STAT 5170: Applied Time Series

Course notes for part B of learning unit 5

Section 5A.1: Differencing.

Thus far in the learning unit, our focus has been on using regression to "de-trend" a time series in order to identify components that describe long-term trends and stationary residual deviations from those trends, and to separate them for component-specific modeling considerations. Models for stationary time series, such as autoregressive, moving average, and ARMA models, would apply specifically to the component describing stationary residual deviations. In this portion of the unit, we explore an alternative route to arranging our data such that, from a certain angle, it would be amenable to description by a stationary time series model. The basic device is not decomposition followed by the application regression techniques, but an operation called *differencing*.

The differencing technique is defined through a notational device, similar to the backshift operator. Specifically, the *difference operator*, denoted ∇, calculates the differences between neighboring values of a time series; when applied to a time series (x_t), the *first difference* is $\nabla x_t = x_t - x_{t-1}$. As with the backshift operator, an exponentiated difference operator signifies that differencing is to be applied multiple times; thus, for instance,

$$\nabla^2 x_t = \nabla x_t - \nabla x_{t-1} = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}.$$

It is sometimes helpful to recognize that difference operator can be written in the notation of the backshift operator: the formula is $\nabla = 1 - B$, so that, $\nabla x_t = x_t - x_{t-1} = (1 - B)x_t$. With this in mind, the *d'th difference* is $\nabla^d x_t = (1 - B)^d x_t$, from which it is straightforward to deduce that the operator ∇^d expands to a d'th order polynomial in B.

An elementary property of the difference operator is that when it is applied to a stationary time series, the time series that results is also stationary, though with different mean and autocovariance functions, typically. To see this, first note that the mean function of the difference series is constant,

$$\mu_t \;=\; E[\nabla x_t] = E[x_t - x_{t-1}] = E[x_t] - E[x_{t-1}] = 0$$

since $E[x_t] = E[x_{t-1}]$, by the stationarity of (x_t). The autocovariance function is

$$\begin{aligned}
\gamma(t+h, t) &= Cov(x_{t+h} - x_{t+h-1}, x_t - x_{t-1}) \\
&= Cov(x_{t+h}, x_t) - Cov(x_{t+h}, x_{t-1}) - Cov(x_{t+h-1}, x_t) + Cov(x_{t+h-1}, x_{t-1}) \\
&= \gamma_x(h) - \gamma_x(h+1) - \gamma_x(h-1) + \gamma_x(h),
\end{aligned}$$

which depends only on h. We have shown that the differenced series (∇x_t) is stationary. The analogous results also holds in the case of strong stationarity.

Importantly, differencing can also produce a a stationary time series from a non-stationary one. As an example, consider the following. Suppose the time series (x_t) is defined by the decomposition $x_t = \mu_t + \epsilon_t$, where $\mu_t = \beta_1 + \beta_2 t$ and (ϵ_t) is a zero-mean, stationary time series. The first difference is

$$\nabla x_t \;=\; \{\beta_1 + \beta_2 t + \epsilon_t\} - \{\beta_1 + \beta_2(t-1) + \epsilon_{t-1}\} = \beta_2 + \epsilon_t - \epsilon_{t-1} = \beta_2 + \nabla \epsilon_t.$$

Observe that β_1 has disappeared, $\nabla\epsilon_t$ is stationary, and β_2 is the mean of the differenced series; hence, (∇x_t) is stationary.

Additional examples are as follows.

Example: Differencing a random walk

Suppose the time series (x_t) is a random walk with initial value $x_0 = 0$, defined by the relationship

$$x_t = \delta + x_{t-1} + w_t, \text{ for } t = 1, 2, \ldots$$

The first-difference series, (∇x_t), is defined according to $\nabla x_t = x_t - x_{t-1}$. It follows that

$$\nabla x_t = x_t - x_{t-1} = \delta + x_{t-1} + w_t - x_{t-1} = \delta + w_t$$

□

Example: Eliminating a quadratic trend

The first difference, ∇, is a type of "linear filter." As we have seen, applying it to a time series will eliminate a linear trend. The second difference eliminates a quadratic trend. To see this, suppose $x_t = \mu_t + \epsilon_t$, where $\mu_t = \beta_1 + \beta_2 t + \beta_3 t^2$. The first difference is

$$\begin{aligned} \nabla x_t &= x_t - x_{t-1} \\ &= \{\beta_1 + \beta_2 t + \beta_3 t^2 + \epsilon_t\} - \{\beta_1 + \beta_2(t-1) + \beta_3(t-1)^2 + \epsilon_{t-1}\} \\ &= \beta_2 + \beta_3\{t^2 - (t-1)^2\} + \epsilon_t - \epsilon_{t-1}. \end{aligned}$$

By expanding $t^2 - (t-1)^2 = t^2 - t^2 + 2t - 1 = -1 + 2t$, and rewriting $\epsilon_t - \epsilon_{t-1} = \nabla\epsilon_t$ (which defines a stationary time series), this becomes

$$\nabla x_t = \beta_2 - \beta_3 + 2\beta_3 t + \nabla\epsilon_t.$$

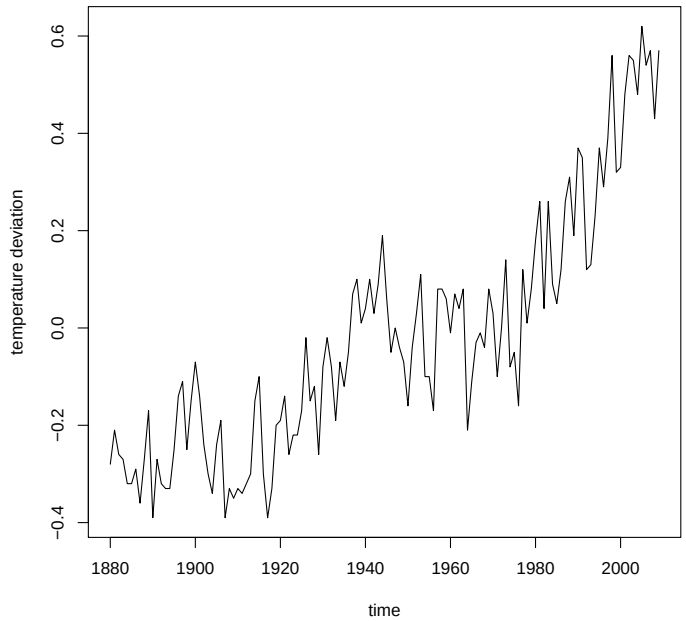
The second difference is therefore

$$\begin{aligned} \nabla^2 x_t &= \{\beta_2 - \beta_3 + 2\beta_3 t + \nabla\epsilon_t\} - \{\beta_2 - \beta_3 + 2\beta_3(t-1) + \nabla\epsilon_{t-1}\} \\ &= 2\beta_3 + \nabla\epsilon_t - \nabla\epsilon_{t-1} \\ &= 2\beta_3 + \nabla^2\epsilon_t. \end{aligned}$$

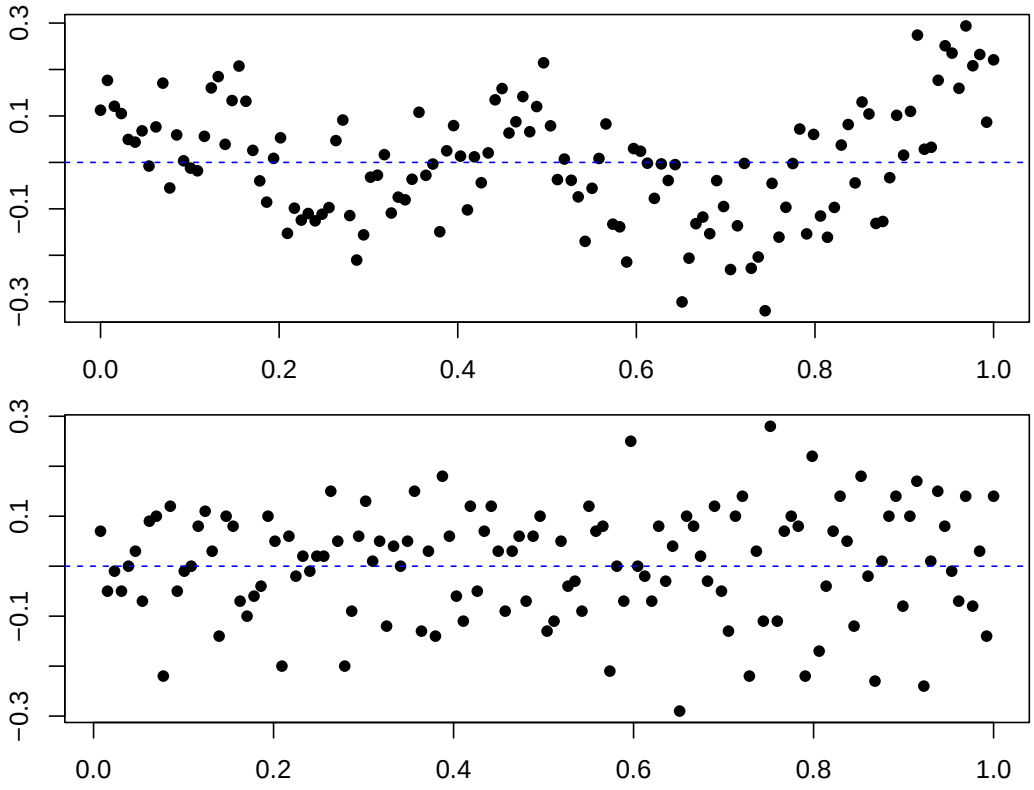
Both β_1 and β_2 have disappeared, $(\nabla^2\epsilon_t)$ is stationary since it is the difference of the stationary series $(\nabla\epsilon_t)$, and $2\beta_3$ is the mean of the twice-differenced series; the quadratic trend has thus been eliminated and the twice-differenced series, $(\nabla^2 x_t)$, is stationary. □

