2

Time Series Regression and Exploratory Data Analysis

2.1 Introduction

The linear model and its applications are at least as dominant in the time series context as in classical statistics. Regression models are important for time domain models discussed in Chapters 3, 5, and 6, and in the frequency domain models considered in Chapters 4 and 7. The primary ideas depend on being able to express a response series, say x_t, as a linear combination of inputs, say $z_{t1}, z_{t2}, \ldots, z_{tq}$. Estimating the coefficients $\beta_1, \beta_2, \ldots, \beta_q$ in the linear combinations by least squares provides a method for modeling x_t in terms of the inputs.

In the time domain applications of Chapter 3, for example, we will express x_t as a linear combination of previous values $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, of the currently observed series. The outputs x_t may also depend on lagged values of another series, say $y_{t-1}, y_{t-2}, \ldots, y_{t-q}$, that have influence. It is easy to see that forecasting becomes an option when prediction models can be formulated in this form. Time series smoothing and filtering can be expressed in terms of local regression models. Polynomials and regression splines also provide important techniques for smoothing.

If one admits sines and cosines as inputs, the frequency domain ideas that lead to the periodogram and spectrum of Chapter 4 follow from a regression model. Extensions to filters of infinite extent can be handled using regression in the frequency domain. In particular, many regression problems in the frequency domain can be carried out as a function of the periodic components of the input and output series, providing useful scientific intuition into fields like acoustics, oceanographics, engineering, biomedicine, and geophysics.

The above considerations motivate us to include a separate chapter on regression and some of its applications that is written on an elementary level and is formulated in terms of time series. The assumption of linearity, stationarity, and homogeneity of variances over time is critical in the regression

context, and therefore we include some material on transformations and other techniques useful in exploratory data analysis.

2.2 Classical Regression in the Time Series Context

We begin our discussion of linear regression in the time series context by assuming some output or dependent time series, say, x_t, for $t = 1, \ldots, n$, is being influenced by a collection of possible inputs or independent series, say, $z_{t1}, z_{t2}, \ldots, z_{tq}$, where we first regard the inputs as fixed and known. This assumption, necessary for applying conventional linear regression, will be relaxed later on. We express this relation through the linear regression model

$$x_t = \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t, \tag{2.1}$$

where $\beta_1, \beta_2, \ldots, \beta_q$ are unknown fixed regression coefficients, and $\{w_t\}$ is a random error or noise process consisting of independent and identically distributed (iid) normal variables with mean zero and variance σ_w^2; we will relax the iid assumption later. A more general setting within which to embed mean square estimation and linear regression is given in Appendix B, where we introduce Hilbert spaces and the Projection Theorem.

Example 2.1 Estimating a Linear Trend

Consider the global temperature data, say x_t, shown in Figure 1.2 and Figure 2.1. As discussed in Example 1.2, there is an apparent upward trend in the series that has been used to argue the global warming hypothesis. We might use simple linear regression to estimate that trend by fitting the model

$$x_t = \beta_1 + \beta_2 t + w_t, \quad t = 1880, 1857, \ldots, 2009.$$

This is in the form of the regression model (2.1) when we make the identification $q = 2$, $z_{t1} = 1$ and $z_{t2} = t$. Note that we are making the assumption that the errors, w_t, are an iid normal sequence, which may not be true. We will address this problem further in §2.3; the problem of autocorrelated errors is discussed in detail in §5.5. Also note that we could have used, for example, $t = 1, \ldots, 130$, without affecting the interpretation of the slope coefficient, β_2; only the intercept, β_1, would be affected.

Using simple linear regression, we obtained the estimated coefficients $\widehat{\beta}_1 = -11.2$, and $\widehat{\beta}_2 = .006$ (with a standard error of .0003) yielding a highly significant estimated increase of .6 degrees centigrade per 100 years. We discuss the precise way in which the solution was accomplished after the example. Finally, Figure 2.1 shows the global temperature data, say x_t, with the estimated trend, say $\widehat{x}_t = -11.2 + .006t$, superimposed. It is apparent that the estimated trend line obtained via simple linear regression does not quite capture the trend of the data and better models will be needed.

To perform this analysis in R, use the following commands:

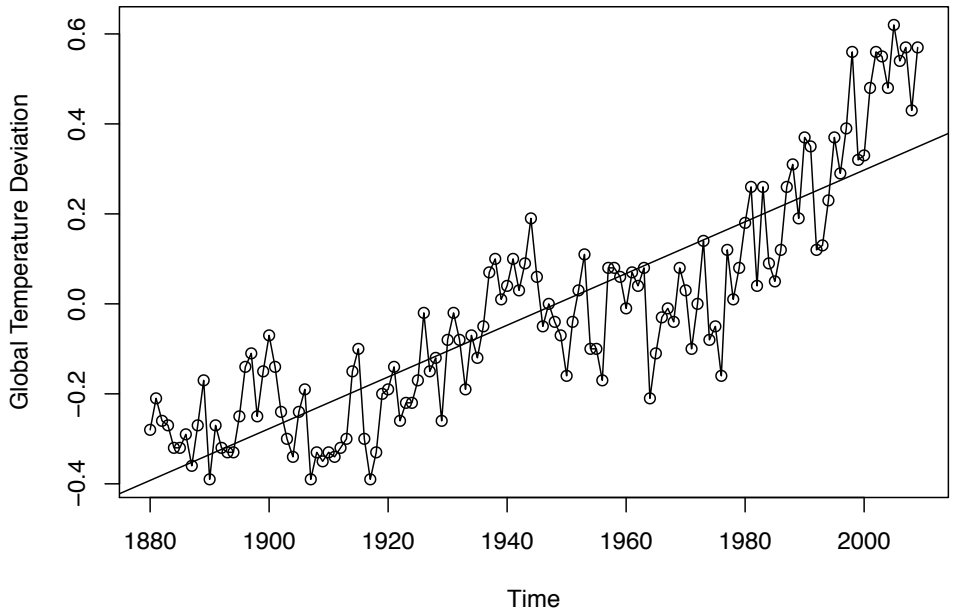

Fig. 2.1. Global temperature deviations shown in Figure 1.2 with fitted linear trend line.

```
summary(fit <- lm(gtemp~time(gtemp)))  # regress gtemp on time
plot(gtemp, type="o", ylab="Global Temperature Deviation")
abline(fit)  # add regression line to the plot
```

The linear model described by (2.1) above can be conveniently written in a more general notation by defining the column vectors $\boldsymbol{z}_t = (z_{t1}, z_{t2}, \ldots, z_{tq})'$ and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_q)'$, where $'$ denotes transpose, so (2.1) can be written in the alternate form

$$x_t = \boldsymbol{\beta}' \boldsymbol{z}_t + w_t. \tag{2.2}$$

where $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$. It is natural to consider estimating the unknown coefficient vector $\boldsymbol{\beta}$ by minimizing the error sum of squares

$$Q = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2, \tag{2.3}$$

with respect to $\beta_1, \beta_2, \ldots, \beta_q$. Minimizing Q yields the ordinary least squares estimator of $\boldsymbol{\beta}$. This minimization can be accomplished by differentiating (2.3) with respect to the vector $\boldsymbol{\beta}$ or by using the properties of projections. In the notation above, this procedure gives the normal equations

$$\left(\sum_{t=1}^{n} \boldsymbol{z}_t \boldsymbol{z}_t' \right) \widehat{\boldsymbol{\beta}} = \sum_{t=1}^{n} \boldsymbol{z}_t x_t. \tag{2.4}$$

The notation can be simplified by defining $Z = [\boldsymbol{z}_1 \,|\, \boldsymbol{z}_2 \,|\, \cdots \,|\, \boldsymbol{z}_n]'$ as the $n \times q$ matrix composed of the n samples of the input variables, the observed $n \times 1$ vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ and the $n \times 1$ vector of errors

$\boldsymbol{w} = (w_1, w_2, \ldots, w_n)'$. In this case, model (2.2) may be written as

$$\boldsymbol{x} = Z\boldsymbol{\beta} + \boldsymbol{w}. \tag{2.5}$$

The normal equations, (2.4), can now be written as

$$(Z'Z)\,\widehat{\boldsymbol{\beta}} = Z'\boldsymbol{x} \tag{2.6}$$

and the solution

$$\widehat{\boldsymbol{\beta}} = (Z'Z)^{-1}Z'\boldsymbol{x} \tag{2.7}$$

when the matrix $Z'Z$ is nonsingular. The minimized error sum of squares (2.3), denoted SSE, can be written as

$$\begin{aligned} SSE &= \sum_{t=1}^{n}(x_t - \widehat{\boldsymbol{\beta}}'\boldsymbol{z}_t)^2 \\ &= (\boldsymbol{x} - Z\widehat{\boldsymbol{\beta}})'(\boldsymbol{x} - Z\widehat{\boldsymbol{\beta}}) \\ &= \boldsymbol{x}'\boldsymbol{x} - \widehat{\boldsymbol{\beta}}'Z'\boldsymbol{x} \\ &= \boldsymbol{x}'\boldsymbol{x} - \boldsymbol{x}'Z(Z'Z)^{-1}Z'\boldsymbol{x}, \end{aligned} \tag{2.8}$$

to give some useful versions for later reference. The ordinary least squares estimators are unbiased, i.e., $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and have the smallest variance within the class of linear unbiased estimators.

If the errors w_t are normally distributed, $\widehat{\boldsymbol{\beta}}$ is also the maximum likelihood estimator for $\boldsymbol{\beta}$ and is normally distributed with

$$\mathrm{cov}(\widehat{\boldsymbol{\beta}}) = \sigma_w^2\left(\sum_{t=1}^{n}\boldsymbol{z}_t\boldsymbol{z}_t'\right)^{-1} = \sigma_w^2(Z'Z)^{-1} = \sigma_w^2 C, \tag{2.9}$$

where

$$C = (Z'Z)^{-1} \tag{2.10}$$

is a convenient notation for later equations. An unbiased estimator for the variance σ_w^2 is

$$s_w^2 = MSE = \frac{SSE}{n-q}, \tag{2.11}$$

where MSE denotes the *mean squared error*, which is contrasted with the maximum likelihood estimator $\widehat{\sigma}_w^2 = SSE/n$. Under the normal assumption, s_w^2 is distributed proportionally to a chi-squared random variable with $n-q$ degrees of freedom, denoted by χ_{n-q}^2, and independently of $\widehat{\beta}$. It follows that

$$t_{n-q} = \frac{(\widehat{\beta}_i - \beta_i)}{s_w\sqrt{c_{ii}}} \tag{2.12}$$

has the t-distribution with $n-q$ degrees of freedom; c_{ii} denotes the i-th diagonal element of C, as defined in (2.10).

