Published on *STAT 510* (https://onlinecourses.science.psu.edu/stat510)

Home > Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values

Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values

Assignments:

- Reading Assignment
  - Sections 3.1 and 3.2 (pages 83-96) are worth a skim, but don't get too bogged down in the theory.
  - Section 3.7 (pages 141-144).
  - Section 3.8 (pages 144-153) is a core section about building models.
  - Sections 3.3-3.6 are quite mathematical and can be skipped or lightly skimmed unless you have a deep interest in theory. Pages 106-108 are of interest.
  - Pages 52-53 in Chapter 2 might also be helpful.
- Read through the Lesson 3 online notes that follow.
- Complete Lesson 3 Assignment.

Overview:

This week we'll learn some techniques for identifying and estimating non-seasonl ARIMA models. We'll also look at the basics of using an ARIMA model to make forecasts. We'll look at seasonal ARIMA models next week. Lesson 3.1 gives the basic ideas for determining a model and analyzing residuals after a model has been estimated. Lesson 3.2 gives a test for residual autocorrelations. Lesson 3.3 gives some basics for forecasting using ARIMA models. We'll look at other forecasting models later in the course.

This all relates to Chapter 3 in the book, although the authors give quite a theoretical treatment of the topic(s).

Learning Objectives:

After successfully completing this lesson, you should be able to:

- Identify and interpret a non-seasonal ARIMA model
- Distinguish ARIMA terms from simultaneously exploring an ACF and PACF
- Test that all residual autocorrelations are zero
- Convert ARIMA models to infinite order MA models
- Forecast with ARIMA models
- Create and interpret prediction intervals for forecasts

3.1 Non-seasonal ARIMA Models

ARIMA models, also called Box-Jenkins models, are models that may possibly include autoregressive terms, moving average terms, and differencing operations. Various abbreviations are used:

- When a model only involves autoregressive terms it may be referred to as an AR model. When a model only involves moving average terms, it may be referred to as an MA model.
- When no differencing is involved, the abbreviation ARMA may be used.

Note: This week we're only considering non-seasonal models. We'll expand our toolkit to include seasonal models next week.

Specifying the Elements of the Model

In most software programs, the elements in the model are specified in the order (AR order, differencing, MA order). As examples,

A model with (only) two AR terms would be specified as an ARIMA of order (2,0,0).

A MA(2) model would be specified as an ARIMA of order (0,0,2).

A model with one AR term, a first difference, and one MA term would have order (1,1,1).

For the last model, ARIMA (1,1,1), a model with one AR term and one MA term is being applied to the variable $z_t = x_t - x_{t-1}$. A first difference might be used to account for a linear trend in the data.

The differencing order refers to successive first differences. For example, for a difference order = 2 the variable analyzed is $z_t = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$, the first difference of first differences. This type of difference might account for a quadratic trend in the data.

Identifying a Possible Model

Three items should be considered to determine a first guess at an ARIMA model: a time series plot of the data, the ACF, and the PACF.

Time series plot of the observed series.

In Lesson 1.1, we discussed what to look for: possible trend, seasonality, outliers, constant variance or nonconstant variance.

- You won't be able to spot any particular model by looking at this plot, but you will be able to see the need for various possible actions.
- If there's an obvious upward or downward linear trend, a first difference may be needed. A quadratic trend might need a 2nd order difference (as described above). We rarely want to go much beyond two. In those cases, we might want to think about things like smoothing, which we will cover later in the course. Over-differencing can cause us to introduce unnecessary levels of dependency (difference white noise to obtain a MA(1)–difference again to obtain a MA(2), etc.)
- For data with a curved upward trend accompanied by increasing variance, you should consider transforming the series with either a logarithm or a square root.

  Note: Nonconstant variance in a series with no trend may have to be addressed with something like an ARCH model which includes a model for changing variation over time. We'll cover ARCH models later in the course.

ACF and PACF

The ACF and PACF should be considered together. It can sometimes be tricky going, but a few combined patterns do stand out. (These are listed in the Table 3.1 of the book on page 108).

- AR models have theoretical PACFs with non-zero values at the AR terms in the model and zero values elsewhere. The ACF will taper to zero in some fashion. (Example [1]) An AR(1) model has an ACF with a pattern

$$\rho_k = \rho_1^k$$

- An AR(2) has a sinusoidal ACF that converges to 0. (Example [2])
- MA models have theoretical ACFs with non-zero values at the MA terms in the model and zero values elsewhere. (Example [3])
- ARMA models (including both AR and MA terms) have ACFs and PACFs that both tail off to 0. These are the trickiest because the order will not be particularly obvious. Basically you just have to guess that one or two terms of each type may be needed and then see what happens when you estimate the model. (Example [4])
- If the ACF and PACF do not tail off, but instead have values that stay close to 1 over many lags, the series is non-stationary and differencing will be needed. Try a first difference and then look at the ACF and PACF of the differenced data.
- If all autocorrelations are non-significant, then the series is random (white noise; the ordering matters, but the data are independent and identically distributed.) You're done at that point.
- If you have taken first differences and all autocorrelations are non-significant, then the series is called a random walk and you are done. (A possible model for a random walk is $x_t = \delta + x_{t-1} + w_t$. The data are dependent and are not identically distributed; in fact both the mean and variance are increasing through time.)

Note: You might also consider examining plots of x_t versus various lags of x_t.

