TIME SERIES STUDENT PROJECTS: TIME SERIES TECHNIQUES

(The attached PDF file has better formatting.)

Updated: May 1, 2008

The SOA requires independent student projects for the regression analysis and time series courses.

- Statistics is not book knowledge alone. Knowing the characteristics of ARIMA models does not make one a statistician.
- Candidates must show they can apply the statistical techniques to empirical data to receive VEE credit for the statistics courses.

The student projects use

- Correlograms, Durbin-Watson statistics, Box-Pierce Q statistics, and Bartlett's test to specify, estimate, and diagnose ARIMA models.
- Multiple linear regression, residual plots, F tests, and dummy variables, plus
- Analyses of serial correlation, heteroscedasticity, and multicollinearity to formulate and test hypotheses.

The on-line courses do not require candidates to buy a statistical package, such as SAS or Minitab, which has these techniques. Buying a statistics package would add several hundred dollars to your cost. You may use Excel or similar spread-sheet software.

Excel is a powerful modeling package that handles number crunching, statistics, data base queries, and graphing. Its built-in functions and add-in features (data analysis and solver) provide most of the functions needed for the statistics courses. VBA macros can handle all the remaining statistical techniques you are likely to use. Most candidates use Excel at work, and they can immediately apply the techniques to data sets from public web sites.

Experienced statisticians often use *R*, a free on-line statistics package, which two of our faculty advisors use in their university courses. *R* is not user-friendly and has a steep learning curve. If you work with statistics in your actuarial jobs, the three to four weeks needed to become familiar with *R* are worthwhile. For most candidates, a good knowledge of Excel, its add-ins, and its VBA capabilities provide sufficient statistical software.

We do not provide VBA macros that do all the work for the student projects. ARIMA modeling is highly subjective. Based on the attributes of the time series, you may specify periods, adjust for seasonality, take first differences, or otherwise re-form the data. The SOA wants to ensure that candidates receiving VEE credit for statistics can apply the techniques to real data, not just run VBA macros that NEAS provides.

Take heed: For some items, a VBA macro is much more efficient. Forming a correlogram from 50,000 records of daily temperature using simple cell formula may run slowly on your machine. If you are familiar with VBA, do the computations in a macro.

We provide a VBA macro for sample autocorrelations (correlograms) and the Box-Pierce Q statistic. Learn first the cell formulas, so you can modify the statistical tests for your student project. The explanation of the VBA macro is in the project template for weather student projects (daily temperature), which use time series of 40,000+ values.

STATISTICAL TECHNIQUES

The time series student project requires linear regression, correlograms, residuals, and the Box-Pierce Q statistic. We provide Excel work-sheets with the time series techniques.

We use the *REGRESSION* section of the *DATA ANALYSIS* add-in for the linear regression and the residuals. We explain this add-in for the regression analysis student project.

Excel has a *CORREL* built-in function (correlation) but no sample autocorrelation function. This posting explains the difference between them. We do not require candidates to code the cell formulas for the sample autocorrelation function. The correlogram sheet on the illustrative work provide the cell formulas.

- Read this documentation so you understand the cell formulas.
- Copy the cell formulas to the work-sheet with your time series.
- You must replace the parameter showing the number of observations.

We have kept the cell formulas simple. If you have never used the *OFFSET* built-in function, read this documentation carefully.

Copy code from these worksheets for your student project. This course is statistics, not programming. You must know enough Excel to use the techniques, form graphs, and copy the output to your written document. If you had to write the Excel functions yourself, you would spend hours writing code. This is not an efficient use of your time.

If you know VBA or if you have worked with Excel data bases, use macros and data base techniques to simplify your project work. You may also use SAS or MINITAB or any other package. You do not have to write or create any functions from scratch.

ARIMA uses partial autocorrelation function and nonlinear regression for processes with a moving average component. We do *not* provide functions or macros for these tasks. Nonlinear regression is not covered in the regression analysis or time series course.

- If the objective of the student project were to find the optimal ARIMA model for a time series, you need nonlinear regression to optimize the moving average part.
- The student project shows that you can apply the course concepts to real data. The courses do not cover nonlinear regression, so we do not use it in the student project.

The course does not require knowledge of Excel. You can use another spread-sheet facility or a statistical package (SAS, Minitab, R) for the student project. You must know some package. We assume that all candidates know how to use Excel or similar packages.

Take heed: This documentation does not fully explain Excel built-in functions or VBA macros. If the cell formulas or the VBA macro is not clear to you, use the on-line help or Excel manuals for more explanation. Then post a question on the discussion forum.

SAMPLE AUTOCORRELATIONS AND CORRELOGRAMS

The time series illustrative worksheets use sample autocorrelation functions, correlograms, the Durbin-Watson statistic, and the Box-Pierce Q statistic. None of these are built into Excel, but all of them are easily coded.

Take heed: The Excel built-in function *CORREL* computes the correlation of two random variables. The sample autocorrelations are not the same as the correlations used by Excel's *CORREL* built-in function. The differences are subtle, but they are important for statistical testing. The illustrative worksheet show the correlogram in three steps to clarify the differences between statistical correlations and sample autocorrelations.

~ The CORREL built-in function shows the shape of the correlogram, but it does *not adjust for degrees of freedom* and has *random fluctuations* that distort the statistical tests.
~ We can adjust the CORREL built-in function for degrees of freedom.
~ We use the SUMPRODUCT built-in function and the exact formula. This corrects for degrees of freedom and the random fluctuations.

The student project should use the exact formula. The two simpler versions help you grasp the intuition: why the correl built-in function is biased and how to correct.

Take heed: The sample autocorrelation formula still has a slight bias. The bias is too small to affect your results, and you need not concern yourself with it.

For your student project, copy the code using the SUMPRODUCT built-in function and the exact formula. Change the code to adjust for the number of observations.

- The code is general, and you can copy the code from one cell to any other cell. It uses Excel's *OFFSET* built-in function and a combination of relative and absolute references.
- If you have never used Excel's *OFFSET* built-in function and combinations of relative and absolute references, review the on-line help facility for these items. You can complete the student project even if you do not know how these items work, but you are likely to make errors. Ten minutes reading the on-line help may save you hours.