Table 2.1. Analysis of Variance for Regression

Source	df	Sum of Squares	Mean Square
$z_{t,r+1}, \ldots, z_{t,q}$	$q-r$	$SSR = SSE_r - SSE$	$MSR = SSR/(q-r)$
Error	$n-q$	SSE	$MSE = SSE/(n-q)$
Total	$n-r$	SSE_r	

Various competing models are of interest to isolate or select the best subset of independent variables. Suppose a proposed model specifies that only a subset $r < q$ independent variables, say, $\boldsymbol{z}_{t:r} = (z_{t1}, z_{t2}, \ldots, z_{tr})'$ is influencing the dependent variable x_t. The reduced model is

$$\boldsymbol{x} = Z_r \boldsymbol{\beta}_r + \boldsymbol{w} \tag{2.13}$$

where $\boldsymbol{\beta}_r = (\beta_1, \beta_2, \ldots, \beta_r)'$ is a subset of coefficients of the original q variables and $Z_r = [\boldsymbol{z}_{1:r} \,|\, \cdots \,|\, \boldsymbol{z}_{n:r}]'$ is the $n \times r$ matrix of inputs. The null hypothesis in this case is H_0: $\beta_{r+1} = \cdots = \beta_q = 0$. We can test the reduced model (2.13) against the full model (2.2) by comparing the error sums of squares under the two models using the F-statistic

$$F_{q-r,n-q} = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-q)}, \tag{2.14}$$

which has the central F-distribution with $q-r$ and $n-q$ degrees of freedom when (2.13) is the correct model. Note that SSE_r is the error sum of squares under the reduced model (2.13) and it can be computed by replacing Z with Z_r in (2.8). The statistic, which follows from applying the likelihood ratio criterion, has the improvement per number of parameters added in the numerator compared with the error sum of squares under the full model in the denominator. The information involved in the test procedure is often summarized in an Analysis of Variance (ANOVA) table as given in Table 2.1 for this particular case. The difference in the numerator is often called the regression sum of squares

In terms of Table 2.1, it is conventional to write the F-statistic (2.14) as the ratio of the two mean squares, obtaining

$$F_{q-r,n-q} = \frac{MSR}{MSE}, \tag{2.15}$$

where MSR, the *mean squared regression*, is the numerator of (2.14). A special case of interest is $r = 1$ and $z_{t1} \equiv 1$, when the model in (2.13) becomes

$$x_t = \beta_1 + w_t,$$

and we may measure the proportion of variation accounted for by the other variables using

$$R^2 = \frac{SSE_1 - SSE}{SSE_1}, \tag{2.16}$$

where the residual sum of squares under the reduced model

$$SSE_1 = \sum_{t=1}^{n} (x_t - \bar{x})^2, \tag{2.17}$$

in this case is just the sum of squared deviations from the mean $\bar{x}$. The measure R^2 is also the *squared multiple correlation* between x_t and the variables $z_{t2}, z_{t3}, \ldots, z_{tq}$.

The techniques discussed in the previous paragraph can be used to test various models against one another using the F test given in (2.14), (2.15), and the ANOVA table. These tests have been used in the past in a stepwise manner, where variables are added or deleted when the values from the F-test either exceed or fail to exceed some predetermined levels. The procedure, called stepwise multiple regression, is useful in arriving at a set of useful variables. An alternative is to focus on a procedure for model selection that does not proceed sequentially, but simply evaluates each model on its own merits. Suppose we consider a normal regression model with k coefficients and denote the maximum likelihood estimator for the variance as

$$\widehat{\sigma}_k^2 = \frac{SSE_k}{n}, \tag{2.18}$$

where SSE_k denotes the residual sum of squares under the model with k regression coefficients. Then, Akaike (1969, 1973, 1974) suggested measuring the goodness of fit for this particular model by balancing the error of the fit against the number of parameters in the model; we define the following.[1]

Definition 2.1 Akaike's Information Criterion (AIC)

$$\text{AIC} = \log\, \widehat{\sigma}_k^2 + \frac{n + 2k}{n}, \tag{2.19}$$

where $\widehat{\sigma}_k^2$ *is given by* (2.18) *and* k *is the number of parameters in the model.*

The value of k yielding the minimum AIC specifies the best model. The idea is roughly that minimizing $\widehat{\sigma}_k^2$ would be a reasonable objective, except that it decreases monotonically as k increases. Therefore, we ought to penalize the error variance by a term proportional to the number of parameters. The choice for the penalty term given by (2.19) is not the only one, and a considerable literature is available advocating different penalty terms. A corrected

[1] Formally, AIC is defined as $-2 \log L_k + 2k$ where L_k is the maximized likelihood and k is the number of parameters in the model. For the normal regression problem, AIC can be reduced to the form given by (2.19). AIC is an estimate of the Kullback-Leibler discrepency between a true model and a candidate model; see Problem 2.4 and Problem 2.5 for further details.

form, suggested by Sugiura (1978), and expanded by Hurvich and Tsai (1989), can be based on small-sample distributional results for the linear regression model (details are provided in Problem 2.4 and Problem 2.5). The corrected form is defined as follows.

Definition 2.2 AIC, Bias Corrected (AICc)

$$\text{AICc} = \log\ \widehat{\sigma}_k^2 + \frac{n+k}{n-k-2}, \tag{2.20}$$

where $\widehat{\sigma}_k^2$ *is given by* (2.18)*,* k *is the number of parameters in the model, and* n *is the sample size.*

We may also derive a correction term based on Bayesian arguments, as in Schwarz (1978), which leads to the following.

Definition 2.3 Bayesian Information Criterion (BIC)

$$\text{BIC} = \log\ \widehat{\sigma}_k^2 + \frac{k \log n}{n}, \tag{2.21}$$

using the same notation as in Definition 2.2.

BIC is also called the Schwarz Information Criterion (SIC); see also Rissanen (1978) for an approach yielding the same statistic based on a minimum description length argument. Various simulation studies have tended to verify that BIC does well at getting the correct order in large samples, whereas AICc tends to be superior in smaller samples where the relative number of parameters is large; see McQuarrie and Tsai (1998) for detailed comparisons. In fitting regression models, two measures that have been used in the past are adjusted R-squared, which is essentially s_w^2, and Mallows C_p, Mallows (1973), which we do not consider in this context.

Example 2.2 Pollution, Temperature and Mortality

The data shown in Figure 2.2 are extracted series from a study by Shumway et al. (1988) of the possible effects of temperature and pollution on weekly mortality in Los Angeles County. Note the strong seasonal components in all of the series, corresponding to winter-summer variations and the downward trend in the cardiovascular mortality over the 10-year period.

A scatterplot matrix, shown in Figure 2.3, indicates a possible linear relation between mortality and the pollutant particulates and a possible relation to temperature. Note the curvilinear shape of the temperature mortality curve, indicating that higher temperatures as well as lower temperatures are associated with increases in cardiovascular mortality.

Based on the scatterplot matrix, we entertain, tentatively, four models where M_t denotes cardiovascular mortality, T_t denotes temperature and P_t denotes the particulate levels. They are

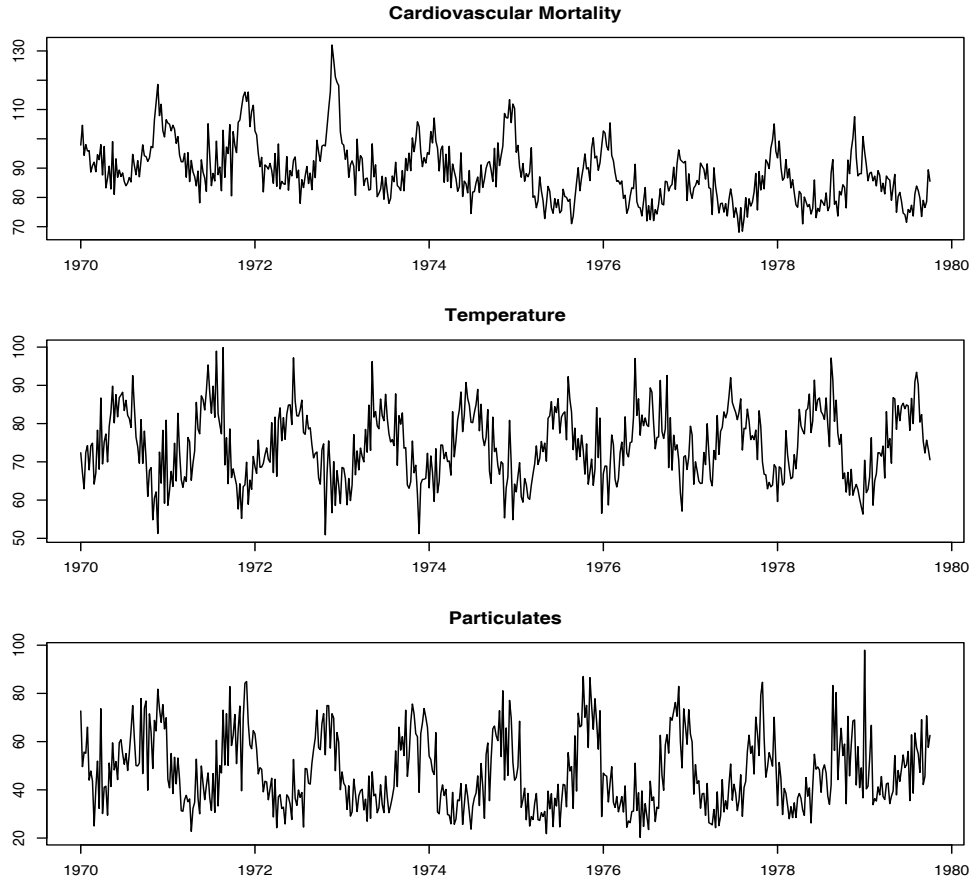

Fig. 2.2. Average weekly cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.

$$M_t = \beta_1 + \beta_2 t + w_t \tag{2.22}$$
$$M_t = \beta_1 + \beta_2 t + \beta_3 (T_t - T.) + w_t \tag{2.23}$$
$$M_t = \beta_1 + \beta_2 t + \beta_3 (T_t - T.) + \beta_4 (T_t - T.)^2 + w_t \tag{2.24}$$
$$M_t = \beta_1 + \beta_2 t + \beta_3 (T_t - T.) + \beta_4 (T_t - T.)^2 + \beta_5 P_t + w_t \tag{2.25}$$

where we adjust temperature for its mean, $T. = 74.6$, to avoid scaling problems. It is clear that (2.22) is a trend only model, (2.23) is linear temperature, (2.24) is curvilinear temperature and (2.25) is curvilinear temperature and pollution. We summarize some of the statistics given for this particular case in 2.2. The values of R^2 were computed by noting that $SSE_1 = 50,687$ using (2.17).

We note that each model does substantially better than the one before it and that the model including temperature, temperature squared, and particulates does the best, accounting for some 60% of the variability and with the best value for AIC and BIC (because of the large sample size, AIC and AICc

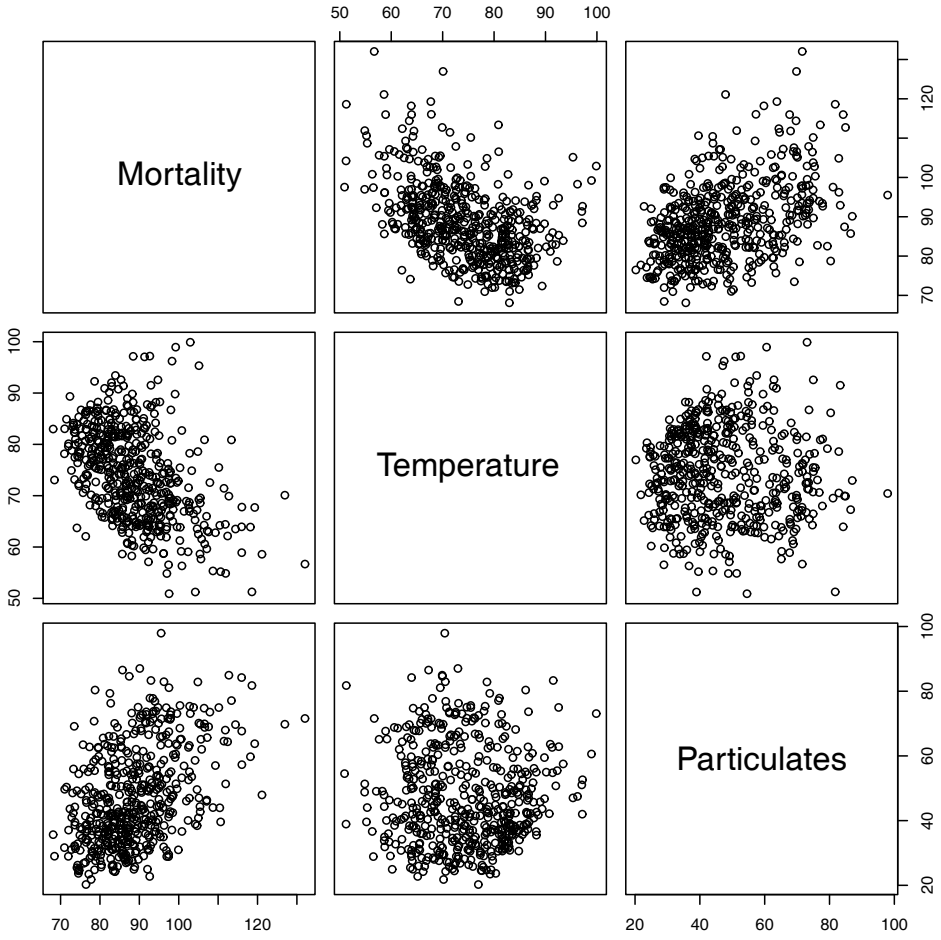

Fig. 2.3. Scatterplot matrix showing plausible relations between mortality, temperature, and pollution.