Example: Exploring average global temperatures

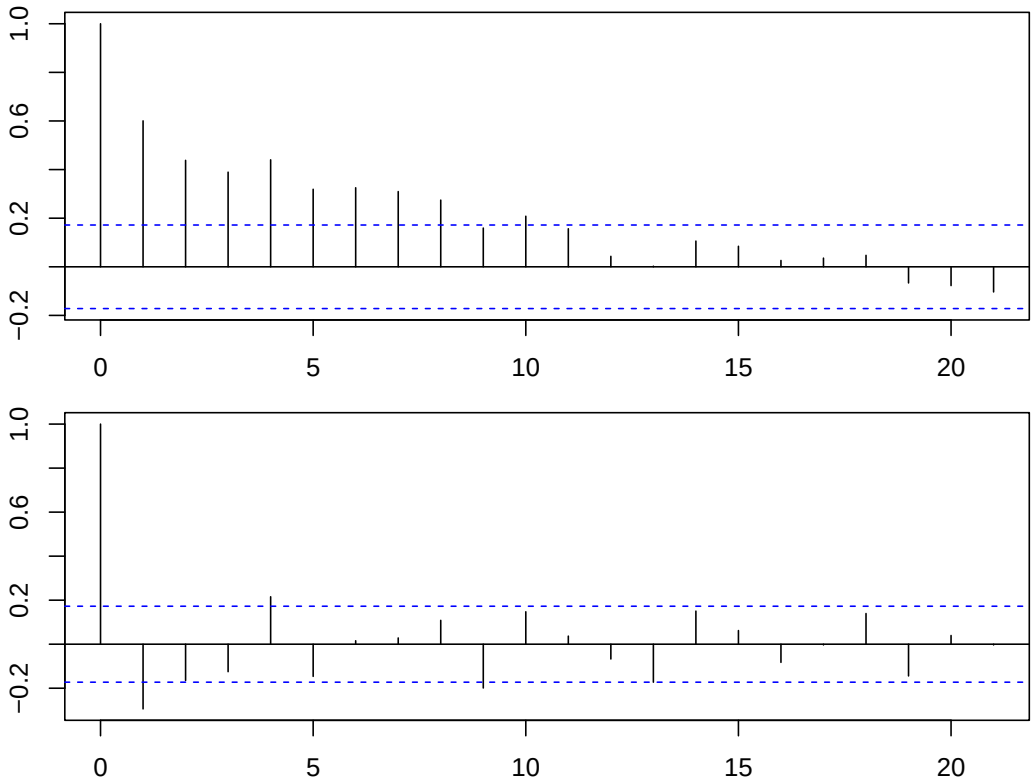
The data shown in the plot below records average global temperature deviations from 1880-2009.



We have worked with this time series before, in demonstrating the use of regression to de-trend a time series. Presently, we also consider the possibility of using differencing to manage the increasing trend. The figure below displays the results of de-trending and differencing the data. Specifically, the top panel displays the estimated residuals, $e_t = x_t - \hat{x}_t$, under the regression model that includes a linear term only, $x_t = \beta_1 + \beta_2 t + w_t$. The bottom panel displays the differenced sequence, ∇x_t.



Both time series appear stationary, but the differenced series is perhaps better behaved as it deviates less frequently from its center. Plots of the sample ACFs offer closer examination of dependency patterns. These are shown below, for the detrended series in the top panel and the differenced series in the bottom panel.



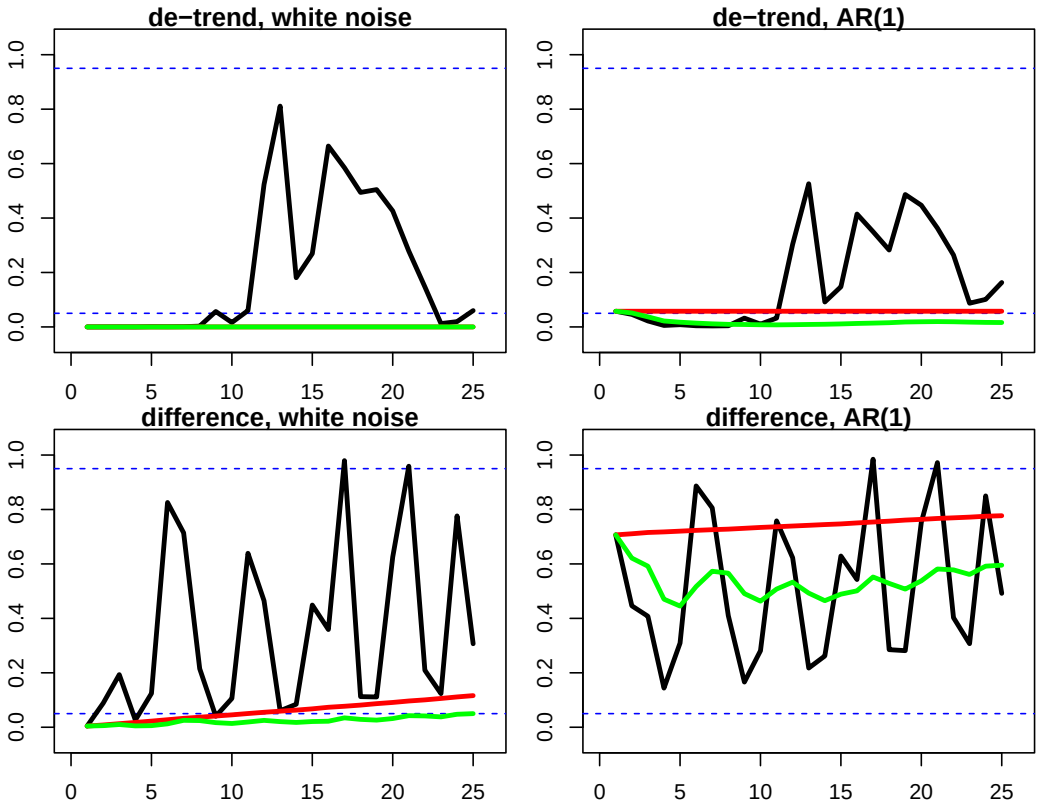
Though stationarity may be inferred in either series, ACFs suggest a simpler dependency pattern in the differenced series than that of the detrended series. On the argument for simplicity, differencing may be the preferred approach for identifying a stationarity aspect to these data.

When exploring this time series, model checking and diagnostics for assessing predictive performance may also be helpful to select between approaches. The following table lists the DIC and WAIC statistics obtained from several versions of the polynomial regression model, under uncorrelated and $AR(1)$ residual deviations, and, in the last row of the table, those of differencing model in which the time series of differences is either white-noise or an $AR(1)$ time series.

	DIC		WAIC	
μ_t	white noise	$AR(1)$	white noise	$AR(1)$
linear	-171.4	-229.6	-171.6	-169.4
quadratic	-204.1	-240.4	-204.3	-203.3
cubic	-215.1	-244.1	-215.4	-214.4
difference	-203.2	-204.1	-213.4	-203.9

In these numbers, we see some of the limitations of predictive assessment for selecting a model. Although our previous graphical explorations of estimated residual deviations suggest that a differencing model more adequately accounts for autocorrelations, relative to the differencing model, the corresponding predictive diagnostics suggest that the differencing models' predictive performance mostly falls somewhere between that of the linear and cubic regression models. That is, a complicated regression model seems to offer superior predictive performance, despite the indication that it does not adequately account for autocorrelations.