The correlogram is a chart. Most candidates do not need instructions for Excel charts. Use the chart wizard (if necessary) to create correlograms.

- We do not require a fancy chart.
- A line chart or a bar chart using the Excel chart wizard is sufficient.
- Create labels using the chart wizard. Edit the labels to make sure they are clear.
- If you have never used charts in Excel, you can document the chart in your write-up.

If you know how, add features to make your chart clear. The guidance here is a minimum.

Illustration: To show seasonality in a correlogram, use markers or arrows for the seasonal autocorrelations. If you can't do so, explain in the write-up how the correlogram shows the seasonality.

THE ILLUSTRATIVE WORKSHEET

Illustration: We use the last 42 months (3½ years) of 90 day Treasury bill interest rates on the NEAS web site: January 1997 – June 2000. This example shows how to use the Excel *CORREL*, *SUMPRODUCT*, and *OFFSET* built-in functions and the chart wizard.

- Copy Treasury bill interest rates for January 1997 through June 2000 from the spreadsheet on the NEAS web site to a blank spreadsheet.
- Place the interest rates in cells B11:B52. Leave one column on the left for line labels and ten rows on top for headers and documentation.

Enter Jan 1997 in cell A11 and Feb 1997 in cell A12. Select both cells and drag down to cell A52. Excel's *AUTOFILL* puts the proper months in these cells. Column A documents the work.

Take heed: The NEAS faculty members review the student projects. The relevant graphs and charts should be copied to your write-up, with references to the Excel spread-sheet where they are formed. If the reviewer can not understand what the worksheet does, we send it back to you for better documentation. This adds 4 - 6 weeks to the grading process. For your own sake, keep your work-sheets clean, and provide a clear write-up.

Format the cells with two decimal places. The formatting aligns the rates at the decimal place and makes them easier to read.

Enter the values 1 and 2 in cells C11 and C12. Select both cells, place the cursor at the lower right corner o f cell C12, and drag down to cell C52. This forms a column with the value 1 through 42, using Excel's *AUTOFILL* procedure. Use this index, not the month names, for Excel's *OFFSET* function.

Place column headers in cells B10, C10, D9 and D10, as

~ B10: interest rate
~ C10: lag
~ D9: sample
~ D10: autocorrelation

(If you know how, join cells D9 and D10. Joining the cells makes the column headers look better; it is not necessary.)

We create correlograms three ways.

- Use the *third* method for the student project.
- The first two methods explain why the Excel built-in *CORREL* function does not give the sample autocorrelations. Follow the steps here, and you will avoid errors in your code.

Note: Degrees of freedom are discussed in the regression analysis course. You do not have to understand the statistical theory in this posting. But you must copy the cell function for the sample autocorrelation to your work-sheet, and you must change the parameter for the number of observations in your time series.

Method 1: Built-in *CORREL* Function

Begin by entering the full correlation formula separately for two cells, so you see how the lag affects the cell formula. Then use the *OFFSET* function to simplify.

In Cell D11, enter the code =*CORREL(B11:B51,B12:B52)*. This is the correlation with a one period lag: the correlation of the first 41 values with the 41 values lagged one month. Use the first ten rows for comments: document your steps as you do them, and then copy the documentation to your write-up.

In Cell D12, enter the code =*CORREL(B11:B50,B13:B52)*. This is the correlation with a two period lag: the correlation of the first 40 values with the 40 values lagged 2 months.

Format the values in column D with four decimal places.

This procedure uses separate code for each cell. The final method writes the code once.

Take heed: Write the code by hand or select the cells. Write the name of the function or use the function wizard to select the function.

WRITING THE CODE ONCE: THE OFFSET BUILT-IN FUNCTION

A student project on daily temperature may have 50,000 observations. A student project on over-night LIBOR rates may have 3,500 observations. We can't write separate code for each cell.

Take heed: If you are not familiar with the *OFFSET* function, jot down on scrap paper the values in cells D11 and D12. Compare these two values with the values from the final version to make sure you have not made an error.

Erase the formula in cell D12. Change the formula in cell D11 to

=*CORREL(B$11:B51,B12:B$52)*.

The dollar sign makes the row number absolute. The formula asks for the correlation of the 41 values starting cell B11 with the 41 values ending cell B52

Copy this formula from cell D11 to cell D12. We get =*CORREL(B$11:B52,B13:B$52)*. This asks for the correlation of the *42* values starting in cell B11 with the 40 values ending in cell

B52. This is incorrect. We want the formula =*CORREL(B$11:B50,B13:B$52)*: the correlation of the *40* values starting in cell B11 with the 40 values ending in cell B52.

We use the *OFFSET* function. Write the formula in cell D11 as

=CORREL(OFFSET(B$11,0,0,42-C11,1),B12:B$52)

The *OFFSET* function has five parameters. The formula here selects a range that

1. Begins in cell B$11 with an offset of 0 rows and 0 columns
2. Has a height of 42 – C11 rows and a width of 1 column.

The height of 42 – C11 is 42 – 1 = 41 for the correlation of lag 1. (Column C has the lags.) The formula is a relative reference. It changes to 42 – C12 for the next row = 42 – 2 = 40.

Copy this formula from cell D11 to cells D12 through D49. We don't use the last three cells:

- The correlation of lag N-2 (lag 40) is 1 or –1. This is the correlation of (x_1, x_2) with (y_1, y_2). The *CORREL* function gives a value, but this figure has no meaning. [Use the definition of the correlation from Module 1 of this course to find this value.]
- The autocorrelation of lag N-1 (lag 41) has a division by zero. This is the correlation of a scalar *x* with a scalar *y*.
- The correlations of lag N (lag 42) and higher are undefined. There are no values in the time series with lags this great.

The relative cell references adjust to the proper values for each correlation. Examine the formulas in the first several cells, so you see how the formulas change.

Compare your values with those on the illustrative worksheet. If they differ, review this documentation to find the error.

