BUILDING ARIMA MODELS: A STEP BY STEP GUIDE

Updated: April 22, 2008

(The attached PDF file has better formatting.)

The postings on the discussion forums provide guidance for your student project. We describe what each posting covers, and we suggest an order for your initial review.

The student projects are independent projects. The NEAS web site has hundreds of data sets and various project templates that you may use for the student project. You may use any time series with enough observations, as long as it is not a random walk or white noise.

This posting is a step-by-step guide to ARIMA modeling. Separate postings

- explain the requirements for the student project and the learning objectives
- outline the written documentation that accompanies the statistical work
- review the statistical techniques for student projects on time series analysis
- compare common ARIMA processes and suggest which ones to explore
- document the illustrative worksheets for the project template on interest rates
- clarify the balance between faculty guidance and independent work
- answer questions from candidates about the time series student projects

The instructions summarizes questions and answers in past semesters. They provide more guidance than most candidates need. You have wide latitude: you may choose the data, topics, and statistical procedures, and write a student project that differs from the project templates on the discussion forums.

Your student project applies statistical techniques to real data. It is not a full statistical study to select the optimal ARIMA model. You need not compare all ARIMA processes or test all structural models. We review if your project properly uses modeling techniques, not if your solution is optimal.

This posting is discursive. It discusses each step with many examples. It refers to other discussion forum postings for further explanation. It is not a cookbook to a rigid sequence of statistical techniques

Some candidates want to know an exact order for the student project, and they are frustrated by the subjectivity of ARIMA modeling. As the course textbook says, ARIMA modeling is a second-best alternative. We search for patterns in the data. We don’t find the true causes of the time series patterns, but the ARIMA pattern helps our forecasts.

Your student project is successful if it sheds light on the pattern over time of the data. Read the suggestions in this step-by-step guide and apply them to your time series.

ARIMA MODELING

Chapter 19 of the textbook has five examples of ARIMA models. The text assumes you

- are familiar with nonlinear regression and partial autocorrelation functions
- have the needed statistical software
- understand well the time series that you are modeling.

The text gives an outline of ARIMA models: specification, diagnosis, and validation.

The time series on-line course assumes

- you know linear regression and sample autocorrelation functions
- you have Excel, but not other statistical software
- you have not worked with the data except for the student project

The ARIMA models for the student project can be built with basic Excel functions. The illustrative work-sheets on the discussion forum provide code for the statistical techniques. If you understand the concepts, you can complete the student project without difficulty.

This step-by-step guide to building ARIMA models is geared to candidates who have taken the time series on-line course but who have never worked with ARIMA models.

~ We use plain language for most of the explanation.
~ When we use a statistical procedure, we explain the steps you need.

A step-by-step guide is not a cookbook. You make decisions at each step. We give enough guidance that you won't get stuck, but you form the model.

- We do specify exactly what to do at each step. The empirical data are stochastic, and they do not fit any model exactly. Differences from the model may reflect random fluctuation or a poor model.
- We explain what to consider at each step. Your analysis depends on the empirical relations and your statistical judgment.

The student project is an educational process, not a consulting job.

- Working through the techniques helps you learn them.
- We examine if you can apply techniques to actual data, not if your solution is correct.

Read this step-by-step guide when you begin the ARIMA modeling part of the time series course, so you know the techniques you must master.

Illustration: The Box-Pierce Q statistic seems vague in the textbook, but it is a basic tool to validate ARIMA models. Validation is subjective, since ARIMA models rarely fit the data exactly. The student project uses the Box-Pierce Q statistic, Bartlett's test, intuition, and parsimony to select an ARIMA model.

See also: The illustrative work-sheets provide templates for correlograms, Durbin-Watson statistic, Box-Pierce Q statistic, Yule-Walker equations, and statistical tools. You need not have statistical software or Excel expertise to complete the student project.

Take heed: Statisticians differ in their approaches.

- The text uses complex ARIMA processes, often with six to eight parameters.
- For the student project, use simpler processes, with two or three parameters.

The project demonstrates that you understand the concepts, not that you can forma model with many parameters.

Step #1: *Knowledge*

The student project assumes you know time series analysis and regression techniques.

- ARIMA modeling seems complex at first, but the modeling sequence is logical.
- Each step requires a decision; a limited set of decisions leads to many combinations.

Below are topics in the time series on-line course that are used in the student project.

1. Identify *stationary* time series. *Most time series that actuaries use are homogeneous non-stationary*. A time series with a trend or drift, or a random walk even with no drift, is not stationary. Examine the correlogram, check for unit roots, and graph your results.
2. Form a stationary time series: take first differences, using logarithms if appropriate; divide the time series into periods; and correct for seasonality.
3. The *autocorrelation function* of an ARIMA process is not the *sample autocorrelation function* of a time series. Use the *sample autocorrelation function to specify the model* and the *autocorrelation function to validate the model*.
4. The textbook discusses partial autocorrelation functions as well. You do not have the statistical software for the partial autocorrelation function, and the discussion in the text is weak. You need not use the partial autocorrelation function for the student project.
5. The pattern of the correlogram reflects the type of ARIMA process. Focus on *geometric decay* vs *sudden drops*. Low ARIMA parameters and high standard errors of short time series obscure the pattern. A correlogram is hard to analyze if stochasticity is high. If a pattern is not obvious, explain the pros and cons of alternative models. We judge if you understand the reasoning, not if you choose the optimal model
6. Understand the *intuition* for moving average vs autoregressive models. Much practical statistical work is subjective. A non-intuitive ARIMA process with good in-sample fits may have poor out-of-sample results. Explain in lay terms what each parameter implies.
7. Chapter 19 builds ARIMA models for several time series. The student project does the same with less complex time series and ARIMA processes.

As you work through your project, review the course modules for these topics.

Step #2: *TOOLS*

You need statistical software for the student project. The illustrative Excel worksheets

- Have cell formulas and examples for statistical functions not built into Excel.
- Have VBA macros and custom functions that simplify your work.

The items below are provided on the illustrative work-sheets.

- *Sample autocorrelation function:* Excel has a *CORREL* built-in function but no built-in function for the sample autocorrelation. An illustrative worksheet explains the difference and gives the cell formulas, using the Excel *SUMPRODUCT* and *OFFSET* built-in functions.
- The *SUMPRODUCT* with *OFFSET* is slow for large time series, such as 40,000 or 50,000 days. We provide a VBA macro that is quicker and simpler.
- *Correlogram:* Use the chart wizard to construct correlograms. Label your axes so our faculty can review your work. Copy and paste the correlograms into your write-up.
- *Durbin-Watson statistic* is taught in the regression analysis course and used in the time series student projects. An illustrative worksheet gives you the code.
- Use the *Box-Pierce Q statistic* to test the model, using the sample cell formulas in the illustrative worksheet. Compare the Q statistic with the critical values for the χ-squared distribution, using Excel's built-in function or the tables in the textbook.
- Use linear regression to form autoregressive models. Use the Excel *REGRESSION* add-in and use the residual output for the Box-Pierce Q statistic.

Take heed: We explain the code in the illustrative worksheets and the rationale for each procedure. You may copy the cell formulas from the illustrative worksheets.

