably simple way to calculate the partial autocorrelations in Excel. The value of this method is that it actually shows you what is happening. However, we can automate this even more and bring the calculations of the partial autocorrelation coefficients to a single formula.

Fig. 5.6.5 below is identical to Fig. 5.6.1, except that we added three new columns. Column S contains just sequential numbers, representing the lags for the partial autocorrelation coefficients. Columns T and V have identical values, and they are the partial autocorrelation coefficients, but calculated using a single formula in two different ways. Let's explain.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	V
1	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	-0.12	-0.11	-0.10	-0.09	-0.06	-0.01	-0.01	1	0.800	0.800
2	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	-0.12	-0.11	-0.10	-0.09	-0.06	-0.01	2	0.066	0.066
3	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	-0.12	-0.11	-0.10	-0.09	-0.06	3	0.076	0.076
4	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	-0.12	-0.11	-0.10	-0.09	4	0.050	0.050
5	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	-0.12	-0.11	-0.10	5	-0.170	-0.170
6	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	-0.12	-0.11	6	-0.072	-0.072
7	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	-0.12	7	-0.245	-0.245
8	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	-0.07	8	0.042	0.042
9	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	-0.04	9	-0.049	-0.049
10	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	0.04	10	0.085	0.085
11	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	0.12	11	0.012	0.012
12	-0.12	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	0.28	12	0.105	0.105
13	-0.11	-0.12	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	0.39	13	0.016	0.016
14	-0.10	-0.11	-0.12	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	0.51	14	-0.026	-0.026
15	-0.09	-0.10	-0.11	-0.12	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	0.58	15	0.035	0.035
16	-0.06	-0.09	-0.10	-0.11	-0.12	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	0.66	16	0.013	0.013
17	-0.01	-0.06	-0.09	-0.10	-0.11	-0.12	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	0.80	17	-0.063	-0.063
18	-0.01	-0.01	-0.06	-0.09	-0.10	-0.11	-0.12	-0.07	-0.04	0.04	0.12	0.28	0.39	0.51	0.58	0.66	0.80	1	18	0.051	0.051
19	0.03																				

Fig. 5.6.5 Calculating the partial autocorrelations (PACF) with a single Excel formula

Excel formula for inverse matrix, as we just learned, is: =MINVERSE(array). The red font refers to an array that represents the original matrix from which an inverse matrix will be created. So, if we say =MINVERSE($A$1:C3), this means that the original matrix is in cells A1:C3.

Excel formula for multiplying two matrices, as we also learned, is: =MMULT(array1, array2). The red and green are the two matrices that will be multiplied. So, if we say =MMULT(MINVERSE($A$1:C3),$A$2:$A4), we are multiplying an inverse of the square matrix ($A$1:C3) with a single column matrix ($A$2:$A4). As we know the result will be a single column matrix.

And finally, we'll also use Excel formula for index: =INDEX(aray, row_num, [column_num]), or the other form, which is: =INDEX(reference, row_num, [column_num], [area_num]). The segments in square brackets are optional. Let's break down this formula using an example.

If, for example, we say =INDEX(A5:C10,3,2), the first part A5:C10 is an array from which we are going to extract the value. To extract the value, we find the intersection between the third row in this area and the second column. The result is the value that is in cell C8. Why? The first cell in this example array is A5, so we go 3 rows down (which is to row 8). Again, we start with A5 and go two columns to the right, which is the column C. So, the intersection between row 8 and column C is the cell C8. Therefore, in response to the function =INDEX(), Excel returns the value from this cell.

We can combine all the previous formulas into one single formula that will give us the partial autocorrelation coefficients. For example:
=INDEX(MMULT(MINVERSE($A$1:C3),$A$2:$A4),COUNT($A$2:$A4)).

What we are doing here is exactly what we did before (see Fig. 5.6.4), which is: multiplying the matrix A2:A4 with the inverse matrix of A1:C3. The =INDEX() function is saying that from the result of this multiplication (which will be a column of numbers), we want just the last value in

the column. The part that says COUNT($A$2:$A4) gives us this. It counts how many elements are in this column and only the last one is displayed.