We use the *OFFSET* function for several of the statistical techniques. Excel has several alternatives to this function as well as VBA code that replicates it.

Take heed: This sample autocorrelation function has the proper shape, but the values differ slightly from those in the exact function. We form a correlogram and explain the problems with this sample autocorrelation function.

The cells in this column use the Excel built-in functions. Make sure you understand the *OFFSET* function and the correlation used for each lag.

Use the exact formula for your student project (Method #3). We use several columns for Method #3, but if you proceed through the steps here, you should understand the logic. Copy the code for Method #3 from the illustrative worksheet to your student project.

CORRELOGRAM

The illustrative work-sheet has three methods to form correlograms. Use the third method for your student project. The first two methods explain the logic of correlograms.

~ One correlogram is formed directly from the Excel *CORREL* built-in function. This is the easiest correlogram to form, but it does not adjust for degrees of freedom and it has large random fluctuations at late periods. It is biased and can not be used for Bartlett's test or the Box-Pierce Q statistic.
~ A second correlogram adjusts for degree of freedom. It uses the *CORREL* built-in function and multiplies by (N – k) / N. It removes most of the bias in the first method. These first two correlograms show the difference between the sample autocorrelation and the correlation. Use the third correlogram for the student project.
~ A third correlogram uses the formula needed for the Box-Pierce Q statistic. The Excel work-sheet uses the SUMPRODUCT built-in function, not the CORREL built-in function.

Take heed: When you copy the cell functions, be sure to adjust the number of observations in the time series.

Form correlograms from the sample autocorrelation function. You may form half a dozen correlograms in your student project, corresponding to different versions of the time series.

The illustrative worksheet uses Excel's chart wizard. You may prefer to form charts directly, without the chart wizard. These instructions are for candidates not familiar with Excel.

Select cells C11:D49. Click on the *CHART WIZARD* and choose a *LINE GRAPH*. (You may also use bar graphs for the correlogram. Use whatever seems clearest.)

On the second wizard menu, the data range should be D11:D49, not C11:D49. Make this change manually. Alternatively, select cells D11:D49. Cells C11:C49 are your x-axis, which are used in other charts.

You can re-format any parts of the graph in the chart wizard or after making the chart.

- Eliminate the *LEGEND* on the right hand side, or rename it as sample autocorrelations.
- Give a title to the correlogram, such as *Correlogram of Interest Rate Time Series*.
- Label the axes, such as *Month Lag* for the horizontal axis and *Sample Autocorrelation* for the vertical axis.

Recommendation: These instructions form a simple chart. If you use Excel regularly, add titles, legends, markers, and other documentation to your chart before copying it to Word.

DEGREES OF FREEDOM

The degrees of freedom affects the denominator of the correlation formula. A higher lag means fewer data points and fewer degrees of freedom.

Copy the lags from Column D to Column F. This is not essential, but new Excel users may find it easier to format the chart if the lags are next to the autocorrelations.

Recommendation: Learn Excel's Chart options. Your student project uses many charts and graphs, and graphics are equally useful for other actuarial reports. An hour spent learning to form charts will save you many hours of later work.

Enter the formula =D11*(42-F11)/42 in Cell G11. Copy cell G11 to Cells G12:G49.

Each cell in Column G (the revised correlations) is the Column D value × (N – k) / N. This correlogram does not have the large random fluctuations at high lags and the decay is closer to a straight line. The *shape* of the correlogram remains the same.

EXACT SAMPLE AUTOCORRELATIONS

The exact sample autocorrelation function differs slightly from the correlations above.

Take heed: The illustrative work-sheet uses simple Excel functions. The cells formulas can be made more efficient, but the formulas here are easier to understand.

Replace the interest rates by their deviations.

- Place the average interest rate in cell B9. Cell I11 has the formula =B11-B$9.
- Copy cell I11 to the rest of this column: cells I12:I52.

Use the *SUMPRODUCT* built-in function for Column J, not the *CORREL* built-in function. Use the *OFFSET* function in the same fashion as for the *CORREL* built-in function.

- Cell J11 has =*SUMPRODUCT(OFFSET(I$11,0,0,42-C11,1),I12:I$52)*.
- Copy this formula to cells J12:J52.

Take heed: It is easy to make an error with the *OFFSET* function. To avoid errors, compute the *SUMPRODUCT* function for two or three cells by explicitly referencing all the cells, and compare this value with the results using the *OFFSET* function.

Enter the cell formula =I11^2 in Cell K11. Copy this formula to Cells K:12:K52. Enter the cell formula =SUM(K11:K52) in Cell K9.

Column K has squares of the interest rate deviations. Cell K9 has the sum of the squares.

Enter the formula =J11/K$9 in Cell L11.

The sample autocorrelations in column L are the *SUMPRODUCT* in Column J divided by the sum of the squares in cell K9.

Form a correlogram from the sample autocorrelations in Column L. This correlogram is smoother than the other correlograms.

- The correlogram using the CORREL built-in function has fewer terms in the denominator, so it is more distorted by random fluctuations.
- The adjustment for degrees of freedom adjusts the magnitude of the sample autocorrelations, but keeps the distortions from random fluctuations.
- The exact formula has more terms in the denominator, and it is less distorted by random fluctuations.

Summary: This worksheet explains how to form the correlogram and why it differs from ordinary correlations. For your student project, use the code for the exact autocorrelation formula. You do not have to explain the code in your student project.

Take heed: This code is written for new Excel users. Experienced Excel users may prefer other cell formulas. You may write a VBA macro to form a correlogram from any sequence of data points. If you know VBA, this macro will save time and prevents typos.

The SOA wants candidates to show they can work with the statistical techniques. We don't automate the steps of the student project with macros, so that you work with the figures.

USE OF THE CORRELOGRAM

Your student project forms a correlogram for each time series. Know how to interpret the sample autocorrelation function and the correlogram.