Estimating and Diagnosing a Possible Model

After you've made a guess (or two) at a possible model, use software such as R, Minitab, or SAS to estimate the coefficients. Most software will use maximum likelihood estimation methods to make the estimates. Once the model has been estimated, do the following.

- Look at the significance of the coefficients. In R, p-values aren't given. For each coefficient, calculate z = estimated coeff. / std. error of coeff. If |z| > 1.96, the estimated coefficient is significantly different from 0.
- Look at the ACF of the residuals. For a good model, all autocorrelations for the residual series should be non-significant. If this isn't the case, you need to try a different model.
- Look at Box-Pierce (Ljung) tests for possible residual autocorrelation at various lags (see Lesson 3.2 for a description of this test).
- If non-constant variance is a concern, look at a plot of residuals versus fits and/or a time series plot of the residuals.

If something looks wrong, you'll have to revise your guess at what the model might be. This might involve adding parameters or re-interpreting the original ACF and PACF to possibly move in a different direction.

What if More Than One Model Looks Okay?

Sometimes more than one model can seem to work for the same dataset. When that's the case, some things you can do to decide between the models are:

- Possibly choose the model with the fewest parameters.
- Examine standard errors of forecast values. Pick the model with the generally lowest standard errors for predictions of the future.
- Compare models with regard to statistics such as the MSE (the estimate of the variance of the w_t), AIC, AICc, and SIC (also called BIC). Lower values of these statistics are desirable.
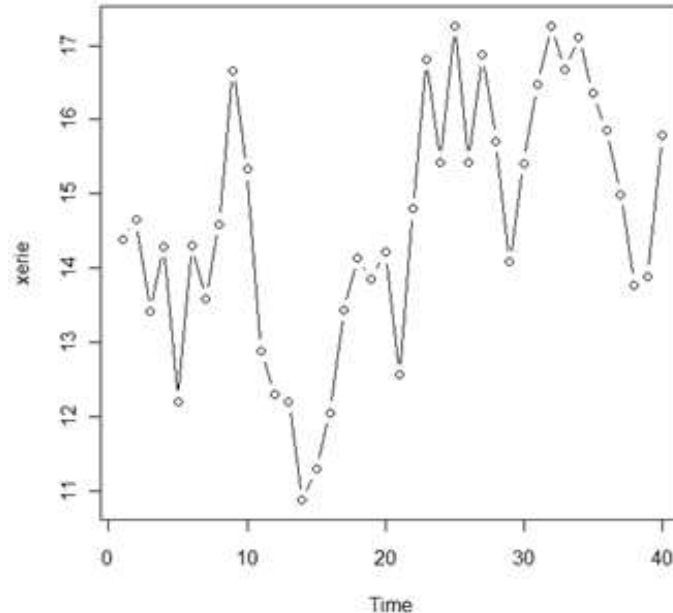
  AIC, AICc, and SIC (or BIC) are defined and discussed on pages 52-53 of our book. The statistics combine the estimate of the variance with values of the sample size and number of parameters in the model.

One reason that two models may seem to give about the same results is that, with the certain coefficient values, two different models can sometimes be nearly equivalent when they are each converted to an infinite order MA model. [Every ARIMA model can be converted to an infinite order MA – this is useful for some theoretical work, including the determination of standard errors for forecast errors.] We'll see more about this in Lesson 3.2.

Example 1: The Lake Erie data from Week 1 assignment. The series is n = 40 consecutive annual measurements of the level of Lake Erie in October.
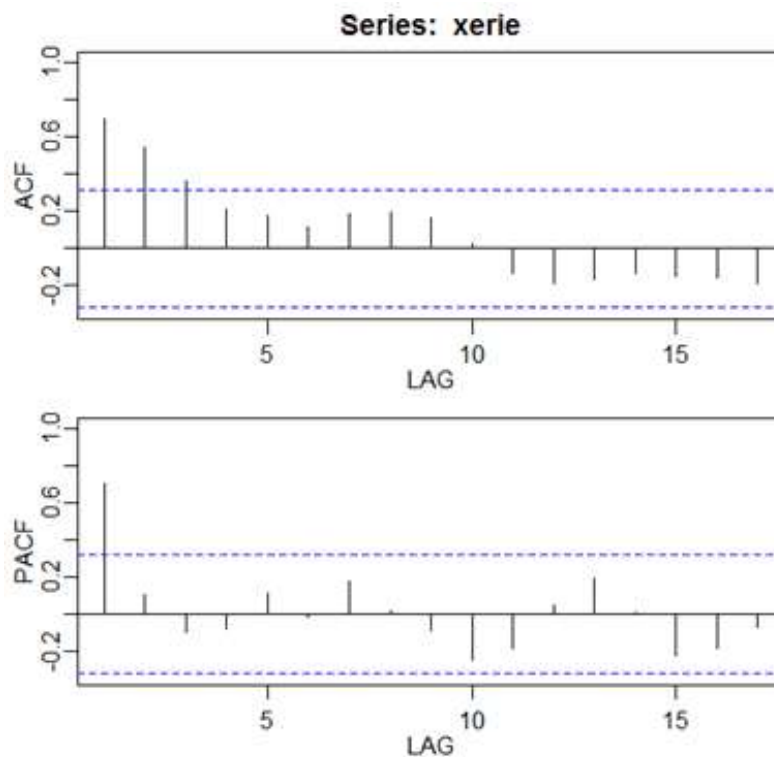
Identifying the model.