The cell formulas in the illustrative worksheets are simple. Experienced Excel users will find *SOLVER* and VBA to be more efficient tools. Nothing in the student project requires more advanced Excel knowledge than in the simple cell formulas.

Your write-up states the results in your worksheet. *Our faculty can not figure out what you have done from the Excel workbook alone.* State the techniques you use and the results. Explain what the results imply and how you test them for significance (when appropriate).

You can use any statistical software or any spread-sheet package. If you use SAS at work, you can save time by using it for the student project. If you know VBA and Excel built-in functions, they can save you time as well.

SAS, MINITAB, and "R" have all the built-in functions you might use in a student project. You may use these software packages; they are not required.

Step #3: *Choose the Time Series*

You can use any time series you want. We show ARIMA models for interest rates and daily temperature, and structural models for various macroeconomic indices.

Many project templates suggest a variety of student projects. Use the project templates to generate ideas for your own student project.

Choose a topic that interests you. The web has dozens of sites with statistics on almost any topic.

- If climate change intrigues you, do a project on daily temperature or rainfall.
- If you are a sports fan, do a project on won-loss ratios of your home team.
- If you like music, do a project on monthly DVD sales by genre.

We suggest numerous topics for your student project.

Illustration: Interest rates have a hundred flavors. We show sample work with 90 day Treasury bill rates, and extracts from student projects on other rates. You can choose

- short rates (three month bills, over-night rates) vs long rates (twenty year bonds).
- private rates (Moody's corporate bond index, the bank prime rate) or government rates (Treasury securities, bank discount rates).
- spot rates, forward rates, futures rates, or other derivatives.
- nominal interest rates or real interest rates (structural models)

The type of rate should reflect the analysis.

- For interest rate seasonality, use short rates, not long rates, such as over-night LIBOR.
- For the relation of interest rates and budget deficits, use real interest rates.
- For the relation of interest rates and recessions, use corporate spreads.

The NEAS discussion forum has many interest rate time series. Read the project templates and the other discussion forum postings. Feel free to choose another type of rate, such as risk-free rates in other currencies.

Take heed: It is often easier to model real interest rates, the residuals of interest rates on inflation rates, corporate spreads, or the spread between long and short rates with ARIMA processes. Spend an extra half hour setting up the data; you save hours in your analysis.

Suggestion: You read dozens of theories about interest rates and other macroeconomic indices: real interest rates are higher or lower in recessions, higher or lower when the U.S. runs a deficit vs a surplus, and so forth. Choose a hypothesis, form a structural model, and fit an ARIMA process to the residuals.

Step #4: *STRUCTURAL VS TIME SERIES MODELS*

If the time series is a by-product of *clear and easily accessible* causes, we use regression analysis. For a stochastic time series with its own internal logic, we use ARIMA models.

Illustration: Unemployment rates depend on economic, demographic, and legislation, such as hiring practices, restraints on firing, unemployment benefits, and minimum wages.

- If the legislature raises the minimum wage, teen-age employment drops.
- If the state raises unemployment benefits or mandates employer provided health insurance, unemployment rises.
- During recessions, unemployment rises.

The macroeconomics on-line course reviews these effects. Barro's textbook is an excellent source of ideas for time series and regression analysis student projects. Government web sites have extensive data on economic variables.

Take heed: The residuals from structural models are easier to fit with ARIMA processes, and the time series are more meaningful.

We use regression analysis if the explanatory factors change frequently. We use ARIMA models for the *residuals* of the regression.

Illustration: Inflation rates affect interest rates. We may regress interest rates on inflation and use an ARIMA model to forecast the residuals. We provide both interest rates and inflation rates on the web site so you can model either nominal interest rates or real interest rates. Chapter 19 of the textbook has a similar example.

Take heed: Many macroeconomic indices are functions of other indices. Use differences, periods, or structural models.

Illustration: Nominal interest rates are a function of inflation. Inflation is not mean reverting, so nominal interest rates are not stationary. Using first differences and dividing the time series into periods creates stationary time series, but your student project will be better if you use real interest rates and adjust for economic activity.

Illustration: Suppose you want to model interest rates.

- Divide the one month LIBOR by the CPI for the previous period to get the real LIBOR.
- Real GDP is the detrended GDP.
- Regress the one month LIBOR on real GDP.

Your student project may show a sequence of models.

- Fit nominal interest rates by taking first differences. Use the mean squared error over the next 12 months to estimate goodness-of-fit.
- Convert to real interest rates and fit a new ARIMA process. It is more difficult to decide if first differences are needed. Re-compute the mean squared error.
- Use a regression on real GDP.

Your student project may explain the advantages and drawbacks of a structural model. You may have a good model of real interest rates, but if you don't know future inflation and real GDP, you can't forecast future interest rates.

You may determine real interest rates three ways:

- Rate A minus Rate B.
- Rate A divided by Rate B.
- Rate A regressed on rate B.

The regression is the best procedure, since it combines additive and multiplicative models. The *REGRESSION* add-in does the regression and gives the residuals. But you can use any of the three methods.

The regression can be done using different inflation rates as the explanatory variable. Choose one and explain the rationale. This is a statistics course, not an economics course. You are not graded on the choice of the inflation rate.

We use residuals for real interest rates, maturity spreads, and corporate spreads. The definitions below use *Rate A minus Rate B*, but you can use any of the definitions above.

- Real interest rate = nominal interest rate minus expected inflation.
- Maturity spread = long risk-free rate minus short risk-free rate.
- Corporate spread = corporate bond rate minus Treasury bond rate.

The NEAS web site has many time series in Excel format. Form the time series you want. Use simplifications, even if they are not perfectly accurate.

Illustration: Use last month's actual inflation as a proxy for expected inflation. The ratio of the CPI in the two previous months as the expected inflation for the current month.

Don't worry that your time series is not perfect. Construct the time series you want and fit an ARIMA process. But be consistent. To analyze corporate spreads, use the average rate in each month, or the corporate bond rate at the start or middle of each month.

Take heed: Do not worry that a time series on the residuals of a regression is too complex for the student project. The opposite is true. Many macroeconomic and demographic time series are too complex to fit with an ARIMA process. The residuals of a regression analysis are easier to fit, and they are more likely to have an intuitive relation.

The project templates discuss structural model for many of the time series on the NEAS discussion forum. For some models, you must find an appropriate explanatory variable on the internet. Spend half an hour or an hour looking for time series on the internet that fit the hypothesis you want to examine. We do not grade your success in finding the right explanatory variables for a structural model.

Step #5: *Time Periods*

We fit an ARIMA process to model a time series over a given period. If the mean, drift, or variance of the time series changes because of *external causes*, we use different models for the different parts of the time series. Statisticians speak of *interventions*, or exogenous events that change the ARIMA process.

- If changes in the time series are random fluctuations, we use a single process to model the underlying structure. If we change models each year, we can't forecast future rates.
- If the time series itself changes, we need separate models. Forcing a single model to cover all years gives an ARIMA model that is does not fit well in any period.