Table 2.2. Summary Statistics for Mortality Models

Model	k	SSE	df	MSE	R^2	AIC	BIC
(2.22)	2	40,020	506	79.0	.21	5.38	5.40
(2.23)	3	31,413	505	62.2	.38	5.14	5.17
(2.24)	4	27,985	504	55.5	.45	5.03	5.07
(2.25)	5	20,508	503	40.8	.60	4.72	4.77

are nearly the same). Note that one can compare any two models using the residual sums of squares and (2.14). Hence, a model with only trend could be compared to the full model using $q = 5, r = 2, n = 508$, so

$$F_{3,503} = \frac{(40,020 - 20,508)/3}{20,508/503} = 160,$$

which exceeds $F_{3,503}(.001) = 5.51$. We obtain the best prediction model,

$$\widehat{M}_t = 81.59 - .027_{(.002)}t - .473_{(.032)}(T_t - 74.6) + .023_{(.003)}(T_t - 74.6)^2 + .255_{(.019)}P_t,$$

for mortality, where the standard errors, computed from (2.9)–(2.11), are given in parentheses. As expected, a negative trend is present in time as well as a negative coefficient for adjusted temperature. The quadratic effect of temperature can clearly be seen in the scatterplots of Figure 2.3. Pollution weights positively and can be interpreted as the incremental contribution to daily deaths per unit of particulate pollution. It would still be essential to check the residuals $\widehat{w}_t = M_t - \widehat{M}_t$ for autocorrelation (of which there is a substantial amount), but we defer this question to to §5.6 when we discuss regression with correlated errors.

Below is the R code to plot the series, display the scatterplot matrix, fit the final regression model (2.25), and compute the corresponding values of AIC, AICc and BIC.[2] Finally, the use of `na.action` in `lm()` is to retain the time series attributes for the residuals and fitted values.

```
par(mfrow=c(3,1))
plot(cmort, main="Cardiovascular Mortality", xlab="", ylab="")
plot(tempr, main="Temperature", xlab="", ylab="")
plot(part, main="Particulates", xlab="", ylab="")
dev.new()  # open a new graphic device for the scatterplot matrix
pairs(cbind(Mortality=cmort, Temperature=tempr, Particulates=part))
temp = tempr-mean(tempr)  # center temperature
temp2 = temp^2
trend = time(cmort)  # time
fit = lm(cmort~ trend + temp + temp2 + part, na.action=NULL)
summary(fit)  # regression results
summary(aov(fit))  # ANOVA table   (compare to next line)
summary(aov(lm(cmort~cbind(trend, temp, temp2, part)))) # Table 2.1
num = length(cmort)  # sample size
AIC(fit)/num - log(2*pi)  # AIC
BIC(fit)/num - log(2*pi)  # BIC
(AICc = log(sum(resid(fit)^2)/num) + (num+5)/(num-5-2)) # AICc
```

As previously mentioned, it is possible to include lagged variables in time series regression models and we will continue to discuss this type of problem throughout the text. This concept is explored further in Problem 2.2 and Problem 2.11. The following is a simple example of lagged regression.

[2] The easiest way to extract AIC and BIC from an `lm()` run in R is to use the command `AIC()` or `BIC()`. Our definitions differ from R by terms that do not change from model to model. In the example, we show how to obtain (2.19) and (2.21) from the R output. It is more difficult to obtain AICc.

Example 2.3 Regression With Lagged Variables

In Example 1.25, we discovered that the Southern Oscillation Index (SOI) measured at time $t-6$ months is associated with the Recruitment series at time t, indicating that the SOI leads the Recruitment series by six months. Although there is evidence that the relationship is not linear (this is discussed further in Example 2.7), we may consider the following regression,

$$R_t = \beta_1 + \beta_2 S_{t-6} + w_t, \tag{2.26}$$

where R_t denotes Recruitment for month t and S_{t-6} denotes SOI six months prior. Assuming the w_t sequence is white, the fitted model is

$$\widehat{R}_t = 65.79 - 44.28_{(2.78)} S_{t-6} \tag{2.27}$$

with $\widehat{\sigma}_w = 22.5$ on 445 degrees of freedom. This result indicates the strong predictive ability of SOI for Recruitment six months in advance. Of course, it is still essential to check the the model assumptions, but again we defer this until later.

Performing lagged regression in R is a little difficult because the series must be aligned prior to running the regression. The easiest way to do this is to create a data frame that we call `fish` using `ts.intersect`, which aligns the lagged series.

```
fish = ts.intersect(rec, soiL6=lag(soi,-6), dframe=TRUE)
summary(lm(rec~soiL6, data=fish, na.action=NULL))
```

2.3 Exploratory Data Analysis

In general, it is necessary for time series data to be stationary, so averaging lagged products over time, as in the previous section, will be a sensible thing to do. With time series data, it is the dependence between the values of the series that is important to measure; we must, at least, be able to estimate autocorrelations with precision. It would be difficult to measure that dependence if the dependence structure is not regular or is changing at every time point. Hence, to achieve any meaningful statistical analysis of time series data, it will be crucial that, if nothing else, the mean and the autocovariance functions satisfy the conditions of stationarity (for at least some reasonable stretch of time) stated in Definition 1.7. Often, this is not the case, and we will mention some methods in this section for playing down the effects of nonstationarity so the stationary properties of the series may be studied.

A number of our examples came from clearly nonstationary series. The Johnson & Johnson series in Figure 1.1 has a mean that increases exponentially over time, and the increase in the magnitude of the fluctuations around this trend causes changes in the covariance function; the variance of the process, for example, clearly increases as one progresses over the length of the series. Also, the global temperature series shown in Figure 1.2 contains some

evidence of a trend over time; human-induced global warming advocates seize on this as empirical evidence to advance their hypothesis that temperatures are increasing.

Perhaps the easiest form of nonstationarity to work with is the trend stationary model wherein the process has stationary behavior around a trend. We may write this type of model as

$$x_t = \mu_t + y_t \tag{2.28}$$

where x_t are the observations, μ_t denotes the trend, and y_t is a stationary process. Quite often, strong trend, μ_t, will obscure the behavior of the stationary process, y_t, as we shall see in numerous examples. Hence, there is some advantage to removing the trend as a first step in an exploratory analysis of such time series. The steps involved are to obtain a reasonable estimate of the trend component, say $\widehat{\mu}_t$, and then work with the residuals

$$\widehat{y}_t = x_t - \widehat{\mu}_t. \tag{2.29}$$

Consider the following example.

Example 2.4 Detrending Global Temperature

Here we suppose the model is of the form of (2.28),

$$x_t = \mu_t + y_t,$$

where, as we suggested in the analysis of the global temperature data presented in Example 2.1, a straight line might be a reasonable model for the trend, i.e.,

$$\mu_t = \beta_1 + \beta_2\, t.$$

In that example, we estimated the trend using ordinary least squares[3] and found

$$\widehat{\mu}_t = -11.2 + .006\, t.$$

Figure 2.1 shows the data with the estimated trend line superimposed. To obtain the detrended series we simply subtract $\widehat{\mu}_t$ from the observations, x_t, to obtain the detrended series

$$\widehat{y}_t = x_t + 11.2 - .006\, t.$$

The top graph of Figure 2.4 shows the detrended series. Figure 2.5 shows the ACF of the original data (top panel) as well as the ACF of the detrended data (middle panel).

[3] Because the error term, y_t, is not assumed to be iid, the reader may feel that weighted least squares is called for in this case. The problem is, we do not know the behavior of y_t and that is precisely what we are trying to assess at this stage. A notable result by Grenander and Rosenblatt (1957, Ch 7), however, is that under mild conditions on y_t, for polynomial regression or periodic regression, asymptotically, ordinary least squares is equivalent to weighted least squares.

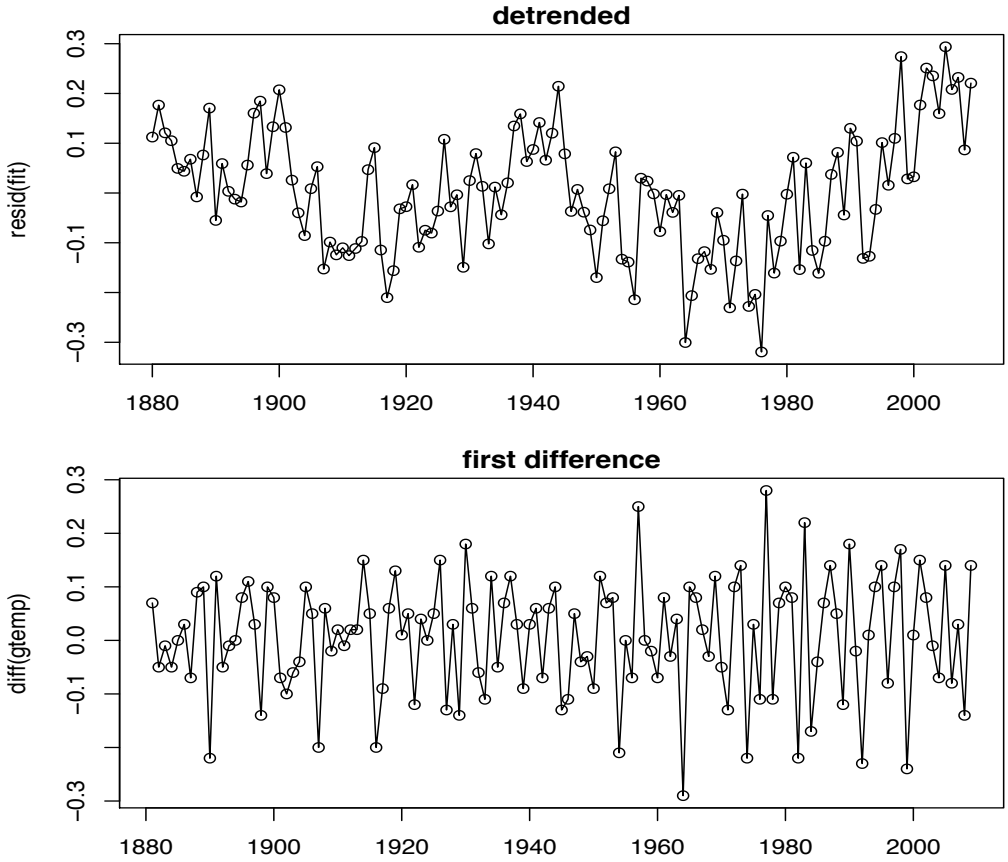

Fig. 2.4. Detrended (top) and differenced (bottom) global temperature series. The original data are shown in Figure 1.2 and Figure 2.1.

