Model Assessment 9

In the previous chapters we discussed techniques for fitting a univariate time series model to an observed realization. We discussed two basic types of univariate models: (1) strictly correlation-based models and (2) signal-plus-noise models. Strictly correlation-based models include AR and ARMA models for stationary data along with ARIMA and seasonal models for nonstationary data. Signal-plus-noise models are nonstationary models that include a deterministic signal. We fit these models making decisions about unit roots, seasonal or trend behavior, and obtained final models based on AIC-type methods. If you have carefully followed the model fitting steps, it is tempting to accept the fitted model as the "best we can do" and use it as the final model. However, experienced time series analysts know that *fitting a model is just the beginning point* in obtaining a "final model". After fitting a model to a realization, you should take steps to assure that the model is *suitable.* The first step in this process is to examine the residuals. If the model is appropriate, the residuals should be uncorrelated. This topic will be discussed in Section 9.1. Suppose you have obtained a model by using a model identification technique (such as AIC), estimated the parameters, and examined the residuals to satisfy yourself that they are sufficiently white (that is, uncorrelated).[1] The next step is to determine whether the model *makes sense*. As we examine models, we will evaluate them with regard to the following Key Questions which guide us in assessing whether a given model is appropriate.

Key Questions for Assessing a Model Fit:

1. Does the model "whiten" the residuals?
2. Do realizations and their characteristics behave like the data?
3. Do forecasts reflect what is known about the physical setting?

Section 9.1 discusses methods for testing residuals for white noise and normality. In Section 9.2 we use the analysis strategy outlined above to analyze and compare models fit to the global temperature data. Section 9.3 examines models fit to the sunspot data.

9.1 RESIDUAL ANALYSIS

We recommended the use of AIC and its variations for ARMA, ARIMA, seasonal, and signal-plus-noise model identification in Chapters 6–8. AIC and its variations select a model for the stationary component of the model *based on reducing the estimated white noise variance,* $\hat{\sigma}_a^2$, while controlling the number of parameters required to do so. The estimate, $\hat{\sigma}_a^2$, is output from **`est.ar.wge`** and **`est.arma.wge`** in the variable **`$avar`**. In this section we consider residuals from three types of models

1 Refer to Section 7.3 regarding uncorrelated noise for which you might suspect conditional heteroskedasticity.

DOI: 10.1201/9781003089070-9

- ARMA models
- ARIMA or seasonal models
- Signal-plus-noise models

(a) Residuals from a Fitted ARMA(p,q) model
For the ARMA(p,q) model

$$X_t - \hat{\phi}_1 X_{t-1} - \cdots - \hat{\phi}_p X_{t-p} = \hat{a}_t - \hat{\theta}_1 \hat{a}_{t-1} - \cdots - \hat{\theta}_q \hat{a}_{t-q} + \bar{x}\left(1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p\right),$$

the residuals were previously given in (6.5) as

$$\hat{a}_t = X_t - \hat{\phi}_1 X_{t-1} - \cdots - \hat{\phi}_p X_{t-p} + \hat{\theta}_1 \hat{a}_{t-1} + \cdots + \hat{\theta}_q \hat{a}_{t-q} - \bar{x}\left(1 - \hat{\phi}_1 - \cdots - \hat{\phi}_p\right). \tag{9.1}$$

In Section 6.1.1.3 we discussed the calculation of the residuals, the concern about dependence on starting values, and the technique of calculating residuals using *backcasting*. The following are points discussed in Section 6.1.1.3.

Key Points:

1 We recommend computing the residuals using *backcasting*.
   - The backcast residuals provide a "full set" of ***n*** residuals that do not have undue dependence on starting values.
2. All ***tswge*** routines that calculate residuals use the backcast procedure.

(b) Residuals from an ARIMA or Seasonal Model
The analysis of ARIMA and seasonal models involves two steps:

(1) Identify the nonstationary components of the data and transform the data to stationarity
(2) Model the transformed data using an ARMA model

As an example, for a fitted model, $\left(1-B\right)\hat{\phi}\left(B\right)(X_t - \bar{X}) = \hat{\theta}\left(B\right)\hat{a}_t$, we

(1) Transform the data to obtain $Y_t = \left(1-B\right)X_t$
(2) Fit an ARMA model to the Y_t data
   - The backcast residuals in question are those obtained from the modeling of Y_t

(c) Residuals from a Signal-plus-Noise Model
The procedure is similar to the nonstationary case (b) above

(1) Fit a signal (for example, a line) to the data and remove the estimated signal by transforming the data
(2) Fit an ARMA model to the transformed data

Consider the line+noise model $X_t = a + bt + Z_t$, where Z_t is a zero-mean, stationary process. The analysis steps are as follows:

(1) Find the least squares estimates of a and b, call them $\hat{a}$ and $\hat{b}$, and transform the data to obtain $\hat{Z}_t = X_t - \hat{a} - \hat{b}t$
(2) Fit an AR model to the transformed data, $\hat{Z}_t$
   − The backcast residuals are those obtained from modeling $\hat{Z}_t$

Key Points:

1. Case (a) above involves fitting an ARMA model to a stationary time series
2. In cases (b) and (c), the final step is fitting an ARMA model to the transformed "stationary" data.

9.1.1 Checking Residuals for White Noise

Recall that the models discussed in Chapters 6–8 all assume that the noise, a_t, is *white* noise. Consequently, if the fitted model is appropriate, then the residuals should be well modeled as white noise.[2] If the residuals of the fitted model contain significant autocorrelation, then the model has not suitably accounted for the correlation structure in the data, regardless of the size of the estimated white noise variance.[3] Another assessment that may be of interest is whether the residuals appear to be *normally distributed*, an assumption of the MLE which is based on the normal likelihood (see Woodward et al. (2017). Note that the YW and Burg estimation procedures have no underlying distributional assumptions.

Key Points:

1. An appropriate model should "whiten" the residuals.
2. The backcast residuals should have the appearance of white noise.
3. If there is correlation structure remaining in the residuals, then this is an indication that the model is inadequate**.**

9.1.1.1 Check Residual Sample Autocorrelations against 95% Limit Lines

One method of testing residuals for white noise has already been discussed in Section 6.1.2.1. Specifically, when fitting an ARMA model to a set of data, we recommend first checking the data for white noise. For white noise, the autocorrelations satisfy $\rho_k = 0$, $k \neq 0$, and the method discussed in Section 6.1.2.1 is designed to test $H_0 : \rho_k = 0$ by plotting the sample autocorrelations, $\hat{\rho}_k, k = 1,2,\ldots,K$, against the 95% limits, $\pm 2/\sqrt{n}$, for some K. For a specific lag k, if the data are white noise, there is a 5% chance that $\hat{\rho}_k$,

2 We focus the discussion here on the residuals from a *stationary* ARMA(p,q)fit. This discussion is applicable to the fitting of nonstationary models in cases (b) and (c) above, because the final step in those estimation procedures is to fit a stationary ARMA (p,q) model to some transformed version of the original data.

3 This is analogous to the examination of residuals in multiple (or simple) linear regression analysis to assess whether the residuals are uncorrelated.

will fall outside the limits $\pm 2/\sqrt{n}$. Consequently, if the data are white noise then it would not be unusual for about 5% of the sample autocorrelations to fall outside the 95% limit lines. If substantially more than 5% of the sample autocorrelations of the data fall outside these lines, this is evidence against white noise. See Section 6.1.2.1 for a more complete discussion of the use of these limit lines for white noise checking including the advice to always plot the data.

While we previously used the above method for checking an "original" dataset for white noise, it can also be used to check the (backcast) residuals from a fitted model for white noise. That is, we can compare the *residual* sample autocorrelations against $\pm 2/\sqrt{n}$ to help decide whether the residuals appear to be "white".

9.1.1.2 Ljung-Box Test

One problem with checking the sample autocorrelations against the 95% limit lines is that the "5% chance of exceeding these lines" applies to the sample autocorrelations separately for each lag k. It would be desirable to have a single *portmanteau*[4] procedure that tests the first K sample autocorrelations "as a group" rather than individually. That is, we the test for white noise by testing the hypotheses

$$\begin{aligned} &H_0 : \rho_1 = \rho_2 = \cdots = \rho_K = 0 \\ &H_1 : \text{at least one of the } \rho_k \neq 0 \text{ for } 1 \le k \le K. \end{aligned} \tag{9.2}$$

Box and Pierce (1970) and Ljung and Box (1978) developed tests of the null hypothesis in (9.2). The test developed by Ljung and Box (1978) is widely used for this purpose. The Ljung-Box test statistic is given by

$$L = n(n+2)\sum_{k=1}^{K} \frac{\hat{\rho}_k^2}{n-k}. \tag{9.3}$$

Ljung and Box (1978) show that L in (9.3) is approximately distributed as χ^2 with K degrees of freedom when the data are white noise. When the data to be tested are residuals from a fitted ARMA(p,q) model, then the test statistic is approximately χ^2 with $K - p - q$ degrees of freedom. Examination of (9.3) shows that the test statistic measures the size of the first K sample autocorrelations as a group, so that large values of L suggest that the data are not white noise. Thus, the null hypothesis of white noise residuals is rejected when L is sufficiently large, and in particular if $L > \chi^2_{1-\alpha}(K-p-q)$ where $\chi^2_{1-\alpha}(m)$ denotes the $(1-\alpha)\times 100\%$ percentile of the χ^2 distribution with m degrees of freedom. The Ljung-Box test can be implemented by using the ***tswge*** command **`ljung.wge`**. Although other values of K could be used, in the examples that follow we will use $K = 24$ and $K = 48$, which is consistent with the values of K used by Box et al. (2008). For small sample sizes (n ≤ 100), Box et al. (2008) and Ljung (1986) recommend the use of smaller values of K.

QR 9.1
Ljung-Box Test

4 Box and Jenkins used the uncommon term "portmanteau" which can be defined to mean "combining two or more aspects or qualities" to describe a test that considers the sample autocorrelations "as a group".

Example 9.1 Checking Residuals for White Noise
In this example we check the whiteness of residuals for several models fit to datasets in Chapters 6–8.

(a) AR(4) Realization Figure 6.22(a)
Realization 1 in Figure 5.16 and Figure 5.18(g) show a realization from the AR(4) model previously referred to as Model (B). The realization and sample autocorrelations are shown in Figures 9.1(a) and (b), respectively. The code below generates the data, plots the data, sample autocorrelations, and Parzen spectral density estimate. It uses AIC(which selects an AR(4)), and obtains the ML estimates of the AR(4) model fit.

```
modelB=gen.arma.wge(n=150,phi=c(2.6,-3.34,2.46,-.9024),sn=3233)
plotts.sample.wge(modelB)
aic.B=aic.wge(modelB,p=0:10,q=0:4)# AIC selects p=4,q=0
fit40=est.arma.wge(modelB,p=4,q=0)
```

The backcast residuals from the fitted model are contained in **`fit40$res`**. The residuals are plotted in Figure 9.1(c) and the residual sample autocorrelations are shown in Figure 9.1(d) along with the 95% limit lines. The command

```
plotts.sample.wge(fit40$res,lag.max=48,arlimits=TRUE)
```

will produce Figures 9.1(c) and 9.1(d) along with the Parzen spectral density estimate (not shown).

The residuals appear to be random, and the sample autocorrelations stay within the 95% limit lines, with the exception of $\hat{\rho}_{24} = -.180$ which falls just outside the horizontal lower limit line (−.163). The Ljung-Box test results for $K = 24$ and $K = 48$ are obtained using the commands

```
ljung.wge(fit40$res,p=4,q=0)  # K=24 is the default
ljung.wge(fit40$res,p=4,q=0,K=48)
```

Note that p and q are inputs to the **`ljung.wge`** command because they are involved in calculating the degrees of freedom needed for the chi-square test.[5] The output from these two commands follows.

```
$K 24
$chi.square 25.79192
$df 20
$pval 0.1727946

$K 48
$chi.square 46.69656
$df 44
$pval 0.3622047]
```

5 If the Ljung-Box test is applied to "raw data" (that are not residuals from a fitted model), then use $p = q = 0$.

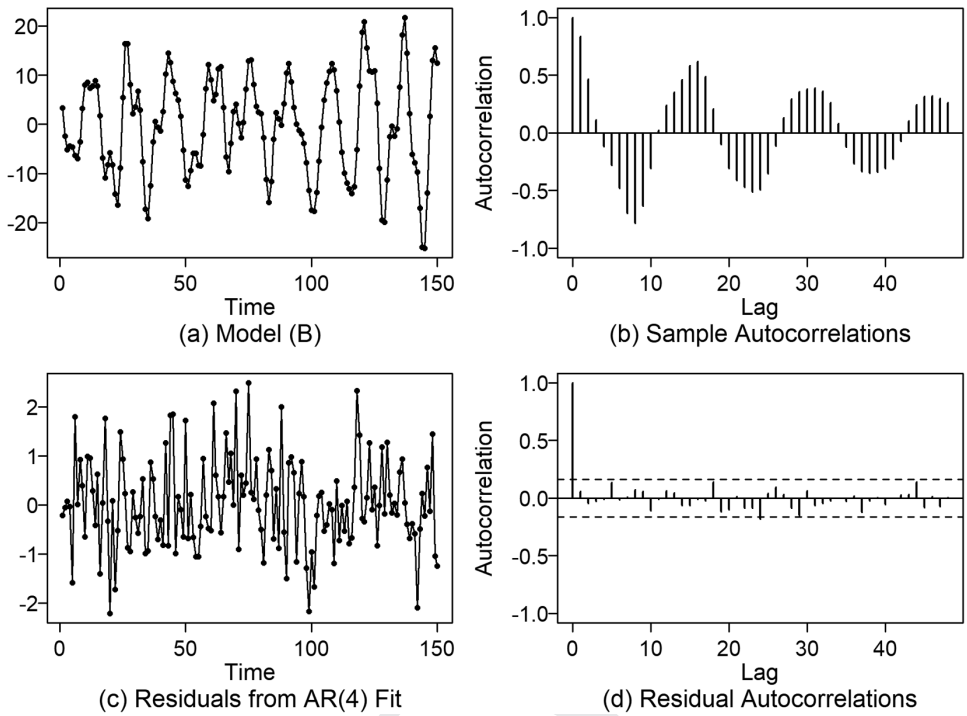

FIGURE 9.1 (a) Realization from Model (B) previously shown in Figure 5.16, (b) sample autocorrelations of the data in (a), (c) residuals from the AR(4) fit to data in Figure 6.22(a), (d) Sample autocorrelations of the residuals in (c) along with 95% limit lines.