To see how this was applied, look at the formulae in column T. You can see the following:

T1=INDEX(MMULT(MINVERSE($A$1:A1),$A$2:$A2),COUNT($A$2:$A2))
T2=INDEX(MMULT(MINVERSE($A$1:B2),$A$2:$A3),COUNT($A$2:$A3))
T3=INDEX(MMULT(MINVERSE($A$1:C3),$A$2:$A4),COUNT($A$2:$A4))
.
.
T17 =INDEX(MMULT(MINVERSE($A$1:Q17),$A$2:$A18),COUNT($A$2:$A18))

So, column T gives us the partial autocorrelation coefficients. This is exactly the same set of formulae that we used in our previous example that helped us create the partial autocorrelation coefficients in the range B23:R39 (Fig. 5.6.2). The only difference is that we do not see all the intermediate coefficients $r_{k,j}$, but only the final ones $r_{k,k}$, which is what we want. The only problem with the formulae here are the red marked parts. If you copy the formula down, it will copy everything correctly, except the red parts. You will need to manually change A1, to B2, to C3, etc.

A solution to this problem is to make the matrices relative, rather than use the explicit ranges. To do this, we need Excel formula =OFFSET(reference, rows, cols, [height], [width]). The last two elements in square brackets are optional. Let's use an example to explain this function.

If we say =OFFSET(D3,3,-2,1,1), this means that the starting point is the cell D3, which is the reference cell. From there, we go three rows down (which is to row 6) and then move 2 columns to the left (which is to column B). The result is B6. In response to the function =OFFSET(), Excel returns the value from this cell.

We can now use this format to create a one single formula as follow:
=INDEX(MMULT(MINVERSE(OFFSET($A$1,0,0,S3,S3)),OFFSET($A$1,1,0,S3,1)),S3).

What we are saying is that we first multiply the matrix that is defined by the OFFSET function as ($A$1,1,0,S3,1), which is the same as ($A$2:$A4), with an inverse matrix in ($A$1,0,0,S3,S3), which is the same as $A$1:C3. The result will be a column of numbers, but we want the last one which is actually the last value in column S, so in our case the value from cell S3. Instead of S3 in the above formula, we could have used COUNT($A$2:$A4). Either way it gives us the same number for the number of lags.

To see how this was applied, look at the formulae in column V. You can see the following:

V1=INDEX(MMULT(MINVERSE(OFFSET($A$1,0,0,S1,S1)),OFFSET($A$1,1,0,S1,1)),S1)
V2=INDEX(MMULT(MINVERSE(OFFSET($A$1,0,0,S2,S2)),OFFSET($A$1,1,0,S2,1)),S2)
V3=INDEX(MMULT(MINVERSE(OFFSET($A$1,0,0,S3,S3)),OFFSET($A$1,1,0,S3,1)),S3)
.
.

V17=INDEX(MMULT(MINVERSE(OFFSET($A$1,0,0,S17,S17)),OFFSET($A$1,1,0,S17,1)), S17)

Now it is clear that if you copied cell V1 down, there will be no need for manual corrections. This is now fully automated formula for calculating the partial autocorrelation coefficients with a single line and a Copy/Paste action.

5.7 Standard errors for the partial autocorrelations

Just like with the autocorrelations, we need to calculate the confidence interval around the partial autocorrelation coefficients indicating that all the coefficients inside this interval will be treated as zero value partial autocorrelations. We already indicated that for 95% confidence interval, which is effectively an interval covered by two standard errors (in fact, 1.96 standard errors), the equation is:

$$95\%\,\mathrm{CI} = \pm\frac{1.96}{\sqrt{n}} \qquad (5.7.1)$$

Just as with equation (5.3.2) the value of 1.96 represents the area under the normal curve, or the z-score, for a 95% confidence interval, we can assume with 95% certainty that all sample partial autocorrelations that are within ±1.96 standard errors are virtually of zero value. Fig. 5.7.1 shows the DJI partial autocorrelation function and the corresponding confidence interval.

