

## Autocorrelation

A large number of measurements are collected over a period of time. Stock prices, quantities sold, student enrollments, grade point averages, etc. may vary systematically across time periods. Variations may reflect trends which repeat by week, month or year. For example, a grocery item may sell at a fairly steady rate on Tuesday through Thursday but increase or decrease on Friday, Saturday, Sunday and Monday. If we were examining product sales variations for a product across the days of a year, we might calculate the correlation between units sold over consecutive days. The data might be recorded simply as a series such as “units sold” each day. The observations can be recorded across the columns of a grid or as a column of data in a grid. As an example, the grid might contain:

CASE/VAR	Day	Sold
Case 1	1	34
Case 2	2	26
Case 3	3	32
Case 4	4	39
Case 5	5	29
Case 6	6	14
...		
Case 216	6	15
Case 217	7	12

If we were to copy the data in the above “Sold” column into an adjacent column but starting with the Case 2 data, we would end up with:

CASE/VAR	Day	Sold	Sold2
Case 1	1	34	26
Case 2	2	26	32
Case 3	3	32	39
Case 4	4	39	29
Case 5	5	29	14
Case 6	6	14	11
...			
Case 216	6	15	12
Case 217	7	12	-

In other words, we repeat our original scores from Case 2 through case 217 in the second column but moved up one row. Of course, we now have one fewer case with complete data in the second column. We say that the second column of data “lags” the first column by 1. In a similar fashion we might create a third, fourth, fifth, etc. column representing lags of 2, 3, 4, 5, etc.. Creating lag variables 1 through 6 would result in variables starting with sales on days 1 through 7, that is, a week of sale data. If we obtain the product-moment correlations for these seven variables, we would have the correlations among Monday sales, Tuesday Sales, Wednesday Sales, etc. We note that the mean and variance are best estimated by the lag 0 (first column) data since it contains all of the observations (each lag loses one additional observation.) If the sales from day to day represent “noise” or simply random variations then we would expect the correlations to be close to zero. If, on the other hand, we see an systematic increase or decrease in sales between say, Monday and Tuesday, then we would observe a positive or negative correlation.

In addition to the inter-correlations among the lagged variables, we would likely want to plot the average sales for each. Of course, these averages may reflect simply random variation from day to day. We may want to “smooth” these averages to enhance our ability to discern possible trends. For example, we might want the average of day three to be a weighted average of that day plus the previous two day sales. This “moving average” would tend to smooth random peaks and valleys that occur from day to day.

It is also the case that an investigator may want to predict the sales for a particular day based on the previous sales history. For example, we may want to predict day 8 sales given the history of previous seven day sales.

Now let us look at an example of auto-correlation. We will use a file named strikes.tab. The file contains a column of values representing the number of strikes which occurred each month over a 30 month period. Select the auto-correlation procedure from the Correlations sub-menu of the Statistics main menu. Below is a representation of the form as completed to obtain auto-correlations, partial auto-correlations, and data smoothing using both moving average smoothing and polynomial regression smoothing:

**Autocorrelation**

Directions: Select a variable to analyze. You may analyze series from either a column (default) variable or a "Case" row. You may elect to analyze all values in a column (or row) as desired. Click the buttons for any desired smoothing options. The program will automatically "split" the list of row values (or column values) for that variable into two sub-sets of X and Y scores with each Y score being the value which "lags" behind the X score in the list by k lag values. All possible lags which yield a sample as large as 3 or more are computed and plotted in a "Correlogram". You may optionally print the lag, correlation, means, standard deviations and confidence interval for each correlation. The differences between original and smoothed values (residuals) may be plotted. The smoothed points replace the original values in the analysis if smoothing is elected.

The Series is coded in:  
☒ A Grid Column    ☐ A Row of the Grid:

Available Variables:  
 z  
 VAR3

Selected Variable:  
 VAR00001

Alpha Level: 0.05  
 Maximum Lag: 12

Include Cases:  
☒ All Cases  
☐ Only Cases From:  
 To:

Projection Option:  
☒ Project 5 Points.