The degrees of freedom for the $K = 24$ case, for example, are $K - p - q = 24 - 4 = 20$. For both $K = 24$ and $K = 48$, the p-values are considerably greater than $\alpha = .05$, so we do not have evidence to reject the null hypothesis of white noise. Consequently, based on the plots in Figure 9.1 and the Ljung-Box test, it appears that the fitted model does a good job of "whitening" the residuals.

(b) Log Air Passengers Data

In Example 7.2 we fit a seasonal "airline model" to the log Air Passengers data shown in Figure 7.20(a) (among other places throughout the book). The fitted model using Burg estimates is

$$(1-B)\left(1-B^{12}\right)\phi_{13}(B)X_t = a_t, \tag{9.4}$$

where $\hat{\sigma}_a^2 = 0.0013$ and $\phi_{13}(B)$ is given in (7.15). Recall that the analysis procedure was to difference the data, seasonally difference the "differenced data", and then model the remaining stationary data using an AR(13) model. The residuals to be analyzed and checked for white noise are those from the AR(13) fit. The following commands produce the desired residuals.

```
data(airlog)
d1=artrans.wge(airlog,phi.tr=1) # d1 is the differenced data
d1.s12=artrans.wge(d1,phi.tr=c(rep(0,11),1)) #seasoanlly difference d1
air.est=est.ar.wge(d1.s12,p=13,type= 'burg')
plotts.sample.wge(air.est$res,lag.max=48,arlimits=TRUE)
ljung.wge(air.est$res,p=13)
ljung.wge(air.est$res,p=13,K=48)
```

The residuals are in **air.est$res** and are plotted in Figure 9.2(a) along with their sample autocorrelations and limit lines in Figure 9.2(b).

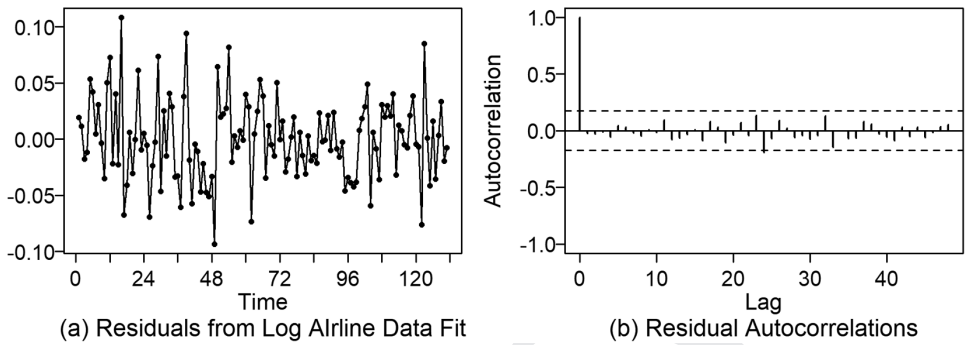

FIGURE 9.2 (a) Residuals from the "airline model" fit in (9.4) to the log airline data and (b) Sample autocorrelations of the residuals in (a) along with 95% limit lines.

The residuals look fairly white with some wandering behavior beginning at about $t = 45$. The sample autocorrelations stay within the limit lines, and the Ljung-Box statistics at $K = 24$ and 48 have p-values of 0.0724 and 0.3504, respectively.[6] The sample autocorrelations are fairly large for k at lags $k = 23$ and $k = 24$. These sample autocorrelations are responsible for the rather small p-value of 0.0724 for $K = 24$.[7] All sample autocorrelations greater than $k = 24$ are inside the limits and most are quite small, which is consistent with the p-value of 0.3504 for $K = 48$. Overall, the model appears to sufficiently whiten the residuals.

(c) DFW Temperature Data

The Dallas-Ft. Worth monthly temperature data from January 2000 through December 2020 are shown in Figure 7.22(a). In Section 7.3 these data were modeled using the nonstationary model

$$\left(1-1.732B+B^2\right)\hat{\phi}_{11}(B)\left(X_t-67.36\right)=\left(1-.95B\right)a_t, \tag{9.5}$$

where $\hat{\sigma}_a^2 = 9.41$. The model above along with the coefficients of $\hat{\phi}_{11}(B)$ are given in Example 7.3. The residuals from the fitted model along with the plots in Figure 9.3 can be obtained using the following code.

```
data(dfw.mon)
dfw.2000=window(dfw.mon,start=c(2000,1))
tr2.temp=artrans.wge(dfw.2000,phi.tr=c(1.732,-1))
tr2.est=est.arma.wge(tr2.temp,p=11,q=1)
plotts.sample.wge(tr2.est$res,lag.max=48,arlimits=TRUE)
ljung.wge(tr2.est$res,p=11,q=1)
ljung.wge(tr2.est$res,p=11,q=1,K=48)
```

The residuals look fairly white and the sample autocorrelations stay within the limit lines although the sample autocorrelation at $k = 13$ is quite close to the lower limit. The Ljung-Box test at $K = 24$ and 48 has p-values of 0.1044 and 0.1374, respectively. The fact that the sample autocorrelations seem to have a pseudo-cyclic pattern is of some concern. However, the model appears to sufficiently whiten the residuals.[8]

6 The degrees of freedom for the Ljung-Box test in this case are $K - 13$ based on the 13th-order stationary AR factor, $\hat{\phi}_{13}(B)$, in (8.4). The factors $1 - B$ and $1 - B^{12}$ were chosen by the investigator to obtain stationarity and are not counted as estimated parameters.

7 If you use the command **ljung.wge(tr2.est$res,p=11,K=22)** the p-value is .372.

8 The commands **ljung.wge(temp.est$res,p=9)** and **ljung.wge(temp.est$res,p=9,K=48)** were used because the factor $1 - 1.732 + B^2$ is a nonstationary factor that we "selected" based on examination of the factor tables and our knowledge of the 12-month cycle in the data. It was not "estimated". So, $p = 11$ is based on the 11th-order stationary factor, $\hat{\phi}_{11}(B)$ in (9.5).

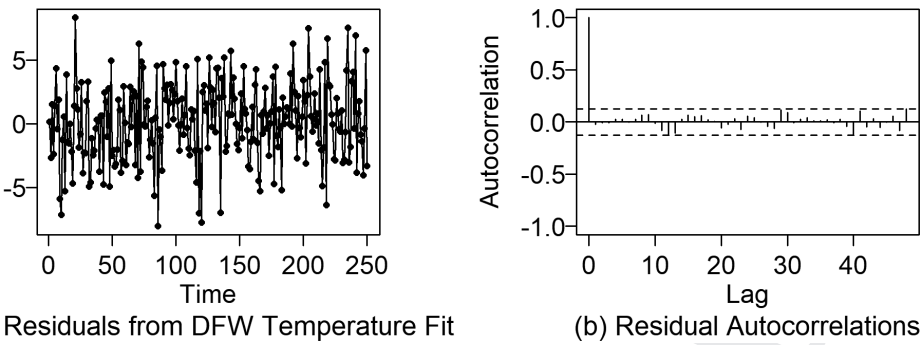

FIGURE 9.3 (a) Residuals from the Example 7.3 model fit to **DFW.2000** temperature data and (b) Sample autocorrelations of the residuals in (a) along with 95% limit lines.

In Example 8.6 we fit a cosine signal+noise model to the **DFW.2000** data.

(d) Sunspot Data

In Example 6.15 two models were considered for the **sunspot2.0** data:

- an AR(2) model
- an AR(9) model

(i) In Example 6.15 we considered an AR(2) model. Box, Jenkins, and Reinsel (2008) suggested an AR(2) model for a shorter sunspot series based on the partial autocorrelations (see Problem 6.13(a)). As mentioned in Example 6.15 the partial autocorrelations show "some" support for an AR(2) fit to the **sunspot2.0** data. The AR(2) model fit to the **sunspot2.0** data is

$$\left(1-1.38B+.69B^2\right)\left(X_t-78.97\right)=a_t,$$

where $\hat{\sigma}_a^2 = 659.39$.

The following code calculates the residuals and places the backcast residuals in **ss2$res**.

```
data(sunspot.new)
ss2=est.ar.wge(sunspot.new,p=2)
ljung.wge(ss2$res,p=2)
ljung.wge(ss2$res,p=2,K=48)
```

The backcast residuals in **ss2$res** are plotted in Figure 9.4(a) and they appear to be "somewhat white". Figure 9.4(b) shows the sample autocorrelations where it is seen that 6 of the 48 sample autocorrelations (that is, 12.5%) fall outside the 95% limit lines, suggesting that the AR(2) residuals are not white noise. Further evidence against white noise is the fact that the Ljung-Box p-values at $K = 24$ and $K = 48$ are less than 0.001. Consequently, it does not appear that the AR(2) is a satisfactory fit based on an analysis of the residuals.

(ii) In Example 6.8 we found that AIC selected the AR(9) model $\phi_9(B)\left(X_t-78.97\right)=a_t$, where

$$\phi_9(B)=1-1.16B+.41B^2+.13B^3-.10B^4+.07B^5-.01B^6-.02B^7+.05B^8-.22B^9,$$

and $\hat{\sigma}_a^2 = 546.82$. The following code calculates the residuals and places the backcast residuals in **ss9$res**.

```
data(sunspot2.0)
ss9=est.ar.wge(sunspot2.0,p=9)
ljung.wge(ss9$res,p=9)
ljung.wge(ss9$res,p=9,K=48)
```

The backcast residuals in **ss9$res** are plotted in Figure 9.4(c), and these are very similar in appearance to the AR(2) residuals. The sample autocorrelations are shown in Figure 9.4(d) along with the 95% limit lines. The sample autocorrelations stay within the 95% limit lines; however, a few are close to the limits, and the p-values associated with the Ljung-Box test at $K = 24$ and $K = 48$ are 0.0789 and 0.0626, respectively. The plots in Figure 9.4 and the Ljung-Box tests give us some concern about the "whiteness" of the residuals for the AR(9) fit, but we do not "reject white noise" at the $\alpha = .05$ level.

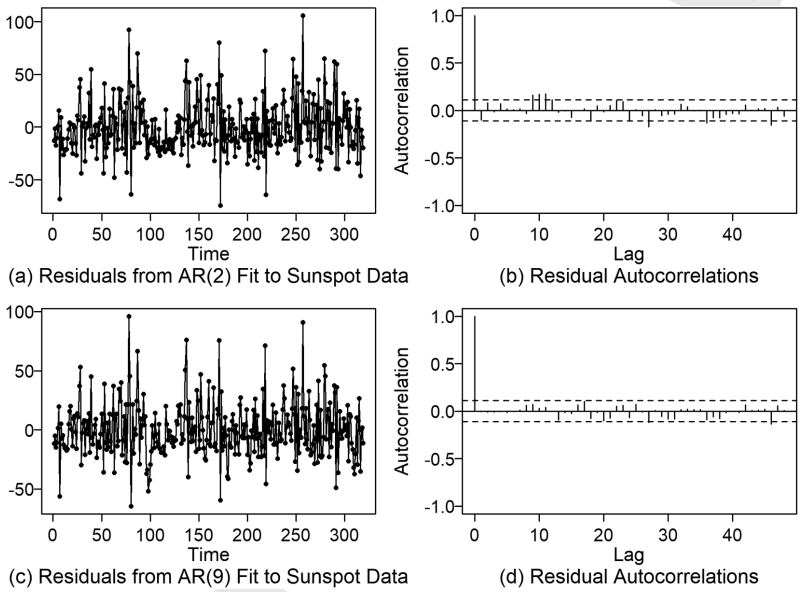

FIGURE 9.4 (a) Residuals from the AR(2) model fit to the sunspot data, (b) sample autocorrelations of the residuals in (a) along with 95% limit lines, (c) residuals from the AR(9) model fit to the sunspot data, and (d) sample autocorrelations of the residuals in (c) along with 95% limit lines.