Take heed: The sample autocorrelation function and the correlogram have many uses. Statisticians differ on the implications of a sample autocorrelation function. We explain how we might interpret the sample autocorrelations of these 42 time series observations:

~ The statistical techniques are used in combination. We use these 42 points for the correlogram, an AR(1) process, the Durbin-Watson statistic, and the Box-Pierce Q statistic. Some implications are ambiguous.
~ The analysis in this workbook is incomplete. We explain what else to do for the student project, but we don't show every part. Using the data of 42 months, we should examine first differences, second differences, an AR(2) model, and perhaps an ARMA(1,1) process. An ARMA(1,1) process is not easy to fit using standard Excel functions, so this process is not required for the student project.
~ The textbook tries several complex models for interest rates, such as ARIMA(8,1,4). The student project does not require you to construct complex ARIMA models.
~ Statisticians differ in their interpretations of the statistical techniques, results, and plots. We give several explanations for the results here.
~ The results differ by time period. You may compare ARIMA models for one time period or one ARIMA model for two or three time periods. Dividing these 42 observations into two time series gives a different result.

Recommendation: When you begin the student project, do not worry about ARMA(1,1) processes. You can fit them several ways:

- Use a statistical package, such as SAS, MINITAB, or "R"
- Use the Yule-Walker equations to get an estimate of θ_1.
- Use the *SOLVER* add-in and have Excel iterate for the solution.

After fitting AR(1) and AR(2) processes (or ARIMA(1,1,0) and ARIMA(2,1,0) processes), fit an MA(1) or ARIMA(0,1,1) process, using the Yule-Walker equations for an estimate of θ_1. If you find this easy, fit an ARMA(1,1) or ARIMA(1,1,1) process. If this is not easy, do not worry about the ARMA(1,1) and ARIMA(1,1,1) processes. The student project gives you experience with using statistical techniques. New Excel users are not expected to code the sophisticated formulas needed for these processes.

STATIONARITY

The correlogram drops to zero by a lag of 7 months. Similar correlograms appear in many student projects. Some candidates mistakenly infer that the time series is a stationary AR(1) process. Know the principles:

- The sample autocorrelations from a stationary AR(1) process decline geometrically to zero. It stays zero at subsequent lags, with random fluctuations that depend on the length of the time series.
- The sample autocorrelation here drops steadily to –47% by lag 14. The sample autocorrelation does not remain close to zero until lag 22.

Take heed: Some correlograms have the following form:

- Decline steadily to zero by lag N.
- Continue declining to a minimum by lag 2N.
- Rise to zero by lag 3N.

A stationary AR(1) process does not do this. A stationary process has autocorrelations close to zero after a few lags.

Illustration: An AR(1) process with ϕ_1 of 50% has an autocorrelation of $0.5^{10} = 0.10\%$, or a tenth of a percent after ten lags. With random fluctuations of perhaps 5% at each lag, we do not see a pattern in the autocorrelations after 4 or 5 lags.

Large ϕ_1 parameters do not give the pattern in this correlogram.

- An AR(1) process with ϕ_1 of 95% has an autocorrelation of $0.95^{14} = 48.77\%$, not –47%, at 14 lags.
- An AR(1) process with ϕ_1 of –95% has an oscillatory pattern about the mean, which is not the pattern in this time series.

The pattern here reflects a change in the mean or trend of the time series. A change in the trend is a change in the mean of the first differences.

Exercise 1.2: Change in Mean of Time Series

We simulate a time series of 100 observations.

- Observations 1 – 50 are random draws from a normal distribution with a mean of 2 and a standard deviation of 1.
- Observations 51 – 100 are random draws from a normal distribution with a mean of –2 and a standard deviation of 1.

A. Graph the time series.

B. For each era, form a correlogram. The sample autocorrelations have a normal distribution, with a mean of zero and a standard deviation of $1/\sqrt{100} = 0.100$.
C. Form the correlogram from the entire series. The mean is zero, and the standard deviation is about 2. The distribution is not normal.
D. Consider the sample autocorrelation of lag 1. The values for the pairs (1,2), (2,3), …, (49, 50), (51, 52), (52, 53), …, (99,100) are highly positive. The value for the pair (50, 51) is highly negative. The average for these 99 pairs is highly positive.
E. Consider the sample autocorrelation of lag 2. The values for the pairs (1,3), (2,4), …, (48, 50), (51, 53), (52, 54), …, (98,100) are highly positive. The values for the pairs (49, 51), (50, 52) are highly negative. The average for these 98 pairs is highly positive, but lower than the average for the 99 pairs of lag 1.

Graphs, Intuition, and Correlograms (Sample Autocorrelations)

The student project shows you can apply statistical techniques to empirical data and interpret the results to model a time series. Your written report shows three elements:

- The correlogram for the initial time series and any adjusted time series (first and second differences, seasonal averages, moving averages, logarithms).
- Graphs of the initial and adjusted time series.
- The reasoning linking the time series and its correlogram, explaining why the time series has the correlogram and what the correlogram implies for ARIMA modeling.

In this example, the student project would show

- The graph of the time series with a different mean or trend in two periods. A different trend in two periods causes a different mean of the first differences.
- The correlogram of the initial series or the first differences showing the pattern described here.
- The possible reason(s) for the change, if you are aware of them. The change in interest rates is explained in other postings.

You may not know why the time series changed. The student project does not require you to research the change.

Illustration: A project on crime rates in City Z may say: “The frequency of violent crimes increased from 1971 to 1992 and decreased from 1992 to 2006. The first differences show means of +0.06% in the first period and –0.02% in the second period. The correlogram shows decreasing sample autocorrelations for the first 14 lags, reaching a minimum of –22%, and then increasing toward zero in subsequent lags. I chose this time series to examine if the 1992 mayoral election on a crime fighting platform had any effect on the time series.

OSCILLATION VS CHANGE IN MEAN OR TREND

{This sub-section comments on the shape of this correlogram. We include this sub-section so you see how to analyze a correlogram. Several other postings on this discussion forum discuss the shapes of correlograms.}

In this illustration, the interest rates decline in the first half of the series and rise in the second half. The decline and rise are not smooth because interest rates are stochastic, but a three month moving average shows a clear pattern. The sample autocorrelations for lags 8 through 26 are less than zero. This is a common pattern, which we model by dividing the time series into two periods and taking first differences.