On the other hand, rather than base our choice of model on predictive performance, if our concern is to find a model that accounts for autocorrelations, it can be helpful to check our models by calculating posterior predictive p-values. The following plot graphs posterior predictive p-values across lags for each of four models, the linear regression model with uncorrelated and $AR(1)$ residual deviations (top left and top right panels, respectively), and the differencing model with white-noise and $AR(1)$ differences (bottom left and bottom right panels, respectively).



In each panel, the solid black line graphs the p-values associated with the absolute predicted autocorrelation, $S(\boldsymbol{x}, \boldsymbol{\theta}) = |\tilde{\rho}(h)|$ across lags h; the solid red line graphs the p-value associated with the maximum absolute predicted autocorrelation $S(\boldsymbol{x}, \boldsymbol{\theta}) = \max_{h=1,\ldots,H} |\tilde{\rho}(h)|$ across maximum lags H; and, the solid green line graphs the p-value associated with the sum of absolute predicted autocorrelations $S(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{h=1}^{H} |\tilde{\rho}(h)|$ across maximum lags H. Here we see that only for the differencing model with $AR(1)$ differences do the p-values generally remain far from their extreme possible values near zero or one. □

Section 5A.2: ARIMA models.

The differencing concept inspires a new class of time-series models, which extends the ARMA models to accommodate a certain type of non-stationarity.

Definition: The time series (x_t) is $ARIMA(p,d,q)$ *autoregressive integrated moving average* if the time series defined by its d'th difference, $(\nabla^d x_t)$, is $ARMA(p,q)$.

Recalling the examples, above, we see, *e.g.*, that a random walk might be described as $ARIMA(0,1,0)$ time series, since its first difference is a constant mean-value plus white noise.

Understanding a time series through differencing can complicate the task of forecasting. Suppose (x_t) is $ARIMA(p,1,q)$, or, more generally, that its first-difference series, (∇x_t), is stationary. Suppose further that from the measured values $x_1,\ldots,x_n$ our aim is to predict future values x_{n+1}, x_{n+2}, *etc.* For convenience, let us write $y_t = \nabla x_t = x_t - x_{t-1}$. Rearranging terms provides the formula $x_t = y_t + x_{t-1}$, which suggests that a prediction of x_{n+1} would be determined as the value x_n plus a prediction of the difference $y_{n+1} = x_{n+1} - x_n$. Continuing the deduction provides the prediction formula

$$x^n_{n+m} \;=\; y^n_{n+m} + x^n_{n+m-1},$$

for $m \geq 1$, where $x^n_n = x_n$ defines a starting value. Parallel formulas for higher-order differencing, where $d > 1$, can also be derived in a fairly straightforward way, though they become more complex as the order of differencing increases.

The ARIMA relationship may be expressed in backshift operator notation as

$$\phi(B)(1-B)^d x_t = \theta(B) w_t$$

for the autoregressive and moving-average operators

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p \quad \text{and} \quad \theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q.$$

This implies that the d'th difference $y_t = (1-B)^d x_t$ is an ARMA time series, whose relationship is $\phi(B) y_t = \theta(B) w_t$. If it is causal, then it has the representation

$$y_t \;=\; \sum_{j=0}^{\infty} \psi^*_j w_{t-j},$$

where, through its characteristic polynomial, the infinite moving average operator $\psi^*(B)$ solves

$$\phi(z)\psi^*(z) = \theta(z);$$

that is, $\psi^*(B)$ is defined from the coefficients of the power-series expansion

$$\psi^*(z) = \frac{\theta(z)}{\phi(z)} = \sum_{j=0}^{\infty} \psi^*_j z^j.$$

It follows that the original time series, (x_t), may expressed through the causal representation

$$x_t = \sum_{j=0}^{\infty} \psi_j w_{t-j},$$

where the operator $\psi(B)$ solves

$$\phi(z)(1-z)^d \psi(z) = \theta(z),$$

and may be defined from the power-series expansion

$$\psi(z) = \frac{\theta(z)}{\phi(z)(1-z)^d} = \sum_{j=0}^{\infty} \psi_j z^j.$$

Returning now to the task of forecasting, if the prediction of y_{n+m} is defined by conditional expectation, as $y^n_{n+m} = E[y_{n+m}|y_2, \ldots, y_n]$, then its m-step-ahead mean squared prediction error is

$$P^{*n}_{n+m} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^{*2},$$

where $\sigma_w^2 = Var[w_t]$. The m-step-ahead mean squared prediction error for predicting x_{n+m}, based on prediction formulas of the type deduced, above, is

$$P^n_{n+m} = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2.$$

The examples below offer further practice with these ideas.

Example: Differencing a random walk (continued)

Suppose the time series (x_t) is a random walk with initial value $x_0 = 0$, and so satisfies

$$x_t = \delta + x_{t-1} + w_t$$

for $t = 1, 2, \ldots$, where (w_t) is Gaussian white noise; write $\sigma_w^2 = Var(w_t)$. From previous discussion, we know that (x_t) is non-stationary, and that

$$\nabla x_t = x_t - x_{t-1} = \delta + x_{t-1} + w_t - x_{t-1} = \delta + w_t$$

is stationary. As we know, this shows that (∇x_t) is $ARMA(0,0)$, possibly with a non-zero mean (*i.e.*, (∇x_t) is shifted white noise), from which it follows that (x_t) is $ARIMA(0,1,0)$.

Corresponding prediction formulas deduced through conditional expectation are easy to derive by direct manipulation. To do this, first recall the representation

$$x_n = n\delta + \sum_{j=1}^{n} w_j$$

and note that it implies that w_{n+m} is independent of $x_1, \ldots, x_n$, when $m \geq 1$. Next, observe

$$\begin{aligned}
x_{n+m}^n &= E[x_{n+m}|x_1, \ldots, x_n] \\
&= E\left[(n+m)\delta + \sum_{j=1}^{n+m} w_j \middle| x_1, \ldots, x_n\right] \\
&= E\left[m\delta + \sum_{j=n+1}^{n+m} w_j \middle| x_1, \ldots, x_n\right] \\
&\quad + E\left[n\delta + \sum_{j=1}^{n} w_j \middle| x_1, \ldots, x_n\right] \\
&= m\delta + \sum_{j=n+1}^{n+m} E(w_j|x_1, \ldots, x_n) + E(x_n|x_1, \ldots, x_n) \\
&= m\delta + \sum_{j=n+1}^{n+m} E(w_j) + x_n \\
&= m\delta + x_n.
\end{aligned}$$

This provides the desired prediction formula. A formula for mean squared prediction error is derived as

$$\begin{aligned}
P_{n+m}^n &= E[(x_{n+m} - x_{n+m}^n)^2] \\
&= E\left[\left\{(n+m)\delta + \sum_{j=1}^{n+m} w_j - \left(m\delta + n\delta + \sum_{j=1}^{n} w_j\right)\right\}^2\right] \\
&= E\left[\left(\sum_{j=n+1}^{n+m} w_j\right)^2\right] \\
&= \sum_{j=n+1}^{n+m}\sum_{k=n+1}^{n+m} E(w_j w_k) \\
&= m\sigma_w^2.
\end{aligned}$$

Example: Exponentially weighted moving average

The time series (x_t) is an *exponentially weighted moving average* time series if it satisfies

$$x_t = \delta + x_{t-1} + w_t - \lambda w_{t-1}$$

for $t = 1, 2, \ldots$, where $|\lambda| < 1$ and (w_t) is Gaussian white noise with $\sigma_w^2 = Var[w_t]$. To complete the description of this time series model, suppose x_t is constant and $w_t = 0$ for $t \leq 0$. Observe that the form of this time series model reflects that of an $ARMA(1,1)$ time series, but the implied autoregressive coefficient has the value $\phi_1 = 1$.

Induction techniques that we worked with when studying random walks and autoregressive time series may be used in a straightforward manned to deduce the representation

$$x_t = t\delta + x_0 + w_t + (1-\lambda)\sum_{j=1}^{t-1} w_j - \lambda w_0.$$

This representation implies

$$\mu_t = E[x_t] = t\delta + x_0$$

and

$$\begin{aligned} Var[x_t] &= Var[w_t] + (1-\lambda)^2 \sum_{j=1}^{t-1} Var[w_j] + \lambda^2 Var[w_0] \\ &= \{1 + (t-1)(1-\lambda)^2 + \lambda^2\}\sigma_w^2, \end{aligned}$$

and furthermore that (x_t) is causal. From these formulas it is also seen that (x_t) is non-stationary, even when $\delta = 0$. However, it is not difficult to see that the first-difference series, (∇x_t), is stationary. Observe that

$$\nabla x_t = x_t - x_{t-1} = \delta + x_{t-1} + w_t - \lambda w_{t-1} - x_{t-1} = \delta + w_t - \lambda w_{t-1},$$

which identifies a shifted $MA(1)$ relationship, across $t = 1, 2, \ldots$ It follows that the model for (x_t) is $ARIMA(0, 1, 1)$. Given that the autoregressive component has order $p = 0$, the notation and terminology for this model is sometimes shortened to $IMA(1, 1)$, in reference to an *integrated moving-average* time series.