A time series plot of the data is the following:

There's a possibility of some overall trend, but it might look that way just because there seemed to be a big dip around the 15th time or so. We'll go ahead without worrying about trend.

The ACF and the PACF of the series are the following. (They start at lag 1).

Series: xerie

The PACF shows a single spike at the first lag and the ACF shows a tapering pattern. An AR(1) model is indicated.

Estimating the Model

We used an R script written by one of the authors of our book (Stoffer) to estimate the AR(1) model. Here's part of the output:

```
Coefficients:
```

	ar1	xmean
Estimate	0.6909	14.6309
Std. Error	0.1094	0.5840

```
sigma^2 estimated as 1.447: log likelihood = -64.47,
$AIC
[1] 1.469364
$AICc
[1] 1.536030
$BIC
[1] 0.5538077
```
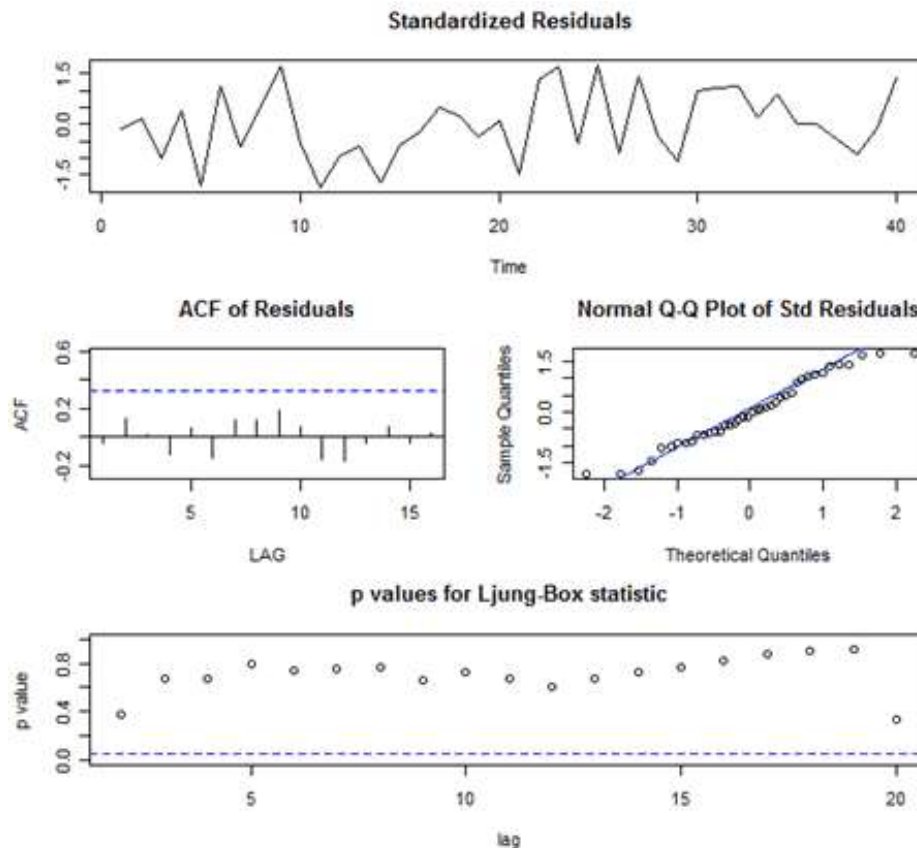
Where the coefficients are listed, notice the heading "xmean." This is giving the estimated mean of the series based on this model, not the intercept. The model used in the software is of the form $(x_t - \mu) = \phi_1(x_{t-1} - \mu) + w_t$.

The estimated model can be written as $(x_t - 14.6309) = 0.6909(x_{t-1} - 14.6309) + w_t$.

This is equivalent to $x_t = 14.6309 - (14.6309 \ast 0.6909) + 0.6909x_{t-1} + w_t = 4.522 + 0.6909x_{t-1} + w_t$.

The AR coefficient is statistically significant (z = 0.6909/0.1094 = 6.315). It's not necessary to test the mean coefficient. We know that it's not 0.

The author's routine also gives residual diagnostics in the form of several graphs. Here's that part of the output:

Standardized Residuals

ACF of Residuals

Normal Q-Q Plot of Std Residuals

p values for Ljung-Box statistic

Interpretations of the Diagnostics

The time series plot of the standardized residuals mostly indicates that there's no trend in the residuals, no outliers, and in general, no changing variance across time.

The ACF of the residuals shows no significant autocorrelations – a good result.

The Q-Q plot is a normal probability plot. It doesn't look too bad, so the assumption of normally distributed residuals looks okay.

The bottom plot gives p-values for the Ljung-Box-Pierce statistics for each lag up to 20. These statistics consider the accumulated residual autocorrelation from lag 1 up to and including the lag on the horizontal axis. The dashed blue line is at .05. All p-values are above it. That's a good result. We want non-significant values for this statistic when looking at residuals. Read Lesson 3.2 of this week for more about the Ljung–Box-Pierce statistic.

All in all, the fit looks good. There's not much need to continue, but just to show you how things looks when incorrect models are used, we will present another model.

Example 1 Continued: Output for a Wrong Model

Suppose that we had misinterpreted the ACF and PACF of the data and had tried an MA(1) model rather than the AR(1) model.