Take heed: The sample autocorrelation functions in your student project may differ from the one in this correlogram.

- For a stationary AR(1) process, the sample autocorrelations begin at 1 and decline to 0.
- For a random walk, the sample autocorrelations begin at 1 and stay high.

Do not confuse this with an oscillatory model, where the sample autocorrelations alternate about the mean. The sample autocorrelations here decline and then rise.

- If this occurs repeatedly, the correlogram is cyclical; the authors say *sinusoidal*.
- This pattern occurs once here. It is a change in the trend, not an oscillatory model.

HIGH VALUES AT END

The *CORREL* function gives large values for the long lags, where the correlation is based on few values. The high correlations are random fluctuations.

Excel's *CORREL* built-in function shows the shape of the sample autocorrelation function.

~ An autoregressive model shows geometric decay
~ A moving average model shows a sharp drop
~ A white noise process shows small fluctuations

All stochastic time series show the white noise process about the expected values.

~ An autoregressive model shows random fluctuations about the geometric decay
~ A moving average model shows random fluctuations about the sharp drop
~ A white noise process shows small fluctuations about the mean of zero

The sample autocorrelation function has two changes.

~ The correlation here is not adjusted for degrees of freedom; it has the same number of terms in the numerator as the denominator. The sample autocorrelation of lag *k* has *k* more terms in the denominator than in the numerator.
~ The correlation divides the sum of the cross-products by the product of the standard deviations of each series. The sample autocorrelation divides the sum of the cross-products, which has N – k terms, by the sum of the squares of the elements, which has N terms. This second adjustment smooths much of the random fluctuations.

Jacob: For the student project, do we just explain this reasoning or do we use graphs?

Rachel: Suppose we want to test whether the time series is an AR(1) model with a coefficient of 95%. See the worksheet with the AR(1) model for this coefficient. We form the autocorrelation function and the associated correlogram for an AR(1) process with a coefficient of 95% and compare that autocorrelation with the one in this worksheet. The two correlograms look different, indicating the time series is not an AR(1) process. The student project also uses other statistical techniques, as we explain below.

Jacob: The correlogram goes to zero eventually; does that mean the series is stationary?

Rachel: We have only 42 observations. The sample autocorrelations at high lags, such as the last 25% or 30% of the lags, may be small even for a non-stationary series.

Jacob: How would we deal with this series in the student project?

Rachel: We have several methods. They are used in combination. Statisticians have their preferred methods; no method is necessarily right or wrong.

~ Examine first and second differences of the observations. The pattern of this sample autocorrelation function suggests second differences may be stable. For the student project, take first and second differences. Examine the correlograms, explain what the correlograms imply, and explain in English what the first and second differences mean. We almost always examine the first differences of interest rates. Economists differ of whether we should use first or second differences; your student project can decide.
~ Deflate the interest rate time series (to real interest rates), adjust for business activity (GDP) and interest rate cycles (if any exist). The textbook speaks of structural models, or fitting an ARIMA process to the residuals of a regression analysis on another index.
~ Use higher order ARIMA processes. The textbook authors examine the correlogram and infer the *maximum* order of the ARIMA process. If the sample autocorrelations are near zero after 4 lags, they infer a maximum order of 4 and test various models. Most statisticians have the opposite perspective. They begin with an AR(1) model and work up to more complex models *only if needed*. We recommend the later approach for the student project. Your student project can decide if an AR(1), AR(2), or higher order model is appropriate.
~ We divide the time series into segments and fit ARIMA processes to each segment. These interest rates decline for about a year and a half and then increase. We might fit different random walk models to each segment. Your student project can decide if we should use one ARIMA process for the entire time series or different processes for different parts.

Take heed: Structural models and homogeneous time periods (segments) are preferred methods, but they require knowledge of the time series. Other postings on this discussion

forum explain that first or second differences and ARIMA processes with more parameters may complicate a simple time series. Do not worry that your model may not be optimal. We review if you use the statistical techniques correctly, not if you form the optimal model.

EXCEL REGRESSION BUILT-IN FUNCTION

The regression analysis on-line course shows how to fit regression lines with ordinary least squares estimators. For the student project, use the Excel REGRESSION built-in function.

Jacob: Where is the Excel REGRESSION built-in function?

Rachel: Choose the TOOLS menu from the menu bar. From the menu, choose DATA ANALYSIS. You may have to include the DATA ANALYSIS add-in to your version of Excel. From DATA ANALYSIS, choose REGRESSION.

Jacob: Simpler Excel built-in functions determine a linear trend line using regression. Can we use those built-in functions?

Rachel: We need the Excel *REGRESSION* add-in to get the table of residuals.

Jacob: How do we include the DATA ANALYSIS add-in?

Rachel: Check to see if the add-in is already installed. Some actuarial departments have the add-in installed. If the add-in is not installed, choose ADD-INS… from the tools menu. From the menu that appears, choose ANALYSIS TOOLPAK. To work with VBA, include also the ANALYSIS TOOLPAK VBA.

Jacob: What does the ANALYSIS TOOLPAK VBA give that the plain add-in doesn't have?

Rachel: You need the VBA version to invoke the add-in from VBA code. Most candidates do not need this facility.

Your version of Excel may differ. If you can't find the REGRESSION built-in function, post a question on the discussion forum, listing your version of Excel and of windows.

DURBIN-WATSON STATISTIC

The Durbin-Watson statistic is covered in the regression analysis on-line course. It is a simple test of serial correlation. Coding and using the Durbin-Watson statistic is a good prelude to the Box-Pierce Q statistic. The Durbin-Watson statistic by itself is not a valid statistical measure for a lagged regression, as we use for autoregressive processes. Use it as a prelude to the Box-Pierce Q statistic.

Jacob: How do we form the Durbin-Watson statistic? Is there an Excel built-in function?

Rachel: Excel has no built-in function for this; we write the formula.