When $m \geq 2$, a prediction formula for an m-step ahead prediction is readily deduced from the conditional expectation formula:

$$\begin{aligned} x_{n+m}^n &= E[x_{n+m}|x_1, \ldots, x_n] \\ &= E[\delta + x_{n+m-1} + w_{n+m} - \lambda w_{n+m-1}|x_1, \ldots, x_n] \\ &= \delta + x_{n+m-1}^n, \end{aligned}$$

having written $x_{n+m-1}^n = E[x_{n+m-1}|x_1, \ldots, x_n]$ and noted that causality implies $E[w_{n+m} \mid x_1, \ldots, x_n] = 0$ and $E[w_{n+m-1}|x_1, \ldots, x_n] = 0$. An additional recursive argument connects this formula to the one-step ahead prediction, x_{n+1}^n, as follows:

$$\begin{aligned} x_{n+m}^n &= \delta + x_{n+m-1}^n \\ &= \delta + (\delta + x_{n+m-2}^n) \\ &= \delta + (\delta + (\delta + \cdots (\delta + x_{n+1}^n) \cdots)) \\ &= (m-1)\delta + x_{n+1}^n. \end{aligned}$$

When $m = 1$, a prediction formula may be deduced from the differenced series's invertible representation. Writing $y_t = \nabla x_t$, the invertible representation is

$$w_t = \sum_{j=0}^{t-1} \lambda^j (y_{t-j} - \delta),$$

for $t = 1, 2, \ldots$ Note the parallel with the invertible representation of a shifted $MA(1)$ time series, for which

$$w_t = \sum_{j=0}^{\infty} \lambda^j (y_{t-j} - \delta),$$

which is valid due to the assumption $|\lambda| < 1$. Now substitute $y_t = x_t - x_{t-1}$ and set $w_0 = 0$ to deduce the formula

$$\begin{aligned} w_t &= \sum_{j=0}^{t-1} \lambda^j (x_{t-j} - x_{t-j-1} - \delta) \\ &= x_t - \sum_{j=1}^{t-1} \lambda^{j-1}(1-\lambda) x_{t-j} - \delta \sum_{j=0}^{t-1} \lambda^j - \lambda^{t-1} x_0, \end{aligned}$$

by which

$$x_t = \sum_{j=1}^{t-1} \lambda^{j-1}(1-\lambda) x_{t-j} + w_t + \delta \sum_{j=0}^{t-1} \lambda^j + \lambda^{t-1} x_0.$$

To make sense of this formula's implications for prediction, let us first simplify its context by setting $\delta = 0$ and $x_0 = 0$, so that the formula reduces to

$$x_t = \sum_{j=1}^{t-1} \lambda^{j-1}(1-\lambda) x_{t-j} + w_t.$$

It follows that

$$x_{n+1} = \sum_{j=1}^{n} \lambda^{j-1}(1-\lambda) x_{n+1-j} + w_{n+1},$$

and the one-step-ahead prediction is

$$\begin{aligned} x_{n+1}^n &= E[x_{n+1} | x_1, \ldots x_n] \\ &= E\left[\sum_{j=1}^{n} \lambda^{j-1}(1-\lambda) x_{n+1-j} + w_{n+1} \,\middle|\, x_1, \ldots x_n \right] \\ &= \sum_{j=1}^{n} \lambda^{j-1}(1-\lambda) E[x_{n+1-j} | x_1, \ldots x_n] + E[w_{n+1} | x_1, \ldots x_n] \\ &= \sum_{j=1}^{n} \lambda^{j-1}(1-\lambda) x_{n+1-j}, \end{aligned}$$

having noted that $E[x_{n+1-j} | x_1, \ldots x_n] = x_{n+1-j}$ for $j = 1, \ldots, n$ and $E[w_{n+1} | x_1, \ldots x_n] = 0$, by causality.

Some have deduced a curious relationship inherent to this formula. Observe

$$\begin{aligned} x_{n+1}^n &= \sum_{j=1}^{n} \lambda^{j-1}(1-\lambda)x_{n+1-j} \\ &= (1-\lambda)x_n + \sum_{j=2}^{n} \lambda^{j-1}(1-\lambda)x_{n+1-j} \\ &= (1-\lambda)x_n + \lambda \sum_{j=1}^{n-1} \lambda^{j-1}(1-\lambda)x_{n-j} \\ &= (1-\lambda)x_n + \lambda x_n^{n-1}, \end{aligned}$$

where x_n^{n-1} is obtained by shifting n to $n-1$ in the formula for x_{n+1}^n. This new formula is curious because it indicates that a new prediction x_{n+1}^n can be understood as a weighted average of a new observation, x_n, and the previous prediction, x_n^{n-1}. That is, it suggests a computationally simple way of updating one's prediction each time a new measurement is made.

In the general context where neither δ nor x_0 is necessarily zero, the one-step-ahead prediction is

$$\begin{aligned} x_{n+1}^n &= E[x_{n+1}|x_1, \ldots x_n] \\ &= E\left[\sum_{j=1}^{n} \lambda^{j-1}(1-\lambda)x_{n+1-j} + w_{n+1} + \delta \sum_{j=0}^{n} \lambda^j + \lambda^n x_0 \middle| x_1, \ldots x_n \right] \\ &= \sum_{j=1}^{n} \lambda^{j-1}(1-\lambda)x_{n+1-j} + \delta \sum_{j=0}^{n} \lambda^j + \lambda^n x_0, \end{aligned}$$

which satisfies a similar, curious relationship:

$$\begin{aligned} x_{n+1}^n &= \sum_{j=1}^{n} \lambda^{j-1}(1-\lambda)x_{n+1-j} + \delta \sum_{j=0}^{n} \lambda^j + \lambda^n x_0 \\ &= (1-\lambda)x_n + \lambda \left\{ \sum_{j=1}^{n-1} \lambda^{j-1}(1-\lambda)x_{n-j} + \delta \sum_{j=0}^{n-1} \lambda^j + \lambda^{n-1} x_0 \right\} \\ &\qquad - \lambda \left\{ \delta \sum_{j=0}^{n-1} \lambda^j + \lambda^{n-1} x_0 \right\} + \delta \sum_{j=0}^{n} \lambda^j + \lambda^n x_0 \\ &= (1-\lambda)x_n + \lambda x_n^{n-1} + \delta \left\{ \sum_{j=0}^{n} \lambda^j - \lambda \sum_{j=0}^{n-1} \lambda^j \right\} + \{\lambda^n x_0 - \lambda^n x_0\} \\ &= (1-\lambda)x_n + \lambda x_n^{n-1} + \delta, \end{aligned}$$

having noted that $\sum_{j=0}^{n} \lambda^j - \lambda \sum_{j=0}^{n-1} \lambda^j = 1 + \sum_{j=1}^{n} \lambda^j - \sum_{j=0}^{n-1} \lambda^{j+1} = 1$.

When n is large, a good approximation to mean squared prediction error may be calculated from the formula

$$P_{n+m}^n = \sigma_w^2 \sum_{j=0}^{m-1} \psi_j^2,$$

wherein the ψ_j are defined from the power-series expansion of

$$\psi_j(z) = \frac{1-\lambda z}{1-z} = \sum_{j=0}^{\infty} \psi_j z^j.$$

The solution is $\psi_0 = 1$ and $\psi_j = 1 - \lambda$ for $j \geq 1$, which provides