A time series may follow different ARIMA process in different periods. If exogenous factors change the mean, variance, or drift of the time series, the time series is not stationary and can not be modeled by a single ARIMA process. Examples:

- We model an insurer's premium volume in 1985 - 2005. For 1985 - 1995, the insurer has a monopoly; in 1996, the market becomes competitive. Premium *volume* may be high when the insurer has a monopoly and lower when the insurer competes. The *variance* of the premium volume may be low when the insurer has a monopoly and high when it competes.
- We model airline passenger volume before and after deregulation. Greater competition and lower fares after deregulation raise the industry's passenger volume. The variance of any carrier's passenger volume increases: some new carriers rapidly gain market share and some established carriers fail.
- We model sales, profitability, and cash flow of firms differently in their start-up phases and their mature phases.
- Oil prices have different time series for the pre-OPEC era (before 1973) and the OPEC era (1973 onwards).

Interest rates have both types of changes.

~ Rates are stochastic, varying from month to month. The ARIMA process models these fluctuations.

~ Federal Reserve Board policy (monetary policy), federal budget deficits (fiscal policy), domestic and foreign capital investment, economic growth, and perhaps trade balances affect the mean, drift, and variance of interest rates. An ARIMA process for the 1960's may not be a good model for the 1980's.

It is not always easy to identify external causes. Even simple questions, such as warming vs cooling of the earth, are much debated.

We use two methods of dividing a time series into periods:

- We examine the means, drifts, and variances of the time series itself.

- We examine the exogenous events that might change the ARIMA process.

Illustration: An actuary examines a time series of personal auto written premium from 1980 to 2008. An acquisition of a personal auto subsidiary in 1995 writing in different states changes the time series. We use separate models for 1980-1994 and 1996-2008.

Examine the time series you choose. You may divide it into two or three periods based on the attributes of the time series, even if you have no explanation.

We don't expect you to know post-World War II Federal Reserve Board policy or other events that affect interest rates. For the student project, examine the means, drifts, and variances of the time series. Select periods that have reasonably stable attributes.

We provide some basic information about post-World War II Federal Reserve Board policy to explain the differences observed in the time series. Just as actuaries examine policy provisions, distribution systems, and market competition to set optimal rates, a statistician should know the attributes of the time series to fit an ARIMA process. But the student project focuses on the statistics, just as the SOA and CAS exams focus on the actuarial procedures. You are not graded on your knowledge of economics.

- From the end of World War II (1945) through the mid-1970's, the U.S. economy expanded briskly. Government officials worried about Depression-era deflation, not the mild inflation of an expanding economy. Inflation was thought to be an antidote to unemployment, which had been high during the Depression. The federal government and the Federal Reserve Board believed that mild inflation was beneficial, in that it restrained unemployment and did not hamper economic prosperity.

  This presumed relation of inflation and unemployment was an error, but it was the prevailing macroeconomic policy in the 1960's and 1970's. Interest rates had a steady upward trend; the time series is not stationary.

- From the late 1970's through early 1980's, inflation and interest rates were high and volatile, resulting from (i) the mistaken macroeconomic policies of these times and perhaps (ii) the supply shocks of OPEC oil price increases.

- Paul Volcker, who became chairman of the FED in 1981, adopted a monetarist perspective (Milton Friedman's views). The money supply grew at a steady, slow rate. Interest rates and inflation rates declined steadily (downward drift). Greenspan continued Volcker's policies. The interest rate patterns in the first and second periods should not recur if the FED continues a monetarist policy.

No single ARIMA process is an appropriate model for all three periods. *For the first part of the student project, you select appropriate periods.*

- Global daily temperature has long periods (thousands of years). We have had ice ages and warm eras; different models are appropriate for each. A student project on daily temperature over the past 130 years may examine if a trend exists and if the trend rate has changed. Many web sites on global warming have information about changes in the daily temperature. A student project may compare the daily temperature time series in a period of no trend vs a period of trend.

- If the time series is unemployment rates, the time periods depend on legislation for hiring practices, restraints on firings, unemployment benefits, and minimum wages. U.S. law stayed relatively constant over the past fifty years; European law provided increasing liberal benefits. You might compare U.S. and European unemployment.
- Interest and inflation rates depend on monetary and fiscal policies. In Europe, policies changed with entry into the European Union. Use European, Asian, or Latin American rates to make your student project different.

We use separate models for each period; we don't take differences are use a single model. We use two or three *interest rate eras* because of different Federal Reserve Board policies.

For the student project, you can rely on internal characteristics of the time series. Inspect the graphs for the time series and choose the time periods. We show *illustrative* graphs for three interest rate periods:

- Period 1: January 1945 – December 1978
- Period 2: January 1979 – December 1982
- Period 3: January 1983 – June 2000

These are *illustrative* periods. For your project, examine the data and choose periods.

The periods need not be contiguous. You can leave gaps. You can choose January 1945 – December 1978 as Period 1 and July 1979 – June 1982 as Period 2. This leaves an 18 month gap between the periods. During the gap, policies are changing, and no ARIMA process may properly model interest rates.

If the drift is steady for several years and then reverses for several years, a single ARIMA process doesn't work. Taking second differences is not proper, since the first differences form stationary time series in adjoining periods. Instead, you can

- Use separate ARIMA processes for the two periods.
- Form *real interest rates* and see if a single ARIMA process works for both periods.

The second method is a structural model: form a regression and use the ARIMA process on the *residuals*. Use this method if an exogenous factor affects the time series.

Take heed: Shifts in daily temperature are not well understood.

A shift in the mean *suggests* an exogenous intervention. Unemployment rates may be 6% for several decades followed by a rise to 11% for a decade, as in France and Germany. The cause may be higher unemployment benefits and restrictions on work terminations. Unemployment rates have since declined in Europe, and you may have three periods.

The discussion board graphs three month Treasury bills and suggests rough interest rate eras. For the student project, graph the time series and choose time periods.

- Do not just copy the three periods on the discussion board. Examine the graph of your time series and explain whether two or more time periods are needed.
- A project *comparing* time periods may use two or more periods. A project *fitting* an ARIMA process may focus on one period.
- You can leave gaps between periods or ignore some periods. A student project can compare ARIMA processes for the first and third periods on the discussion board.

A time series with different means in different segments is not stationary. Separating the time periods is necessary to create stationary time series, but it is generally not enough. For several reasons, a time series may not be stationary.

- A time series with an upward or downward drift is not stationary. A moving average graph of the time series reveals most drifts.
- A random walk (an autoregressive time series with a unit root) is not stationary.

    - The graph doesn't show whether the time series is a random walk.
    - The correlogram shows sample autocorrelations that do not decline rapidly.
    - Fitting an AR(1) process shows a β of about one (a unit root).

A time series can have a drift and also be a random walk. In both scenarios, we test for stationarity and take first differences; see the later steps of this guide.

Take heed: As an alternative to first differences, you may detrend the time series. If daily temperature increases 0.03% a year, detrend the time series.

Common Errors

As you construct ARIMA models, check if periods are needed If you divide a time series into two periods and the ARIMA process is similar for both, you don't need two periods.

Illustration: Daily temperature may have cycles or long-term trends, but the ARIMA process may be similar to each period.

An unusual pattern in a correlogram may reflect a changing trend or drift. Suppose the correlogram shows sample autocorrelations

- declining from 25% to zero over the first 20 lags
- declining from zero to –15% from lags 20 to 30
- rising back to zero by lag 40.