To detrend in the series in R, use the following commands. We also show how to difference and plot the differenced data; we discuss differencing after this example. In addition, we show how to generate the sample ACFs displayed in Figure 2.5.

```
fit = lm(gtemp~time(gtemp), na.action=NULL) # regress gtemp on time
par(mfrow=c(2,1))
plot(resid(fit), type="o", main="detrended")
plot(diff(gtemp), type="o", main="first difference")
par(mfrow=c(3,1))  # plot ACFs
acf(gtemp, 48, main="gtemp")
acf(resid(fit), 48, main="detrended")
acf(diff(gtemp), 48, main="first difference")
```

In Example 1.11 and the corresponding Figure 1.10 we saw that a random walk might also be a good model for trend. That is, rather than modeling trend as fixed (as in Example 2.4), we might model trend as a stochastic component using the random walk with drift model,

$$\mu_t = \delta + \mu_{t-1} + w_t, \tag{2.30}$$

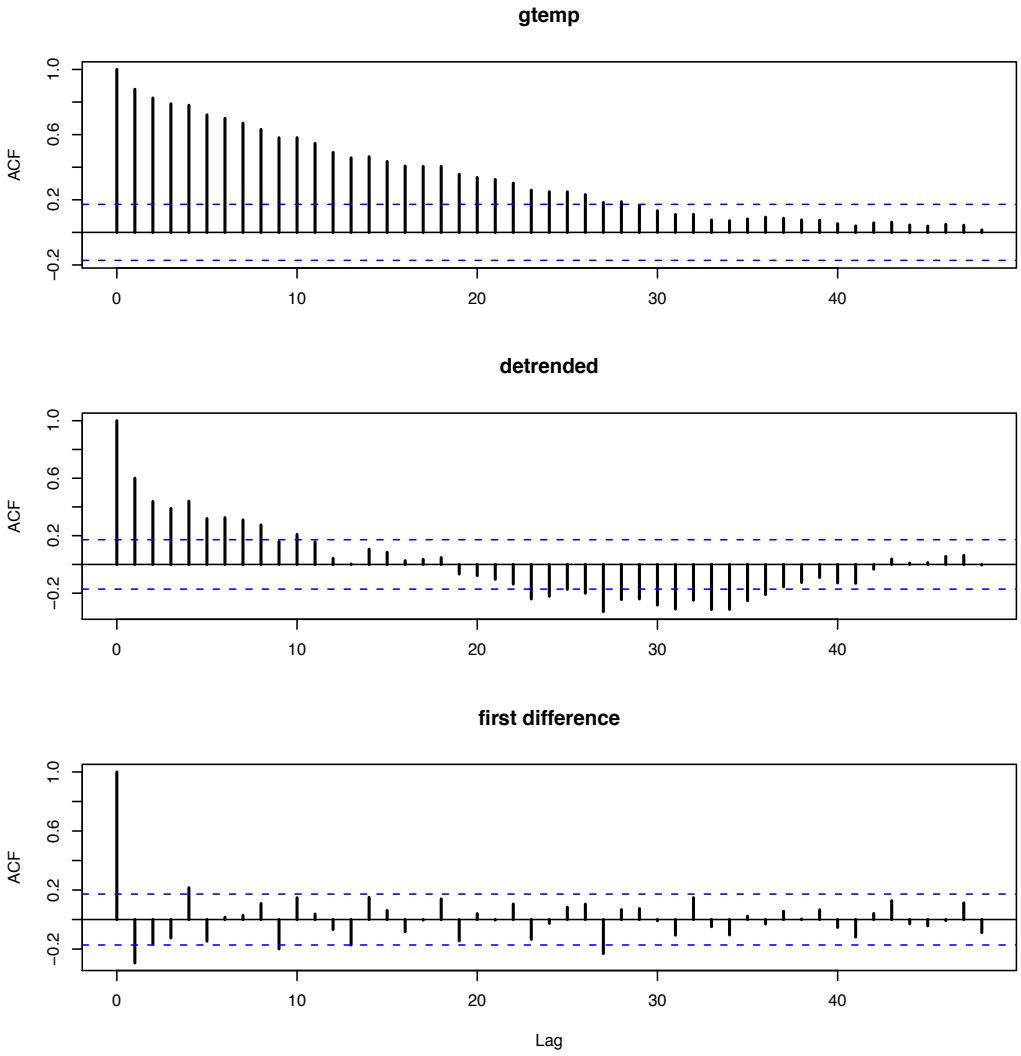

Fig. 2.5. Sample ACFs of the global temperature (top), and of the detrended (middle) and the differenced (bottom) series.

where w_t is white noise and is independent of y_t. If the appropriate model is (2.28), then differencing the data, x_t, yields a stationary process; that is,

$$\begin{aligned} x_t - x_{t-1} &= (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) \\ &= \delta + w_t + y_t - y_{t-1}. \end{aligned} \tag{2.31}$$

It is easy to show $z_t = y_t - y_{t-1}$ is stationary using footnote 3 of Chapter 1 on page 20. That is, because y_t is stationary,

$$\begin{aligned} \gamma_z(h) &= \mathrm{cov}(z_{t+h}, z_t) = \mathrm{cov}(y_{t+h} - y_{t+h-1}, y_t - y_{t-1}) \\ &= 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1) \end{aligned}$$

is independent of time; we leave it as an exercise (Problem 2.7) to show that $x_t - x_{t-1}$ in (2.31) is stationary.

One advantage of differencing over detrending to remove trend is that no parameters are estimated in the differencing operation. One disadvantage, however, is that differencing does not yield an estimate of the stationary process y_t as can be seen in (2.31). If an estimate of y_t is essential, then detrending may be more appropriate. If the goal is to coerce the data to stationarity, then differencing may be more appropriate. Differencing is also a viable tool if the trend is fixed, as in Example 2.4. That is, e.g., if $\mu_t = \beta_1 + \beta_2 t$ in the model (2.28), differencing the data produces stationarity (see Problem 2.6):

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_2 + y_t - y_{t-1}.$$

Because differencing plays a central role in time series analysis, it receives its own notation. The first difference is denoted as

$$\nabla x_t = x_t - x_{t-1}. \tag{2.32}$$

As we have seen, the first difference eliminates a linear trend. A second difference, that is, the difference of (2.32), can eliminate a quadratic trend, and so on. In order to define higher differences, we need a variation in notation that we will use often in our discussion of ARIMA models in Chapter 3.

Definition 2.4 *We define the* **backshift operator** *by*

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

and extend it to powers $B^2 x_t = B(Bx_t) = Bx_{t-1} = x_{t-2}$*, and so on. Thus,*

$$B^k x_t = x_{t-k}. \tag{2.33}$$

It is clear that we may then rewrite (2.32) as

$$\nabla x_t = (1 - B)x_t, \tag{2.34}$$

and we may extend the notion further. For example, the second difference becomes

$$\begin{aligned} \nabla^2 x_t &= (1 - B)^2 x_t = (1 - 2B + B^2)x_t \\ &= x_t - 2x_{t-1} + x_{t-2} \end{aligned}$$

by the linearity of the operator. To check, just take the difference of the first difference $\nabla(\nabla x_t) = \nabla(x_t - x_{t-1}) = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$.

Definition 2.5 **Differences of order d** *are defined as*

$$\nabla^d = (1 - B)^d, \tag{2.35}$$

where we may expand the operator $(1-B)^d$ *algebraically to evaluate for higher integer values of* d*. When* $d = 1$*, we drop it from the notation.*

The first difference (2.32) is an example of a linear filter applied to eliminate a trend. Other filters, formed by averaging values near x_t, can produce adjusted series that eliminate other kinds of unwanted fluctuations, as in Chapter 3. The differencing technique is an important component of the ARIMA model of Box and Jenkins (1970) (see also Box et al., 1994), to be discussed in Chapter 3.

Example 2.5 Differencing Global Temperature

The first difference of the global temperature series, also shown in Figure 2.4, produces different results than removing trend by detrending via regression. For example, the differenced series does not contain the long middle cycle we observe in the detrended series. The ACF of this series is also shown in Figure 2.5. In this case it appears that the differenced process shows minimal autocorrelation, which may imply the global temperature series is nearly a random walk with drift. It is interesting to note that if the series is a random walk with drift, the mean of the differenced series, which is an estimate of the drift, is about .0066 (but with a large standard error):

```
mean(diff(gtemp))    # = 0.00659 (drift)
sd(diff(gtemp))/sqrt(length(diff(gtemp))) # = 0.00966 (SE)
```

An alternative to differencing is a less-severe operation that still assumes stationarity of the underlying time series. This alternative, called fractional differencing, extends the notion of the difference operator (2.35) to fractional powers $-.5 < d < .5$, which still define stationary processes. Granger and Joyeux (1980) and Hosking (1981) introduced long memory time series, which corresponds to the case when $0 < d < .5$. This model is often used for environmental time series arising in hydrology. We will discuss long memory processes in more detail in §5.2.

Often, obvious aberrations are present that can contribute nonstationary as well as nonlinear behavior in observed time series. In such cases, transformations may be useful to equalize the variability over the length of a single series. A particularly useful transformation is

$$y_t = \log x_t, \tag{2.36}$$

which tends to suppress larger fluctuations that occur over portions of the series where the underlying values are larger. Other possibilities are power transformations in the Box–Cox family of the form

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log x_t & \lambda = 0. \end{cases} \tag{2.37}$$

Methods for choosing the power λ are available (see Johnson and Wichern, 1992, §4.7) but we do not pursue them here. Often, transformations are also used to improve the approximation to normality or to improve linearity in predicting the value of one series from another.

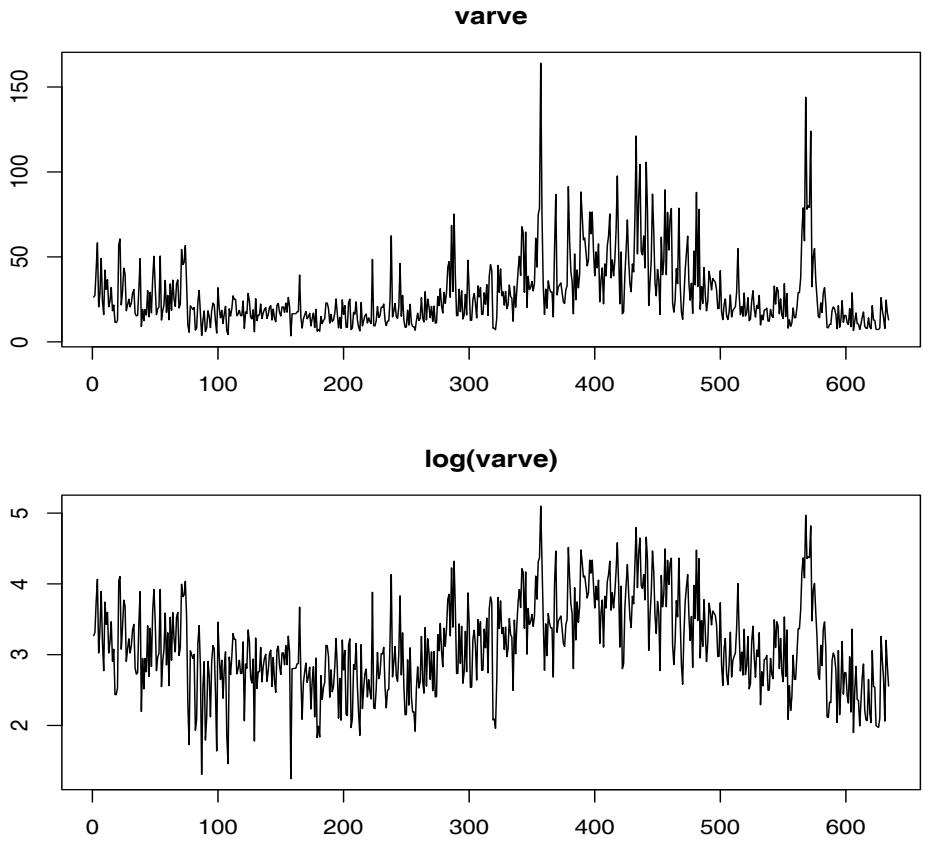

Fig. 2.6. Glacial varve thicknesses (top) from Massachusetts for $n = 634$ years compared with log transformed thicknesses (bottom).