Jacob: The formula uses the residuals. How do we calculate the residuals? We can do this from the equations in the textbook, but it would take a while. Is there a simple method?

Rachel: The Excel *REGRESSION* add-in calculates the residuals. The add-in computes the ordinary least squares estimators and the residuals. Copy the formula for the Durbin-Watson statistic and the Box-Pierce Q statistic from the illustrative spreadsheet on the NEAS web site and use it with the residual output from the Excel *REGRESSION* add-in.

Form an AR(1) model from the last 3½ years of Treasury bill interest rates: January 1997 through June 2000.

Jacob: Must the time series be stationary?

Rachel: Even if the time series is not stationary, we can form the Durbin-Watson statistic for the residuals from an AR(1) model.

Start with the *REGRESSION* add-in. Copy the January 1997 – June 2000 Treasury bill rates to a new worksheet. Place these in cells B11:B52 and also in cells C12:C53. Column B is the Y values and Column C is the X values. We don't use the values in B11 or C53.

On the illustrative worksheet, we eliminate rows 53 and 11, getting rid of the original cell B11 and cell C53. This gives a matrix of B11:C51 for the regression analysis.

Jacob: If we use an AR(2) model, can we still use the *REGRESSION* add-in?

Rachel: To use an AR(2) model, place these rates in cells D13:D54 as well. Column B is the Y values, Column C is the X_1 values, and Column D is the X_2 values. We don't use the values in B11, B12, C12, C53, D53, or D54. We have 40 triplets (observations).

Jacob: What do we choose on the *REGRESSION* menu for the AR(1) model?

Rachel: The dependent variable is in cells B11:B51, after eliminating the original cell B11. The independent variable is in cells C11:C51.

Ask for *RESIDUALS*. You don't need the *STANDARDIZED RESIDUALS* since the interest rates are all about the same size in this illustration. If interest rates change greatly over the time series, we would examine *STANDARDIZED RESIDUALS*.

Take heed: For time series analysis, we use first differences to eliminate trends. The time series values should not change materially. Use *RESIDUALS* for your analysis.

You can place the output on the same sheet or a new sheet. In a new worksheet, the residual output is in Columns A, B, and C. We place the output on the same worksheet starting in cell A61, so the output is in Columns A, B, and C, rows 85 through 125.

Take heed: The Excel default is a new worksheet. To use the same worksheet, over-ride the default and enter the upper-left cell of the output region on the *REGRESSION* screen.

The residual output shows the observation number, the fitted Y value, and the residual.

Take heed: The observation number is not the X value. For some analyses, like conditional heteroscedasticity, you may copy it the X values.

In column D, place the square of the residual. For cell D85, write =*C85*^*2*. Copy this formula to cells D86:D125.

In column E, place the difference of successive residuals. For cell E86, write =*D86-D85*. Copy this formula to cells E87:E125. Column E has one less figure than Column D has; this reflects the degrees of freedom in the regression.

In Column F, place the square of the difference of successive residuals. For cell F86, write =*E86*^*2*. Copy this formula to cells F87:F125.

Use Excel's quick sum function to get totals for Columns D and F. Place the cursor in cell D126 and click on the quick sum icon. Do the same for cell F126.

The Durbin-Watson statistic is the sum in cell F126 divided by the sum in cell D126. Place the Durbin-Watson statistic in cell G126 as the formula =F126/D126.

Jacob: What do we expect to find?

Rachel: A Durbin-Watson statistic of 2 indicates no serial correlation. This example gives a Durbin-Watson statistic of 2.10, which is not significantly different from 2.

- The correlation of lag 1 on the residuals, using the *CORREL* built-in function, is –0.06289 (cell C127), which is not significantly different from zero.
- The autocorrelation of lag 1 on the residuals, using the exact formula, is –0.06195 (cell C128).

Other time series show serial correlation. Your student project should explain the meaning of your results.

Jacob: The slope coefficient is 95%. Is the *t* statistic significantly different from one?

Rachel: Be careful that you read the regression output correctly. This regression shows $\beta = 0.95$, its standard error = 0.07, and the *t* statistic is 13.66.

- This *t* statistic means we reject the null hypothesis that $\beta = 0$, not the null hypothesis that $\beta = 1$.
- For the null hypothesis of $\beta = 1$, the *t* statistic is $(0.95315 - 1) / 0.06977 = -0.480$. We do *not* reject the null hypothesis that $\beta = 1$.

BOX-PIERCE Q STATISTIC

Your student project uses the Box-Pierce Q statistic

- To select among ARIMA models and
- To verify that a particular model fits the empirical data.

The textbook explanation of the Box-Pierce Q statistic is abstract.

- Use 15 to 40 sample autocorrelations.
- Ignore the first several sample autocorrelations.

The textbook does not specify precisely how many sample autocorrelations to use and how many to ignore. You understand the use of this technique after seeing its application.

The dialogue below explains how to use the Box-Pierce Q statistic. The illustrative Excel worksheet provides the code. Be sure you understand how to use the technique before applying it to the time series in your student project.

Jacob: The Box-Pierce Q statistic tests whether the time series is a white noise process. How do we use the Box-Pierce Q statistic for ARIMA processes, which are not white noise?

Rachel: We illustrate with the file of residuals from the AR(1) process.

Jacob: Do we apply the Box-Pierce Q statistic to the values of the time series (such as the interest rates) or to the residuals from an ARIMA model?

Rachel: If the time series values (interest rates) themselves show no pattern, we test if they are a white noise process. As an example, form the Box-Pierce Q statistic for the sample autocorrelations from the interest rates themselves on the *CORRELOGRAM* worksheet.

- The sample autocorrelations do not remain close to zero until lag 22.
- The Box-Pierce Q statistic is too high to reflect a white noise process.

Interest rates have random walks, trends, or mean reversion; we do not expect the interest rates themselves to be a white noise process. If an ARIMA model fits well, the residuals from the model may be a white noise process.

Take heed: Your student project may use the Box-Pierce Q statistic several ways. You may use the Box-Pierce Q statistic to identify a white noise process or a random walk.