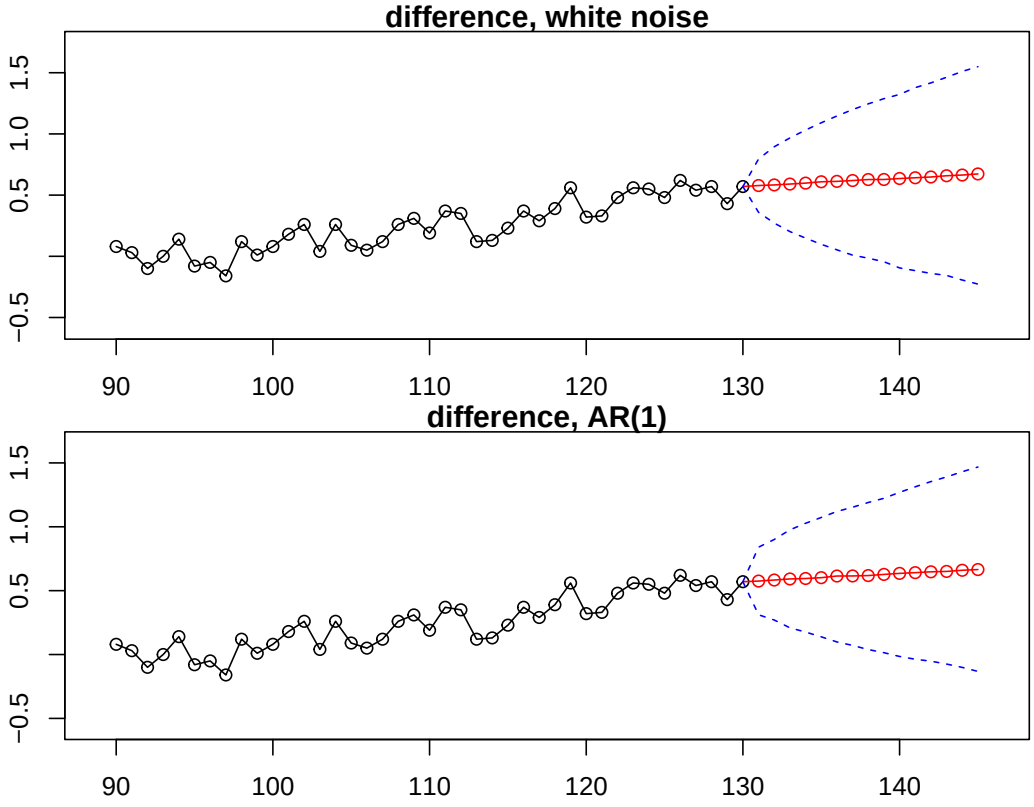
$$P_{n+m}^n = \sigma_w^2 \{1 + (m-1)(1-\lambda)^2\}.$$

□

The Bayesian technique of simulating predictions of the time series offers an alternative to calculating forecasts of an ARIMA model. This is demonstrated in the following example.

Example: Forecasting average global temperatures

The following plots display forecasts of the average global temperatures time series for up to $m = 15$ years ahead, under two differencing models, the first with uncorrelated differences (top panel) and the other with autoregressive AR(1) residuals (bottom panel). The dotted blue lines are 95% prediction bands.



Both forecasts depict a continuing increasing trend, and any differences between the two sets of forecasts are very small. □

Section 5A.3: Seasonal ARIMA models.

Substantial extension of the already broad ARIMA class of time series models is possible by folding in the possibility of seasonal dependencies. Building from the simplest to most complex models, several definition are given below.

The first definition describes a class of *pure* seasonal models, where interesting dependencies reach back an entire season, while the dependencies of time-series values within the same season are of those of white noise.

Definition: The time series (x_t) is $ARMA(P,Q)_s$ *seasonal autoregressive moving-average* if

$$\Phi_P(B^s)x_t = \Theta_Q(B^s)w_t$$

for operators

$$\begin{aligned}\Phi_P(B^s)x_t &= 1-\Phi_1B^s-\cdots-\Phi_PB^{Ps}\\ \Theta_Q(B^s)x_t &= 1+\Theta_1B^s+\cdots+\Theta_QB^{Qs}\end{aligned}$$

where (w_t) is Gaussian white noise.

As an example, consider an $ARMA(1,1)_4$ model for yearly period in quarterly data, which is defined by the relationship

$$x_t = \Phi_1 x_{t-4} + w_t + \Theta_1 w_{t-4}.$$

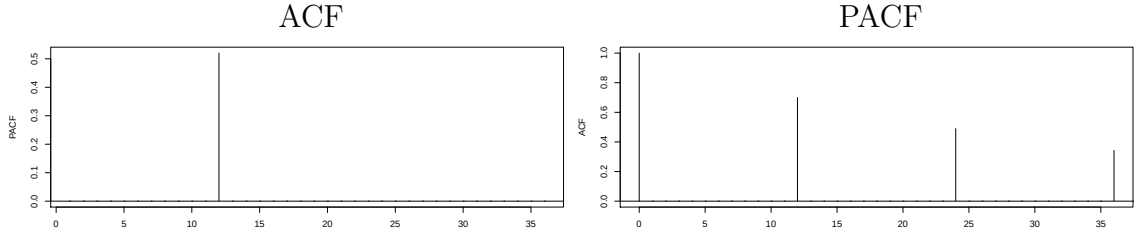
As another example, consider the $MA(1)_{12}$ model for yearly a period in monthly data, defined by

$$x_t = w_t + \Theta_1 w_{t-12}$$

Direct adaptation of previous mathematical techniques may be used to derive the corresponding autocorrelation function, which is given by

$$\rho(h) = \begin{cases} 1 & \text{if } h=0\\ 0 & \text{if } h=1,\ldots,11\\ \dfrac{\Theta_1}{1+\Theta_1^2} & \text{if } h=12\\ 0 & \text{if } h=13,14\ldots\end{cases}$$

Plots of the autocorrelation and partial autocorrelation functions for hypothetical parameter values are



As a third example, consider the $AR(1)_{12}$ model for yearly a period in monthly data, which is defined by

$$x_t = \Phi_1 x_{t-12} + w_t.$$

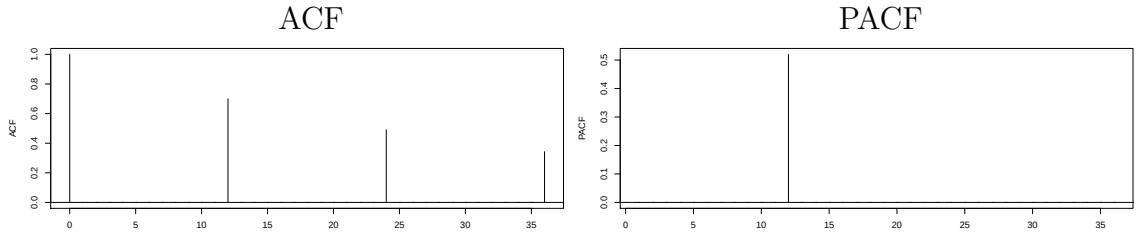
When $|\Phi_1| < 1$, this model has the causal representation

$$x_t = \sum_{j=0}^{\infty} \Phi_1^j w_{t-12j},$$

from which it is not difficult to deduce that the model's autocorrelation function is

$$\rho(12h) = \Phi_1^j$$

Plots of the autocorrelation and partial autocorrelation functions for hypothetical parameter values are



An extension of the seasonal ARMA models incorporates the possible presence of interesting dependencies among time-series values within the same season.

Definition: The time series (x_t) is $ARMA(p,q) \times (P,Q)_s$ *multiplicative seasonal autoregressive moving-average* if

$$\Phi_P(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t$$

for operators

$$\begin{aligned}
\Phi_P(B^s) &= 1 - \Phi_1 B^s - \cdots - \Phi_P B^{Ps} \\
\phi(B) &= 1 - \phi_1 B - \cdots - \phi_p B^p \\
\Theta_Q(B^s) &= 1 + \Theta_1 B^s + \cdots + \Theta_Q B^{Qs} \\
\theta(B) &= 1 + \theta_1 B + \cdots + \theta_q B^q
\end{aligned}$$

where (w_t) is Gaussian white noise.