9.1.2 Checking the Residuals for Normality

The ML estimation procedure assumes that the residuals are normal. Also note that the functions **gen.arma.wge, gen.arima.wge,** and **gen.sigplusnoise.wge** generate realizations based on normal (Gaussian) residuals. For real datasets, you may need to check for normality of the residuals (for example, if you are using ML estimation). Methods for checking normality include the use of histograms, Q-Q plots, or formal tests for normality such as the Shapiro-Wilk test (see Shapiro and Wilk, 1965). In the following we will not examine the AR(2) residuals for normality because the AR(2) model was determined to be a poor fit. The histograms in Figure 9.5 were obtained using the base R command **hist**. Examination of the histograms shows that the sunspot model residuals are somewhat skewed to the right and are the furthest from normality.

Base R provides the Shapiro-Wilk test of the null hypothesis of normality via the command **shapiro.test**. Small p-values are an indication of non-normal data. For the residuals from the AR(4) fit that we named **fit40$res**, the command

```
shapiro.test(fit40$res)
```

yields the p-value 0.415, so that normality is not rejected. This is not surprising because the data in Figure 6.22(a) were generated with normal residuals. The *p*-values obtained by applying the Shapiro-Wilk test to the other three datasets are given below:

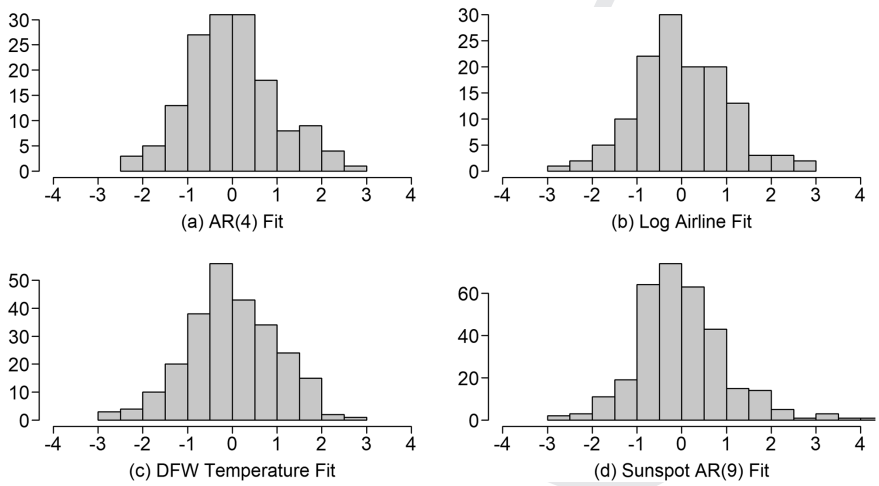

FIGURE 9.5 Histograms of the residuals from the models obtained in Example 9.1: (a) AR(4) fit to data in Figure 6.22, (b) model (9.4) fit to the log airline data, (c) model (9.5) fit to the DFW monthly temperature data, and (d) AR(9) model fit to the sunspot data.

Log airline residuals: **`0.7404`**
`dfw.2000` temperature residuals: **`0.4587`**
Sunspot residuals: **`9.216e-07`**

Consequently, we have reason to doubt the normality of the residuals from the AR(9) fit to the sunspot data.[9] In such cases, because of the normality assumption of the ML estimates, we could use the Burg estimates which make no such assumption. We have previously mentioned that realizations generated (with normal white noise) from the AR(9) sunspot model do not have the asymmetric appearance of the original data (more variability in the heights of the peaks than the troughs).[10] Consequently, nonlinear models (see Tong, 1990) have been proposed for the sunspot data. Another option is to model the log(**`sunspot2.0`**+10) data as we have done in Example 8.6(b). The log sunspot data will also be analyzed in Section 9.3.

9.2 MODELING THE GLOBAL TEMPERATURE DATA

In this section we consider modeling the global temperature data as a case study in which we will fit a variety of models to the data and evaluate them with regard to the *Key Questions for Assessing a Model Fit* given in the introduction to this chapter. The temperature data in Figure 9.6(a) are the global average temperature anomalies from the base period 1901−2000 (ncdc.noaa.gov).

9 Other tests for normality, including the Anderson-Darling, Cramer-von Moses, and Lilliefors (Kolmogorov-Smirnov), can be obtained using functions in the CRAN package **`nortest`**. See Stephens (1974).

10 Even if realizations are generated using a right-skewed noise distribution, the realizations still do not have the dramatic asymmetric appearance observed in the sunspot data.

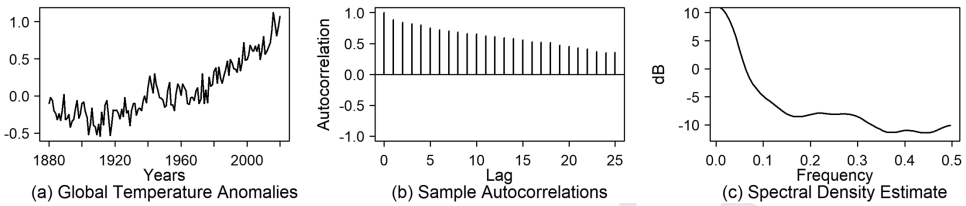

FIGURE 9.6 (a) Global average temperature anomalies for the years 1880–2020 from the base period 1901–2000 (ncdc.noaa.gov), (b) sample autocorrelations of the data in (a), and (c) the Parzen spectral density estimate.

The temperature data in Figure 9.6(a) show that there has been a rise in temperatures over the past 60 years. It is well known using proxy data such as tree rings, ocean sediment data, etc., that temperatures have not remained static over time leading up to the rise over the past 60 years, but have tended to vary. There is some historical (non-thermometer-based) evidence that at least in much of the northern hemisphere, the temperatures during the medieval warm period (950–1100) may have been as high as they are today. The medieval warm period was followed by a period known as the Little Ice Age. While not a true ice age, this was a period during which available temperature data indicate there was cooling that lasted into the nineteenth century.

However, the recent increases in global temperature have led many climatologists to conclude that the warming is likely due to man-made influences, such as increased atmospheric CO_2, and is a cause for concern. Others have argued that much of the increasing temperatures can be attributed to the fact that we are coming out of the Little Ice Age, and the man-made contribution is not as substantial as others might claim.

CAVEAT:

1. The topic of global warming is highly charged and political. The analysis and interpretation of the temperature data is a "tricky" and emotionally charged topic.
2. Our interpretation of the controversy is that climatologists and other scientists believe that the current rise in temperatures is due to two factors:
   - a natural increase as we come out of the "Little Ice Age" period in the 1800s
   - warming due to man-made influences such as increased atmospheric CO_2.

The controversy arises concerning the relative contribution of the two sources.[11]

Please Note:

In this example, we avoid any controversial issues and simply let the data speak for themselves as we demonstrate the challenges involved in finding an appropriate model for the temperature data and the implications of the selected models for forecasting.

9.2.1 A Stationary Model

In this section we will discuss the topic of modeling the global temperature data using a stationary ARMA model. The exponentially damping sample autocorrelations and wandering (upward) behavior of the data cause us to consider an ARMA model, and we can conjecture that the factored form of the model will have a factor $1 - \alpha B$, where α is close to but less than one. The data are most certainly not white noise, so using

11 Our interpretation may be oversimplistic.

the strategies discussed in Chapter 6, we use AIC-type measures to assist in model identification, and estimate the parameters using ML estimates. We use the following ***tswge*** commands.

```
data(global2020)
aic.wge(global2020,p=0:10,q=0:4) #AIC selects p=4,q=1
aic.wge(global2020,p=0:10,q=0:4,type='bic') #BIC selects p=3,q=0
global41.est=est.arma.wge(global2020,p=4,q=1)
```

The ARMA(4,1) model using ML estimates is

$$\left(1-.900B+.102B^2-.077B^3-.116B^4\right)\left(X_t-.072\right)=\left(1-.406B\right)a_t, \tag{9.6}$$

where $\hat{\sigma}_a^2 = 0.018$. The associated factor table is given in Table 9.1.

TABLE 9.1 Factor Table for $\left(1-.900B+.102B^2-.077B^3-.116B^4\right)\left(X_t-.072\right)=\left(1-.406B\right)a_t$,

AR-FACTOR	*ROOTS*	$\lvert r\rvert^{-1}$	f_0
$1-.994B$	1.010	.994	.000
$1-0.289B+.305B^2$	$.473\pm1.748i$	.552	.208
$1+.382B$	−2.618	.382	.500
MA-FACTOR	*ROOTS*	$\lvert r\rvert^{-1}$	f_0
$1-.406B$	2.460	.406	.000

The AR factor $1-.994B$ is close to the nonstationary factor $1-B$, but the MA factor $1-.406B$ cancels some of the near stationarity of the model.

9.2.1.1 Checking the Residuals

The residuals are in vector **`global41.est$res`**, and these are plotted along with sample autocorrelations in Figure 9.7. There we see no strong evidence against white noise. Additionally, the Ljung-Box p-values for $K = 24$ and $K = 48$ are .55 and .37, respectively. These are strongly favorable to a decision to not reject white noise. The Shapiro-Wilk p-value is 0.69 suggesting that a normality assumption is reasonable. The ***tswge*** commands to plot the residuals, their sample autocorrelations with limit lines, and to calculate the Ljung-Box and Shapiro-Wilk tests are

```
plotts.sample.wge(global41.est$res,lag.max=48,arlimits=TRUE)
ljung.wge(global41.est$res,p=4,q=1)
ljung.wge(global41.est$res,p=4,q=1,K=48)
shapiro.test(global41.est$res)
```

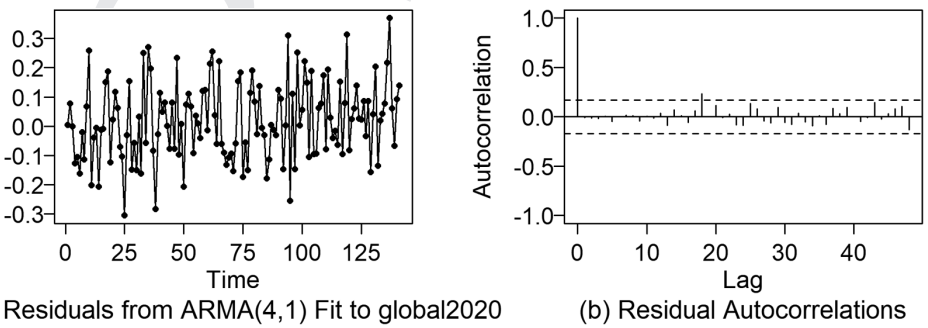

FIGURE 9.7 (a) Residuals from an ARMA(4,1) fit to the global temperature data and (b) sample autocorrelations of the residuals in (a).

9.2.1.2 Realizations and their Characteristics

Figure 9.8(a) shows the global temperature data and Figures 9.8(b)−(f) show five randomly selected realizations of length $n = 141$ from the ARMA(4,1) model in (9.6). The temperature data have a (mostly upward) "wandering or trending behavior along with some high frequency wiggle". The realizations in Figures 9.8(b)−(f) show a variety of behaviors but all could also be described as "wandering or trending behavior along with some high frequency wiggle". Notice that Realization (f) has a strong upward trend while Realization (d) trends down for about 100 time periods and then trends up. However, most of the realizations do not have the primarily monotonic trending behavior of the temperature data in Figure 9.8(a).

Note:
When fitting a correlation-based model (ARMA or ARIMA) to the data, we are tacitly assuming that any observed trends are "random trends" that will eventually abate. The behavior in Figure 9.8(e) is typical in that there is an initial short-lived downward trend, followed by an upward that last from about $t = 20$ to $t = 100$, followed by a short downward trend. These simulations are designed to help assess whether trending behavior in realizations from the ARMA(4,1) model is likely to be as dramatic as that in the actual temperature data. For this reason, we are interested in downward trends as well as upward trends.

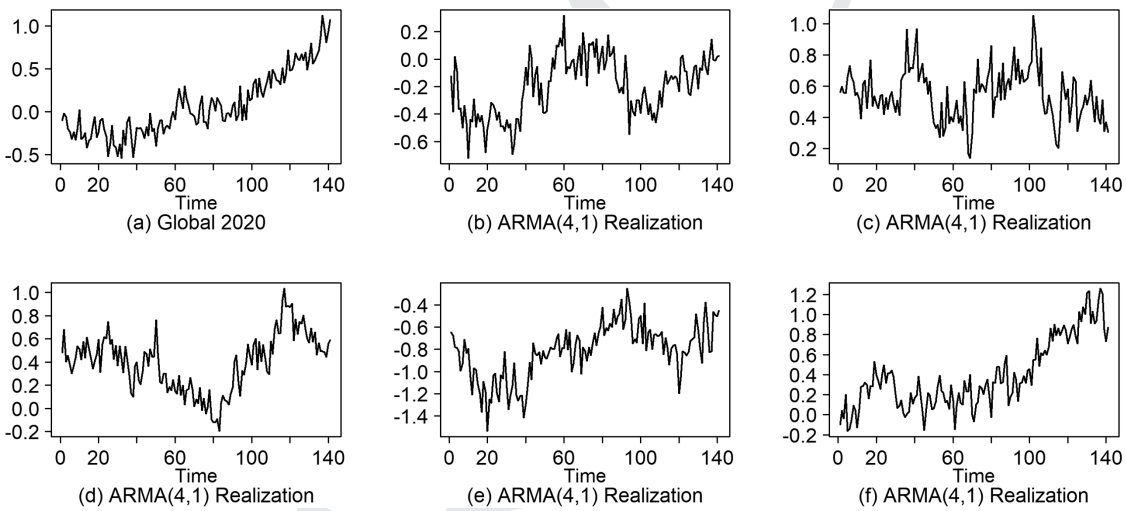

FIGURE 9.8 (a) Global temperature data and (b–f) five realizations of length $n = 141$ from the ARMA(4,1) model in (9.6) which was fit to the global temperature data.