Example 2.6 Paleoclimatic Glacial Varves

Melting glaciers deposit yearly layers of sand and silt during the spring melting seasons, which can be reconstructed yearly over a period ranging from the time deglaciation began in New England (about 12,600 years ago) to the time it ended (about 6,000 years ago). Such sedimentary deposits, called varves, can be used as proxies for paleoclimatic parameters, such as temperature, because, in a warm year, more sand and silt are deposited from the receding glacier. Figure 2.6 shows the thicknesses of the yearly varves collected from one location in Massachusetts for 634 years, beginning 11,834 years ago. For further information, see Shumway and Verosub (1992). Because the variation in thicknesses increases in proportion to the amount deposited, a logarithmic transformation could remove the nonstationarity observable in the variance as a function of time. Figure 2.6 shows the original and transformed varves, and it is clear that this improvement has occurred. We may also plot the histogram of the original and transformed data, as in Problem 2.8, to argue that the approximation to normality is improved. The ordinary first differences (2.34) are also computed in Problem 2.8, and we note that the first differences have a significant negative correlation at lag $h = 1$. Later, in

Chapter 5, we will show that perhaps the varve series has long memory and will propose using fractional differencing.

Figure 2.6 was generated in R as follows:

```
par(mfrow=c(2,1))
plot(varve, main="varve", ylab="")
plot(log(varve), main="log(varve)", ylab="" )
```

Next, we consider another preliminary data processing technique that is used for the purpose of visualizing the relations between series at different lags, namely, scatterplot matrices. In the definition of the ACF, we are essentially interested in relations between x_t and x_{t-h}; the autocorrelation function tells us whether a substantial linear relation exists between the series and its own lagged values. The ACF gives a profile of the linear correlation at all possible lags and shows which values of h lead to the best predictability. The restriction of this idea to linear predictability, however, may mask a possible nonlinear relation between current values, x_t, and past values, x_{t-h}. This idea extends to two series where one may be interested in examining scatterplots of y_t versus x_{t-h}

Example 2.7 Scatterplot Matrices, SOI and Recruitment

To check for nonlinear relations of this form, it is convenient to display a lagged scatterplot matrix, as in Figure 2.7, that displays values of the SOI, S_t, on the vertical axis plotted against S_{t-h} on the horizontal axis. The sample autocorrelations are displayed in the upper right-hand corner and superimposed on the scatterplots are locally weighted scatterplot smoothing (lowess) lines that can be used to help discover any nonlinearities. We discuss smoothing in the next section, but for now, think of lowess as a robust method for fitting nonlinear regression.

In Figure 2.7, we notice that the lowess fits are approximately linear, so that the sample autocorrelations are meaningful. Also, we see strong positive linear relations at lags $h = 1, 2, 11, 12$, that is, between S_t and $S_{t-1}, S_{t-2}, S_{t-11}, S_{t-12}$, and a negative linear relation at lags $h = 6, 7$. These results match up well with peaks noticed in the ACF in Figure 1.14.

Similarly, we might want to look at values of one series, say Recruitment, denoted R_t plotted against another series at various lags, say the SOI, S_{t-h}, to look for possible nonlinear relations between the two series. Because, for example, we might wish to predict the Recruitment series, R_t, from current or past values of the SOI series, S_{t-h}, for $h = 0, 1, 2, \ldots$ it would be worthwhile to examine the scatterplot matrix. Figure 2.8 shows the lagged scatterplot of the Recruitment series R_t on the vertical axis plotted against the SOI index S_{t-h} on the horizontal axis. In addition, the figure exhibits the sample cross-correlations as well as lowess fits.

Figure 2.8 shows a fairly strong nonlinear relationship between Recruitment, R_t, and the SOI series at $S_{t-5}, S_{t-6}, S_{t-7}, S_{t-8}$, indicating the SOI series tends to lead the Recruitment series and the coefficients are negative, implying that increases in the SOI lead to decreases in the Recruitment. The

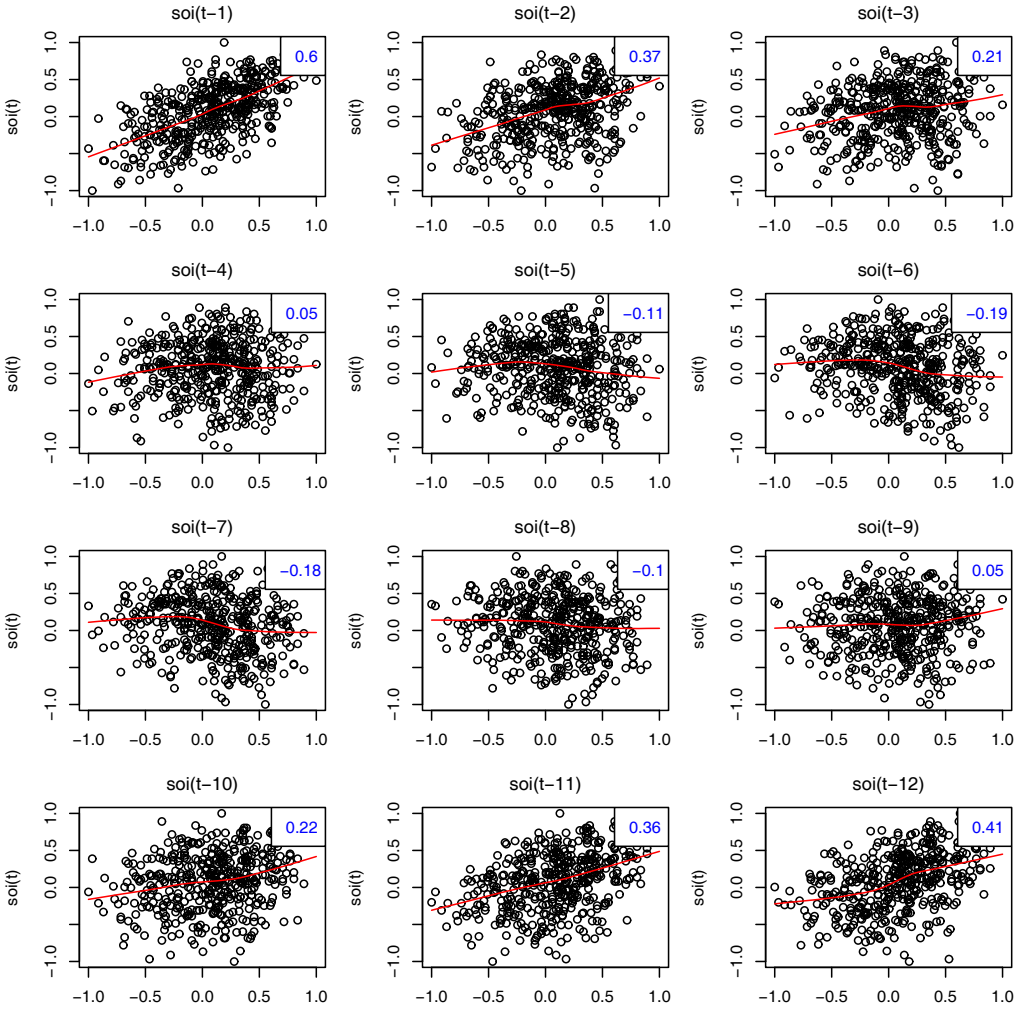

Fig. 2.7. Scatterplot matrix relating current SOI values, S_t, to past SOI values, S_{t-h}, at lags $h = 1, 2, ..., 12$. The values in the upper right corner are the sample autocorrelations and the lines are a lowess fit.

nonlinearity observed in the scatterplots (with the help of the superimposed lowess fits) indicate that the behavior between Recruitment and the SOI is different for positive values of SOI than for negative values of SOI.

Simple scatterplot matrices for one series can be obtained in R using the `lag.plot` command. Figure 2.7 and Figure 2.8 may be reproduced using the following scripts provided with `astsa` (see Appendix R for details):

```
lag1.plot(soi, 12)      # Figure 2.7
lag2.plot(soi, rec, 8)  # Figure 2.8
```

As a final exploratory tool, we discuss assessing periodic behavior in time series data using regression analysis and the periodogram; this material may be thought of as an introduction to spectral analysis, which we discuss in

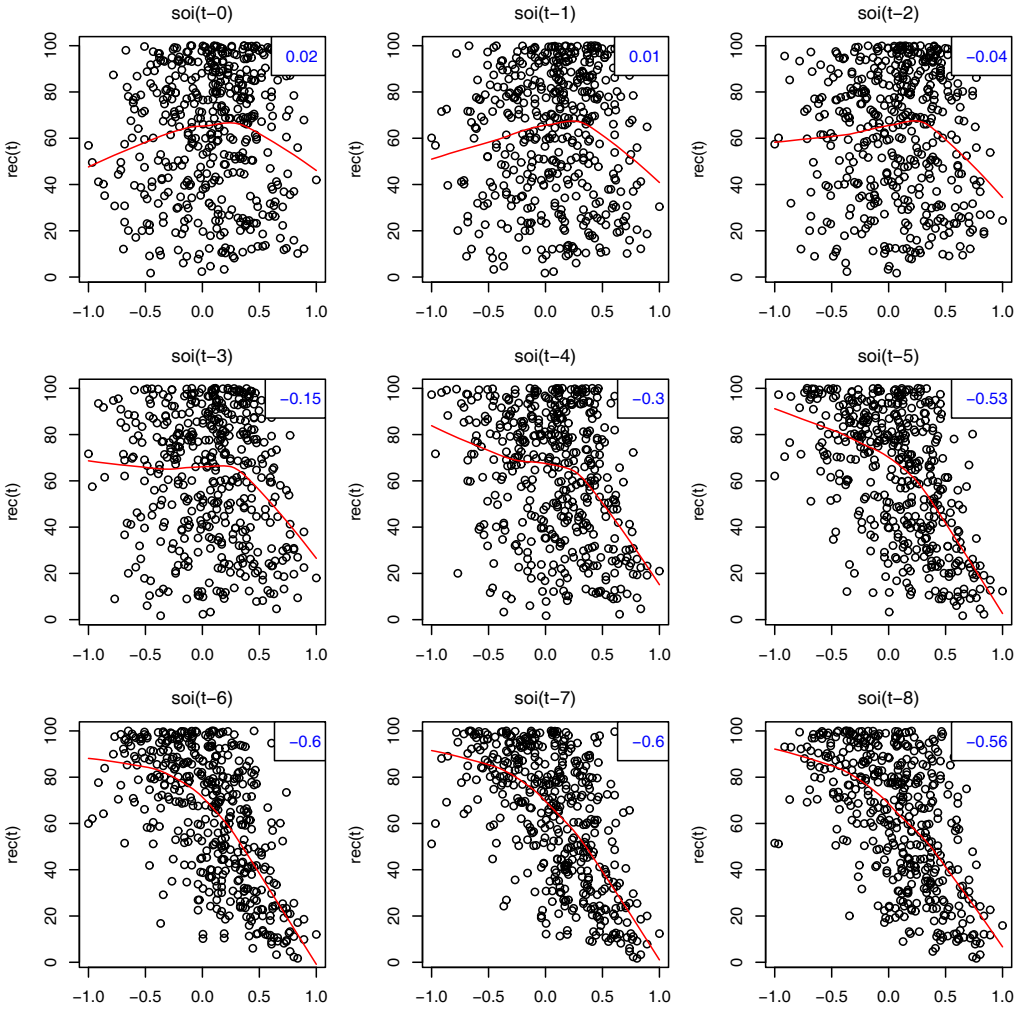

Fig. 2.8. Scatterplot matrix of the Recruitment series, R_t, on the vertical axis plotted against the SOI series, S_{t-h}, on the horizontal axis at lags $h = 0, 1, \ldots, 8$. The values in the upper right corner are the sample cross-correlations and the lines are a lowess fit.

detail in Chapter 4. In Example 1.12, we briefly discussed the problem of identifying cyclic or periodic signals in time series. A number of the time series we have seen so far exhibit periodic behavior. For example, the data from the pollution study example shown in Figure 2.2 exhibit strong yearly cycles. Also, the Johnson & Johnson data shown in Figure 1.1 make one cycle every year (four quarters) on top of an increasing trend and the speech data in Figure 1.2 is highly repetitive. The monthly SOI and Recruitment series in Figure 1.6 show strong yearly cycles, but hidden in the series are clues to the El Niño cycle.