Here's the output:

```
Coefficients:
```

	ma1	xmean
Estimate	0.5570	14.5881
Std. Error	0.1251	0.3337

```
sigma^2 estimated as 1.870: log likelihood = -69.46
$AIC
[1] 1.725741
$AICc
[1] 1.792408
$BIC
[1] 0.8101852
```
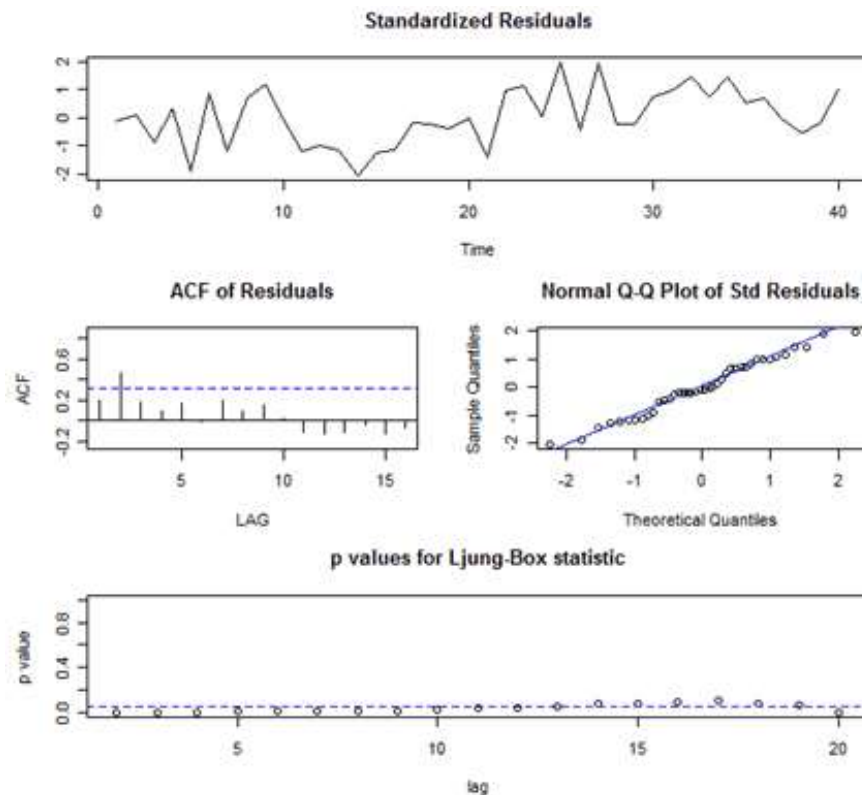
The MA(1) coefficient is significant (you can check it), but mostly this looks worse than the statistics for the right model. The estimate of the variance is 1.87, compared to 1.447 for the AR(1) model. The AIC and BIC statistics are higher for the MA(1) than for the AR(1). That's not good.

The diagnostic graphs aren't good for the MA(1). The ACF has a significant spike at lag 2 and several of the Ljung-Box-Pierce p-values are below .05. We don't want them there. So, the MA(1) isn't a good model.

Standardized Residuals

ACF of Residuals

Normal Q-Q Plot of Std Residuals

p values for Ljung-Box statistic

Example 1 Continued: A Model with One Too Many Coefficients:

Suppose we try a model (still the Lake Erie Data) with one AR term and one MA term. Here's some of the output:

	ar1	ma1	xmean
Estimate	0.7362	−0.0909	14.6307
Std. Error	0.1362	0.1969	0.6142

```
sigma^2 estimated as 1.439: log likelihood = -64.36,
$AIC
[1] 1.514079
$AICc
[1] 1.592650
$BIC
[1] 0.640745
```

Note that the MA(1) coefficient is not significant (z = -0.0909/.1969 is less than 1.96 in absolute value). The MA(1) term could be dropped so that takes us back to the AR(1). Also, the estimate of the variance is barely better than the estimate for the AR(1) model and the AIC and BIC statistics are higher for the ARMA(1,1) than for the AR(1).

A Troublesome Example - Parameter Redundancy

Suppose that the model for your data is white noise. If this is true for every t, then it is true for $t - 1$ as well, in other words:

$$x_t = w_t$$

$$x_{t-1} = w_{t-1}$$

Let's multiply both sides of the second equation by 0.5:

$$0.5x_{t-1} = 0.5w_{t-1}$$

Next, we will move both terms over to one side:

$$0 = -0.5x_{t-1} + 0.5w_{t-1}$$

Because the data is white noise, $x_t = w_t$, so we can add x_t to the left side and w_t to the right side:

$$x_t = -0.5x_{t-1} + w_t + 0.5w_{t-1}$$

This is an ARMA(1, 1)! The problem is that we know it is white noise because of the original equations. If we looked at the ACF what would we see? You would see the ACF corresponding to white noise, a spike at zero and then nothing else. This also means if we take the white noise process and you try to fit in an ARMA(1, 1), R will do it and will come up with coefficients that looks something like what we have above. This is one of the reasons why we need to look at the ACF and the PACF plots and other diagnostics. We prefer a model with the fewest parameters. This example also says that for certain parameter values, ARMA models can appear very similar to one another.

R Code for Example 1

Here's how we accomplished the work for the example in this lesson.

We first loaded the astsa library discussed in Lesson 1. It's a set of scripts written by Stoffer, one of the textbook's authors. If you installed the astsa package during Week 1, then you only need the library command.

Use the command library("astsa"). This makes the downloaded routines accessible.