Figure 9.9 compares sample autocorrelations and Parzen spectral density estimates for the global temperature data and the five realizations in Figure 9.8 from the ARMA(4,1) model. Figure 9.9(a) shows that all sample autocorrelations tend to have an exponential-type damping. The sample autocorrelations for the ARMA(4,1) realizations (indicated by open circles connected by solid lines) typically damp more quickly than those for the global temperature data. However, the sample autocorrelations for two realizations are quite similar to those for the actual temperature data. The Parzen spectral density estimate for the global temperature data is shown with the bold curve in Figure 9.9(b), and the other five curves are the Parzen spectral density estimates for the five ARMA(4,1) realizations. The Parzen spectral density estimates all have a peak at zero and are similar to the spectral density estimate for the global temperature data. Based on the sample autocorrelations and Parzen spectral estimates in Figure 9.9, the ARMA(4,1) model seems to provide a reasonable fit.

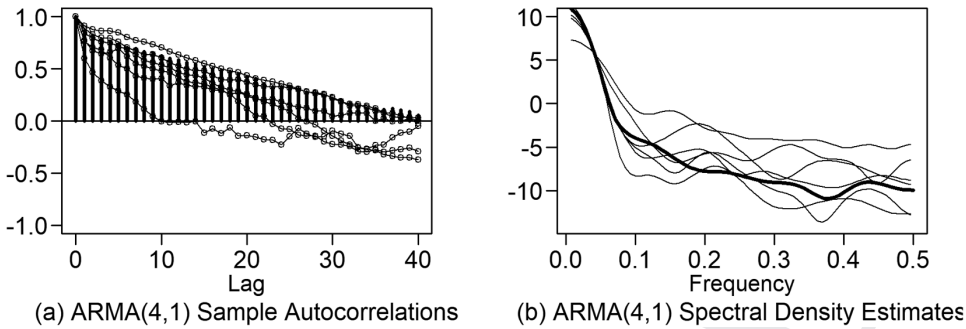

FIGURE 9.9 (a) Sample autocorrelations for the global temperature data (bold vertical bars) along with sample autocorrelations for the five ARMA(4,1) realizations in Figure 9.8. (b) Parzen spectral estimate for global temperature (in bold) along with Parzen spectral density estimates for the five ARMA(4,1) realizations in Figure 9.8.

9.2.1.3 Forecasting Based on the ARMA(4,1) Model

We next use the ARMA(4,1) model to forecast temperatures. Figure 9.10 shows forecasts of the next 50 values for the data in Figure 9.6(a) based on the ARMA(4,1) model.

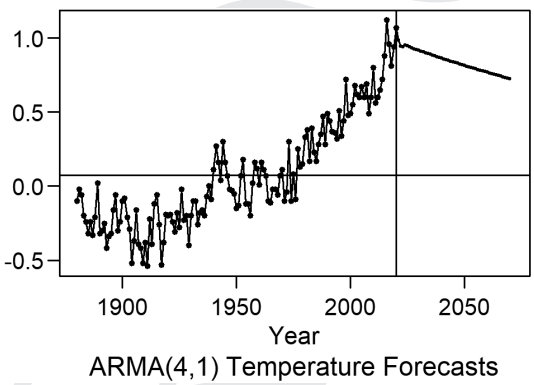

FIGURE 9.10 Forecasts for global temperature for the years 2021–2070 using the stationary ARMA(4,1) model in (9.6).

The forecasts were obtained using the command

```
fore.31=fore.arma.wge(global2020,phi=c(1.185,-.227,.037),theta=.71,
n.ahead=50,limits=FALSE)
```

(1) Generic Discussion of Forecast Performance

We first discuss the forecasts as if the data in Figure 9.6 were simply some generic dataset. Although the recent data values have been increasing, the ARMA(4,1) model in Equation 9.6 is a stationary model which assumes an equilibrium. Under this model, the interpretation is that the values have wandered far above the mean, and we expect the data values to begin decreasing due to the attraction to a mean level. The eventual forecasts tend to a mean level estimated by the sample mean, $\overline{X} = .072$.

(2) Forecasting Results in the Context of the Problem

While the discussion in the previous paragraph was based on a "generic" time series dataset, the data involved are the global temperature anomalies, and these forecasts result in controversial interpretations in relationship to "global warming". First, we note that based on the appearance of the recent temperature data, the forecasts in Figure 9.10 are not very believable. To our knowledge, no one is predicting that temperatures have reached a maximum and will begin a near immediate decline as predicted by the ARMA(4,1) model. Under this assumption, forecasts that immediately trend downward toward some mean level would be inconsistent with the view of most scientists.

Conclusions Concerning the Stationary ARMA(4,1) Fit

1. The ARMA(4,1) model selected by AIC, AICC, and BIC did a good job of whitening the residuals.
2. Some realizations from the ARMA(4,1) model behave similarly to the temperature data, but most realizations lack the sustained trending behavior. Sample autocorrelations tend to damp more rapidly than in the actual temperature data.
3. Forecasts are quite poor. Because of the equilibrium assumption associated with stationary ARMA models, forecasts tend to trend downward toward an overall mean level.

9.2.2 A Correlation-Based Model with a Unit Root

As an alternative to the ARMA(4,1) model fit, we next consider a model with a unit root. The data in Figure 9.6(a) and the slowly damping sample autocorrelations in Figure 9.6(b) are suggestive of a unit root. Additionally, application of the ADF and KPSS tests discussed in Section 7.1.2.1 conclude that there is a unit root. Therefore, despite our cautions about making decisions solely on the basis of a unit root test, we consider a model for the global temperature data with a unit root. That is, we fit an ARIMA(p,d,q) model by first differencing the data to calculate $Y_t = X_t - X_{t-1} = (1-B)X_t$, and then modeling Y_t as an ARMA(p,q) process. The differenced data are shown in Figure 9.11(a) along with the sample autocorrelations of the differenced data in Figure 9.11(b). The differenced data actually appear to be white, but we note that the sample autocorrelation at $k = 1$ is outside the limits. AIC and AICC select an MA(2) model, while BIC selects an MA(1). The MA(1) model is more consistent with Figure 9.11(b) because only the first sample autocorrelation appears to be "strongly nonzero".

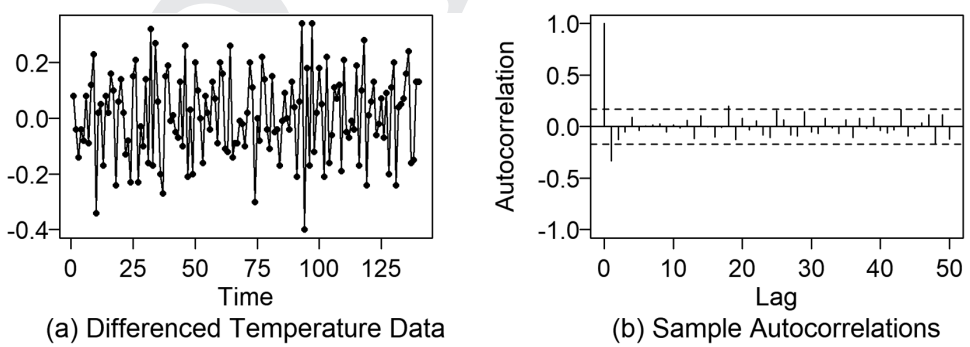

FIGURE 9.11 (a) Differenced global temperature data in Figure 9.6(a) and (b) sample autocorrelations of the data in (a).

The final model is

$$(1-B)X_t = (1-.629B)a_t, \tag{9.7}$$

where $\hat{\sigma}_a^2 = 0.018$.

The following are commands that take the first difference of the **global2020** data and plot the first difference data and sample autocorrelations shown in Figure 9.11. The code also uses BIC to select an MA(1) and obtains the fitted model in (9.7).

```
data(global2020)
y=artrans.wge(global2020,phi.tr=1)
aic.wge(y,p=0:10,q=0:4,type='bic')
# AIC and AICC select an MA(2)while BIC picks MA(1)
# As discussed, we select an MA(1)
y.est=est.arma.wge(y,p=0,q=1)
```

9.2.2.1 Checking the Residuals

Figure 9.12(a) is a plot of the residuals from the MA(1) fit in (9.7) to the differenced data, Y_t. The associated sample autocorrelations are shown in Figure 9.12(b). The residuals appear to be white, and the sample autocorrelations stay sufficiently within the limit lines. The Ljung-Box test (based on $\boldsymbol{p} = 0$ and $\boldsymbol{q} = 1$) reports *p*-values of 0.598 and 0.367 at $\boldsymbol{k} = 24$ and 48, respectively. The Shapiro-Wilk *p*-value is 0.252, giving no reason for concern about non-normality of the residuals. Based on analysis of the residuals, the ARIMA(0,1,1) model in (9.7) appears to be a reasonable fit.

The residual plots, their sample autocorrelations with limit lines, the Ljung-Box test, and Shapiro-Wilk tests can be obtained using the commands

```
plotts.sample.wge(y.est$res,arlimits=TRUE)
ljung.wge(y.est$res,p=0,q=1)
ljung.wge(y.est$res,p=0,q=1,K=48)
shapiro.test(y.est$res)
```

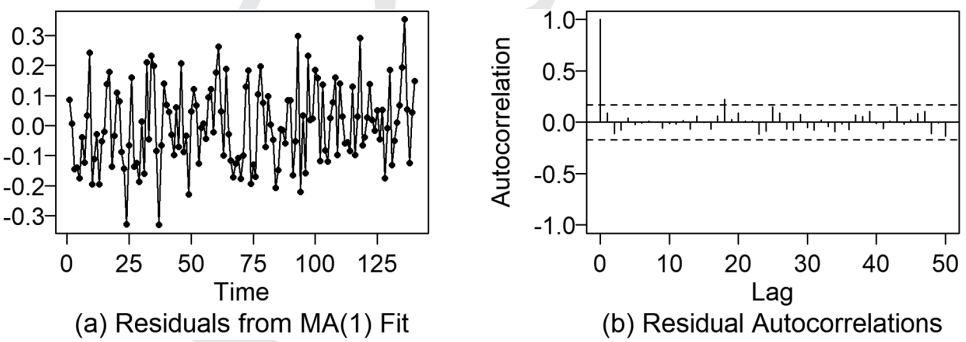

FIGURE 9.12 (a) Residuals from an MA(1) fit to the differenced global temperature data and (b) sample autocorrelations of the residuals in (a).

Key Points:

1. Models (9.6) and (9.7) both whiten the global temperature data.
2. As discussed in Chapter 7, the decision concerning which model to adopt must be made by the investigator.
3. The decision in (2) will have a major impact on forecasts.

9.2.2.2 Realizations and their Characteristics

Similar to the plots in Figure 9.8 of realizations from the stationary ARMA(4,1) model in (9.6), Figure 9.13(a) shows the **`global2020`** data and Figures 9.13(b)–(f) are plots of five randomly selected realizations of length $n = 141$ from the ARIMA(0,1,1) model in (9.7). The realizations in Figure 9.13 are similar to those in Figure 9.8 in that they show a variety of behaviors that could all be described as "wandering or trending" along with some high frequency behavior. While showing somewhat longer "random" trending behavior than the realizations from the ARMA(4,1) model, as a group they do not have trends that hold up as long as the global temperature data.

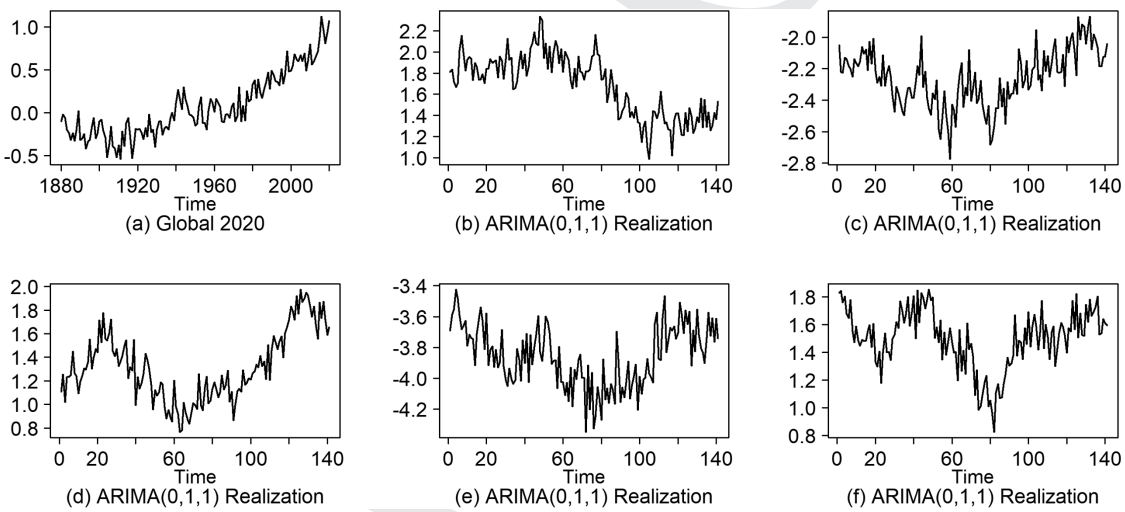

FIGURE 9.13 (a) Global temperature data and (b–f) five realizations of length $n = 141$ from the ARIMA(0,1,1) model in (9.7) which was fit to the global temperature data.