This pattern indicates two periods with different drifts. Many time series have this pattern. Taking second differences to eliminate the pattern loses the information in the time series.

Second differences: First differences may remove stable trends in the time series. If the first differences are not stationary, we have three alternatives:

- If the trend is exponential, take logarithms and then first differences. *Do not take second differences instead of logarithms.* The resulting process is not stationary, but it might *seem* stationary because the trend is small.
- If the first differences have a stable trend, take second differences. This is uncommon. See if you can explain why this occurs.

Illustration: If we invest $1,000 in a common stock portfolio, the value of the portfolio is a non-stationary random walk. Taking logarithms and first differences creates a stationary time series. If we invest $1,000 in a common stock portfolio each month, we take first differences, subtract $1,000, then logarithms, and then second differences.

- Graph the first differences. If the graph shows different means by period, we separate into two time periods. If one part of the time series has an upward trend and another part has a downward trend, taking first differences may not make it stationary.

Illustration: If the time series is {1, 2, …, 99, 100, 99, 98, …, 2, 1}, the first differences are {+1, +1, …, +1, +1, –1, –1, …, –1, –1}. The first differences have a mean of +1 in the first half of the time series and –1 in the second half. This is not a stationary process.

Be careful about taking second differences. If the time series comprises two eras with different means, variances, or drifts, taking higher order differences to make a stationary series obscures the true relation. If different models are appropriate for one part of a time series versus another, taking first and second differences obscures the problem.

Recommendation: Comparing two eras of a time series makes a good student project. Divide the time series into two parts and fit two ARIMA processes. An intuitive break is best, such as movie ticket sales before and after home DVD players. If the explanatory variables are not obvious, separate the eras by their means, drifts, or variances.

Step #6: *Robust Models*

A robust model doesn’t change much if we make small changes in the scenario. Examine if the periods chosen create robust models.

Illustration: If a 20 year period has a drift of +2% per annum, each 10 year sub-period should have a drift of about 2% per annum. If the first ten years have a drift of +5% and the second ten years have a drift of –1%, the overall drift of +2% is not meaningful.

Illustration: For a period of a few months and volatility of 0.1% a month, even a time series with no drift may show a small drift. Do not mistake volatility for drift.

The three interest rate periods on the discussion forum illustrate this concept.

- The observed interest rate drifts are +0.02%, –0.03%, and –0.01% per month for the three periods.
- The interest rate volatility is much higher in the second period.

A statistician would say that the first period has an upward drift, the second period is volatile, and the third period has a downward drift.

- The drift in the first and third periods is stable. If we divide the periods in half, we get the same drifts for each half.
- The second period is volatile and short. The absolute value of the drift is high, but it is *not robust*. Changing the period by a few months changes the drift.

To measure the drift, consider also the volatility of the rates and the length of the period.

The +0.02% and –0.01% drifts in the first and third periods reflect FED policy. The –0.03% drift in the middle period is an artifact of the short time period and the high volatility.

Step #7: *SCALING, INTERVAL LENGTH, STOCHASTICITY, AND MOVING AVERAGES*

Choose an appropriate interval length. Some time series specify the interval length. Others allow you to choose the intervals. For stock prices, you may use daily, weekly, or monthly intervals. One might reason: Shorter intervals give more observations.

~ A monthly Treasury bill rate give 12 observations a year.
~ A daily corporate bond index gives 242 to 250 observations a year (business days).

More data points increase the accuracy of the analysis, so a Box-Pierce Q statistic that is not significant with 12 observations may be significant with 250 observations. But intervals that are too short hide the relations. Daily intervals may complicate the analysis.

Illustration: Take first differences of the interest rates and graph the results. The horizontal axis is the month and the vertical axis is the change in the interest rate. The graph looks like white noise. It is hard to see the upward or downward trends.

Monthly data in a stable series do not show the drifts well. The average monthly drifts are

- Period 1: +0.0215% ≈ +0.02%
- Period 2: –0.0283% ≈ –0.03%
- Period 3: –0.0099% ≈ –0.01%

The monthly drifts are too small to see, unless we use narrow markers for the vertical axis. The *annual* drifts of 0.258%, –0.340%, and –0.119% are clear.

If we change just the scale of the vertical axis to use 0.01% as the marker, the interest rate stochasticity overwhelms the drift. We must change the scale *and* use a 12 month moving average to reduce the stochasticity. To observe the drift in your time series, examine

- a line graph of moving averages, which eliminates the stochasticity
- the 12 month first differences (the year to year change in the monthly rate)

Even monthly intervals are short. Monthly intervals give enough data to test hypotheses. But monthly intervals might make a stationary time series seem like a random walk.

Illustration: Suppose <u>*annual*</u> interest rates are a stationary AR(1) time series with ϕ_1 = 80% and δ = 2%, so the mean interest rate is 2% / (1 – 80%) = 10%. <u>*Monthly*</u> interest rates might have an AR(1) process with ϕ_1 = 98% and δ = 0.2%, so the mean interest rate is still 10%. This looks like a random walk, but it is not. Similarly, the daily corporate bond spread looks like a random walk, but it is mean reverting process using longer periods.

If the interest rate per annum is 12% in January 20X7, the forecast for January 20X8 is 80% × 12% + 2% = 11.60%. Using the monthly model, we get

February 20X7: 98% × 12% + .2% = 11.96%

Daily or weekly intervals of annual interest rates obscure the process. One time series on the web site is daily values of the Moody's 30 year corporate bond rate. A time series with daily intervals obscures the process. You choose monthly values as the average value in the month or as the first value in the month. The monthly time series is easier to work with. With the monthly values, use the techniques in this guide to fit a model.

Take heed: The first steps are preparation for your analysis. Your write-up should explain the periods and intervals. Show that you understand the change in the ARIMA process.

- If you choose a robust time series with proper periods and intervals, your ARIMA analysis proceeds smoothly.
- If you don't choose reasonable periods, you may waste much time in your analysis.

Take heed: Contrast overnight LIBOR and corporate bond spreads:

- Overnight LIBOR is a one day rate, and it changes rapidly; use daily periods.
- The corporate bond spread is a twenty year rate; use monthly periods.

Step #8: *Seasonality*

Examine your time series for seasonality, even if you do not expect seasonality. The write-up should explain what you examined, what you found, and the adjustments you made.

- Daily temperature and rainfall have smooth seasonality.
- Children's toys (high sales in December) or group health insurance, reinsurance, and workers' compensation (high sales in January) have strong, discrete seasonality.
- If stochasticity obscures the seasonality in the graph, use monthly or quarterly averages over several years. The hurricane season may not be clear in any one year, but a 20 year average by month shows the pattern. The textbook has several ways of identifying seasonality, such as dividing monthly rates by a 12 month centered moving average.
- Correlograms identify even weak seasonality. GDP, unemployment rates, and inflation have weak seasonality. Check the 12 month sample autocorrelations.
- The hypothesized relations must make sense. Don't follow numbers blindly. A high sample autocorrelation for a lag of 7 months is random fluctuation, not seasonality.
- Annual figures smooth seasonality. Daily and weekly rainfall is seasonal; semi-annual rainfall may not be seasonal.
- Many macroeconomic indices are adjusted for seasonality. CPI (inflation indices), price levels, unemployment rates, and Gross National Product are seasonally adjusted.