The session for creating Example 1 then proceeds as follows:

```
xerie = scan("eriedata.dat") #reads the data
xerie = ts (xerie) # makes sure xerie is a time series object
plot (xerie, type = "b") # plots xerie
acf2 (xerie) # author's routine for graphing both the ACF and the PACF
sarima (xerie, 1, 0, 0) # this is the AR(1) model estimated with the author's
routine
sarima (xerie, 0, 0, 1) # this is the incorrect MA(1) model
sarima (xerie, 1, 0, 1) # this is the over-parameterized ARMA(1,1) model
```

In Lesson 3.3, we'll discuss the use of ARIMA models for forecasting. Here's how you would forecast for the next 4 times past the end of the series using the author's source code and the AR(1) model for the Lake Erie data.

```
sarima.for (xerie, 4, 1, 0, 0) # four forecasts from an AR(1) model for the erie
```

```
data
```

You'll get forecasts for the next four times, the standard errors for these forecasts, and a graph of the time series along with the forecasts. More details about forecasting will be given in Lesson 3.3.

Some useful information about the author's scripts is at www.stat.pitt.edu/stoffer/tsa3 [5].

3.2 Diagnostics

Analyzing possible statistical significance of autocorrelation values

The Ljung-Box statistic, also called the modified Box-Pierce statistic, is a function of the accumulated sample autocorrelations, r_j, up to any specified time lag m. As a function of m, it is determined as

$$Q(m) = n(n+2) \sum_{j=1}^{m} \frac{r_j^2}{n-j},$$

where n = number of usable data points after any differencing operations. (Please visit forvo.com [6] for the proper pronunciation of Ljung.)

As an example,

$$Q(3) = n(n+2) \left(\frac{r_1^2}{n-1} + \frac{r_2^2}{n-2} + \frac{r_3^2}{n-3} \right).$$

Use of the Statistic

This statistic can be used to examine residuals from a time series model in order to see if all underlying population autocorrelations for the errors may be 0 (up to a specified point).

For nearly all models that we consider in this course, the residuals are assumed to be "white noise," meaning that they are identically, independently distributed (from each other). Thus, as we saw last week, the ideal ACF for residuals is that all autocorrelations are 0. This means that $Q(m)$ should be 0 for any lag m. A significant $Q(m)$ for residuals indicates a possible problem with the model.

(Remember $Q(m)$ measures accumulated autocorrelation up to lag m.)

Distribution of $Q(m)$

There are two cases:

1. When the r_j are sample autocorrelations for residuals from a time series model, the null hypothesis distribution of $Q(m)$ is approximately a χ^2 distribution with df = $m - p$, where p = number of coefficients in the model.

(Note: m = lag to which we're accumulating, so in essence the statistic is not defined until $m > p$).

2. When no model has been used, so that the ACF is for raw data, $p = 0$ and the null distribution of $Q(m)$ is approximately a χ^2 distribution with $df = m$.

p-Value Determination

In both cases, a p-value is calculated as the probability past $Q(m)$ in the relevant distribution. A small p-value (for instance, p-value < .05) indicates the possibility of non-zero autocorrelation within the first m lags.

Example 1

Below there is Minitab output for the Lake Erie level data that was used for homework 1 and in Lesson 3.1. A useful model is an AR(1) with a constant. So, p = 2.

```
Final Estimates of Parameters
```

Type	Coef	SE Coef	T	P
AR 1	0.7078	0.1161	6.10	0.000
Constant	4.2761	0.1953	21.89	0.000
Mean	14.6349	0.6684		

```
Modified Box-Pierce (Ljung-Box) Chi-Square statistic
```

Lag	12	24	36	48
Chi-Square	9.4	23.2	30.0	*
DF	10	22	34	*
P-Value	0.493	0.390	0.662	*

Minitab gives p-values for accumulated lags that are multiples of 12. The R sarima command will give a graph that shows p-values of the Ljung-Box-Pierce tests for each lag (in steps of 1) up to a lag that seems in some way to be a function of the samples size (not sure of the function).

Interpretation of the Box-Pierce Results

Notice that the p-values for the modified Box-Pierce all are well above .05, indicating "non-significance." This is a desirable result. Remember that there only 40 data values, so there's not much data contributing to correlations at high lags. Thus, the results for $m = 24$ and $m = 36$ may not be meaningful.

Graphs of ACF values

When you request a graph of the ACF values, "significance" limits are shown by R and by Minitab. In general, the limits for the autocorrelation are placed at 0 ± 2 standard errors of r_k. The formula used for standard error depends upon the situation.

- Within the ACF of residuals as part of the ARIMA routine, the standard errors are

determined assuming the residuals are white noise. The approximate formula for any lag is that s.e. of $r_k = 1/(n)^{1/2}$.

- For the ACF of raw data (the ACF command), the standard error at a lag k is found as if the right model was an MA(k-1). This allows the possible interpretation that if all autocorrelations past a certain lag are within the limits, the model might be an MA of order defined by the last significant autocorrelation.