- A time series may be a white noise process. Weekly rainfall in a tropical rain forest may show no seasonality or cycles. The weekly rainfall may be a white noise process.
- A time series may be a random walk, and its first differences are a white noise process. Stock prices are often assumed to be a random walk.

If your chosen time series is a white noise process or a random walk, pick another time series for your student project. Use the Box-Pierce Q statistic to confirm white noise.

Sometimes changing the length of the periods gives a better model.

Take heed: Using periods that are too long causes may cause a time series to seem like white noise. Daily rainfall in a tropical rain forest may have strong autoregressive process (if it rains much on Monday, it may rain much on Tuesday as well) or a moving average process (if the rainfall was more than expected on Monday, it may be less than expected on Tuesday). If the effects die out after a few days, the weekly rainfall is white noise. Choose periods that show autoregressive or moving average processes.

Take heed: Use the Box-Pierce Q statistic to verify that your ARIMA model fits.

Jacob: If the residuals are a white noise process, is the ARIMA model correct? If the residuals pass the Box-Pierce Q statistic, have we fit the proper model?

Rachel: An ARIMA process is stochastic. The forecasts are never exact, since each value is a random variable. If the residuals are close to a white noise process, the ARIMA process *may* be a suitable model of the time series. We compare alternative models; we don't solve for an exact answer.

Take heed: The student project has no exact solution. No ARIMA process fits perfectly, and even the best fit may change from one year to the next. Model fitting combines science and art. You have wide discretion in choosing a model.

- One statistician may prefer an AR(2) process and another statistician may prefer an ARIMA(1,1,0) process.
- One statistician may fit an ARIMA(2,1,1) process to the entire time series and another statistician may fit ARIMA(1,1,0) processes separately to two periods.

We examine if you use the statistical tools properly and you proceed correctly through the model-fitting process. We do not judge if you came to the right answer.

CODING THE BOX-PIERCE Q STATISTIC

Jacob: How do we form the Box-Pierce Q statistic in Excel? Does Excel have a built-in function, or do we code the formula ourselves?

Rachel: We provide the cell formulas, which you may copy to your student project. You select the discretionary items:

- How many values of K to use, and how many initial values to ignore.
- What significance level to use.

Take heed: The number of observations in the time series (T) is given. The Box-Pierce Q statistic uses K sample autocorrelations. We use 15 to 40 sample autocorrelations. The illustrative worksheet shows the Box-Pierce Q statistic for values of K from 1 to 40.

The choice of K depends on the number of time series observations. With more observations, we can use a higher value for K.

Illustration: With only 42 observations for the time series in the illustrative worksheet, we use 15 to 20 sample autocorrelations for the Box-Pierce Q statistic.

If the time series has many elements, we can choose a high value of K. But if the Box-Pierce Q statistic suggests the process is (or is not) white noise for K = 40, the indication probably won't change for K = 80.

Illustration: A time series of daily temperature from 1890 through 2005 has 116 × 365.25 = 42,369 observations. We can use as high a value for K as we like.

- If you use the Box-Pierce Q statistic to ensure that the time series is correctly de-seasonalized and then fit with an ARIMA process, you might choose K = 365 (assuming you have a table with the appropriate χ^2 values).
- If you first de-seasonalize the time series and then fit the ARIMA process, K = 40 or 50 is large enough.

Take heed: Be sure to estimate the sample autocorrelation function exactly. The Box-Pierce Q statistic is an approximation, but it is a good approximation with the exact formula. Using the wrong degrees of freedom for the sample autocorrelation function won't change your choice of ARIMA process. But if your time series has few observations, the Box-Pierce Q statistic will not be accurate.

Columns H – K show the autocorrelations formed with the *SUMPRODUCT* built-in function.

~ Column H shows the sum of the cross-products of lag *k*.
~ Column I is the square of Column H.
~ Column J is the sum of the first K terms in Column I.

~ Column K divides Column J × 41 (number of observations) by the sum of all 41 squared residuals.

- Place the formula =SUMPRODUCT(OFFSET(C$85,0,0,41-A85,1),C86:C$125) in Cell H85. Copy the formula to cells H86:H122. We don't use the last three lags, so we don't copy the formula to the last three cells. The *OFFSET* built-in function ensures that the *SUMPRODUCT* uses the correct autocorrelations. Be sure that your relative and absolute references are correct and that you use the proper number of time series observations.
- Place the formula =H85^2 in Cell I85. Copy the formula to Cells I85:I122.
- Place the formula =SUM(I$85:I85) in Cell J85. Note the combination of relative and absolute references for a downward sum. Copy the formula to Cells J86:J122.
- Place the formula =J85*41/D$126^2 in Cell K85. Copy the formula to Cells K86:K122. The figures should increase down the column at a decreasing rate.

Jacob: We have 42 interest rates; why do we have only 41 residuals?

Rachel: We are using an AR(1) model, so we have 42 – 1 = 41 residuals. The degrees of freedom for the Box-Pierce Q statistic is K – p – q.

Jacob: To what do we compare the figures in Column K?

Rachel: The Box-Pierce Q statistic is distributed approximately like a χ-squared distribution with K – p – q degrees of freedom. If we use the sum of the first 15 terms, we compare to a χ-squared distribution with 14 degrees of freedom. If we use the sum of the first 20 terms, we compare to a χ-squared distribution with 19 degrees of freedom.

Jacob: How do we get these figures?

Rachel: Table 2 on page 604 of the textbook has the χ-squared distribution. At a 10% significance level, the critical χ-squared value is 21.06 for 14 degrees of freedom and about 27.2 for 19 degrees of freedom.

Take heed: For your student project, use the Excel built-in functions.

- We placed the formula =CHIINV(0.1,14) in Cell L99. This is the critical χ-squared value for 14 degrees of freedom at a 10% significance level. You can make the formula relative by writing =CHIINV(0.1,A99 – 1). You can copy the formula to every cell of column L. The illustrative worksheet shows the relative formula for two cells. You don't have to look up critical χ-squared values in the textbook chart.
- We placed the formula =CHIDIST(21.06414,14) in Cell M99.