As an example of a multiplicative seasonal ARMA, consider the $ARMA(0,1)\times(1,0)_{12}$ model, which is defined by

$$x_t = \Phi_1 x_{t-12} + w_t + \theta_1 w_{t-1}$$

Under this model it is quickly deduced that the lag-zero autocovariance (*i.e.*, the variance of x_t) satisfies the relationship

$$\begin{aligned}\gamma(0) &= Var(x_t) \\ &= Var(\Phi_1 x_{t-12}) + Var(w_t) + Var(\theta_1 w_{t-12}) \\ &= \Phi_1^2\gamma(0) + \sigma_w^2 + \theta_1^2\sigma_w^2;\end{aligned}$$

hence, upon solving, the lag-zero autocovariance itself is deduced as

$$\gamma(0) = \frac{1+\theta_1^2}{1-\Phi_1^2}\sigma_w^2.$$

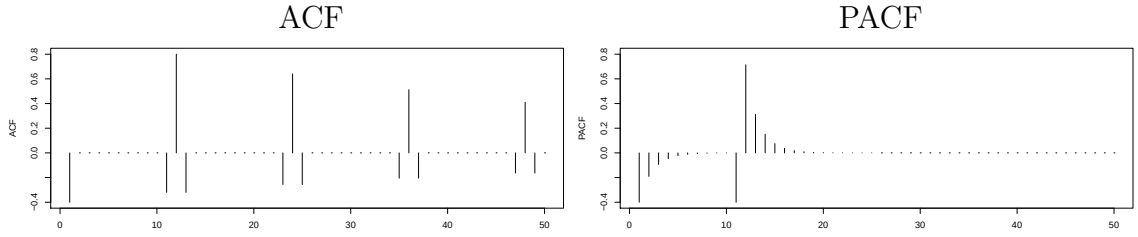
Various patterns in the autocovariance function may be deduced similarly. For instance,

$$\begin{aligned}\gamma(1) &= E[x_t x_{t-1}] \\ &= \Phi_1 E[x_{t-12}x_{t-1}] + E[w_t x_{t-1}] + \theta_1 E[w_{t-12}x_{t-1}] \\ &= \Phi_1 E[x_{t-12}x_{t-1}] + E[w_t]E[x_{t-1}] + \theta_1 E[w_{t-12}]E(x_{t-1}] \\ &= \Phi_1 E[x_{t-12}x_{t-1}] = \Phi_1\gamma(11).\end{aligned}$$

Using the same approach, the reader is encouraged to check that, furthermore,

$$\begin{aligned}\gamma(h) &= \gamma(h-12) \text{ for } h > 2 \\ \rho(12h) &= \Phi_1^h \\ \rho(12h-1) &= \rho(12h+1) = \frac{\theta_1}{1+\theta_1^2}\Phi_1^h \\ \rho(h) &= 0 \text{ for } h \notin \{12h^*-1, 12h^*, 12h^*+1 : h^* = 0, 1, 2, \ldots\}.\end{aligned}$$

These patterns are clearly depicted in the following plot of the autocorrelation function for hypothetical parameter values; a plot of the partial autocorrelation function is also shown:



A third extension of the ARIMA idea additionally brings in both ordinary and seasonal differencing, to accommodate the possibility of non-stationarity. It is defined using the *seasonal difference operator*, which has the notation

$$\nabla_s = (1-B^s).$$

The extended class is defined as follows.

Definition: The time series (x_t) is $ARIMA(p,d,q) \times (P,D,Q)_s$ *multiplicative seasonal autoregressive integrated moving-average*, or *SARIMA*, if

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t \;=\; \Theta_Q(B^s)\theta(B)w_t$$

for operators $\Phi_P(B^s)$, $\phi(B)$, $\Theta_Q(B^s)$, and $\theta(B)$ as in the $ARMA(p,d,q) \times (P,D,Q)_s$ model,

$$\nabla_s^D = (1-B^s)^D, \quad \text{and} \quad \nabla^d = (1-B)^d,$$

where (w_t) is Gaussian white noise.

As an example of a multiplicative seasonal ARIMA, consider the $ARMA(0,1,1) \times (0,1,1)_{12}$ model, which is defined by the relationship

$$(1-B^{12})(1-B)x_t \;=\; (1+\Theta_1 B^{12})(1+\theta_1 B)w_t.$$

Expanding this expression specifies

$$(1-B-B^{12}+B^{13})x_t \;=\; (1+\theta_1 B+\Theta_1 B^{12}+\Theta_1\theta_1 B^{13})w_t,$$

or, upon resolving the backshift operations and rearranging,

$$x_t \;=\; x_{t-1}+x_{t-12}-x_{t-13}w_t+\theta_1 w_{t-1}+\Theta_1 w_{t-12}+\Theta_1\theta_1 w_{t-13}.$$

Section 5A.4: Computational software.

Computational software for time series analysis typically includes functionality for calculating parameter estimates and other statistics under a variety of time series models, including seasonal ARIMA models. The choices of available software are fewer for implementing a Bayesian time series analysis, compared to classical approaches such as those devised from maximum likelihood estimation. Bayesian methods are extraordinarily flexible, and our exploration of them has offered a computational perspective, rather than one emphasizing mathematical deduction, that some may find helpful for understanding time series models and associated inference procedures. However, we have seen that Bayesian methods can be computationally intensive, and can require close monitoring and tuning.

The R software package includes a built-in function called `arima`, which implements time series analysis from the approach of classical inference, using techniques that primarily emphasize maximum likelihood estimation and related methods. This function supports specialized syntax for working with seasonal ARIMA models, and time series models that make use of regression for de-trending. The output to this function includes estimates and standard errors of the model parameters, as well as a value for AIC. In addition, it can produce several statistics for model checking. The model checking statistics are formulated using concepts that bear some resemblance to those underlying the Bayesian posterior predictive p-values we have explored, but require additional explanation.

Whereas the Bayesian model-checking procedures we have discussed for checking autocorrelations are based on predictions of the white noise time series, (w_t), that defines the model. For example, under an autoregressive time series model this is obtained from the innovations,

$e_t = x_t - x_t^{t-1}$. Under an integrated autoregressive model these would be obtained from the innovations of the difference series, $e_t = \Delta x_t - \Delta x_t^{t-1}$. Under a moving average model, they might be calculated using the Innovations Algorithm.

In classical model checking, the innovations time series, (e_t), is replaced with *fitted innovations* time series, $(\hat{e}_t)$, which is determined from the model form having substituted parameter estimates for the parameters themselves. For example, the fitted innovations of an $AR(p)$ time series are $\hat{e}_t = x_t - \hat{x}_t^{t-1}$, where $\hat{x}_t^{t-1} = \hat{\phi}_1 x_{t-1} + \cdots + \hat{\phi}_p x_{t-p}$ is the relevant prediction formula having substituted *estimates* of the autoregressive parameters, $\hat{\phi}_1, \ldots, \hat{\phi}_p$, for $\phi_1, \ldots, \phi_p$.

One approach to model checking is to graph the fitted innovations, as we did with the residuals in regression, and check for patterns that deviate from what we would expect in a white-noise time series. A variation of this is to check the *standardized innovations*, $\hat{z}_t = \hat{e}_t / \sqrt{\hat{\sigma}_w^2}$, where $\hat{\sigma}_w^2$ is an estimate of a white-noise variance σ_w^2.

A classical check on autocorrelations aggregated across multiple lags is derived from what is called the Ljung-Box-Peirce statistic, which is a weighted sum-of-squared fitted autocorrelations. The fitted autocorrelation function is

$$\hat{\rho}_e(h) = \frac{1}{n-h-t_0+1} \sum_{t=t_0}^{n-h} \frac{\hat{e}_{t+h}\hat{e}_t}{\hat{\sigma}_w^2},$$

Here, t_0 is a minimum time point above which it is expected that the fitted innovations are accurate. Typically is set to $t_0 = 1$, but it might be set larger, *e.g.*, when the predictions are calculated from a truncated infinite-sum representation. The fitted autocorrelations are aggregated into the Ljung-Box-Peirce statistic according to

$$Q = n(n+2) \sum_{h=1}^{H} \frac{\hat{\rho}_e^2(h)}{n-h}.$$

Related mathematical theory provides a rule to conclude that autocorrelations are present in the innovations series when the Ljung-Box-Peirce statistic exceeds the $(1-\alpha)$'th quantile of a χ^2 distribution with $H-p-q$ degrees of freedom. That is, the rule is to revise the model if

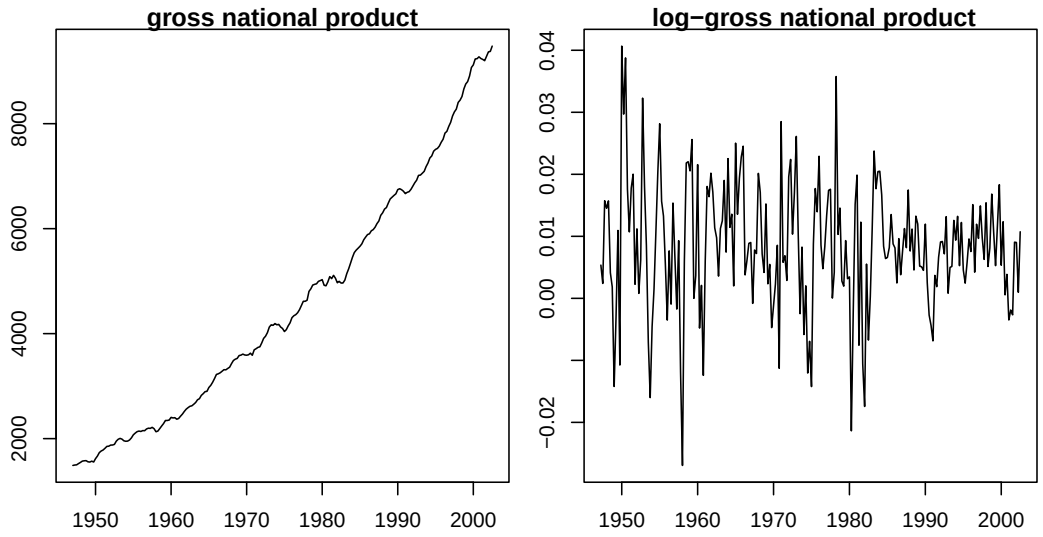
$$Q > \chi^2_{H-p-q,1-\alpha},$$

having denoted bu $\chi^2_{\nu,p}$ the p'th quantile of a χ^2_ν distribution, where α is typically about 0.05.