Appendix: Standardized Residuals

What are standardized residuals in a time series framework? One of the things that we need to look at when we look at the diagnostics from a regression fit is a graph of the standardized residuals.   Let's review what this is for regular regression where the standard deviation is σ. The standardized residual at observation i

$$\frac{y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij}}{\sigma},$$

should be N(0, 1). We hope to see normality when we look at the diagnostic plots. Another way to think about this is:

$$\frac{y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij}}{\sigma} \approx \frac{y_i - \hat{y}_i}{\sqrt{Var(y_i - \hat{y}_i)}}.$$

Now, with time series things are very similar:

$$\frac{x_t - \tilde{x}_t}{\sqrt{P_t^{t-1}}},$$

where

$$\tilde{x}_t = E(x_t | x_{t-1}, x_{t-2}, \dots) \text{and} P_t^{t-1} = E\left[(x_t - \tilde{x}_t)^2\right].$$

This is where the standardized residuals come from. This is also essentially how a time series is fit using R. We want to minimize the sums of these squared values:

$$\sum_{t=1}^n \left(\frac{x_t - \tilde{x}_t}{\sqrt{P_t^{t-1}}}\right)^2$$

(In reality, it is slightly more complicated. The log-likelihood function is minimized, and this is one term of that function.)

3.3 Forecasting with ARIMA Models

Section 3.5 in the textbook gives a theoretical look at forecasting with ARIMA models. That presentation is a bit tough, but in practice it's easy to understand how forecasts are created.

In an ARIMA model, we express x_t as a function of past value of x and/or past errors (as well as a present time error). When we forecast a value past the end of the series, on the right side of the equation we might need values from the observed series or we might, in theory, need values that aren't yet observed.

Example: Consider the AR(2) model $x_t = \delta + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + w_t$. In this model, x_t is a linear function of the values of x at the previous two times. Suppose that we have observed n data values and wish to use the observed data and estimated AR(2) model to forecast the value of x_{n+1} and x_{n+2}, the values of the series at the next two times past the end of the series. The equations for these two values are

$$x_{n+1} = \delta + \varphi_1 x_n + \varphi_2 x_{n-1} + w_{n+1}$$

$$x_{n+2} = \delta + \varphi_1 x_{n+1} + \varphi_2 x_n + w_{n+2}$$

To use the first of these equations, we simply use the observed values of x_n and x_{n-1} and replace w_{n+1} by its expected value of 0 (the assumed mean for the errors).

The second equation for forecasting the value at time $n + 2$ presents a problem. It requires the unobserved value of x_{n+1} (one time past the end of the series). The solution is to use the forecasted value of (the result of the first equation).

In general, the forecasting procedure, assuming a sample size of n, is as follows:

For any w_j with $1 \le j \le n$, use the sample residual for time point j
For any w_j with $j > n$, use 0 as the value of w_j

For any x_j with $1 \le j \le n$, use the observed value of x_j
For any x_j with $j > n$ use the forecasted value of x_j

Notation: Our authors use the notation x_{n+m}^n to represent a forecast m times past the end of the observed series. The "superscript" is to be read as "given data up to time n." Other authors use the notation $x_n(m)$ to denote a forecast m times past time n.

To understand the formula for the standard error of the forecast error, we first need to define the concept of psi-weights.

Psi-weight representation of an ARIMA model

Any ARIMA model can be converted to an infinite order MA model:

$$\begin{aligned} x_t - \mu &= w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \cdots + \psi_k w_{t-k} + \ldots \\ &= \sum_{j=0}^{\infty} \Psi_j w_{t-j} \text{ where} \Psi_0 = 1 \end{aligned}$$

An important constraint so that the model doesn't "explode" as we go back in time is

$$\sum_{j=0}^{\infty} |\Psi_j| < \infty$$

[On page 95 of our book, the authors define a "causal" model as one for which this constraint is in place, along with the additional restraint that we can't express the value of the present x as a function of future values.]

The process of finding the "psi-weight" representation can involve a few algebraic tricks. Fortunately, R has a routine, ARMAtoMA, that will do it for us. To illustrate how psi-weights may be determined algebraically, we'll consider a simple example.

Example: Suppose that an AR(1) model is $x_t = 40 + 0.6x_{t-1} + w_t$

For an AR(1) model, the mean $\mu = \delta/(1 - \varphi_1)$ so in this case, $\mu = 40/(1 - .6) = 100$. We'll define $z_t = x_t - 100$ and rewrite the model as $z_t = 0.6z_{t-1} + w_t$. (You can do the algebra to check that things match between the two expressions of the model.)

To find the psi-weight expression, we'll continually substitute for the z on the right side in order to make the expression become one that only involves w values.

$$z_t = 0.6z_{t-1} + w_t, \text{ so } z_{t-1} = 0.6z_{t-2} + w_{t-1}.$$

Substitute the right side of the second expression for z_{t-1} in the first expression.

This gives $z_t = 0.6(0.6z_{t-2} + w_{t-1}) + w_t = 0.36z_{t-2} + 0.6w_{t-1} + w_t$.

Next, note that $z_{t-2} = 0.6z_{t-3} + w_{t-2}$.

Substituting this into the equation gives $z_t = 0.216z_{t-3} + 0.36w_{t-2} + 0.6w_{t-1} + w_t$.

If you keep going, you'll soon see that the pattern leads to

$$z_t = x_t - 100 = \sum_{j=0}^{\infty} (0.6)^j w_{t-j}$$

Thus the psi-weights for this model are given by $\psi_j = (0.6)^j$ for $j = 0, 1, \ldots, \infty$.

In R, the command ARMAtoMA(ar = .6, ma=0, 12) gives the first 12 psi-weights. This will give the psi-weights ψ_1 to ψ_{12} in scientific notation.