Take heed: If a time series has two peaks at opposite ends of the year, annual seasonality may appear as semi-annual seasonality.

Illustration: A quarterly series may have positive autocorrelations at lags 2, 4, and 8, and negative autocorrelations at lags 1 and 3. If the autocorrelation at lag 4 is greater than the autocorrelation at lag 2, this is annual seasonality. An autoregressive parameter of lag 4 may correct all the autocorrelations.

We correct for seasonality several ways, depending on the time series:

- seasonally adjust the data
- use seasonal differences
- use a seasonal lag in the ARIMA model

Illustration: Youth unemployment is highest in the summer, when school is not in session. Farm and construction work is high in the summer and low in the winter. We seasonally adjust unemployment rates to identify trends, cycles, and other effects.

Seasonal adjustments are covered in the chapter on non-stochastic time series. They are used whenever the value of a series depends on the time of the year, not the value of the series one year back.

Illustration: Suppose we model daily temperature with an ARIMA process. We seasonally adjust the data and then fit the ARIMA process. We *don't* use a 365 day lag, and we *don't* model the year-to-year changes in the daily temperature. See the project template on daily temperature for explanation. A student project may find the optimal method to seasonally adjust the data.

- For seasonal items, such as textbooks, camping equipment, heating oil, and wedding dresses, we examine growth by the year-to-year change in monthly sales.
- If a figure depends on the value 12 months back, ARIMA seasonal lags are best.

Illustration: We model personal auto written premium by month for a direct writer. The policy renewal rate is 90%, so the value 12 months ago is the proper base. We use ARIMA models with a ϕ_{12} of about 90%.

A student project on seasonality may have the following steps.

Daily Averages

Examine average values by day of the year.

- If interest rates have no trend or cycles, compute the average interest rate over the entire period for each day of the year.

- If interest rates have a trend or a cycle, simple averages conflate trend and seasonality.

Distinguish trend from seasonality:

- Compute 365 day centered moving averages. This eliminates seasonality and random fluctuations, leaving trend and cycles.
- To eliminate trend, convert interest rates to their deviations from a 365 day centered moving average.

Take heed: Overnight LIBOR shows business days only, or about 242 to 250 days a year. Use a centered moving average of the 365 calendar days, which cover a variable number of business days.

Take heed: Use the *countif* and *sumif* built-in functions to compute daily averages. The illustrative work-sheet for the project template on daily temperature uses these functions.

- Leap years cause an extra day every fourth year.
- Overlap of holidays with weekends may cause more business days some years.
- Missing values may cause fewer days.

Graph the moving averages. You see a decline followed by a rise in the overnight LIBOR for the period on the discussion forum Excel work-book. A student project might examine the relation of these movements to other macroeconomic indices. The centered moving average smooths the trends.

Take heed: The daily temperature over the past 130 years may have weak trends or cycles (depending on the weather station). See the project template on daily temperature for methods of dealing with temperature trends.

Subtract the centered moving averages from the observed values. This eliminates trends and cycles, leaving seasonality and fluctuations. Each observed value is a deviation from the average in the surrounding year.

Compute long-term averages. This reduces the random fluctuation and leaves seasonality.

Take heed: Daily temperature has a high error term. The daily temperature may fluctuate ±20° because of unexpected weather. The daily temperature may be 30° one day and 65° two days later. Even a 130 year average daily temperature shows random fluctuations. Interest rates fluctuate less, and fluctuations are gradual. Graph the results as a line chart.

- If the line is smooth, the random fluctuations don’t distort the long-term averages.
- If the line is jagged, replace each value by its centered moving average for 3 days or 5 days or 7 days or some other period. Use judgment to select the proper period. See the project template for daily temperature for an example.

Interest Rates and Seasonality

For interest rates, seasonality is much weaker now than in the past and may be seen only in short rates. A student project using rates from the past few decades or rates with durations of one year or more need not discuss seasonality.

Interest rates are seasonal because they depend on the supply and demand for money.

The demand for money varies over the year. It is high for holiday shopping and low in January and February. If the money supply is held constant, interest rates are seasonal.

Eighty years ago, interest rates were seasonal. Now the Federal Reserve Board varies the supply of money to offset the demand for money.

The Federal Reserve Board is not perfect, and some seasonality remains. See if overnight LIBOR has any seasonality. You may examine whether a seasonal autoregressive term improves the model for overnight LIBOR.

Overnight, one week, two week, and one month LIBOR might be seasonal. A rate for one year or longer is not seasonal.

The graphs don’t show much seasonality even for short LIBOR rates. But the correlogram may show a high 12 month sample autocorrelation. The pattern may be even clearer in the first differences.

Recommendation: Decomposing LIBOR, insurance claim costs, and other actuarial items into long-term trends, cycles, seasonality, and stochasticity makes a good student project that can be valuable to your employer.

Take heed: If the time series is a random walk with weak mean reversion or seasonality, the sample autocorrelations show a slow decline for the first 11 months and a slight spike in the twelfth sample autocorrelation.

Stochasticity obscures weak seasonality. Annual seasonality may also cause high sample autocorrelations at 6 months, which further disrupts the pattern. Use the correlogram for the first differences to identify weak seasonality.

Illustration: The 1945-1978 period in the interest rates illustrative worksheet has a 17% correlation for the 12 month lag and lower correlations for the other lags. We have 34 × 12 = 374 observations in the first period. We subtract 1 for the first differences and 12 for the 12 month lag to get 361 observations. A sample autocorrelation higher than $2 \times 1/\sqrt{361} = 10.5\%$ is significant at a 5% level. A 17% sample autocorrelation is significant; most other sample autocorrelations are below 10.5% in absolute value.

Take heed: Don't expect all other sample autocorrelations to be below the critical value. By chance, one or two may be higher.

Many student projects use a correlogram, graph the sample autocorrelations, and conclude that seasonality is not important. Always examine seasonality; you may be surprised.

Illustration: Many candidates are not aware that claim frequency and severity are seasonal in many lines of business. Workers' compensation, group health insurance, reinsurance, are often written on January 1, so written premium is also seasonal.

If you find a significant 12 month sample autocorrelation, try an ARIMA(12,1,0) model: ϕ_1 and ϕ_{12} are non-zero, and the other coefficients are zero. Estimate the parameters with multiple linear regression.

If interest rates are a random walk with annual seasonality, we expect

- ϕ_{12} is low for a non-seasonal product and high for a seasonal product.
- ϕ_1 is low for a white noise process and high for a random walk.

Step #9: *Trends*

A time series with a trend or drift is not stationary. We distinguish the two terms.

- A regression line has a trend.
- A random walk and other autoregressive processes have drifts.

Illustration: Suppose inflation has a drift of 5% per annum. If inflation is 3% in 20X8, we might expect it to be 4% in 20X9: an AR(1) process of the first differences with $\phi_1 = 50\%$.

Illustration: Suppose average claim severity has a trend of 5% per annum. If claim severity increases 3% in 20X8, we might expect it to increase 6% in 20X9. We assume that random fluctuations causes claim severity to be below trend in 20X8, so the increase in 20X9 is higher than trend.