Analysis / Output Options:  
☒ Correlogram  
☒ Statistics  
☒ Print correlation mat.  
☒ Print Partial autocorr.  
☐ Yule-Walker Coef.s  
☒ Residual Plot

Data Smoothing:  
☐ Center on Mean  
☐ Difference Smooth  
☒ Moving Avg. Smooth  
☐ Exponentially Smooth  
☐ Fourier Filter Smooth  
☒ Poly. Reg. Smooth  
☐ Mult. Reg. Smooth

Reset    Cancel    Compute    Return

**Figure 1. The Autocorrelation Dialog**

When we click the Compute button, we first obtain a dialog form for setting the parameters of our moving average.

In that form we first enter the number of values to include in the average from both sides of the current average value. We selected 2. Be sure and press the Enter key after entering the order value. When you do, two theta values will appear in a list box. When you click on each of those thetas, you will see a default value appear in a text box. This is the weight to assign the leading and trailing averages (first or second in our example.) In our example we have accepted the default value for both thetas (simply press the Return key to accept the default or enter a value and press the Return key.) Now press the Apply button. When you do this, the weights for all of the values (the current mean and the 1, 2, ... order means) are recalculated. You can then press the OK button to proceed with the process.

**MoveAvgFrm**

Directions: Enter the order of the moving average.  
 The order is the number of values on each side of a point to be included in the average. When you enter a value, a list of corresponding thetas will appear in the list. Click on each theta of the list for entry of the desired weight (default 1.0)  
 Enter a weight in the theta value box and press the return key. Repeat for each theta in the list.  
 Click the Apply button when ready. The theta values will be re-proportioned to sum to 1.0 across all values. Click the OK button to continue.

Order:

Theta Value:

Cancel

Reset

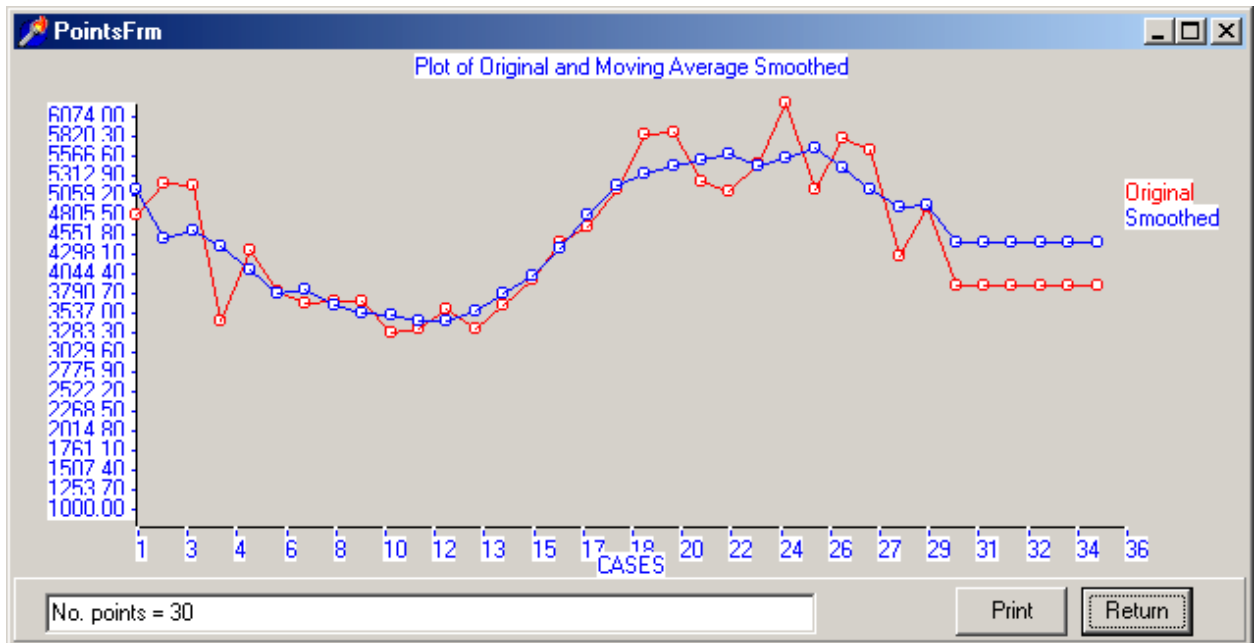
Apply

OK

Theta(1) = 0.2  
 Theta(2) = 0.2  
 Theta(3) = 0.2

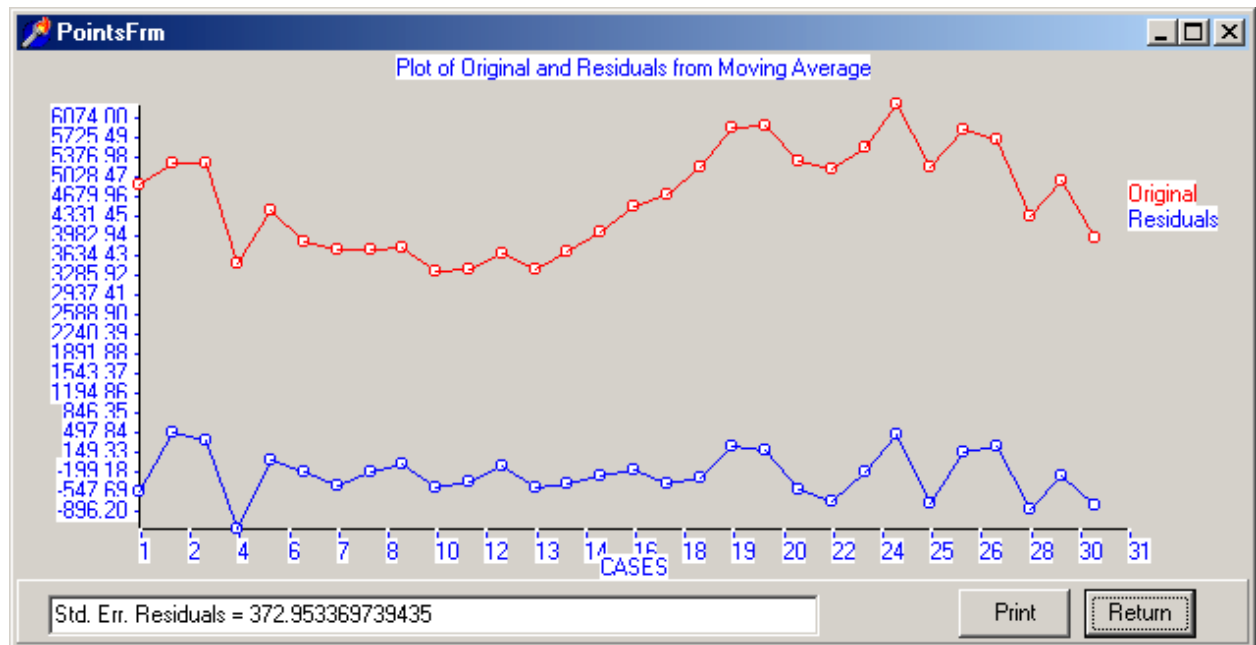
**Figure 2. The Moving Average Dialog**

The procedure then plots the original (30) data points and their moving average smoothed values. Since we also asked for a projection of 5 points, they too are plotted. The plot should look like that shown below:



**Figure 3. Plot of Smoothed Points Using Moving Averages**

We notice that there seems to be a “wave” type of trend with a half-cycle of about 15 months. When we press the Return button on the plot of points we next get the following:



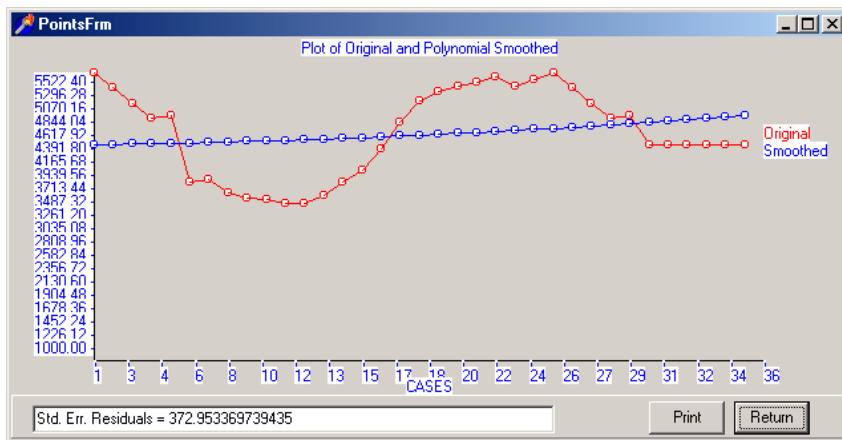
**Figure 4. Plot of Residuals Obtained Using Moving Averages**

This plot shows the original points and the difference (residual) of the smoothed values from the original. At this point, the procedure replaces the original points with the smoothed values. Press the Return button and you next obtain the following:

Figure 5 is a dialog box titled "Polynomial Reg. Smoothing". It contains a text box with the following text: "Directions: In polynomial regression smoothing, the value of a point y at a given time t is estimated by the sum of regression weights times t raised to a power of 1, 2, etc. up to the order specified. Enter the order and click the OK button." Below this text is a label "Polynomial order:" followed by a text box containing the number "2". At the bottom of the dialog are "Cancel" and "OK" buttons.

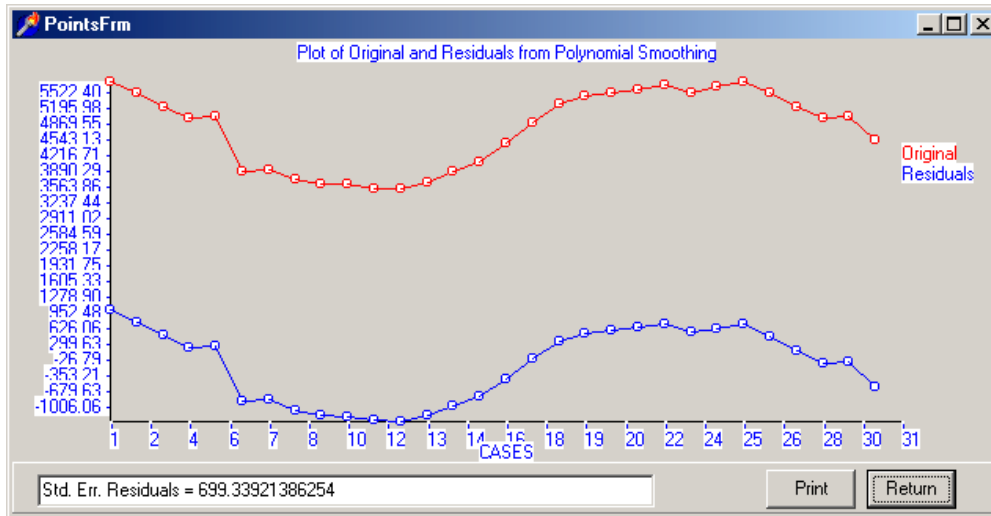
**Figure 5. Polynomial Regression Smoothing Form**

This is the form for specifying our next smoothing choice, the polynomial regression smoothing. We have elected to use a polynomial value of 2 which will result in a model for a data point  $Y_{t-1} = B * t^2 + C$  for each data point. Click the OK button to proceed. You then obtain the following result:



**Figure 6. Plot of Polynomial Smoothed Points**

It appears that the use of the second order polynomial has “removed” the cyclic trend we saw in the previously smoothed data points. Click the return key to obtain the next output as shown below:



**Figure 7. Plot of Residuals from Polynomial Smoothing**

This result shows the previously smoothed data points and the residuals obtained by subtracting the polynomial smoothed points from those previous points. Click the Return key again to see the next output shown below:

Overall mean = 4532.604, variance = 11487.241

Lag	Rxy	MeanX	MeanY	Std.Dev.X	Std.Dev.Y	Cases	LCL	UCL
0	1.0000	4532.6037	4532.6037	109.0108	109.0108	30	1.0000	1.0000
1	0.8979	4525.1922	4537.3814	102.9611	107.6964	29	0.7948	0.9507
2	0.7964	4517.9688	4542.3472	97.0795	106.2379	28	0.6116	0.8988
3	0.6958	4510.9335	4547.5011	91.3660	104.6337	27	0.4478	0.8444
4	0.5967	4504.0864	4552.8432	85.8206	102.8825	26	0.3012	0.7877
5	0.4996	4497.4274	4558.3734	80.4432	100.9829	25	0.1700	0.7287
6	0.4050	4490.9565	4564.0917	75.2340	98.9337	24	0.0524	0.6679
7	0.3134	4484.6738	4569.9982	70.1928	96.7340	23	-0.0528	0.6053
8	0.2252	4478.5792	4576.0928	65.3196	94.3825	22	-0.1470	0.5416
9	0.1410	4472.6727	4582.3755	60.6144	91.8784	21	-0.2310	0.4770
10	0.0611	4466.9544	4588.8464	56.0772	89.2207	20	-0.3059	0.4123
11	-0.0139	4461.4242	4595.5054	51.7079	86.4087	19	-0.3723	0.3481
12	-0.0836	4456.0821	4602.3525	47.5065	83.4415	18	-0.4309	0.2852

In the output above we are shown the auto-correlations obtained between the values at lag 0 and those at lags 1 through 12. The procedure limited the number of lags automatically to insure a sufficient number of cases upon which to base the correlations. You can see that the upper and lower 95% confidence limits increases as the number of cases decreases. Click the Return button on the output form to continue the process.

Matrix of Lagged Variable: VAR00001 with 30 valid cases.

Variables	Lag 0	Lag 1	Lag 2	Lag 3
Lag 4				
Lag 0	1.000	0.898	0.796	0.696
Lag 1	0.898	1.000	0.898	0.796
Lag 2	0.796	0.898	1.000	0.898
Lag 3	0.696	0.796	0.898	1.000
Lag 4	0.597	0.696	0.796	0.898
Lag 5	0.500	0.597	0.696	0.796
Lag 6	0.405	0.500	0.597	0.696
Lag 7	0.313	0.405	0.500	0.597
Lag 8	0.225	0.313	0.405	0.500
Lag 9	0.141	0.225	0.313	0.405
Lag 10	0.061	0.141	0.225	0.313
Lag 11	-0.014	0.061	0.141	0.225
Lag 12	-0.084	-0.014	0.061	0.141

Variables

	Lag 5	Lag 6	Lag 7	Lag 8
Lag 9				
Lag 0	0.500	0.405	0.313	0.225
0.141				
Lag 1	0.597	0.500	0.405	0.313
0.225				
Lag 2	0.696	0.597	0.500	0.405
0.313				
Lag 3	0.796	0.696	0.597	0.500
0.405				
Lag 4	0.898	0.796	0.696	0.597
0.500				
Lag 5	1.000	0.898	0.796	0.696
0.597				
Lag 6	0.898	1.000	0.898	0.796
0.696				
Lag 7	0.796	0.898	1.000	0.898
0.796				
Lag 8	0.696	0.796	0.898	1.000
0.898				
Lag 9	0.597	0.696	0.796	0.898
1.000				
Lag 10	0.500	0.597	0.696	0.796
0.898				
Lag 11	0.405	0.500	0.597	0.696
0.796				
Lag 12	0.313	0.405	0.500	0.597
0.696				