Figure 9.14(a) shows that all sample autocorrelations tend to have an exponential-type damping. The sample autocorrelations for the global temperature data (shown using vertical bars) tend to damp more slowly than most ARIMA(0,1,1) realizations. The Parzen spectral estimates for the global temperature data are shown in bold in Figure 9.14(b). The other five lines represent the Parzen spectral estimates for the five ARIMA(0,1,1) realizations. All simulated realizations have spectral density estimates with a peak at zero and are similar to the spectral density estimate for the temperature data. Based on Figure 9.13, along with the sample autocorrelations and Parzen spectral estimates in Figure 9.14, there is again some cause for concern regarding the ARIMA(0,1,1) model.

Key Points:

1. The ARMA(4,1) and ARIMA(0,1,1) models do not have a "trend component" in the model. However, *some* realizations, e.g. Realization (f) in Figure 9.8 and Realization (b) in Figure 9.13 do show lengthy trends somewhat similar to those in the global temperature data.
2. In general, these "random trends" do not cover as long a time span as the trend observed in the global temperature data.

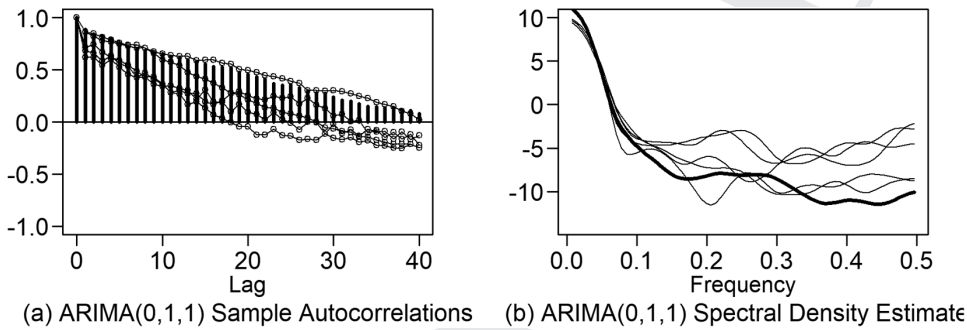

FIGURE 9.14 (a) Sample autocorrelations for the global temperature data (vertical bars) along with sample autocorrelations for the five ARIMA(0,1,1) realizations in Figure 9.13 (b) Parzen spectral estimate for global temperature (in bold) along with Parzen spectral estimates for the five ARIMA(0,1,1) realizations in Figure 9.13.

9.2.2.3 Forecasting Based on ARIMA(0,1,1) Model

We next use the ARIMA(0,1,1) model to forecast the global temperatures. Figure 9.15 shows forecasts of the next 50 values for the data in Figure 9.6(a) based on the ARIMA(0,1,1) model. The forecasts were obtained using the command

```
fore.011=fore.arima.wge(global2020,d=1,phi=0,theta=.661,
n.ahead=50,limits=FALSE)
```

(1) Generic Discussion of Forecasting Results

While the ARMA(4,1) forecasts in Figure 9.10 trended downward toward a mean level, the ARIMA(0,1,1) model in (9.7) is not stationary and, therefore, under this model, the forecasts do not have an attraction to a mean level. The forecasts follow a horizontal line very similar to the last value observed.[12] The ARIMA(0,1,1) does not predict the current trending behavior to continue (nor does it predict that temperatures will decline).

12 The temperature anomaly reading for 2020 is .98 degrees Celsius and the forecasts are a horizontal line at .931 degrees. Recall that forecasts from an ARIMA(0,1,0) model would be equal to the last observed value. However, the moving average term shifts the forecasts slightly.

(2) Forecasting Results in the Context of the Problem

In the context of the problem the forecasts from the ARIMA(0,1,1) model are based on the fact that there is no attraction to a mean level, and the temperatures from 2021 forward are as likely to increase as they are to decrease. This is somewhat in line with those who question global warming although most climatologists believe that there is some "man-made" and some "warming from a colder period" that are impacting global temperatures. This being the case, some continued warming would be expected to persist by most scientists.

Key Points:

1. The forecasts from an ARIMA(0,1,1) fit are the same as those obtained using exponential smoothing (Shumway and Stoffer, 2017).
2. ARMA and ARIMA modeling chooses a model before forecasting, while exponential smoothing always uses the same "model" – an ARIMA(0,1,1).

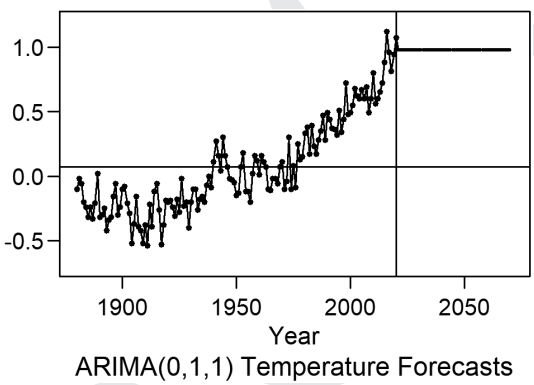

FIGURE 9.15 Forecasts for global temperature for the years 2021–2070 using the stationary ARIMA(0,1,1) model in (9.7).

(3) Is There a Reasonable Correlation-Based Model that Would Predict the Warming to Continue?

Using modeling techniques discussed in Chapters 6–8, we obtained two reasonable correlation-based models which whiten residuals and have characteristics similar to the temperature data. However, the forecasts from these two models are not consistent with "physical expectations". That is, models (9.6) and (9.7) do not forecast the trend to continue, although (9.7) is "silent" on this issue. In Section 7.1.3 it was noted that a model with one unit root does not predict apparent trends to continue, while ARIMA(p,d,q) models with $d=2$ predict trending behavior toward the end of the observed realization to continue.[13] Figure 9.11(a) shows the first-differenced global temperature data. There is *absolutely no indication* of a second unit root, and in fact the differenced data differ very little from white noise.

13 In Example 7.4 we observed that the correlation-based model, $\phi_{13}(B)(1-B)(1-B^{12})X_t = a_t$, fit to the log airline data predicted the observed trend to continue. We noted that the predicted trending is based on the fact that

$1-B$ is a factor of $1-B^{12}$, and, thus, the "airline model" has two factors of $1-B$.

Key Points:

1. Strictly correlation-based models (ARMA and ARIMA) do not predict the warming to continue because there is absolutely no indication of two unit roots.
2. *Do these models provide conclusive evidence that the increasing trend should not be predicted to continue?* The answer to this question is a resounding "No".
3. If there is a "warming signal" then a signal+noise model would be more appropriate than a strictly correlation-based model. This is the topic of the next section.

9.2.3 Line+Noise Models for the Global Temperature Data

Based on the preceding discussion, we next consider a line+noise model for the global temperature data. That is, we consider a model of the form $X_t = s_t + Z_t$ where s_t is the deterministic "warming signal", and Z_t is a zero-mean stationary noise component that we will model as an AR(p). We consider the linear case in which the signal is a line, $s_t = b_0 + b_1 t$. The line+noise (regression) model is given by

$$X_t = b_0 + b_1 t + Z_t. \tag{9.8}$$

Key Point: If there is a "warming signal" it is almost definitely *not linear* because natural trending patterns are not constrained to behave like a straight line or any other particular mathematical curve. We use a line+noise here because it allows us to use the techniques discussed in Section 8.2.3.

The "beauty" of (9.8) is that it allows for the modeling procedures to "decide" whether the appropriate model is purely correlation based (that is, $b = 0$) or whether including the signal (that is using a model with $b \neq 0$) produces a more suitable model.

Recall that **`global2020`** contains 141 annual temperature anomalies from 1880–2020. Using the methods of Section 7.3, we begin to fit the model (9.8) to **`global2020`** using the following ***tswge*** commands:

```
reg=slr.wge(global2020)
t=1:length(global2020)
# residuals from the linear regression line
zhat=global2020-reg$b0hat-reg$b1hat*t
```

$$X_t = -0.4838 + .00782t + \hat{Z}_t \tag{9.9}$$

Figure 9.16(a) is a plot of the **`global2020`** dataset, Figure 9.16(b) shows the temperature data along with the least squares regression line in (9.9), and (c) is a plot of the residuals, $\hat{Z}_t$, from the regression line.

14 t goes from 1 to **`length(global2020)`**. To convert the line in (9.9) to years, the formula becomes $-.4838 + .00782$ (year $- 1879) = -15.1776 + .00782$ * year.

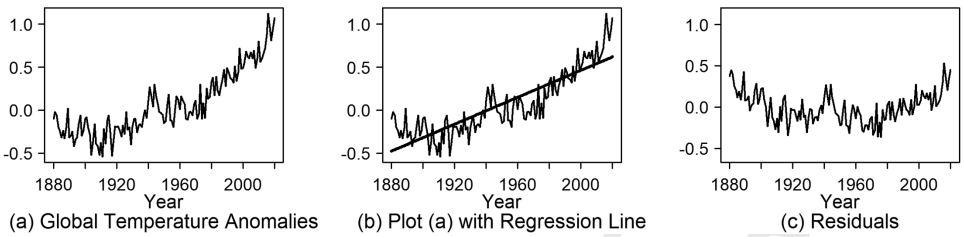

FIGURE 9.16 (a) **global2020** dataset, (b) plot (a) with least squares regression line, and (c) residuals.

The residuals in Figure 9.16(c) are clearly not white noise. Using the command

```
aic.wge(zhat,p=0:10,q=0:0)
```

AIC selects an AR(4) model. Among the output from **aic.wge** are the coefficients of the AR(4) model along with the estimated white noise variance. The fitted model for the residuals is $\hat{\phi}_4(B)\hat{Z}_t = \hat{a}_t$ where

$$\hat{\phi}_4(B) = \left(1 - .476B - .091B^2 - .109B^3 - .196B^4\right), \tag{9.10}$$

and where $\hat{\sigma}_a^2 = .017$.

Thus, the final signal-plus-noise model for the global temperature data, X_t, is

$$X_t = -.4838 + .00782t + \hat{Z}_t. \tag{9.11}$$

where $\hat{\phi}_4(B)\hat{Z}_t = \hat{a}_t$ and $\hat{\sigma}_a^2 = .017$.

(1) Using the Cochrane-Orcutt Method to Assess Significance

Because of the autocorrelation in the residuals shown in Figure 9.16(c) (which we modeled as an AR(4)) we know that it is not appropriate to test $H_0 : b_1 = 0$ using standard linear regression methods. We will use the Cochrane-Orcutt (CO) and the WBG methods discussed in Section 8.2.3 for this purpose.

```
global.co=co.wge(global2020,maxp=10)
global.wgb=wbg.wge(global2020)
```

The p-value for the CO test of $H_0 : b_1 = 0$ is highly significant (less than .0001). Consequently, the CO test indicates that the line in (9.11) should be included in the final model. The WBG test has a p-value about .05[15] also suggesting a significant slope but not as strongly as the CO test.

Key Points: In Equation 9.11 there are two datasets referred to as "residuals".

1. The "residuals" $\hat{Z}_t$ from the regression line are plotted in Figure 9.16(c).
   - These residuals show an autocorrelation structure, and we model them using ML estimates of the AR(4) model selected by AIC.
2. The "residuals", $\hat{a}_t$, plotted in Figure 9.17(a) are the residuals from the AR(4) model fit to the $\hat{Z}_t$ "residuals".
   - These residuals should have the appearance of white noise.

15 The WBG test is based on randomly selected bootstrap replications. As such, the result depends on the seed. Using **sn=0** (random seed) and repeating the test five times, the p-values were .06, .04, .04, .03, .05.