Take heed: Trends in marriage rates, divorce rates, abortion rates, crime rates may change direction. A student project may examine the ARIMA process before and after the change in trend. Graph the data, separate into periods, and fit models to each period.

Stochasticity and seasonality obscure trends. A student project on climate change can be wonderful. Extensive data can be found on public web sites, and the implications are hotly debated. But high weather stochasticity overwhelms small trends. You may fit an ARIMA process to long-term weather indices to see if trends are real.

Illustration: To see trends in home sales, we use 12 month moving averages to remove the seasonality and dampen the stochasticity. Households buy homes in the summer months more than in winter months. A 2% annual trend in real (inflation-adjusted) home sales is obscured by the seasonality and random fluctuations.

Recommendation: Recent economic changes show the difficulty of identifying trends. Economists disagree if the U.S. economy is heading toward recession, if credit problems are a correction of lax lending practices, or if banks gave loans to weaker borrowers to meet federal non-discrimination requirements. A student project on home sales or mortgage rates might examine these issues.

Linear vs Exponential Trends

To distinguish among linear, exponential, and other trends, graph the data.

- A linear trend appears as a straight line.
- An exponential trend appears as a convex (concave upward) curve.

Checking if a trend is linear or exponential is not easy. A non-linear trend is not necessarily exponential. It may be

- A linear trend whose slope coefficient changes over time.
- A linear trend with much random fluctuation.
- A non-linear and non-exponential trend.

Decide if a trend is linear or exponential two ways:

1. Compare the trend of the time series to the trend in the logarithm of the time series.
2. Decide *intuitively* whether a linear or exponential trend makes more sense.

Illustration: If $100 rises to $110,$200 should rise to $220. The relation is multiplicative and the trend is exponential. But if Greenland's temporary rises from 1° Celsius to 2° Celsius in the 20th century, it might rise to 3° Celsius in the 21st century: a linear trend.

To adjust a time series for trend:

- For a linear trend, take first differences.
- For an exponential trend, take logarithms and then take first differences.

For stock prices, financial analysts take logarithms of ratios, which are the first differences of the logarithms. Either method is fine for the student project.

Summary

The initial steps in ARIMA modeling include: separate the time series in homogenous periods, de-seasonalize the data, and adjust for trends.

- If the time series differs in two time periods, separate the periods.
- If the time series is seasonal, de-seasonalize the data, take a seasonal difference, or use a seasonal lag in the ARIMA model.
- If the time series has a linear trend, take first differences.
- If the time series has an exponential trend, take logarithms and first differences.

Step #10: *Stationarity*

Convert the time series to stationary form. Be careful not to take differences unless they are appropriate.

Illustration: The first and third interest rate periods in the interest rate project template have upward or downward drifts. They are not stationary, but their first differences may be stationary. Your student project tests for stationarity and fits an appropriate model.

The middle period is more complex. The interest rates in the middle period are volatile, and the drift may be random fluctuation. Even if the drift is zero and the volatility is constant, the time series could be white noise or a random walk. White noise is stationary and a random walk is not stationary. Your student project may test if the process is white noise, a random walk, or something else.

To *test* if a series is stationary, use sample autocorrelations, unit roots, and correlograms.

(1) Regress the time series on the same values one period back. This is an AR(1) model, which is the most common ARIMA process. If ϕ_1 (the β of the regression equation) is more than 1 or less than –1, the time series is not stationary. We see this in the graph as well.

- If $\phi_1 > 1$, the time series grows continually. Random fluctuations may cause any single value to be smaller (in absolute value) than the preceding one, but the growth is clear over long periods. To correct this, take (logarithms and) first differences.
- If $\phi_1 < -1$, the time series grows continually and oscillates. Random fluctuations may obscure the exact process, but the oscillations are evident. This type of process is rare.

(2) If ϕ_1 is ≈ 1, the time series is a random walk and is not stationary. Because of random fluctuations, the ordinary least squares estimator of the parameter is never exactly one.

- If we estimate ϕ_1 as 0.95 in a time series of 40 observations, we assume it is one and the time series is a non-stationary random walk.
- If we estimate ϕ_1 as 0.80 in a time series of 400 observations, we assume it is less than one and the time series is a stationary AR(1) process.

We rely on judgment, not on hard rules: *t* statistics and *p*-values for the null hypothesis that $\phi_1 = 1$ help us decide, but we don't have rigid rules.

Illustration: In the regression for the AR(1) model on the interest rate project template, the first and third periods have a ϕ_1 of 0.99, and the second period has a ϕ_1 of 0.85. [These are the values in the illustrative worksheet. You may choose a different time series and different periods. You will get different parameters and perhaps different conclusions.]

The volatility is higher in the middle period than in the first or third periods. All three periods may be random walks, but the volatility is so high in the middle period and the length of the period is so short (48 months) that the slope coefficient has a high variance.

(3) The correlogram tests if the time series is stationary. If the autocorrelations do not *decline rapidly*, the time series is not stationary. *Rapid decline* means *at least geometric decline*. The time series is stochastic, and it may be hard to judge if the decline is rapid.

Checking if an interest rate time series is stationary is not easy. Annual interest rates are moving averages of 12 monthly forecasts. Since 11 of the 12 months are the same in adjoining periods, we get a $\phi_1 \approx 1$ in an AR(1) process.

Take heed: Be careful when you examine the stationarity of long duration interest rates. Overnight LIBOR fluctuates rapidly; Moody's 30 year corporate bond rate is steady.

If the time series or its first differences is stationary with a $\phi_1 < 1$, we have a *possible* AR(1) model. We consider three more items:

- Is the model correct? (Are the residuals close to a white noise process?)
- Is the model optimal? (Do other ARIMA processes fit equally well or better?)
- Does the model forecast well? (Do future values fall within a confidence interval?)

COMMON ERRORS: FIRST DIFFERENCES

A random walk is not stationary and an AR(1) process with a high ϕ_1 (but less than one) is stationary. Time series are stochastic, and it hard to distinguish the two scenarios.

- Not taking first differences of the random walk leaves a non-stationary series.
- Taking first differences of the AR(1) process is an error. We lose information about the time series, making it harder to forecast future values.

Take heed: Some candidates reason: the correlogram does not decline to zero quickly. The first differences form a more clearly stationary time series, which is easier to model.

Don't take first differences simply to make the time series easier to model. Take first differences only if the time series is not stationary. First differences lose information.

Illustration: We use monthly interest rates to get enough data points to test hypotheses. But monthly interest rates might make a stationary time series seem like a random walk.

Suppose *annual* interest rates are a stationary AR(1) time series with $\phi_1 = 80\%$ and $\delta = 2\%$, so the mean interest rate is 10%. *Monthly* interest rates might have an AR(1) process with $\phi_1 = 98\%$ and $\delta = 0.2\%$, so the mean interest rate is still 10%. This looks like a random walk, but it is not.

If the interest rate per annum is 12% in January 20X7, the forecast for January 20X8 is 80% × 12% + 2% = 11.60%. Using the monthly model, we get

February 20X7: 98% × 12% + .2% = 11.96%

Repeating this 12 times gives a forecast for January 20X8 of about 11.6%.