#### Variables

	Lag 10	Lag 11	Lag 12
Lag 0	0.061	-0.014	-0.084
Lag 1	0.141	0.061	-0.014
Lag 2	0.225	0.141	0.061
Lag 3	0.313	0.225	0.141
Lag 4	0.405	0.313	0.225
Lag 5	0.500	0.405	0.313
Lag 6	0.597	0.500	0.405
Lag 7	0.696	0.597	0.500
Lag 8	0.796	0.696	0.597
Lag 9	0.898	0.796	0.696
Lag 10	1.000	0.898	0.796
Lag 11	0.898	1.000	0.898
Lag 12	0.796	0.898	1.000

The above data presents the inter-correlations among the 12 lag variables. Click the output form's Return button to obtain the next output:

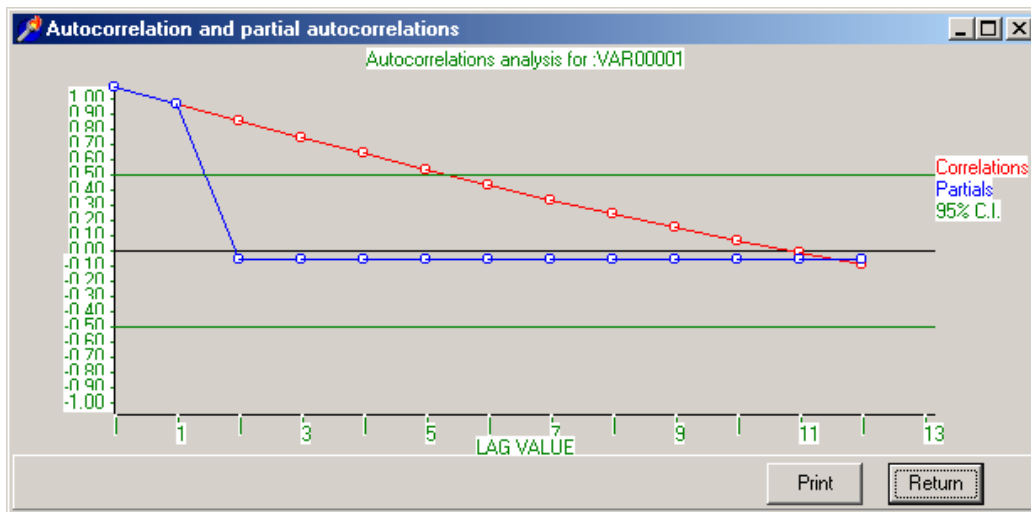
Partial Correlation Coefficients with 30 valid cases.

Variables	Lag 0	Lag 1	Lag 2	Lag 3	
Lag 4	1.000	0.898	-0.051	-0.051	-
0.052					



Variables	Lag 5	Lag 6	Lag 7	Lag 8	
Lag 9	-0.052	-0.052	-0.052	-0.052	-
0.051					
Variables	Lag 10	Lag 11			
	-0.051	-0.051			

The partial auto-correlation coefficients represent the correlation between lag 0 and each remaining lag with previous lag values partialled out. For example, for lag 2 the correlation of -0.051 represents the correlation between lag 0 and lag 2 with lag 1 effects removed. Since the original correlation was 0.796, removing the effect of lag 1 made a considerable impact. Again click the Return button on the output form. Next you should see the following results:



**Figure 8. Auto and Partial Autocorrelation Plot**

This plot or “correlogram” shows the auto-correlations and partial auto-correlations obtained in the analysis. If only “noise” were present, the correlations would vary around zero. The presence of large values is indicative of trends in the data.