For the AR(1) with AR coefficient = 0.6 they are:

```
[1]  0.600000000  0.360000000  0.216000000  0.129600000  0.077760000
0.046656000

[7]  0.027993600  0.016796160  0.010077696  0.006046618  0.003627971
0.002176782
```

Remember that $\psi_0 = 1$. R doesn't give this value. It's listing starts with ψ_1, which equals 0.6 in this case.

MA Models: The psi-weights are easy for an MA model because the model already is written in terms of the errors. The psi-weights = 0 for lags past the order of the MA model and equal the coefficient values for lags of the errors that are in the model. Remember that we always have $\psi_0 = 1$.

Standard error of the forecast error for a forecast using an ARIMA model

Without proof, we'll state a result:

The variance of the difference between the forecasted value at time $n + m$ and the (unobserved) value at time $n + m$ is

$$\text{Variance of } \left(x_{n+m}^n - x_{n+m}\right) = \sigma_w^2 \sum_{j=0}^{m-1} \Psi_j^2.$$

Thus the estimated **standard deviation of the forecast error** at time $n + m$ is

$$\text{Standard error of } \left(x_{n+m}^n - x_{n+m}\right) = \sqrt{\hat{\sigma}_w^2 \sum_{j=0}^{m-1} \Psi_j^2}.$$

Note that the summation of squared psi-weights begins with $(\psi_0)^2 = 1$ and that the summation goes to $m - 1$, one less than the number of times ahead for which we're forecasting.

When forecasting $m = 1$ time past the end of the series, the standard error of the forecast error is

$$\text{Standard error of } \left(x_{n+1}^n - x_{n+1}\right) = \sqrt{\hat{\sigma}_w^2(1)}$$

When forecasting the value $m = 2$ times past the end of the series, the standard error of the forecast error is

$$\text{Standard error of } \left(x_{n+2}^n - x_{n+2}\right) = \sqrt{\hat{\sigma}_w^2(1 + \Psi_1^2)}.$$

Notice that the variance will not be too big when $m = 1$. But, as you predict out farther in the future, the variance will increase. When m is very large, we will get the total variance. In other words, if you are trying to predict very far out, we will get the variance of the entire time series; as if you haven't even looked at what was going on previously.

95% Prediction Interval for x_{n+m}

With the assumption of normally distributed errors, a 95% prediction interval for x_{n+m}, the future value of the series at time $n + m$, is

$$x_{n+m}^n \pm 1.96\sqrt{\hat{\sigma}_w^2 \sum_{j=0}^{m-1} \Psi_j^2}.$$

Example: Suppose that an AR(1) model is estimated to be $x_t = 40 + 0.6x_{t-1} + w_t$. This is the same model used earlier in this handout, so the psi-weights we got there apply.

Suppose that we have n = 100 observations, $\hat{\sigma}_w^2 = 4$ and $x_{100} = 80$. We wish to forecast the values at both times 101 and 102, and create prediction intervals for both forecasts.

First we forecast time 101.

$$\begin{aligned} x_{101} &= 40 + 0.6x_{100} + w_{101} \\ x_{101}^{100} &= 40 + 0.6(80) + 0 = 88 \end{aligned}$$

The standard error of the forecast error at time 101 is

$$\sqrt{\hat{\sigma}_w^2 \sum_{j=0}^{1-1} \Psi_j^2} = \sqrt{4(1)} = 2.$$

The 95% prediction interval for the value at time 101 is 88 ± 2(1.96), which is 84.08 to 91.96. We are therefore 95% confident that the observation at time 101 will be between 84.08 and 91.96. If we repeated this exact process, then 95% of the computed prediction intervals would contain the true value of x at time 101.

The forecast for time 102 is

$$x_{102}^{100} = 40 + 0.6(88) + 0 = 92.8$$

Note that we used the forecasted value for time 101 in the AR(1) equation.

The relevant standard error is

$$\sqrt{\hat{\sigma}_w^2 \sum_{j=0}^{2-1} \Psi_j^2} = \sqrt{4(1 + 0.6^2)} = 2.332$$

A 95% prediction interval for the value at time 102 is 92.8 ± (1.96)(2.332).

To forecast using an ARIMA model in R, we recommend our textbook author's script called sarima.for. (It is part of the astsa library recommended previously.)

Example: In the homework for Week 2, problem 5 asked you to suggest a model for a time series of stride lengths measured every 30 seconds for a runner on a treadmill.

From R, the estimated coefficients for an AR(2) model and the estimated variance are as follows for a similar data set with n = 90 observations:

```
Coefficients:
```

	ar1	ar2	xmean
	1.1480	-0.3359	48.7476
s.e.	0.1009	0.1087	1.8855

```
sigma^2 estimated as 11.47
```