Take heed: If the correlogram shows geometric decline, the time series is stationary, even if the decline is slow.

COMMON ERRORS: MOVING AVERAGES

Use moving averages to find trends, not to test for stationarity. Moving averages are often used to remove seasonality. The moving averages obscure seasonality; they don't test for seasonality. Test for seasonality by examining monthly figures, and adjust for seasonality by one of the other methods in this course.

Don't use 12 month moving averages to form better correlograms. 11 of 12 months are the same in adjoining periods, so the sample autocorrelation function is high.

12 month interest rates are moving averages of 12 monthly forecasts, so be careful when you examine the stationarity of long rates. The correlogram examines the autocorrelation of long range forecasts. The forecasts change slowly; an interval of one month might show a random walk even if the true process is AR(1) with a high ϕ_1 parameter.

Take heed: The NEAS web site shows daily estimates of the investment grade corporate bond rate. A daily correlogram has such short intervals that patterns are hard to observe. You may use monthly intervals to evaluate the correlogram.

Take heed: The length of the time series (number of observations) does not determine the proper length of intervals.

- A time series of daily temperature over 100 years may have 36,524 observations.
- The daily temperature changes so quickly that intervals longer than 1 day do not show ARIMA processes.
- You may use hourly time series, with 24 hour seasonality overlaid on 365 day patterns.

Use quarterly or annual intervals to test if 90 day or 1 year Treasury bills are stationary. This is fine for the first period on the NEAS web site, which has 34 years. The middle period has only 4 years, and we can not use annual rates.

Take heed: If possible, detrend the time series, eliminate cycles and inflation, adjust for seasonality, and use methods besides taking differences to make a time series stationary. Taking differences is the proper adjustment only for random walks. The textbook does not make this clear.

Step #11: *Testing for White Noise*

The objective of ARIMA modeling is to forecast future values. Time series are stochastic, so forecasts do not exactly equal the future values. Ideally, the ARIMA process eliminates everything but the white noise of random fluctuations (the error term).

Check for white noise two places in your student project:

~ Once the time series is stationary, check if it is white noise.
~ After fitting an ARIMA process, check if the residuals are white noise.

Some statisticians consider white noise an ARIMA(0,0,0) process. If the first differences are white noise, the series is an ARIMA(0,1,0) process. Other statisticians say that white noise doesn't require an ARIMA model.

Use three tests for white noise: Durbin-Watson statistic, Bartlett's test, and Box-Pierce Q statistic. If the model is correct and the residuals are white noise:

- The regression of the series on the series lagged one period has no serial correlation, so the Durbin-Watson statistic is ≈ 2.
- The sample autocorrelation of the residuals is normally distributed with a standard deviation of $1/\sqrt{T}$. Test this by examining percentiles. (Excel has a built-in function to test the percentiles. The function is not explained in the textbook, but you may use it.)
- The Box-Pierce Q statistic has a χ-squared distribution with the appropriate degrees of freedom.

If you have taken the regression analysis course, you can check the significance of the test using the Durbin-Watson table in the textbook. Keep in mind two items:

~ We are using a lagged value as the independent variable in the regression. The critical values for significance are not proper in this scenario. We use the Durbin-Watson statistic to help examine the time series, but we do not draw firm conclusions.
~ If the independent variable itself has a high autocorrelation, the Durbin-Watson statistic overstates the correlation of the residuals. The Durbin-Watson statistic may give wrong conclusions for time series modeling, so be careful with your analysis. The textbook mentions the problem, and recommends the Box-Pierce Q statistic instead.

For the student project, you may use the Durbin-Watson statistic. We want to see if you understand how to use the tool and what the results mean. We know that the test is not accurate for time series, and we do not require that you comment on this.

The Durbin-Watson statistic differs from Bartlett's test and the Box-Pierce Q statistic:

- The Durbin-Watson statistic uses autocorrelations of lag 1. Bartlett's test and the Box-Pierce Q statistic use autocorrelations of many lags. Bartlett's test and the Box-Pierce

Q statistic are more robust, but if the sample autocorrelation of lag 1 is close to zero, the time series is probably a white noise process.

- The tests are scaled differently. White noise has a Durbin-Watson statistic of 2, sample autocorrelations that vary normally about zero (Bartlett's test), and a Box-Pierce Q statistic that is lower than the relevant χ-squared statistic.
- The progression of X values affects the autocorrelation of lag 1, so hypothesis testing is more complex for the Durbin-Watson statistic. The regression analysis module on the Durbin-Watson statistic explains how the correlation of the X values obscures the sample autocorrelation of the residuals.

If the Durbin-Watson statistic is 2, the process does not have an autoregressive coefficient of lag 1. It may have moving average coefficients or a seasonal autoregressive coefficient.

- Most ARIMA process have an autoregressive coefficient of lag 1. If the Durbin-Watson statistic is less than 1.600 to 1.700 (depending on the length of the time series), we examine AR(1) and AR(2) processes.
- If the Durbin-Watson statistic is between 1.800 and 2.200, the time series may be a white noise process. We examine Bartlett's test and the Box-Pierce Q statistic.

Bartlett's test and the Box-Pierce Q statistic examine more sample autocorrelations, such as the first 20 values. If they are close to zero, the time series is probably white noise. The *standard deviation* of the sample autocorrelations depends on the length of the time series. We have decades of interest rates, so we use them for the project templates. If you use ten years of annual premium volume for your student project, the data are too sparse for the statistical tests.

We check the percentage of sample autocorrelations with absolute values above various levels. We use judgment to evaluate the significance. This is a strong test. If you are familiar with Excel, you can use built-in functions for most of the work.

Review the discussion forum posting on time series techniques. We provide cell formulas and functions needed to complete the student project.

Step #12: *ARIMA MODELS*

Once the time series is stationary but not white noise, specify an ARIMA process. The most common processes are AR(1), AR(2), MA(1), and ARMA(1,1). Each process has seasonal versions, versions with first differences, and versions with logarithms.

Illustration: An AR(1) process might be ARIMA(1,1,0), AR(12), where the ϕ_{12} parameter is for seasonality, ARIMA (12,1,0), or logarithmic versions of these.

Use simple processes for the student project. We review the student projects to see if you use statistical techniques properly. If you specify and test AR(1), MA(1), and ARIMA(1,1,0) models, and you explain what each model implies, you have completed the student project.

For each process, select parameters.

- For AR(1) and AR(2) processes, fit the model with linear regression.
- For MA(1) processes, use the Yule-Walker equations.

You may use other statistical software to fit the models. You may also use Excel *SOLVER* built-in function to fit a model.

Some time series are white noise after taking differences and logarithms, adjusting for seasonality, and regressing on economic or financial variables. If you begin with a time series that is not a simple random walk, and you take differences and logarithms, adjust for seasonality, regress on economic or financial variables, and convert your time series to a white noise process, your student project is fine.

Illustration: You use interest rates, daily temperatures, unemployment rates, inflation rates, sports won-loss records, sales volume, baby names, claim severity, or claim frequency, and you obtain a white noise process after the adjustments mentioned above:

- Check if a time series with fewer differences is stationary. Some candidates assume a time series with lower sample autocorrelations is better. They use second differences if that gives lower sample autocorrelations than the initial time series or first differences.
- If the time series with fewer differences is not stationary, write up the student project and turn it in. Do not think you erred because your result is a white noise process.