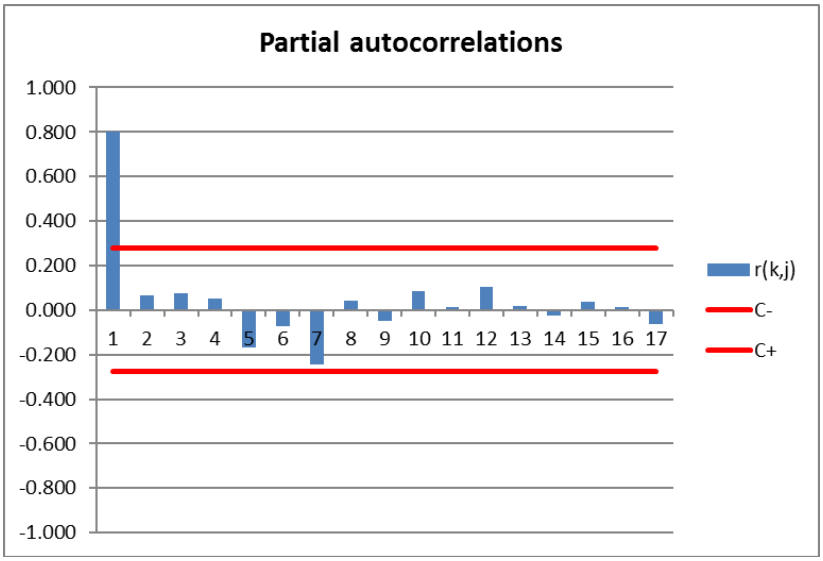


Fig.5.7.1 The DJI partial autocorrelation function and the corresponding confidence interval

Only one partial autocorrelation coefficient is significantly different from zero.

5.8 Autocorrelations evaluation methods

Now we have autocorrelations and partial autocorrelations calculated, and the confidence intervals defined, what else might be useful to know about these two data sets, in particular of autocorrelations? Well, in some textbooks and in numerous software printouts you will invariably get some further statistics associated with autocorrelations, so let's recreate them and provide explanations.

In Fig. 5.8.1 we provided such a typical printout (excluding the confidence interval, just to make the printout less cluttered) that describes various statistics that accompany autocorrelation coefficients.

In column D we have calculated the so called Q-statistic. This statistic is called Ljung-Box statistic and is also explained in Chapter 11. In this chapter we'll refer just to the brief method on how to calculate it. The equation for Q-statistic is:

$$Q_j = n(n+2)\sum_{j=1}^{k}\frac{r_j^2}{n-j} \qquad (5.8.1)$$

Where *n* is the number of observations, *j* is the number of lags and r_j are the autocorrelation coefficients.

The Q-Statistic follows the chi-square distribution and we will explain it fully in Chapter 11. In this chapter it is sufficient to just understand how it is calculated.

See how the equation (5.8.1) was translated into Excel "speak". The part after the summation where the autocorrelations are squared, divided by *(n-j)*, where *j* goes from 1 to *k*, looks a bit complicated, but it is not. We again used the =SUMPRODUCT() function to multiply the squared values of the autocorrelations r_j^2 with 1/(n-j), which just another way to see this part of equation (5.8.1). This enabled us to apply everything in one formula, easy to copy down the column. Fig. 5.8.2 shows this and all other formulae used.

In addition to this, column E gives us another set of values, the so-called probability values. These values are related to the Q-Statistic. They define the cumulative probability, up to every lag, that the calculated Q-stat values are not random. We already said that the Q-stat values are distributed in accordance with the chi-squared distribution, where the lag corresponds with the degrees of freedom. To calculate every value, we just use the =CHIDIST() function. Again, Fig 5.8.2 show how all of these statistics are calculated.