9.2.3.1 Checking the Residuals, $\hat{a}_t$, for White Noise

Using the CO and WBG procedures, we have concluded that the line in (9.9) has a slope that is significantly different from zero. The final line-plus-noise model is shown in (9.11). The residuals, $\hat{a}_t$, in our model fit are found using the ***tswge*** command

```
zhat=est.ar.wge(zhat,p=4)
```

The residuals, **zhat$res,** and sample autocorrelations with limit lines are shown in Figure 9.17. It can be seen that the sample autocorrelation at lag 19 is outside the 95% limits. The Ljung-Box p-values for $K = 24$ and $K = 48$ are .566 and .435, respectively, so there is insufficient evidence to reject white noise. The Shapiro-Wilk p-value is 0.73, which does not reject normality of the residuals. Based on these criteria, the line+noise fit seems plausible. The ***tswge*** commands to plot the residuals, their sample autocorrelations with limit lines, and to calculate the Ljung-Box and Shapiro-Wilk tests are as follows:

```
plotts.sample.wge(zhat$res,lag.max=48,arlimits=TRUE)
ljung.wge(zhat$res,p=4)
ljung.wge(zhat$res,p=4,K=48)
shapiro.test(zhat$res)
```

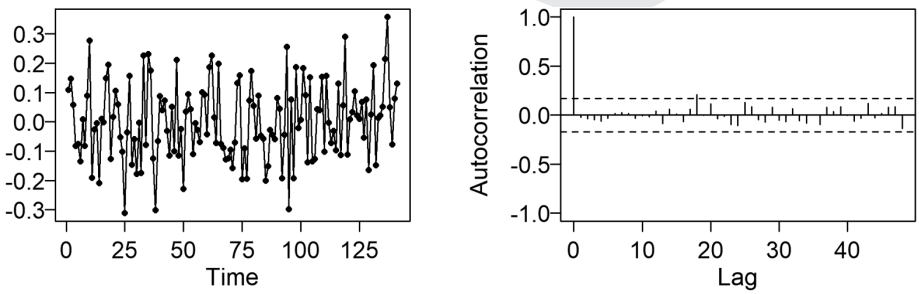

FIGURE 9.17 (a) Residuals from the signal-plus-noise model (9.11) fit to the global temperature data and (b) sample autocorrelations of the residuals in (a).

9.2.3.2 Realizations and their Characteristics

Figure 9.18(a) shows the global2020 data and Figure 9.18(b)−(f) are plots of five randomly selected realizations of length $n = 141$ from the line+noise model in (9.11). As a group, the realizations in Figure 9.18 are more similar to the actual temperature data (shown in Figure 9.18(a)) than are the realizations from the two correlation-based models. Because of the linear trend with positive slope, all realizations tend to increase.

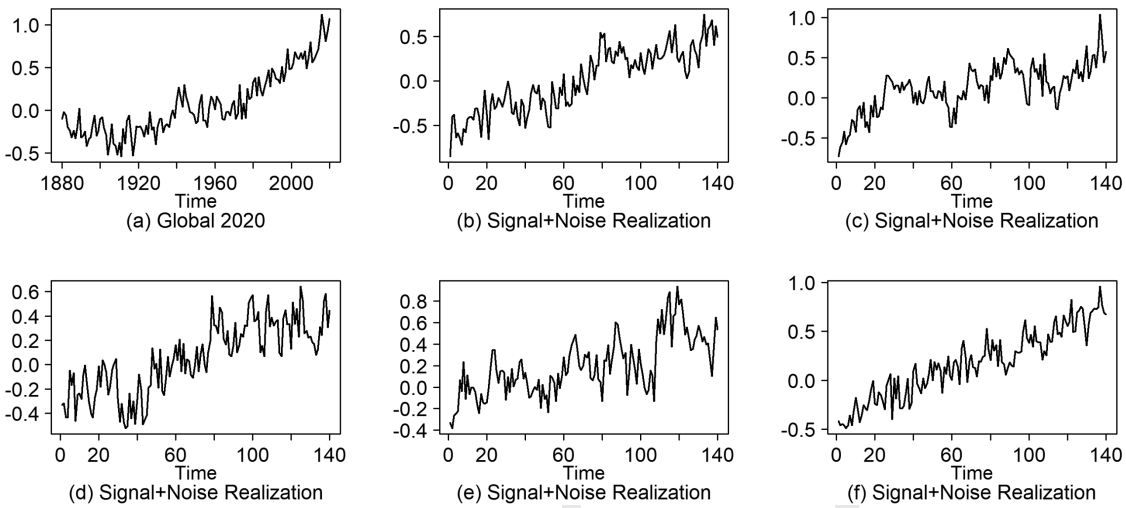

FIGURE 9.18 (a) Global temperature data and (b–f) five realizations of length $n = 141$ from the line+noise model in (9.11) which was fit to the global temperature data.

Figure 9.19(a) plots the sample autocorrelations for the global temperature data (shown using vertical bars) along with the sample autocorrelations shown for the five realizations in Figure 9.18. The sample autocorrelations tend to have an exponential-type damping similar to the sample autocorrelations in Figure 9.6(b) for the actual global temperature data. In fact, three of the five realizations have sample autocorrelations very similar to global temperature sample autocorrelations while the other two damp more quickly. In Figure 9.19(b), the Parzen spectral density estimate for the global temperature data is shown in bold, and the other five curves are the Parzen spectral density estimates for the five signal-plus-noise realizations in Figure 9.18. All spectral estimates have a peak at zero and damp to a similar level by $f = .5$.

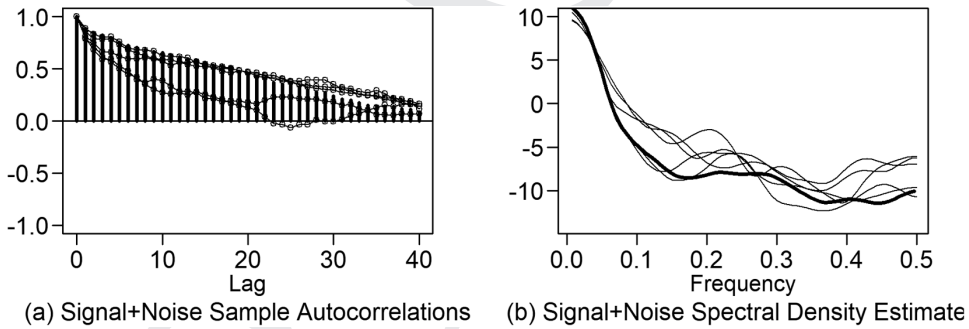

FIGURE 9.19 (a) Sample autocorrelations for the global temperature data (vertical bars) along with sample autocorrelations for the five signal-plus-noise realizations in Figure 9.18. (b) Parzen spectral estimate for global temperature (in bold) along with Parzen spectral estimates for the five signal-plus-noise realizations in Figure 9.18.

9.2.3.3 Forecasting Based on the Signal-plus-Noise Model

Figure 9.20 shows forecasts of the temperature data for the next 50 years using the signal-plus-noise model in (9.11).

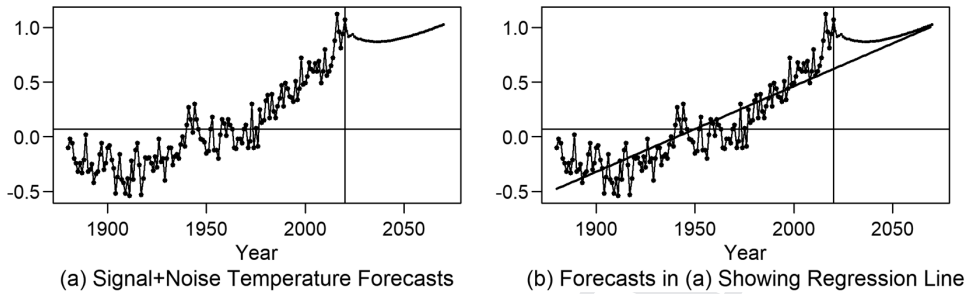

FIGURE 9.20 (a) Signal-plus-noise forecasts for global temperature for the years 2021–2070 and (b) Forecasts in (a) showing fitted regression line.

(1) Generic Discussion of Forecast Performance
The forecasts predict a "near-term" decline with the eventual forecasts following the fitted regression line. The AR(4) model fit to the residuals from the regression line affects the near-term forecasts. However, the eventual forecasts follow the fitted regression line and are not affected by the AR(4) model.

(2) Forecasting Results in the Context of the Problem
As can be seen in Figures 9.20(a) and (b), the forecasts indicate a short decline in temperatures followed by an increase that is at a slower rate than that experienced in the last 45 years. To understand this, note that from 1880−1910 temperatures had a mild decline, followed by a general increase in temperatures from 1910−1940, and this was followed by another mild decline from about 1940−1975. Temperatures have climbed at a relatively high rate after 1975. Consequently, the slope of the regression line fit to the entire dataset is not as steep as the temperature increase for about the last 45 years. The initial dip in the forecasts is reflective of the fact that the forecasts move toward the regression line, which is not as steep as the recent temperature trending behavior.

The following Key Points should be kept in mind.

Key Points Concerning the Line+Noise Fit in (9.11):
1. Realizations from this model show a strong resemblance to the actual temperature data in Figure 9.6(a), and as a whole are more similar to the actual data than are realizations from the strictly correlation-based models.
2. Sample autocorrelations and spectral densities calculated from the realizations are similar to those for the temperature data and are mildly better than those based on realizations from the correlation-based models considered.
3. Forecasts call for an increase in temperatures following a short drop.

This leads to the following.

Conclusion: The line+noise model appears to provide a better fit to the global temperature, and thus, based on our analysis of the data, there is reason to believe that temperatures should be predicted to continue to increase.

Before leaving this discussion, we plot the forecasts from the three models along with their 95% limits in Figure 9.21. There it can be seen that the ARMA(4,1) limits would "accommodate" slight increases in temperature as plausible, but these forecasts indicate a decline in temperatures and it appears to be the

poorest of the three models. The forecasts for the ARIMA(0,1,1) model in Figure 9.21(b) "allow for the fact" that the temperatures may go up or down, and the limits for the eventual forecasts at, say, 50 steps ahead are higher than those for the signal-plus-noise forecast limits in Figure 9.12(c). One thing to notice about the signal-plus-noise forecasts is that the limits tend to be "tight" (even at longer steps ahead). That is, the uncertainty in the forecasts is smaller *because we fit a deterministic signal as a part of the model*. Forecasting (especially long term) from the linear signal-plus-noise model is an example of *extrapolation* which every good data analyst knows to view with caution. The "small uncertainty" may not be justified because the validity of forecasts depends on whether the process driving temperatures continues into the future. Although the ARIMA(0,1,1) forecasts seem to be poor, these forecasts do "admit" to uncertainty and allow for increasing or decreasing temperatures as being plausible.

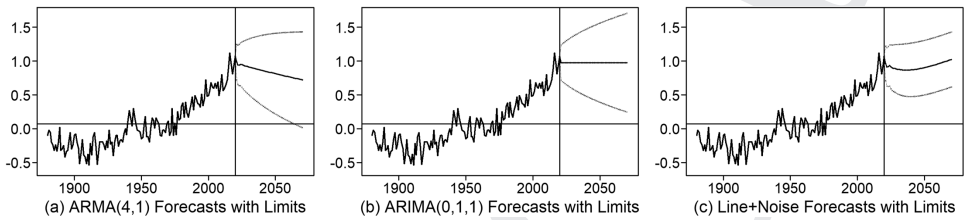

FIGURE 9.21 Forecasts for the years 2021 – 2070 based on (a) the stationary ARMA(4,1) model, (b) the ARIMA(0,1,1) model, and (c) the signal-plus-noise model.

Comments Concerning Temperature Models:

1. Again, if there is a "warming signal", it is almost certain that it is not linear.
2. Beware of extrapolation issues when forecasting with a signal-plus-noise model.
3. Although the ARMA(4,1) and ARIMA(0,1,1) models did not contain a linear trend component, Realization (e) from the ARMA(4,1) model and two of the five realizations from the ARIMA(0,1,1) model had "significant" linear trends according to the CO and WBG tests.
   - Recall that the Cochrane-Orcutt test has inflated significance levels (Section 7.3).
   - The ARIMA(0,1,1) model is "capable" of generating realizations similar to the temperature data and should not be dismissed as a possible model.

9.2.3.4 Other Forecasts

There may be some legitimate other possibilities for forecasting:

1. Holt-Winters
2. Only consider data from 1920 on because that is a better time frame for the beginning of carbon emissions. The line+noise forecasts tend toward the fitted regression line in Figure 9.21. Because there was no tendency for warming in the early part of the 20th century, the regression line is "flatter" than the current visible temperature increase.

We consider these briefly here.

1. Figure 9.22 shows the Holt-Winters forecasts using a trend but not a seasonal component (solid line) and the line+noise forecasts previously shown in Figures 9.20 and 9.21(c). The commands

for the Holt-Winters forecasts follow. The commands to produce the line+noise forecasts are given above.

```
data(global2000)
x.hw=HoltWinters(global2020,gamma=FALSE)
x.pred=predict(x.hw,n.ahead=50)
plot(x.hw,x.pred,lty=1:2)
```

Basically, the trend over the last 10−20 years is extended forward. Quite frankly, the Holt-Winters forecasts are frightening! They stand in stark contrast to the more conservative forecasts using the line+noise model.

2. Considering only global temperature data beginning in 1920: Figure 9.22(b) shows the line+noise forecasts (dotted line) and the Holt-Winters forecasts with a trend (solid line) based only on the global temperature data from 1920 through 2020. These forecasts were simply made using the same commands as for Figure 9.22(a) but are based on the dataset **`global.1=window(global2020,start=1920)`**. In this case, the Holt-Winters and line+noise forecasts are quite similar.