Example 2.8 Using Regression to Discover a Signal in Noise

In Example 1.12, we generated $n = 500$ observations from the model

$$x_t = A\cos(2\pi\omega t + \phi) + w_t, \tag{2.38}$$

where $\omega = 1/50$, $A = 2$, $\phi = .6\pi$, and $\sigma_w = 5$; the data are shown on the bottom panel of Figure 1.11 on page 16. At this point we assume the frequency of oscillation $\omega = 1/50$ is known, but A and ϕ are unknown parameters. In this case the parameters appear in (2.38) in a nonlinear way, so we use a trigonometric identity[4] and write

$$A\cos(2\pi\omega t + \phi) = \beta_1\cos(2\pi\omega t) + \beta_2\sin(2\pi\omega t),$$

where $\beta_1 = A\cos(\phi)$ and $\beta_2 = -A\sin(\phi)$. Now the model (2.38) can be written in the usual linear regression form given by (no intercept term is needed here)

$$x_t = \beta_1\cos(2\pi t/50) + \beta_2\sin(2\pi t/50) + w_t. \tag{2.39}$$

Using linear regression on the generated data, the fitted model is

$$\widehat{x}_t = -.71_{(.30)}\cos(2\pi t/50) - 2.55_{(.30)}\sin(2\pi t/50) \tag{2.40}$$

with $\widehat{\sigma}_w = 4.68$, where the values in parentheses are the standard errors. We note the actual values of the coefficients for this example are $\beta_1 = 2\cos(.6\pi) = -.62$ and $\beta_2 = -2\sin(.6\pi) = -1.90$. Because the parameter estimates are significant and close to the actual values, it is clear that we are able to detect the signal in the noise using regression, even though the signal appears to be obscured by the noise in the bottom panel of Figure 1.11. Figure 2.9 shows data generated by (2.38) with the fitted line, (2.40), superimposed.

To reproduce the analysis and Figure 2.9 in R, use the following commands:

```
set.seed(1000)  # so you can reproduce these results
x = 2*cos(2*pi*1:500/50 + .6*pi) + rnorm(500,0,5)
z1 = cos(2*pi*1:500/50);  z2 = sin(2*pi*1:500/50)
summary(fit <- lm(x~0+z1+z2))  # zero to exclude the intercept
plot.ts(x, lty="dashed")
lines(fitted(fit), lwd=2)
```

Example 2.9 Using the Periodogram to Discover a Signal in Noise

The analysis in Example 2.8 may seem like cheating because we assumed we knew the value of the frequency parameter ω. If we do not know ω, we could try to fit the model (2.38) using nonlinear regression with ω as a parameter. Another method is to try various values of ω in a systematic way. Using the

[4] $\cos(\alpha \pm \beta) = \cos(\alpha)\cos(\beta) \mp \sin(\alpha)\sin(\beta)$.

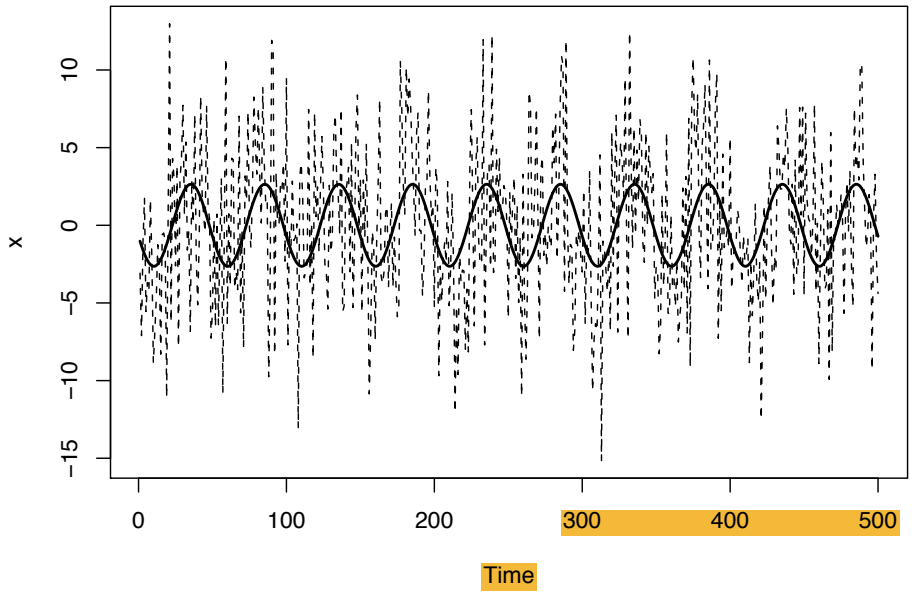

Fig. 2.9. Data generated by (2.38) [dashed line] with the fitted [solid] line, (2.40), superimposed.

regression results of §2.2, we can show the estimated regression coefficients in Example 2.8 take on the special form given by

$$\widehat{\beta}_1 = \frac{\sum_{t=1}^{n} x_t \cos(2\pi t/50)}{\sum_{t=1}^{n} \cos^2(2\pi t/50)} = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(2\pi t/50); \tag{2.41}$$

$$\widehat{\beta}_2 = \frac{\sum_{t=1}^{n} x_t \sin(2\pi t/50)}{\sum_{t=1}^{n} \sin^2(2\pi t/50)} = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(2\pi t/50). \tag{2.42}$$

This suggests looking at all possible regression parameter estimates,[5] say

$$\widehat{\beta}_1(j/n) = \frac{2}{n}\sum_{t=1}^{n} x_t \cos(2\pi t\, j/n); \tag{2.43}$$

$$\widehat{\beta}_2(j/n) = \frac{2}{n}\sum_{t=1}^{n} x_t \sin(2\pi t\, j/n), \tag{2.44}$$

where, $n = 500$ and $j = 1, \ldots, \frac{n}{2} - 1$, and inspecting the results for large values. For the endpoints, $j = 0$ and $j = n/2$, we have $\widehat{\beta}_1(0) = n^{-1}\sum_{t=1}^{n} x_t$ and $\widehat{\beta}_1(\frac{1}{2}) = n^{-1}\sum_{t=1}^{n}(-1)^t x_t$, and $\widehat{\beta}_2(0) = \widehat{\beta}_2(\frac{1}{2}) = 0$.

For this particular example, the values calculated in (2.41) and (2.42) are $\widehat{\beta}_1(10/500)$ and $\widehat{\beta}_2(10/500)$. By doing this, we have regressed a series, x_t, of

[5] In the notation of §2.2, the estimates are of the form $\sum_{t=1}^{n} x_t z_t \,/\, \sum_{t=1}^{n} z_t^2$ where $z_t = \cos(2\pi t j/n)$ or $z_t = \sin(2\pi t j/n)$. In this setup, unless $j = 0$ or $j = n/2$ if n is even, $\sum_{t=1}^{n} z_t^2 = n/2$; see Problem 2.10.

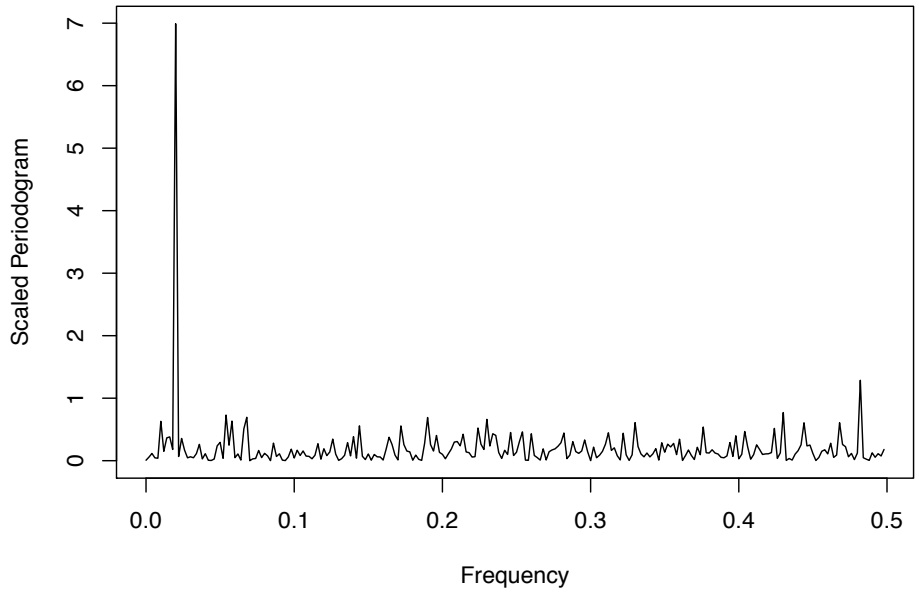

Fig. 2.10. The scaled periodogram, (2.45), of the 500 observations generated by (2.38); the data are displayed in Figure 1.11 and Figure 2.9.

length n using n regression parameters, so that we will have a perfect fit. The point, however, is that if the data contain any cyclic behavior we are likely to catch it by performing these saturated regressions.

Next, note that the regression coefficients $\widehat{\beta}_1(j/n)$ and $\widehat{\beta}_2(j/n)$, for each j, are essentially measuring the correlation of the data with a sinusoid oscillating at j cycles in n time points.[6] Hence, an appropriate measure of the presence of a frequency of oscillation of j cycles in n time points in the data would be

$$P(j/n) = \widehat{\beta}_1^2(j/n) + \widehat{\beta}_2^2(j/n), \tag{2.45}$$

which is basically a measure of squared correlation. The quantity (2.45) is sometimes called the periodogram, but we will call $P(j/n)$ the scaled periodogram and we will investigate its properties in Chapter 4. Figure 2.10 shows the scaled periodogram for the data generated by (2.38), and it easily discovers the periodic component with frequency $\omega = .02 = 10/500$ even though it is difficult to visually notice that component in Figure 1.11 due to the noise.

Finally, we mention that it is not necessary to run a large regression

$$x_t = \sum_{j=0}^{n/2} \beta_1(j/n)\cos(2\pi t j/n) + \beta_2(j/n)\sin(2\pi t j/n) \tag{2.46}$$

to obtain the values of $\beta_1(j/n)$ and $\beta_2(j/n)$ [with $\beta_2(0) = \beta_2(1/2) = 0$] because they can be computed quickly if n (assumed even here) is a highly

[6] Sample correlations are of the form $\sum_t x_t z_t / \left(\sum_t x_t^2 \sum_t z_t^2\right)^{1/2}$.

composite integer. There is no error in (2.46) because there are n observations and n parameters; the regression fit will be perfect. The discrete Fourier transform (DFT) is a complex-valued weighted average of the data given by

$$\begin{aligned} d(j/n) &= n^{-1/2} \sum_{t=1}^{n} x_t \exp(-2\pi i t j/n) \\ &= n^{-1/2} \left(\sum_{t=1}^{n} x_t \cos(2\pi t j/n) - i \sum_{t=1}^{n} x_t \sin(2\pi t j/n) \right) \end{aligned} \tag{2.47}$$

where the frequencies j/n are called the Fourier or fundamental frequencies. Because of a large number of redundancies in the calculation, (2.47) may be computed quickly using the fast Fourier transform (FFT)[7], which is available in many computing packages such as Matlab®, S-PLUS® and R. Note that[8]

$$|d(j/n)|^2 = \frac{1}{n} \left(\sum_{t=1}^{n} x_t \cos(2\pi t j/n) \right)^2 + \frac{1}{n} \left(\sum_{t=1}^{n} x_t \sin(2\pi t j/n) \right)^2 \tag{2.48}$$

and it is this quantity that is called the periodogram; we will write

$$I(j/n) = |d(j/n)|^2.$$

We may calculate the scaled periodogram, (2.45), using the periodogram as

$$P(j/n) = \frac{4}{n} I(j/n). \tag{2.49}$$

We will discuss this approach in more detail and provide examples with data in Chapter 4.