	A	B	C	D	E	F	G	H
1	Number of							
2	observations:		50			0.141421	=SE	
3								
4								
5	**Lag**	**ACF**	**PACF**	**Q-Stat**	**Prob**	**t-value**	**Prob for t**	**Decision**
6	1	0.7995	0.7995	33.9179	5.74875E-09	5.6534	0.000000797	Significant
7	2	0.6631	0.0662	57.7368	2.90151E-13	4.6890	0.000022274	Significant
8	3	0.5752	0.0758	76.0414	2.16717E-16	4.0675	0.000172113	Significant
9	4	0.5114	0.0500	90.8238	8.80033E-19	3.6162	0.000705138	Significant
10	5	0.3876	-0.1701	99.5036	6.72454E-20	2.7407	0.008531259	Significant
11	6	0.2759	-0.0719	104.0031	3.66192E-20	1.9512	0.056760440	Non-significant
12	7	0.1184	-0.2445	104.8501	1.07192E-19	0.8369	0.406715726	Non-significant
13	8	0.0404	0.0421	104.9513	4.1389E-19	0.2859	0.776132593	Non-significant
14	9	-0.0367	-0.0486	105.0369	1.50079E-18	-0.2597	0.796153414	Non-significant
15	10	-0.0695	0.0853	105.3511	4.60391E-18	-0.4916	0.625200542	Non-significant
16	11	-0.1154	0.0116	106.2386	1.0304E-17	-0.8159	0.418515504	Non-significant
17	12	-0.1052	0.1054	106.9957	2.34639E-17	-0.7438	0.460530963	Non-significant
18	13	-0.1046	0.0158	107.7643	5.11447E-17	-0.7395	0.463148512	Non-significant
19	14	-0.0888	-0.0261	108.3336	1.17663E-16	-0.6278	0.533038600	Non-significant
20	15	-0.0649	0.0349	108.6465	2.93189E-16	-0.4590	0.648293369	Non-significant
21	16	-0.0148	0.0129	108.6633	8.04653E-16	-0.1047	0.917026873	Non-significant
22	17	-0.0083	-0.0605	108.6687	2.14707E-15	-0.0586	0.953471151	Non-significant

Fig. 5.8.1 Autocorrelation related statistics for the example from Fig. 5.2.3

Just one brief digression, in cell E6, for example, we see the p-value of 5.748E-9. This means that we must put 9 zeros before 5748, so in fact the number in cell E6 is 0.000000005748. The same applies to all the numbers with the E suffix in notation.