Take heed: Explain in the write-up the procedure you used.

Jacob: The figures in the worksheet are lower. Can we conclude that (i) the residuals are a white noise process, (ii) the interest rates are an AR(1) process, and (iii) we have solved the student project?

Rachel: The Box-Pierce Q statistic says that we can not reject the null hypothesis that the residuals are a white noise process. As the textbook authors note in their examples, many ARIMA models may be close enough fits that we can not reject the null hypothesis that the residuals are a white noise process. We compare the ARIMA processes and decide which one seems best.

A random walk is not stationary, but its residuals are a white noise process. This series is not stationary, so we take first differences and examine an ARIMA(0,1,0) process. We have not shown this on the illustrative workbook so that you can set up your own sheet.

BOX-PIERCE FORMULA, T, K, AND DEGREES OF FREEDOM

Jacob: What is the formula for the Box-Pierce Q statistic?

Rachel: The Box-Pierce Q statistic is $Q = T \times \sum_1^K \hat{r}_k^2$.

Jacob: This formula seems strange. If we have twice as many observations T, won't the Box-Pierce Q statistic be twice as high? If the sample autocorrelations are about 10%, then twice as many of them gives a Box-Pierce Q statistic that is twice as great.

Rachel: If the sample autocorrelations are about 10% for all lags, the time series is not a white noise process. The Box-Pierce Q statistic tests whether the time series is a white noise process; it is not used for non-stationary time series.

If we have twice as many observations, the square of each sample autocorrelation for a white noise process is about half as great, on average. The *autocorrelations* for a white noise process are identically zero. A non-zero *sample* autocorrelation stems from random fluctuation. More observations (a higher T) smooths the random fluctuations.

Jacob: If we use twice as many lags, will the Box-Pierce Q statistic be twice as great?

Rachel: The increase tracks a χ-squared distribution with K – p – q degrees of freedom. The increase is not proportional.

DOF's	1%	10%	90%	99%
10	2.56	4.87	15.99	23.21
20	8.26	12.44	28.41	37.57
30	14.95	20.60	40.26	50.89
40	22.16	29.05	51.81	63.69

Jacob: Why is the increase not proportional?

Rachel: The expected value of the Box-Pierce Q statistic increases proportionally for a white noise process. The probability that the Box-Pierce Q statistic lies outside a Z% confidence interval increases more than proportionally for Z < 50% and less than proportionally for Z > 50%.

QUALITY OF FIT: NON-STATIONARY RANDOM WALK VS STATIONARY WHITE NOISE PROCESS

Jacob: When I regress the *interest rates* on the *lagged interest rates* for the time period in my student project, using an AR(1) model on the monthly interest rates themselves, I get a β coefficient of one, indicating that the model is not stationary. The *t* statistic for β is high, the *p*-value is low, and the R^2 for the regression is high. The fit seems excellent.

When I regress the *first differences* of the interest rates on the *lagged first differences*, using an AR(1) model on the first differences, I get a β coefficient of zero, indicating that the model is stationary. But the *t* statistic for β is low and not significant, the *p*-value is high, and the R^2 for the regression is low. The fit seems poor.

I had expected the opposite results.

→ If $\beta \approx 1$, the time series is a non-stationary random walk; we do not use it for forecasts.
→ If $\beta \approx 0$, the time series is a stationary white noise process; we use it for forecasts.

Why does the random walk have a good fit and the white noise process have a poor fit?

Rachel: These are expected results. Consider what each test implies.

The *t* statistic tests the null hypothesis that $\beta = 0$. If the time series is a random walk, β is 1, not zero. We reject the null hypothesis. When we take first differences, $\beta = 0$; that is exactly correct. We do not reject the null hypothesis.

To test if the time series is a random walk, the null hypothesis should be $\beta_0 = 1$. This gives a *t* statistic close to zero, and we do not reject the null hypothesis.

The R^2 says how much of the variance is explained by the β coefficient. If the time series is a random walk with a low standard error (low σ), the R^2 is high. Values of 98% or 99% are reasonable for a perfect random walk with many observations.

If the β coefficient is zero for the AR(1) model of first differences, we expect the R^2 to be about zero. This says that the β coefficient doesn't explain anything. If the β is zero, it doesn't explain anything.

Forecasts: A random walk is not stationary, but we still use it for forecasts. For a random walk with a drift of zero, the L-period forecast is the current value. For a random walk with a drift of *k*, the L-period forecast is the current value + L × k.

Jacob: For a perfect random walk, with $\beta = 1$ and a drift of zero, I assume the R^2 depends on the stochasticity. If σ is high, the R^2 should be low; if σ is low, the R^2 should be high.

Rachel: That is true for most regression equations, where the values of X are not stochastic. For the AR(1) model, the X values are the Y values lagged one period. The dispersion of the X values varies directly with σ.

If σ is twice as large, σ^2 is four times as large, the sum of squared deviations of the X values ($\sum x_i^2$) is four times as large, the variance of $\hat{\beta}$ does not change, and the *t* statistic does not change. The R^2 is the σ^2 divided by the total sum of squares (TSS) of the Y values. The X values are also the Y values, so the R^2 does not change.

Jacob: That is counter-intuitive. Suppose the starting interest rate is 8%. If σ is 1%, which is high stochasticity for monthly interest rates, I presume the R^2 will be low. If σ is 0.01%, which is low stochasticity, I presume the R^2 will be high.

Rachel: To conceive of the relations, assume the starting interest rate is zero. The deviation is the actual value minus the mean, so we might as well start with a mean of zero.

> If σ = 1%, we have much random fluctuation in the Y values. This makes the X values more dispersed, which lowers the variance of the ordinary least squares estimator.
> If σ = 0.01%, we have little random fluctuation in the Y values. The X values are less dispersed, which raises the variance of the ordinary least squares estimator.

The two effects offset each other.

Jacob: Shouldn't the degree of stochasticity affect the R^2?

Rachel: The change from σ = 1% to σ = 0.01% as a change in the units of measurement. Nothing about the regression has changed.