Figure 2.10 can be created in R using the following commands (and the data already generated in `x`):

```
I = abs(fft(x))^2/500   # the periodogram
P = (4/500)*I[1:250]    # the scaled periodogram
f = 0:249/500           # frequencies
plot(f, P, type="l", xlab="Frequency", ylab="Scaled Periodogram")
```

2.4 Smoothing in the Time Series Context

In §1.4, we introduced the concept of smoothing a time series, and in Example 1.9, we discussed using a moving average to smooth white noise. This method is useful in discovering certain traits in a time series, such as long-term

[7] Different packages scale the FFT differently; consult the documentation. R calculates (2.47) without scaling by $n^{-1/2}$.

[8] If $z = a - ib$ is complex, then $|z|^2 = z\bar{z} = (a - ib)(a + ib) = a^2 + b^2$.

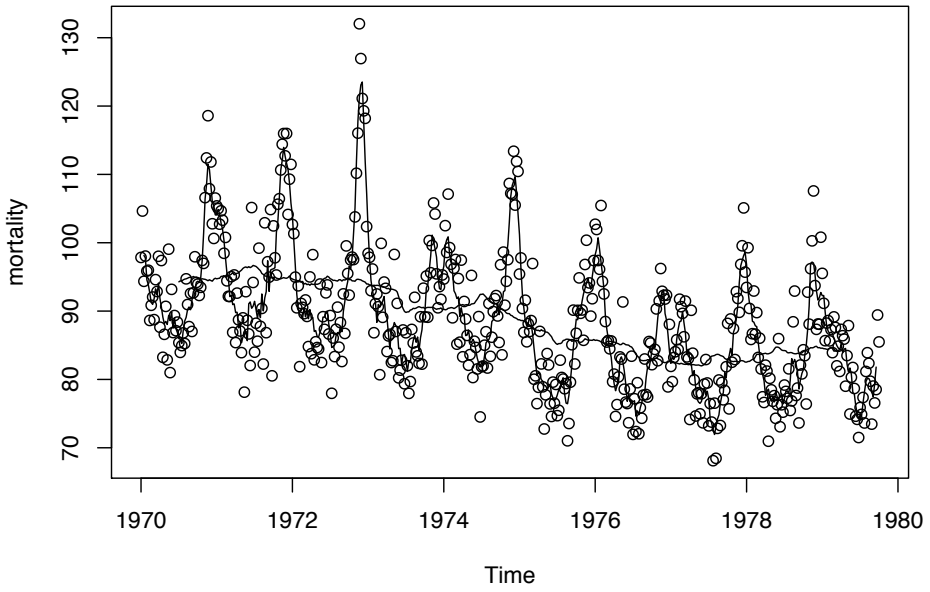

Fig. 2.11. The weekly cardiovascular mortality series discussed in Example 2.2 smoothed using a five-week moving average and a 53-week moving average.

trend and seasonal components. In particular, if x_t represents the observations, then

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}, \tag{2.50}$$

where $a_j = a_{-j} \geq 0$ and $\sum_{j=-k}^{k} a_j = 1$ is a symmetric moving average of the data.

Example 2.10 Moving Average Smoother

For example, Figure 2.11 shows the weekly mortality series discussed in Example 2.2, a five-point moving average (which is essentially a monthly average with $k = 2$) that helps bring out the seasonal component and a 53-point moving average (which is essentially a yearly average with $k = 26$) that helps bring out the (negative) trend in cardiovascular mortality. In both cases, the weights, $a_{-k}, \ldots, a_0, \ldots, a_k$, we used were all the same, and equal to $1/(2k+1)$.[9]

To reproduce Figure 2.11 in R:

```
ma5 = filter(cmort, sides=2, rep(1,5)/5)
ma53 = filter(cmort, sides=2, rep(1,53)/53)
plot(cmort, type="p", ylab="mortality")
lines(ma5); lines(ma53)
```

[9] Sometimes, the end weights, a_{-k} and a_k are set equal to half the value of the other weights.

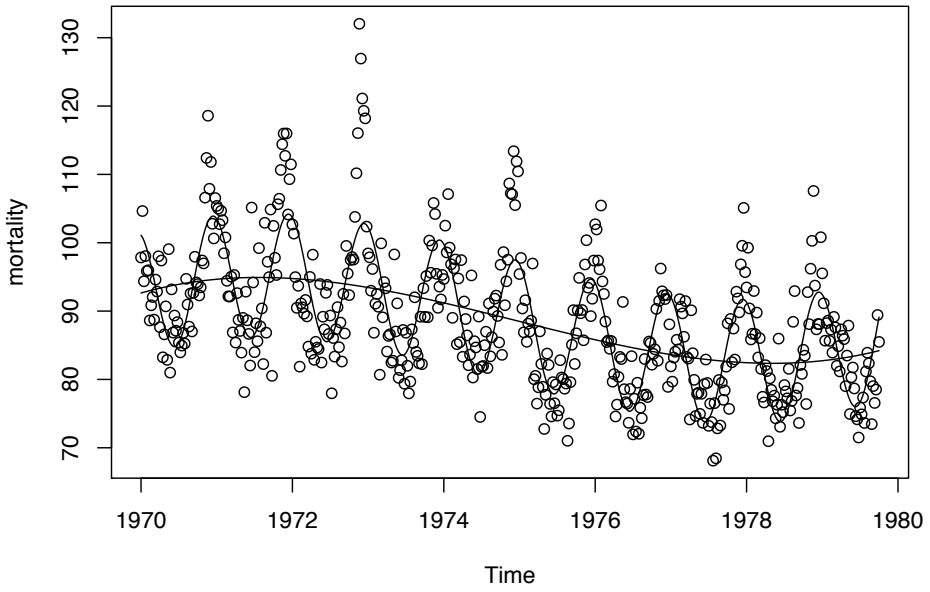

Fig. 2.12. The weekly cardiovascular mortality series with a cubic trend and cubic trend plus periodic regression.

Many other techniques are available for smoothing time series data based on methods from scatterplot smoothers. The general setup for a time plot is

$$x_t = f_t + y_t, \tag{2.51}$$

where f_t is some smooth function of time, and y_t is a stationary process. We may think of the moving average smoother m_t, given in (2.50), as an estimator of f_t. An obvious choice for f_t in (2.51) is polynomial regression

$$f_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p. \tag{2.52}$$

We have seen the results of a linear fit on the global temperature data in Example 2.1. For periodic data, one might employ periodic regression

$$\begin{aligned} f_t = {} & \alpha_0 + \alpha_1 \cos(2\pi\omega_1 t) + \beta_1 \sin(2\pi\omega_1 t) \\ & + \cdots + \alpha_p \cos(2\pi\omega_p t) + \beta_p \sin(2\pi\omega_p t), \end{aligned} \tag{2.53}$$

where $\omega_1, \ldots, \omega_p$ are distinct, specified frequencies. In addition, one might consider combining (2.52) and (2.53). These smoothers can be applied using classical linear regression.

Example 2.11 Polynomial and Periodic Regression Smoothers
Figure 2.12 shows the weekly mortality series with an estimated (via ordinary least squares) cubic smoother

$$\widehat{f}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2 + \widehat{\beta}_3 t^3$$

superimposed to emphasize the trend, and an estimated (via ordinary least squares) cubic smoother plus a periodic regression

$$\widehat{f}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2 + \widehat{\beta}_3 t^3 + \widehat{\alpha}_1 \cos(2\pi t/52) + \widehat{\alpha}_2 \sin(2\pi t/52)$$

superimposed to emphasize trend and seasonality.

The R commands for this example are as follows (we note that the sampling rate is 1/52, so that `wk` below is essentially $t/52$).

```
wk = time(cmort) - mean(time(cmort))
wk2 = wk^2; wk3 = wk^3
cs = cos(2*pi*wk);  sn = sin(2*pi*wk)
reg1 = lm(cmort~wk + wk2 + wk3, na.action=NULL)
reg2 = lm(cmort~wk + wk2 + wk3 + cs + sn, na.action=NULL)
plot(cmort, type="p", ylab="mortality")
lines(fitted(reg1));  lines(fitted(reg2))
```

Modern regression techniques can be used to fit general smoothers to the pairs of points (t, x_t) where the estimate of f_t is smooth. Many of the techniques can easily be applied to time series data using the R or S-PLUS statistical packages; see Venables and Ripley (1994, Chapter 10) for details on applying these methods in S-PLUS (R is similar). A problem with the techniques used in Example 2.11 is that they assume f_t is the same function over the range of time, t; we might say that the technique is global. The moving average smoothers in Example 2.10 fit the data better because the technique is local; that is, moving average smoothers allow for the possibility that f_t is a different function over time. We describe some other local methods in the following examples.

Example 2.12 Kernel Smoothing

Kernel smoothing is a moving average smoother that uses a weight function, or kernel, to average the observations. Figure 2.13 shows kernel smoothing of the mortality series, where f_t in (2.51) is estimated by

$$\widehat{f}_t = \sum_{i=1}^{n} w_i(t) x_i, \tag{2.54}$$

where

$$w_i(t) = K\left(\tfrac{t-i}{b}\right) \Big/ \sum_{j=1}^{n} K\left(\tfrac{t-j}{b}\right). \tag{2.55}$$

are the weights and $K(\cdot)$ is a kernel function. This estimator, which was originally explored by Parzen (1962) and Rosenblatt (1956b), is often called the Nadaraya–Watson estimator (Watson, 1966); typically, the normal kernel, $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$, is used. To implement this in R, use the `ksmooth` function. The wider the bandwidth, b, the smoother the result. In Figure 2.13, the values of b for this example were $b = 5/52$ (roughly weighted two to three

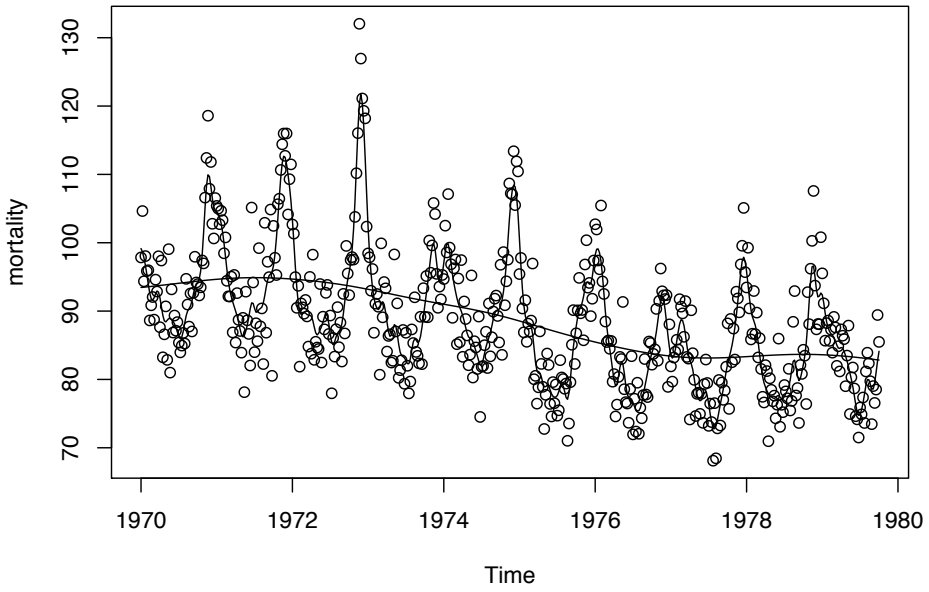

Fig. 2.13. Kernel smoothers of the mortality data.

week averages because $b/2$ is the inner quartile range of the kernel) for the seasonal component, and $b = 104/52 = 2$ (roughly weighted yearly averages) for the trend component.

Figure 2.13 can be reproduced in R (or S-PLUS) as follows.

```
plot(cmort, type="p", ylab="mortality")
lines(ksmooth(time(cmort), cmort, "normal", bandwidth=5/52))
lines(ksmooth(time(cmort), cmort, "normal", bandwidth=2))
```

Example 2.13 Lowess and Nearest Neighbor Regression

Another approach to smoothing a time plot is nearest neighbor regression. The technique is based on k-nearest neighbors linear regression, wherein one uses the data $\{x_{t-k/2}, \ldots, x_t, \ldots, x_{t+k/2}\}$ to predict x_t using linear regression; the result is $\widehat{f}_t$. For example, Figure 2.14 shows cardiovascular mortality and the nearest neighbor method using the R (or S-PLUS) smoother `supsmu`. We used $k = n/2$ to estimate the trend and $k = n/100$ to estimate the seasonal component. In general, `supsmu` uses a variable window for smoothing (see Friedman, 1984), but it can be used for correlated data by fixing the smoothing window, as was done here.