	A	B	C	D	E	F	G	H
1								
2	Number		50			=1/SQRT(C2)	=SE	
3								
4								
5	Lag	ACF	PACF	Q-Stat	Prob	t-value	Prob for t	Decision
6	1	0.799513	0.799513	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B6)^2,1/($C$2-($A$6:A6)))	=CHIDIST(D6,A6)	=B6/$F$2	=TDIST(ABS(F6),$C$2-1,2)	=IF(G6<0.05,"Significant","Non-significant")
7	2	0.663123	0.066249	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B7)^2,1/($C$2-($A$6:A7)))	=CHIDIST(D7,A7)	=B7/$F$2	=TDIST(ABS(F7),$C$2-1,2)	=IF(G7<0.05,"Significant","Non-significant")
8	3	0.575231	0.075757	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B8)^2,1/($C$2-($A$6:A8)))	=CHIDIST(D8,A8)	=B8/$F$2	=TDIST(ABS(F8),$C$2-1,2)	=IF(G8<0.05,"Significant","Non-significant")
9	4	0.511404	0.050002	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B9)^2,1/($C$2-($A$6:A9)))	=CHIDIST(D9,A9)	=B9/$F$2	=TDIST(ABS(F9),$C$2-1,2)	=IF(G9<0.05,"Significant","Non-significant")
10	5	0.387593	-0.17014	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B10)^2,1/($C$2-($A$6:A10)))	=CHIDIST(D10,A10)	=B10/$F$2	=TDIST(ABS(F10),$C$2-1,2)	=IF(G10<0.05,"Significant","Non-significant")
11	6	0.275944	-0.07193	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B11)^2,1/($C$2-($A$6:A11)))	=CHIDIST(D11,A11)	=B11/$F$2	=TDIST(ABS(F11),$C$2-1,2)	=IF(G11<0.05,"Significant","Non-significant")
12	7	0.118354	-0.24451	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B12)^2,1/($C$2-($A$6:A12)))	=CHIDIST(D12,A12)	=B12/$F$2	=TDIST(ABS(F12),$C$2-1,2)	=IF(G12<0.05,"Significant","Non-significant")
13	8	0.040437	0.042140	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B13)^2,1/($C$2-($A$6:A13)))	=CHIDIST(D13,A13)	=B13/$F$2	=TDIST(ABS(F13),$C$2-1,2)	=IF(G13<0.05,"Significant","Non-significant")
14	9	-0.03673	-0.04861	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B14)^2,1/($C$2-($A$6:A14)))	=CHIDIST(D14,A14)	=B14/$F$2	=TDIST(ABS(F14),$C$2-1,2)	=IF(G14<0.05,"Significant","Non-significant")
15	10	-0.06952	0.085303	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B15)^2,1/($C$2-($A$6:A15)))	=CHIDIST(D15,A15)	=B15/$F$2	=TDIST(ABS(F15),$C$2-1,2)	=IF(G15<0.05,"Significant","Non-significant")
16	11	-0.11538	0.011598	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B16)^2,1/($C$2-($A$6:A16)))	=CHIDIST(D16,A16)	=B16/$F$2	=TDIST(ABS(F16),$C$2-1,2)	=IF(G16<0.05,"Significant","Non-significant")
17	12	-0.10519	0.105364	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B17)^2,1/($C$2-($A$6:A17)))	=CHIDIST(D17,A17)	=B17/$F$2	=TDIST(ABS(F17),$C$2-1,2)	=IF(G17<0.05,"Significant","Non-significant")
18	13	-0.10457	0.015824	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B18)^2,1/($C$2-($A$6:A18)))	=CHIDIST(D18,A18)	=B18/$F$2	=TDIST(ABS(F18),$C$2-1,2)	=IF(G18<0.05,"Significant","Non-significant")
19	14	-0.08878	-0.02609	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B19)^2,1/($C$2-($A$6:A19)))	=CHIDIST(D19,A19)	=B19/$F$2	=TDIST(ABS(F19),$C$2-1,2)	=IF(G19<0.05,"Significant","Non-significant")
20	15	-0.06490	0.034894	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B20)^2,1/($C$2-($A$6:A20)))	=CHIDIST(D20,A20)	=B20/$F$2	=TDIST(ABS(F20),$C$2-1,2)	=IF(G20<0.05,"Significant","Non-significant")
21	16	-0.01480	0.012924	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B21)^2,1/($C$2-($A$6:A21)))	=CHIDIST(D21,A21)	=B21/$F$2	=TDIST(ABS(F21),$C$2-1,2)	=IF(G21<0.05,"Significant","Non-significant")
22	17	-0.00829	-0.06052	=$C$2*($C$2+2)*SUMPRODUCT(($B$6:B22)^2,1/($C$2-($A$6:A22)))	=CHIDIST(D22,A22)	=B22/$F$2	=TDIST(ABS(F22),$C$2-1,2)	=IF(G22<0.05,"Significant","Non-significant")

Fig. 5.8.2 Probabilities for the Q-statistics calculated manually in column F

The way to understand the probability for Q-stat is to remind yourself of the hypothesis testing procedure. The H_0, i.e. the null hypothesis for Ljung-Box statistic, is that all the data (in this case the autocorrelations) are random. The H_1, i.e. the alternative hypothesis, is that the data are not random. We can arbitrarily decide that the level of significance is 0.05. In other words, we want to be 95% certain that we can reject the null hypothesis. The general principle that always

applies is: if the calculated probability value is SMALLER than the level of significance (in this case 0.05), we REJECT the null hypothesis.

In some textbooks, and software packages, we can also find another method of checking if the autocorrelations (or the partial autocorrelations) are significantly different from zero. This is achieved by using a simple t-test. The t-test statistic is calculated as:

$$t_{r_k} = \frac{r_k}{SE_{r_k}} \qquad (5.8.2)$$

Once we calculated the t-value for every autocorrelation, we need to test its significance, i.e. decide if the autocorrelation is significant (non-zero) or insignificant (virtually zero). Conventionally we would have to look in the table for the t-values, but in Excel we can just use the =TDIST() function.

Fig.5.8.2 shows how to calculate the t-values and the associated probabilities for every autocorrelation coefficient (columns F and G). In columns H we included a brief formula/descriptor to show us if the specific autocorrelation value is significant or not.

In summary, we just want to say that the autocorrelation coefficients together with the partial autocorrelations provide essential assistance to characterize the time series and select the correct model for forecasting. The chapters that follow will show us the practical value of these two functions and how to use them. This chapter is focused on simple mechanics of how to produce these two functions.