The following example illustrates a model-building procedure proceeding from a classical approach. All calculations are made using the R function `arima`.

Example: GNP growth rate

The time series plotted in the left panel of the figure below is of quarterly measurements of the gross national product (in billions of chained 1996 dollars) from 1947 to 2002. The raw time series is clearly non-stationary, as it exhibits a strong trend.

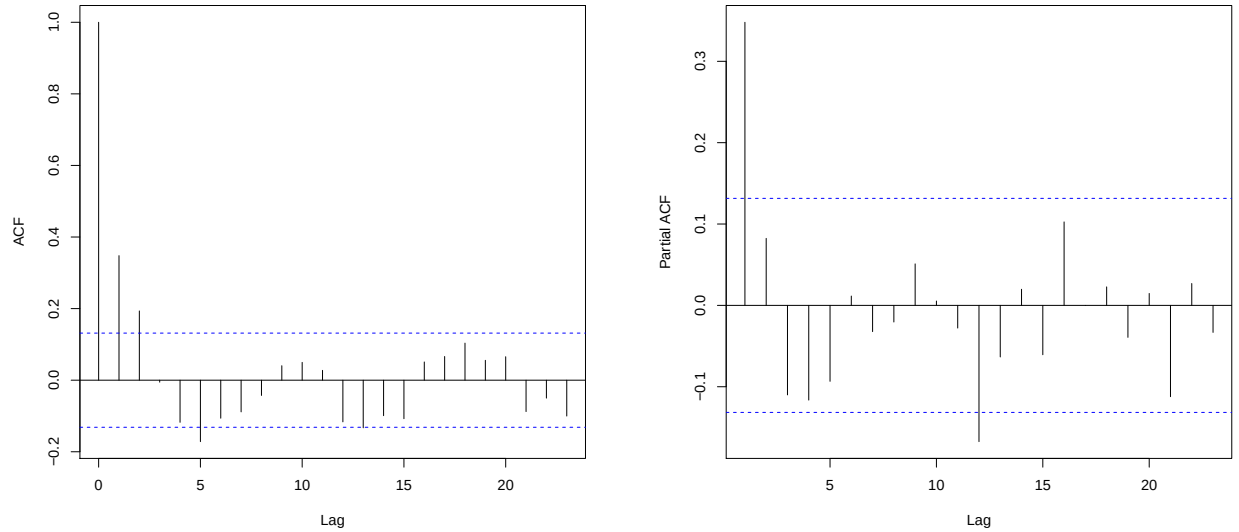


The right panel displays a plot of the GNP growth rate, which is typically defined as

$$y_t \quad = \quad \nabla \log x_t$$

where x_t is the GNP at time t. That is (y_t) is the time series of differences calculated on the logged data time series, $(\log x_t)$. The patterns in the latter plot resemble those of a stationary time series. The following is an exploration of these data, whose goal is to find an $ARIMA(p, d = 1, q)$ model that suitably describes the logged-data time series $(\log x_t)$. This goal may be stated equivalently as that of finding an $ARMA(p, q)$ model that is suitable for the growth-rate time series (y_t).

A next step in the exploration is to identify the orders of the autoregressive and moving-average pieces, p and q. Let us check weather the ACF and PACF offer any indications of what p and q could be. These diagnostics are plotted as follows.



There is clearly a lot of uncertainty in these plots. There are indications in the ACF that an $ARIMA(p = 0, d = 1, q = 2)$ model may be suitable. However, the PACF suggests

an $ARIMA(p = 1, d = 1, q = 0)$ model. These two indications are contradictory, so our approach will be to explore several values of p and q through various diagnostics.

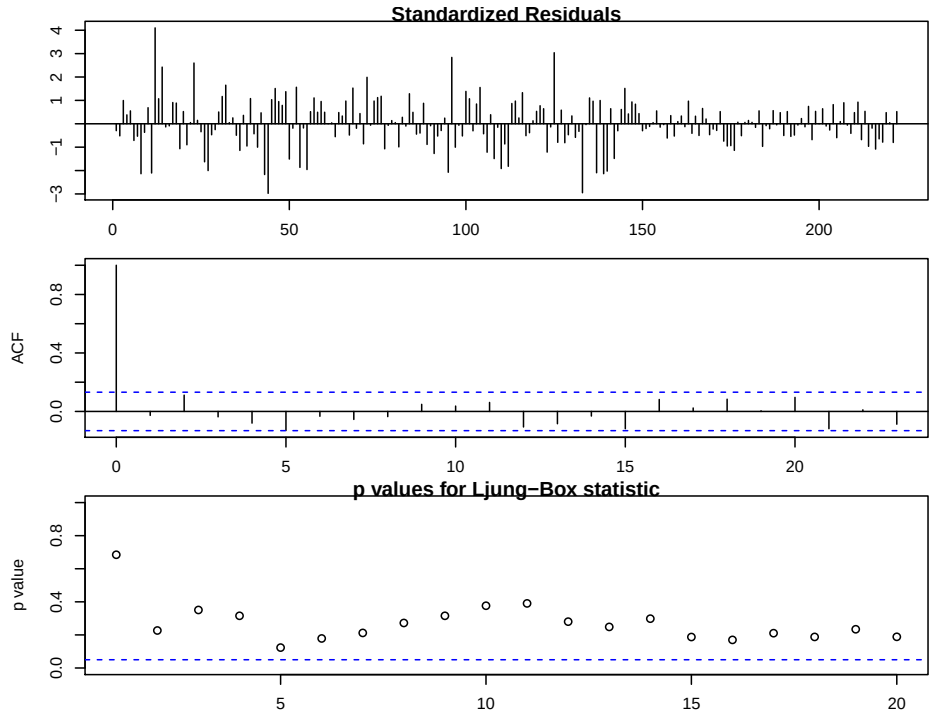
The following table lists parameter estimates and other numerical diagnostics for several possible models. The parameter estimates are listed with standard errors in parentheses

p	q	μ	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\sigma}_w^2$	AIC
1	0	0.008	0.347				0.0001	-8.2944
		(0.001)	(0.063)					
0	2	0.008			0.303	0.204	0.0001	-8.2977
		(0.001)			(0.065)	(0.064)		
1	1	0.008	0.463		0.076		0.0001	-8.2893
		(0.001)	(0.127)		(0.135)			
1	2	0.008	0.241		0.076	0.162	0.0001	-8.2934
		(0.001)	(0.207)		(0.203)	(0.085)		
2	2	0.008	1.346	-0.738	-1.063	0.562	0.0001	-8.3104
		(0.001)	(0.137)	(0.154)	(0.187)	(0.197)		

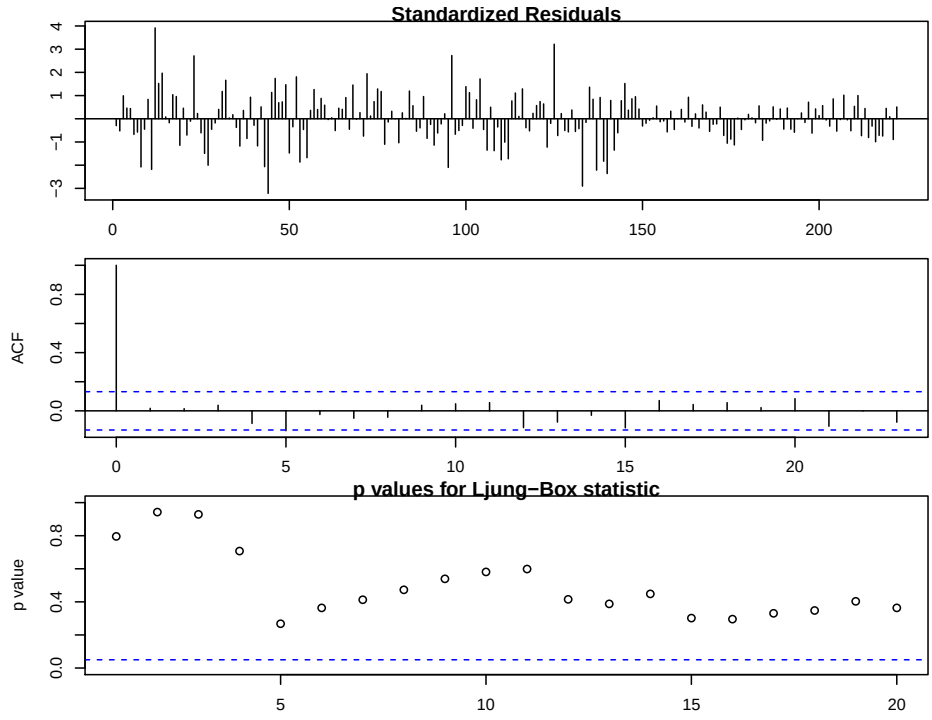
Observe that the $ARIMA(2, 1, 2)$ model is preferred by the AIC criterion. Note also that in both the $ARIMA(1, 1, 1)$ and $ARIMA(1, 1, 2)$ models, the estimated coefficient $\hat{\theta}_1$ is not significant when compared against its standard error, and $\hat{\theta}_2$ is also not significant in the latter model. Here, significance refers to a situation where the parameter within a distance of about twice its standard error from zero.