Lowess is a method of smoothing that is rather complex, but the basic idea is close to nearest neighbor regression. Figure 2.14 shows smoothing of mortality using the R or S-PLUS function `lowess` (see Cleveland, 1979). First, a certain proportion of nearest neighbors to x_t are included in a weighting scheme; values closer to x_t in time get more weight. Then, a robust weighted regression is used to predict x_t and obtain the smoothed estimate of f_t. The larger the fraction of nearest neighbors included, the smoother the estimate

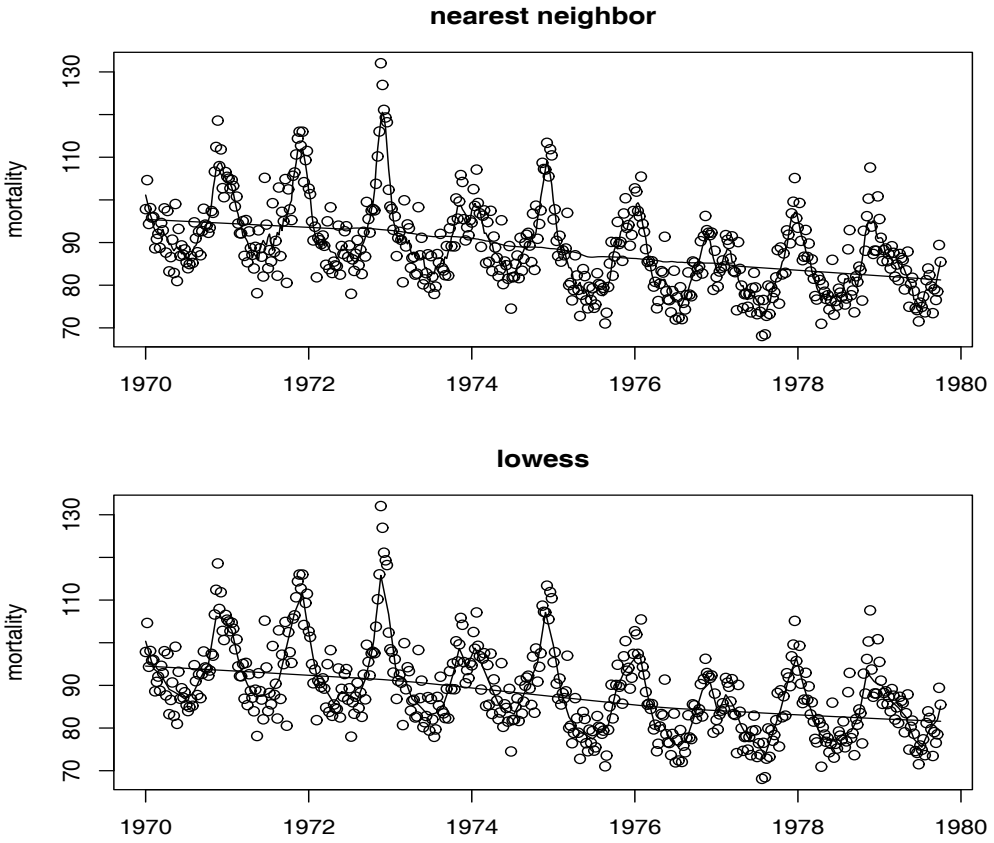

Fig. 2.14. Nearest neighbor (`supsmu`) and locally weighted regression (`lowess`) smoothers of the mortality data.

$\widehat{f_t}$ will be. In Figure 2.14, the smoother uses about two-thirds of the data to obtain an estimate of the trend component, and the seasonal component uses 2% of the data.

Figure 2.14 can be reproduced in R or S-PLUS as follows.

```
par(mfrow=c(2,1))
plot(cmort, type="p", ylab="mortality", main="nearest neighbor")
lines(supsmu(time(cmort), cmort, span=.5))
lines(supsmu(time(cmort), cmort, span=.01))
plot(cmort, type="p", ylab="mortality", main="lowess")
lines(lowess(cmort, f=.02));  lines(lowess(cmort, f=2/3))
```

Example 2.14 Smoothing Splines

An extension of polynomial regression is to first divide time $t = 1, \ldots, n$, into k intervals, $[t_0 = 1, t_1]$, $[t_1 + 1, t_2]$, $\ldots$, $[t_{k-1} + 1, t_k = n]$. The values $t_0, t_1, \ldots, t_k$ are called *knots*. Then, in each interval, one fits a regression of the form (2.52); typically, $p = 3$, and this is called cubic splines.

A related method is smoothing splines, which minimizes a compromise between the fit and the degree of smoothness given by

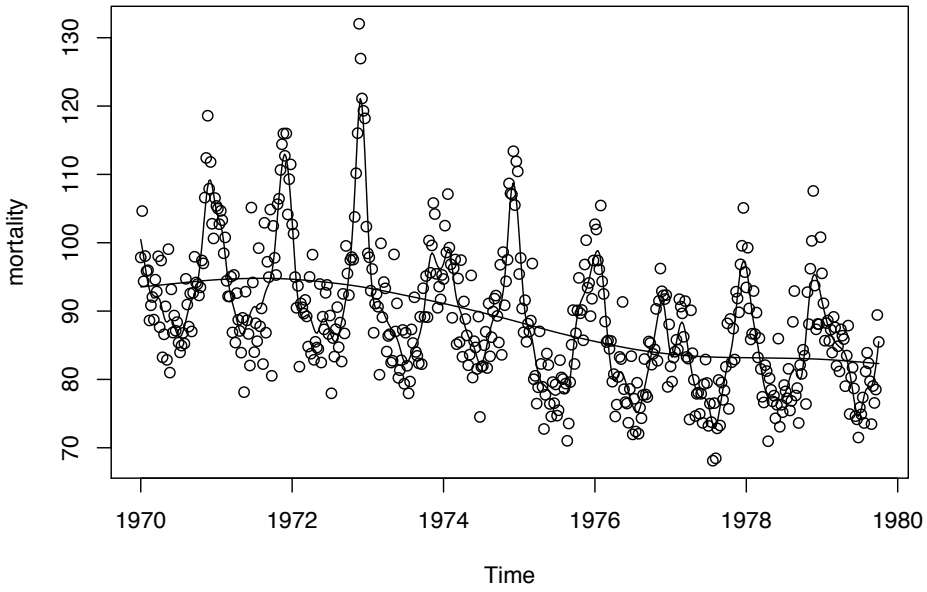

Fig. 2.15. Smoothing splines fit to the mortality data.

$$\sum_{t=1}^{n} [x_t - f_t]^2 + \lambda \int \left(f_t''\right)^2 dt, \tag{2.56}$$

where f_t is a cubic spline with a knot at each t. The degree of smoothness is controlled by $\lambda > 0$. There is a relationship between smoothing splines and state space models, which is investigated in Problem 6.7.

In R, the smoothing parameter is called `spar` and it is monotonically related to λ; type `?smooth.spline` to view the help file for details. Figure 2.15 shows smoothing spline fits on the mortality data using generalized cross-validation, which uses the data to "optimally" assess the smoothing parameter, for the seasonal component, and `spar=1` for the trend. The figure can be reproduced in R as follows.

```
plot(cmort, type="p", ylab="mortality")
lines(smooth.spline(time(cmort), cmort))
lines(smooth.spline(time(cmort), cmort, spar=1))
```

Example 2.15 Smoothing One Series as a Function of Another

In addition to smoothing time plots, smoothing techniques can be applied to smoothing a time series as a function of another time series. In this example, we smooth the scatterplot of two contemporaneously measured time series, mortality as a function of temperature. In Example 2.2, we discovered a nonlinear relationship between mortality and temperature. Continuing along these lines, Figure 2.16 shows scatterplots of mortality, M_t, and temperature, T_t, along with M_t smoothed as a function of T_t using lowess and using smoothing splines. In both cases, mortality increases at extreme

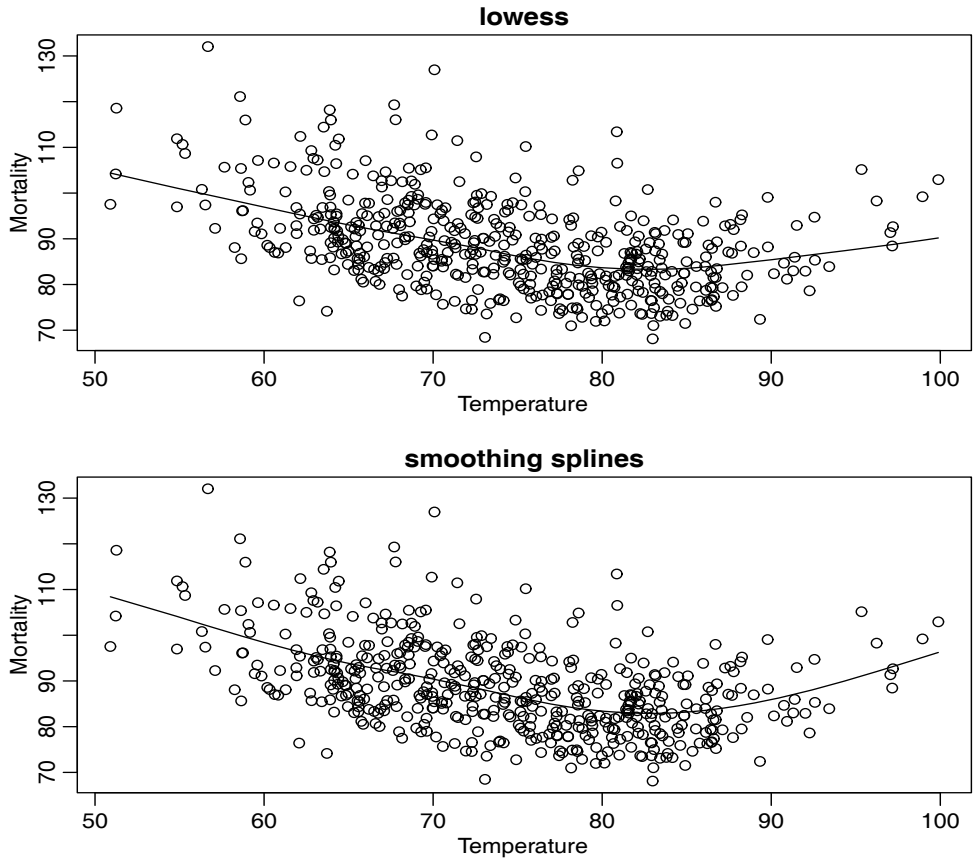

Fig. 2.16. Smoothers of mortality as a function of temperature using lowess and smoothing splines.

temperatures, but in an asymmetric way; mortality is higher at colder temperatures than at hotter temperatures. The minimum mortality rate seems to occur at approximately 80° F.

Figure 2.16 can be reproduced in R as follows.

```
par(mfrow=c(2,1), mar=c(3,2,1,0)+.5, mgp=c(1.6,.6,0))
plot(tempr, cmort, main="lowess", xlab="Temperature",
    ylab="Mortality")
lines(lowess(tempr,cmort))
plot(tempr, cmort, main="smoothing splines", xlab="Temperature",
    ylab="Mortality")
lines(smooth.spline(tempr, cmort))
```

As a final word of caution, the methods mentioned in this section may not take into account the fact that the data are serially correlated, and most of the techniques have been designed for independent observations. That is, for example, the smoothers shown in Figure 2.16 are calculated under the false assumption that the pairs (M_t, T_t), are iid pairs of observations. In addition,