The command

```
sarima.for(stridelength, 6, 2, 0, 0)  # 6 forecasts with an AR(2) model for stridelength
```
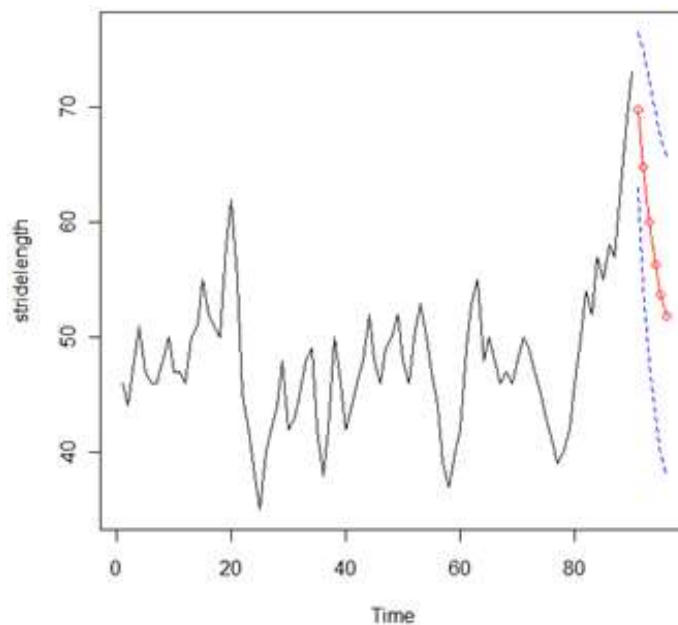
will give forecasts and standard errors of prediction errors for the next six times past the end of the series. Here's the output (slightly edited to fit here):

```
$pred
Time Series:
Start = 91
End = 96
[1]  69.78674 64.75441 60.05661 56.35385 53.68102 51.85633

$se
Time Series:
Start = 91
End = 96
[1]  3.386615 5.155988 6.135493 6.629810 6.861170 6.962654
```

The forecasts are given in the first batch of values under $pred and the standard errors of the forecast errors are given in the last line in the batch of results under $se.

The procedure also gave this graph, which shows the series followed by the forecasts as a red line and the upper and lower prediction limits as blue dashed lines:

Psi-Weights for the Estimated AR(2) for the Stride Length Data

If we wanted to verify the standard error calculations for the six forecasts past the end of the series, we would need to know the psi-weights. To get them, we need to supply the estimated AR coefficients for the AR(2) model to the ARMAtoMA command.

The R command in this case is ARMAtoMA(ar = list(1.148, -0.3359), ma = 0, 5)

This will give the psi-weights to in scientific notation. The answer provided by R is:

[1] 1.148000e+00 9.820040e-01 7.417274e-01 5.216479e-01 3.497056e-01

(Remember that $\psi_0 = 1$ in all cases)

The output for estimating the AR(2) included this estimate of the error variance:

```
sigma^2 estimated as 11.47
```

As an example, the standard error of the forecast error for 3 times past the end of the series is

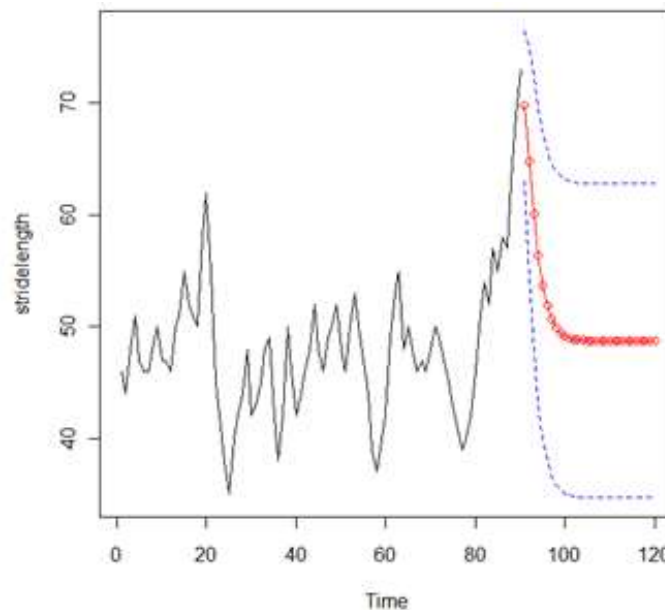$$\sqrt{\hat{\sigma}_w^2 \sum_{j=0}^{3-1} \Psi_j^2} = \sqrt{11.47(1 + 1.148^2 + 0.982^2)} = 6.1357$$

which, except for round off error, matches the value of 6.135493 given as the third standard error in the sarima.for output above.

Where the Forecasts Will End Up?

For a stationary series and model, the forecasts of future values will eventually converge to the mean and then stay there. Note below what happened with the stride length forecasts, when we asked for 30 forecasts past the end of the series. [Command was sarima.for (stridelength, 30, 2, 0, 0)]. The forecast got to 48.74753 and then stayed there.

```
$pred
Time Series:
Start = 91
End = 120
[1]   69.78674 64.75441 60.05661 56.35385 53.68102 51.85633 50.65935 49.89811
[9]   49.42626 49.14026 48.97043 48.87153 48.81503 48.78339 48.76604 48.75676
[17]  48.75192 48.74949 48.74833 48.74780 48.74760 48.74753 48.74753 48.74755
[25]  48.74757 48.74759 48.74760 48.74761 48.74762 48.74762
```

The graph showing the series and the six prediction intervals is the following

Links:
[1]
https://onlinecourses.science.psu.edu/stat510/sites/onlinecourses.science.psu.edu.stat510/files/L03/ar1_example.png
[2]

https://onlinecourses.science.psu.edu/stat510/sites/onlinecourses.science.psu.edu.stat510/files/L03/ar2_example.png

[3] https://onlinecourses.science.psu.edu/stat510/sites/onlinecourses.science.psu.edu.stat510/files/L03/ma_example.png

[4] https://onlinecourses.science.psu.edu/stat510/sites/onlinecourses.science.psu.edu.stat510/files/L03/arma11_example.png

[5] http://www.stat.pitt.edu/stoffer/tsa3

[6] http://www.forvo.com/search/Ljung