Graphical diagnostics are shown below for the individual models. These diagnostics include a plot of the standardized residuals, $\hat{z}_t$, a plot of the fitted autocorrelation function $\hat{\rho}_e(h)$, and a plot of classical p-values from the Ljung-Box-Peirce statistic $Q(H)$ across a range of H.

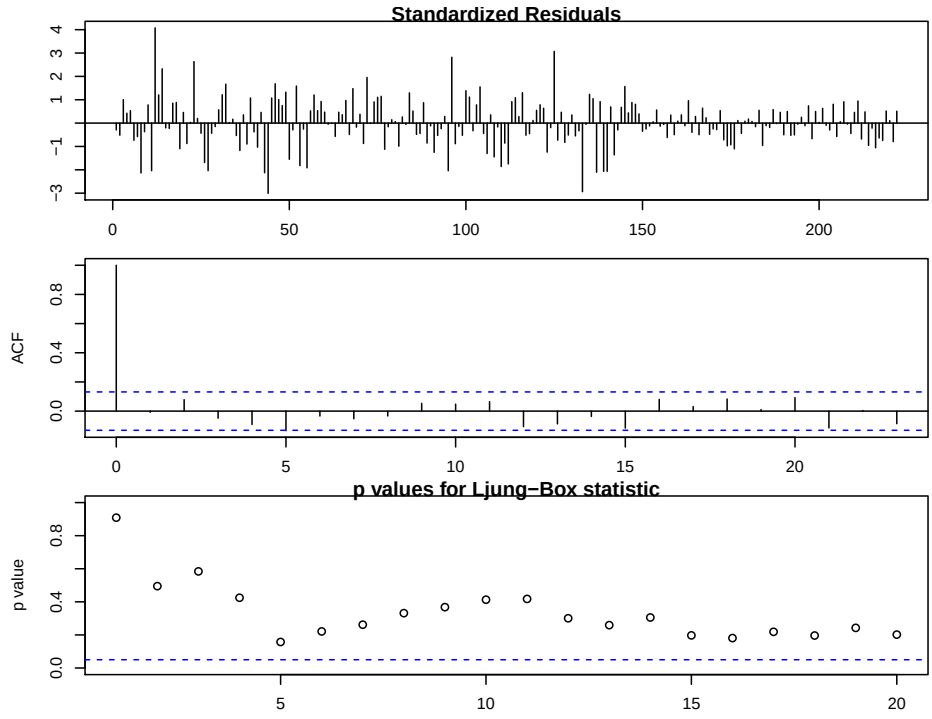
For the $ARIMA(1, 1, 0)$ model the diagnostics are as follows:



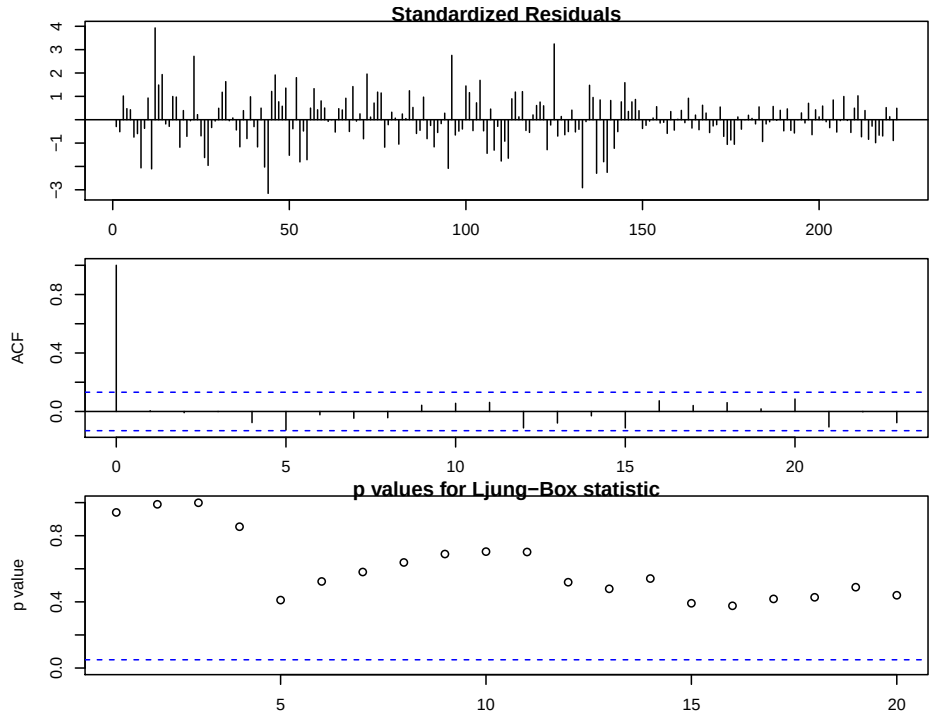
For the $ARIMA(0,1,2)$ model the diagnostics are



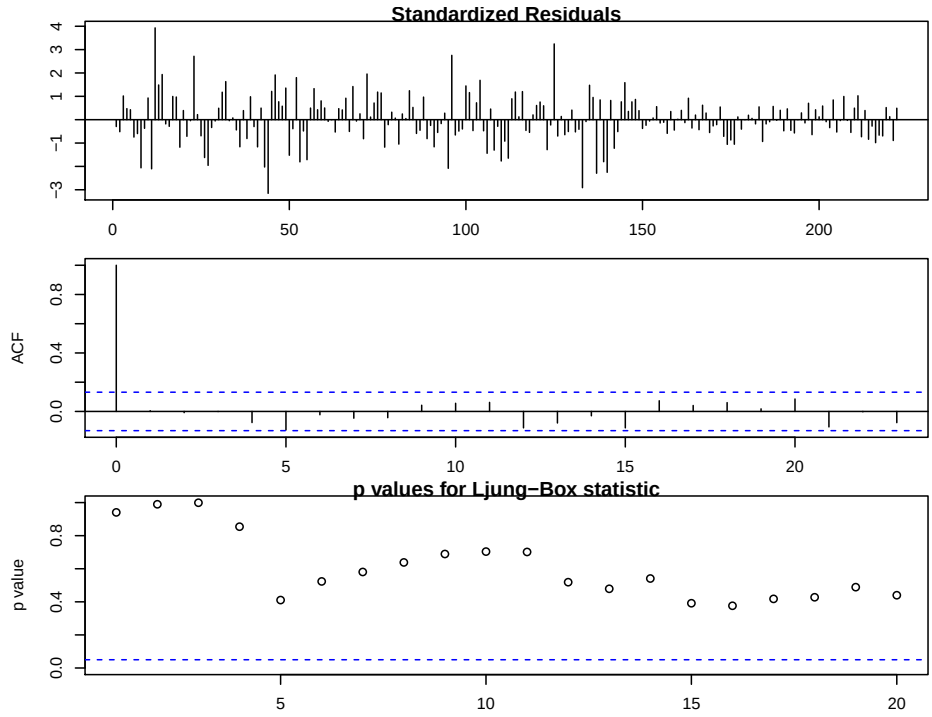
For the $ARIMA(1,1,1)$ model they are



For the $ARIMA(1,1,2)$ model the diagnostics are



For the $ARIMA(2,1,2)$ model they are



Comparing the diagnostics, the classical p-values associated with the Ljung-Box-Peirce statistic are generally larger for the $ARIMA(2,1,2)$ model, compared to the $ARIMA(1,1,0)$ and $ARIMA(0,1,2)$ models, and all of its estimated parameter values are significant when compared against their standard errors.