If you start with a white noise process or a random walk, you don’t use ARIMA modeling.

- Do not use earthquake frequency (a white noise process) and test for white noise. We are not asking if you can form a white noise process.
- Do not use daily stock prices (a random walk), take logarithms and first differences, and say the result is white noise. This does not show that you can use ARIMA processes.

You don't use the statistical techniques in the time series course for the two series above. But both series can be used if you analyze seasonality, cycles, drifts, and trends.

You can begin with daily stock prices and model weekly or annual seasonality. Economists refer to these as the Monday effect and the January effect. Statisticians have spent years modeling these two effects, and we don't know what causes them. Many financial papers analyze these patterns, and they are good topics for student projects.

Illustration: Monday effect

Take logarithms and first differences of daily stock prices. Index the data so that Mondays are 1 mod 5, Tuesdays are 2 mod 5, …, and Fridays are 0 mod 5. Use an AR(5) model with values for ϕ_1 and ϕ_5, which you can fit easily with Excel.

Take heed: Holidays (New Year's, Fourth of July, Thanksgiving) don't have stock prices. To obtain entries for the time series, use the geometric average of the adjoining days.

Illustration: January effect

Take logarithms and first differences of monthly stock prices. Use an AR(12) model with values for ϕ_1 and ϕ_{12}. The monthly stock price may be the average in the month or the value on the 15th of the month.

Illustration: Natural catastrophes

Hurricanes have possible trends and cycles. You can fit an ARIMA model to a hundred year history of hurricane frequency.

Take heed: You don't need complex ARIMA models for the student project. Your student project should show that you understand how a moving average model differs from an autoregressive model and that you know how to test for each model. Use simpler models. If the simple models do not pass the Box-Pierce Q statistic, explain in your write-up what else a statistician might look at.

Illustration: Suppose your student project examines new home sales in Boston, and no simple ARIMA process gives white noise residuals. Your student project may say:

"New home sales are affected by economic conditions and mortgage rates. I regressed new home sales on GDP, but the indices I used were rough. Ideally, we should examine the residuals of new home sales in Boston regressed on per capital real persona income in Boston and on new home mortgage rates. In addition, I looked only at AR(1), AR(2), and MA(1) processes. A more complete analysis would look at ARIMA processes with more parameters."

Step #13: Correlograms

To choose among ARIMA processes, examine the correlograms. The illustrative spreadsheet on the NEAS web site gives the sample autocorrelation function.

- Autoregressive models have geometrically declining sample autocorrelations and spikes in the partial autocorrelation function.
- Moving average models have spikes in the sample autocorrelation function and declining partial autocorrelations.

Use the partial autocorrelation function if you have more sophisticated statistical software. To determine partial autocorrelations, we use nonlinear regression, which we do not cover in the statistics courses. Use the sample autocorrelations in the correlogram.

The autocorrelations from an AR(1) model have a geometric decline beginning with the first lag. The *sample* autocorrelations are the relations in a sample of observations. Ten years of monthly interest rates give 120 observations. The sample autocorrelations are stochastic and do not show a perfect geometric decline.

- If the sample autocorrelations have a reasonably rapid decline (but don't drop to zero immediately), AR(1) is usually the best model. You may test other processes, but the AR(1) solution is fine. Unless the number of observations is high and ϕ_1 is close to 1 (or –1), you can't confirm that the decay is geometric, since stochasticity overwhelms the expected results.
- If the sample autocorrelations are close to zero after the first lag, the indicated model depends on the first sample autocorrelation and the number of elements.
  - If the sample autocorrelation for the first lag is high or negative, the model may be MA(1).
  - If the sample autocorrelation for the first lag is positive but low, such as 15%, the model is probably AR(1). The indicated sample autocorrelation for the second lag of an AR(1) process is $15\%^2 = 2.25\%$, which is overwhelmed by stochasticity. Even if the sample autocorrelation for the first lag is 30% or 40%, the stochasticity is large enough in small or medium-size samples to obscure the sample autocorrelations.
- If the sample autocorrelation for the first lag is high or negative, and the remaining sample autocorrelations have a geometric decline, the model may be ARMA(1,1).
- If the sample autocorrelations for the first two lags are high and geometrically declining afterward, the model may have an AR(2) term and perhaps MA terms of order 1 or 2.

Step #14: *Structural Models*

Some candidates presume that structural models are more complex, so the student project takes more time. The opposite is true: the regression eliminates much of the variability in the original time series, and the ARIMA fitting is easier.

Illustration: Moody's investment grade corporate bond yield shows fluctuations, cycles, and trends. The corporate bond spread (after subtracting the long-term Treasury bond rate) is an ideal time series for ARIMA modeling.

Regress the corporate bond spread on the GDP growth rate. The residuals should be a stationary time series, which you might fit as white noise, AR(1), MA(1), or ARMA(1,1).

If you model the corporate bond yield itself

- The fit is less good and you may have to separate the time series into periods.
- You spend more time analyzing correlograms and choosing among alternative models.

By modeling the residuals of the corporate bond spread regressed on the GDP growth rate, you spend an extra hour getting the time series, but the rest of your project is quick.

Step #15: *ORDER OF MODELS*

The textbook uses correlograms and in-sample tests to select a model. But ARIMA modeling is imprecise, since other factors may affect the time series values.

For the student project, use a sequence of models. First check trends and seasonality.

- De-trend the values. Use first differences and logarithms to determine the type of trend. You get a better ARIMA fit if you de-trend the time series with an inflation index and you de-seasonalize the data than if you take differences.
- Add seasonal lags or de-seasonalize the time series if needed.

Form a correlogram to see if an AR(1) or MA(1) model is appropriate.

Fit an AR(1) model with ordinary least squares for the autoregressive parameter and the seasonal parameter, if any. Compute the residuals from the AR(1) model and check the Durbin-Watson statistic, Barlett's test, and the Box-Pierce Q statistic.

If these tests are not significant, the residuals may be a white noise process and an AR(1) model is reasonable.

Do the same fitting for an AR(2) model, and state whether the better fit justifies the extra parameter. *If the AR(2) model is much better than the AR(1) model*, use it for the diagnostic testing; otherwise, we use the principle of parsimony and stick with AR(1).

Use the Yule-Walker equations to estimate θ_1 for an MA(1) process. Estimate the residuals from this model and apply the in-sample goodness-of-fit tests.

Take heed: Statistical software uses nonlinear regression to estimate an MA(1) process. For the student project, use the Yule-Walker equations. The estimated model is not the best possible, but it is close. We examine whether you correctly form the residuals from the MA(1) model for the Box-Pierce Q statistic and Bartlett's test.

Jacob: When do we use higher order autoregressive models?

Rachel: Some statisticians routinely use high order ARIMA processes; others do not. If you use the simple model correctly, and you explain what each model implies, you don't need more complex models.

Jacob: What ARMA(1,1), as well as its ARIMA and seasonal variants?

Examine the correlogram to see if ARMA(1,1) is indicated. If it is, explain how we see this from the correlogram. You do not have to fit the model.