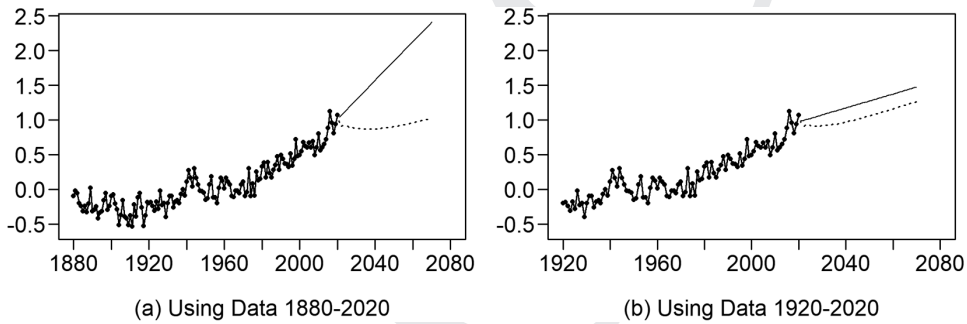

FIGURE 9.22 Holt-Winters forecasts (solid line) and line+noise forecasts (dotted line) (a) based on the data from 1880−2020 and (b) based on the data from 1920−2020.

Let's be honest:

A *problem with forecasts is that they depend on so many factors that can be controlled by the investigator.*

- If for example the investigator wanted to "scare folks" about global warming, then obviously the Holt-Winters forecasts based on the 1880−2020 dataset are the forecasts to use.
- The other forecasts could be used to promote various levels of concern about the global warming issue.

Forecasting Advice: In addition to the comments made regarding forecasts in the Key Points above, we offer the following:

(1) Realize that the model you choose will result in a certain type of forecast, so *carefully choose the type of model* to be used.

Note:
The decision about the "type" of model is more important than the decision to use the AIC or BIC selection of model orders.

(2) It is useful to examine various forecast models before deciding which forecasts to use.
(3) Validate forecasts using techniques described in this book
   - forecast limits
   - RMSE and rolling window RMSE are very useful tools

(4) Don't ever provide a forecast value as "the correct answer".
   - Be sure decision makers are aware of the uncertainties involved and their implications.

(5) As a "consumer" of forecasts, it's important to know "who (or what organization)" produced the forecasts and what assumptions were made.

QR 9.2 Recap

9.3 COMPARING MODELS FOR THE SUNSPOT DATA

In this section we return to the problem of modeling the sunspot data contained in the ***tswge*** dataset **`sunspot2.0`**. As we did with the models for the global temperature data, in this case study we will compare models proposed for the sunspot data based on the three questions posed in the introduction to this chapter:

(1) *Does the model "whiten" the residuals?*
(2) *Do realizations and their characteristics behave like the data?*
(3) *Do forecasts reflect what is known about the physical setting?*

In Chapter 6 we noted two models that have been proposed for the sunspot data: an AR(2) and an AR(9) model. In this case study we will investigate steps involved in a thorough analysis of the data and models selected. Item (2) above is problematic for the sunspot data because we know that realizations from an AR or ARMA model are not going to have the same asymmetric appearance. In Example 8.6(b) we fit a cosine signal+noise model to the data.[16]

```
logss=log(sunspot2.0+10)
```

For convenience we refer to the data calculated above as the *log sunspot data*. Figure 9.23 was previously shown as Figure 8.15, but we repeat it here for completeness. The similarity between the sample

16 Recall that the +10 is added to the sunspot numbers because there were instances of zero sunspots, and adding 10 before taking logs produced symmetric behavior similar to that seen in ARMA realizations. The number 10 was arbitrarily chosen for purpose of illustration. However, adding 1 to each sunspot number did not fully remove the asymmetry

autocorrelations and spectral density estimates of the raw and log sunspot data is striking! In fact, you must look closely to see *any* differences.

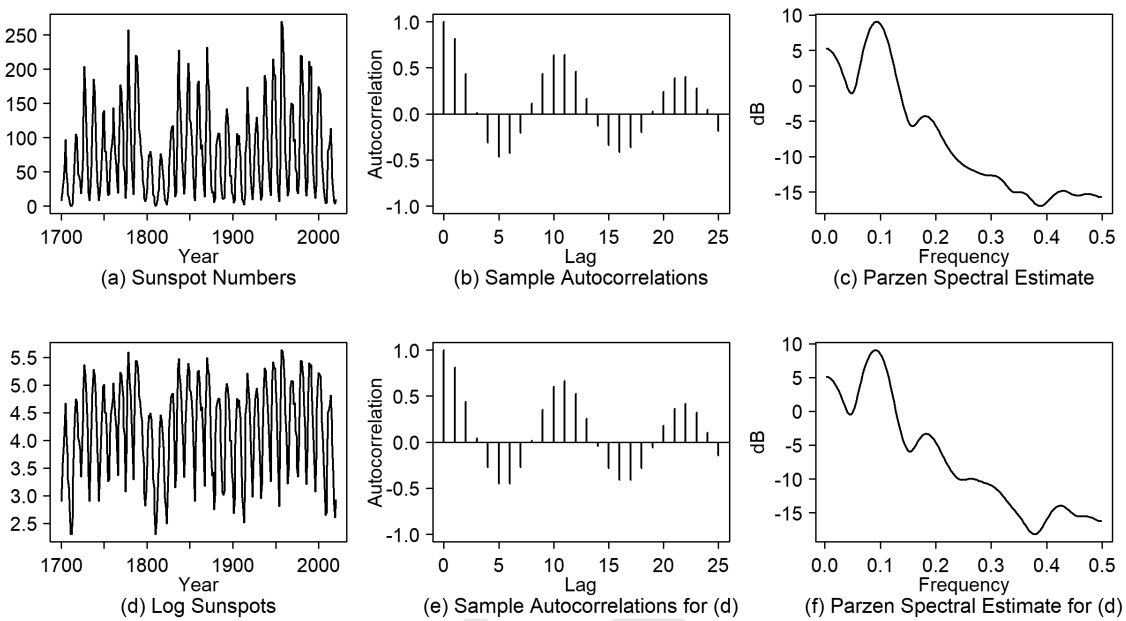

FIGURE 9.23 (a) Sunspot2.0 data, (b) sample autocorrelations and (c) Parzen spectral density estimate of sunspot data in (a); (d) log(sunspot2.0+10), (e) sample autocorrelations and (f) Parzen spectral density estimate of the log sunspot data in (d).

9.3.1 Selecting the Models for Comparison

AR(2) and AR(9) models have been proposed for the "raw" sunspot data, but we will be comparing models for the log sunspot data. Using the **`aic5.wge`** command below

```
aic.wge(logss,p=0:12,q=0:4)
```

AIC selects an AR(9) model. Using AIC-type tools there is very little support for an AR(2). If you restrict the search to AR models and let p range from 0 to 9, BIC selects an AR(2) as the third choice. Box, Jenkins, and Reinsel (2008) choose an AR(2) (for a different version of the sunspot data) on the basis of the partial autocorrelation function. The ***tswge*** command

```
pacf.wge(logss)
```

produces the plot in Figure 9.24 which shows two large sample partial autocorrelations suggesting an AR(2).[17]

17 Not to be missed is the fact that the pacf at lags 6 through 9 are also outside the limits, suggesting an AR(9).

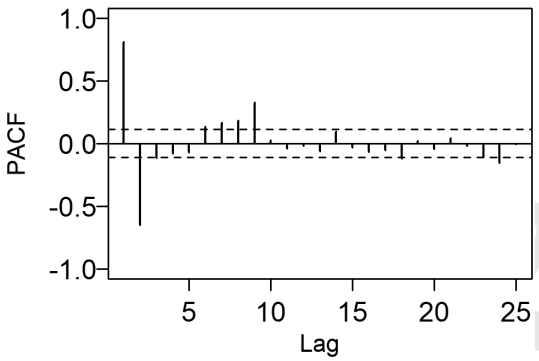

FIGURE 9.24 pacf output for log sunspot data in Figure 9.23(d).

We will use the AR(2) and AR(9) fits for the log sunspot data comparison.

9.3.2 Do the Models Whiten the Residuals?

In Example 9.1 we analyzed the AR(2) and AR(9) models for the raw sunspot data and saw that the AR(9) whitened the residuals while the AR(2) did not. Because we will be comparing models for the log sunspot data, we will repeat the analysis. The following commands estimate the parameters and plot the residuals and their sample autocorrelations for the AR(2) and AR(9) models.

```
data(sunspot2.0)
logss=log(sunspot2.0+10)
ss2=est.ar.wge(logss,p=2)
ss9=est.ar.wge(logss,p=9)
```

The estimated AR(2) model is

$$\left(1-1.37B+.67B^2\right)\left(X_t-4.20\right)=a_t, \tag{9.12}$$

with $\hat{\sigma}_a^2=.117$. The AR(9) model is

$$\left(1-1.18B+.51B^2+.01B^3-.13B^4+.16B^5-.01B^6-.16B^7+.26B^8-.36B^9\right)\left(X_t-4.20\right)=a_t \tag{9.13}$$

where $\hat{\sigma}_a^2=.090$. We leave it as an exercise to find factor tables for AR(2) and AR(9) models for the log data and compare them with the corresponding models for the raw sunspot data.

The residuals and their sample autocorrelations are plotted in Figure 9.25. The AR(2) residuals have a "waviness" that doesn't look "white" and several sample autocorrelations are outside the limit lines. The Ljung-Box tests were obtained using the commands

```
ljung.wge(ss2$res,p=2)
ljung.wge(ss2$res,p=2,K=48)
```

and in each case $p<.001$ indicating rejection of the null hypothesis of white noise.

The AR(9) residuals are shown in Figure 9.25(c). These residuals look like white noise and the sample autocorrelations stay within the limit lines. The Ljung-Box tests were obtained using the commands

```
ljung.wge(ss9$res,p=9)
ljung.wge(ss9$res,p=9,K=48)
```

and the p-values were .10 and .53 for K=24 and 48, respectively. Consequently, the residuals appear to be white.

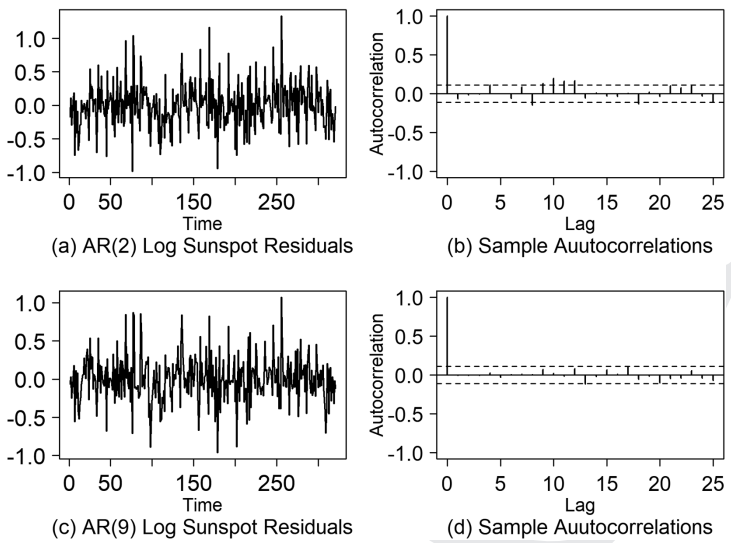

FIGURE 9.25 (a) Residuals from AR(2) fit to log sunspot data, (b) sample autocorrelations of data in (a), (c) residuals from AR(9) fit to log sunspot data, (d) sample autocorrelations of data in (c).

9.3.3 Do Realizations and Their Characteristics Behave Like the Data?

Figures 9.26(a) and 9.27(a) are plots of the log sunspot data. Figures 9.26(b–f) are realizations from the AR(2) model in (9.12) while Figures 9.27(b–f) show five realizations from the AR(9) model in (9.13). The AR(2) model is much more erratic than the log sunspot data and the 10–11 year cycle is barely visible. The AR(9) realizations do a much better job of resembling the log sunspot data. While the behavior is more erratic than the log sunspot data, the 10–11 years cycles are clear.

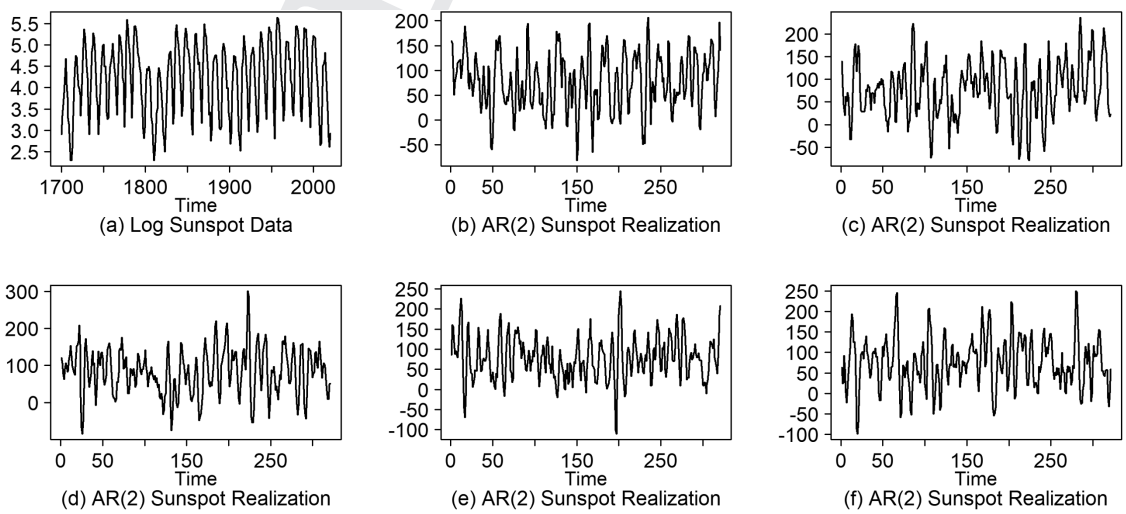

FIGURE 9.26 (a) Log sunspot data and (b–f) realizations from the AR(2) model in (9.13) fit to the log sunspot data.

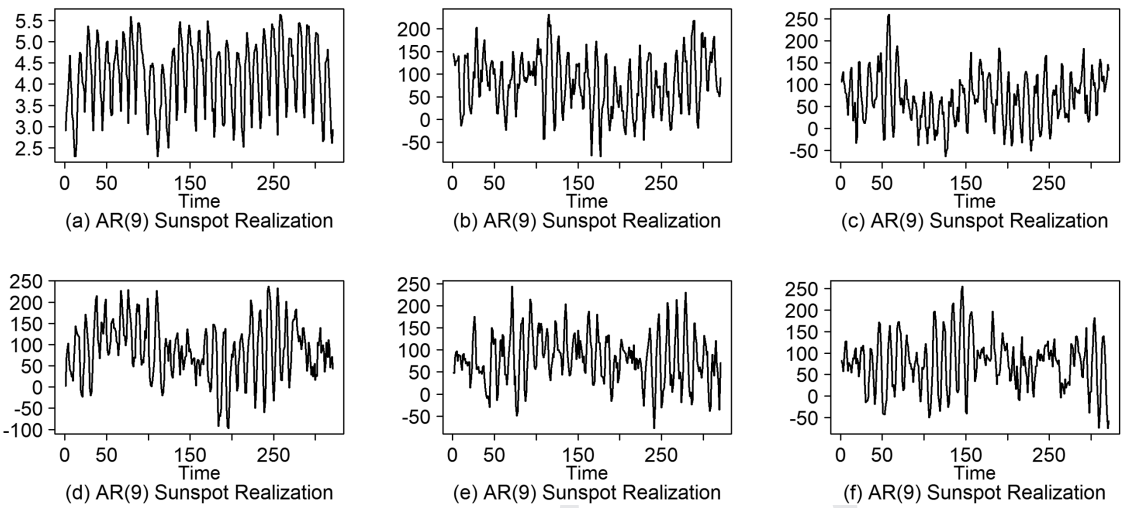

FIGURE 9.27 (a) Log sunspot data and (b–f) realizations from the AR(9) model in (9.13) fit to the log sunspot data.

Figure 9.28(a) shows the sample autocorrelations for the log sunspot data (shown using vertical bars) along with the sample autocorrelations for the five realizations in Figure 9.26 from the AR(2) model. The sample autocorrelations for the realizations have a damped sinusoidal behavior that tend to damp much more quickly than the log sunspot sample autocorrelations. Also, after the first cycle, the sample autocorrelations from the AR(2) model tend to "get off cycle". This is consistent with the lack of distinct cyclic behavior in the AR(2) realizations. The sample autocorrelations in Figure 9.28(b) for the AR(9) realizations in Figures 9.27(b–f) are very similar in appearance to those of the actual log sunspot data.

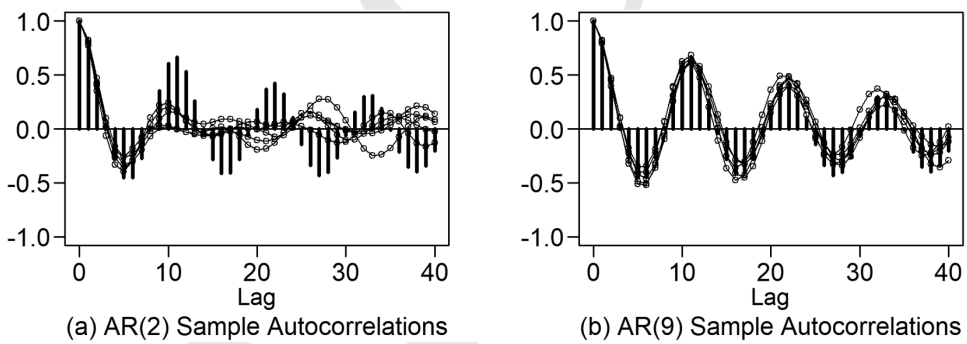

FIGURE 9.28 (a) Sample autocorrelations for log sunspot data (bold vertical bars) and for the five realizations in Figure 9.26 and (b) Sample autocorrelations for actual log sunspot data (bold vertical bars) and for the five realizations in Figure 9.27.

Figure 9.29(a) shows Parzen spectral density estimates for the sunspot data (bold) and for the AR(2) realizations in Figure 9.26. The spectral densities for the AR(2) realizations have peaks at about $f = .1$ that are not as sharp as the $f = .1$ peak for the actual log sunspot data. Noticeably, there is no tendency for the AR(2) spectral density estimates to have a peak at $f = 0$. The spectral density estimates in Figure 9.29(b) for the AR(9) realizations are quite consistent with the spectral density for the sunspot data since they show distinct peaks at $f = 0$ and $f = 0.1$.

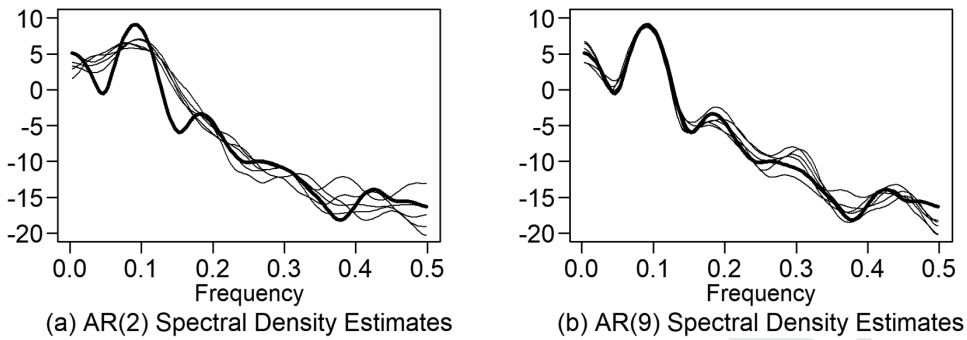

FIGURE 9.29 (a) Parzen spectral estimate for sunspot data (bold) and for the five AR(2) realizations in Figure 9.26 and (b) Parzen spectral estimate for sunspot data (bold) and for the five AR(9) realizations in Figure 9.27.

9.3.4 Do Forecasts Reflect What Is Known about the Physical Setting?

Example 6.15 went into detail concerning the forecasts from the AR(9) and AR(2) models fit to the raw sunspot data. Specifically, the AR(2) forecasts tended to damp very quickly and simply forecast the mean value for a moderate number of steps ahead. On the other hand, the AR(9) forecasts predicted the cyclic behavior to continue, but because of the fact that the sunspot data does not have fixed cycle lengths, forecasts tend to get off cycle. The behavior of forecasts for the log sunspot data will be similar. The cycles are the same in both datasets. Taking logarithms simply took out the asymmetric behavior, but did not alter the cycle lengths.

Conclusion: The AR(2) is a poor model

Although the partial autocorrelations for the sunspot data suggest an AR(2) model:

1. The residuals are not white noise.
2. Realizations from the AR(2) model do not have consistent 10–11 year cyclic behavior, and have sample autocorrelations and spectral estimates that are quite different from the corresponding quantities for the actual log sunspot data.
3. Forecasts do not predict a stable cyclic behavior.

9.3.4.1 Final Comments about the Models Fit to the Sunspot Data

(a) As mentioned previously, no AR (or ARMA) model will be able to sufficiently account for the asymmetric behavior in the sunspot data. The AR(9) model provides a reasonably good fit with this one exception. Consequently, for this example we used the log sunspot data.

(b) The Parzen spectral density estimate strongly indicates the existence of two peaks, an obvious one at about $f = .10$ which accounts for the strong 10–11 year cycle in the data, but also a distinct peak at $f = 0$. The factor table for the AR(9) model fit to the log sunspot data is very similar to the one in Table 5.6 for the raw sunspot data, which shows a system frequency of $f = 0$. The AR(2) model only accounts for the 10-11 year cycle.

Key Points:

1. The fact that the Parzen spectral density estimate had strong peaks at $f = 0$ and at about $f = .10$ ***should have told us immediately that an AR(2) model was insufficient.***
2. In order for a model to account for a peak at $f = 0$ and $f = .10$, it ***must*** be at least of order $p = 3$.
   - Recall that the spectral density for an AR(2) model can have peaks at $f = 0$ and/or $f = .5$, but if it has a peak associated with $0 < f < .5$, then this is the only spectral peak associated with the model. See Section 5.1.3

QR 9.3 Recap and Additional Information

9.4 COMPREHENSIVE ANALYSIS OF TIME SERIES DATA: A SUMMARY

In this chapter, we have discussed issues involved in an analysis of actual time series data. To summarize, given a realization to be modeled, a comprehensive analysis should involve:

1. Examination of the data before fitting a model
   - is it white noise?
2. Obtaining a model
   - correlation-based or signal+noise
   - identifying *p, q, d,* and seasonal components
   - ARCH/GARCH
   - estimating parameters of the signal and of the model for the residuals
3. Checking for model appropriateness
   - checking residuals
   - examining realizations and their characteristics
   - obtaining forecasts, spectral estimates, etc., as dictated by the situation

9.5 CONCLUDING REMARKS

This short (but hopefully rewarding!) chapter has neatly summarized several of the key time series analysis techniques presented in earlier chapters. This chapter alone can be a convenient reference for you when you need to quickly review the procedures we recommend for analyzing time series data.

One of the diagnostics for checking the adequacy of a candidate model is the analysis of the residuals. For example, an appropriate time series model should have residuals that approximate white noise. Another attribute of a good model is that realizations generated from the model should have the overall

characteristics of the original data realization. Finally, forecasts given by the model should "make sense" and be consistent with what is known about the physical setting of the problem.

A variety of examples are provided, and detailed solutions are presented. Interesting comparisons and contrasts are made between various competing models in modeling the popular global temperature data.

It is often the case that you will want to incorporate more than one variable when computing forecasts. This is analogous to the use of several explanatory variables in multiple regression. Such "multivariate" methods will be the topic of Chapters 10 and 11.

APPENDIX 9A

TSWGE FUNCTIONS

The only new ***tswge*** function introduced in this chapter is **`ljung.wge`** which performs the Ljung-Box test on a set of data to test for white noise.

`ljung.wge(x,K,p,q)` performs the Ljung-Box test on the data in vector **`x`**.
`x` is a vector containing the realization to be analyzed for white noise.
`K` is the maximum number of lags to be used in the Ljung–Box test statistic formula in Equation 9.3. Box, Jenkins, and Reinsel (2008) recommend running the test using **`K = 24`** (default) and **`K = 48`**. Other authors give a variety of recommendations.
`p` and **`q`**: If the data are residuals from an ARMA(p,q) model fit to a set of data, then **`p`** and **`q`** specify the model orders of the fit. If **`x`** is simply a dataset which you want to test for white noise, then use the defaults **`p = 0`** and **`q = 0`**.

Output:

`$K` is a reminder of the value of K specified in the input
`$chi.square` is the test statistic calculated using (9.3)
`$df` is the degrees of freedom associated with the test statistic
`$pval` is the p-value of the test $H_0 : \rho_1 = \rho_2 = \cdots \rho_K = 0$.

Base R Function

`shapiro.test(x)` performs a Shapiro-Wilk test of the null hypothesis that the data in **`x`** are normal.

PROBLEMS

9.1 Consider the data generated using the command

```
ar4=gen.arma.wge(n=100,phi=c(2.76,-3.76,2.6,-.89,sn=463)
```

Find the AR(4) model selected by AIC using ML estimates. Examine the appropriateness of the fitted model using the outline below:

(a) Check the whiteness of the residuals.
(b) Examine the appropriateness of the model in terms of forecasting performance by determining how well the model forecasts the last ? steps for whatever value or values you decide ? should be.
(c) Generate realizations from the fitted model to determine whether realizations have similar appearance and characteristics (sample autocorrelations and Parzen spectral densities) as the original realization.

9.2 Seasonal model B was generated using the command

```
xB=gen.arima.wge(n=100,phi=c(1.3,-.65)ms=4,sn=290)
```

The model fit to the data is given in (7.15). Use the outline in Problem 9.1 to examine the appropriateness of the fitted model.

9.3 For each model below use the outline in Problem 9.1 to assess model appropriateness.

(a) The AR(12) model fit to the log lynx data
(b) The AR(2) model fit to the log lynx data
(c) The ARMA(2,3) model fit to the log lynx data

9.4 In Section 9.1.2 we noted the Shapiro-Wilk test rejected normality for the residuals from the AR(9) fit to the sunspot data. Check the residuals of the AR(9) fit to the log sunspot data for normality using the Shapiro-Wilk test. Also, plot a histogram of the residuals and compare it with Figure 9.5(d).