Chapter 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 $x_{t-1}, x_{t-2}, \ldots, x_{tp}$, of the currently observed series. The outputs $x_t$ may also depend on lagged values of another series, say $y_{t-1}, y_{t-2}, \ldots, y_{t-qp}$, 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 re-
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

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

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 $\sigma_w^2$; 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 Trend

Consider the global temperature data, say $x_t$, shown in Figure 1.2. 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

$$x_t = \beta_1 + \beta_2 t + w_t, \quad t = 1900, 1901, \ldots, 1997.$$  

This is in the form of the regression model (2.1) when we make the identification $q = 2$, $z_{t1} = 1$, $z_{t2} = t$. 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, e.g., $t = 0, \ldots, 97$, without affecting the interpretation of the slope coefficient, $\beta_2$; only the intercept, $\beta_1$, would be affected.

Using simple linear regression, we obtained the estimated coefficients $\hat{\beta}_1 = -12.186$, and $\hat{\beta}_2 = .006$ (with a standard error of .0005) yielding a significant estimated increase of .6 degrees centigrade per 100 years. We
discuss the precise way in which the solution was accomplished below. Finally, Figure 2.1 shows the global temperature data, say \( x_t \), with the estimated trend, say \( \hat{x}_t = -12.186 + .006t \), 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, we note that the data file `globtemp.dat` has 142 observations starting from the year 1856. We are only using the final 98 observations corresponding to the years 1900 to 1997.

```r
> gtemp = scan("/mydata/globtemp.dat")
> x = gtemp[45:142]
> t = 1900:1997
> fit=lm(x˜t) # regress x on t
> summary(fit) # regression output
> plot(t,x, type="o", xlab="year", ylab="temp 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 \( z_t = (z_{t1}, z_{t2}, \ldots, z_{tq})' \) and \( \beta = (\beta_1, \beta_2, \ldots, \beta_q)' \), where ' denotes transpose, so (2.1) can be written in the alternate form

\[
x_t = \beta' z_t + w_t. \tag{2.2}
\]

where \( w_t \sim \text{iid}(0, \sigma_w^2) \). It is natural to consider estimating the unknown coef-
2.2: Classical Regression

The vector \( \beta \) by minimizing the residual sum of squares

\[
RSS = \sum_{t=1}^{n} (x_t - \beta z_t)^2,
\]

(2.3)

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

\[
\left( \sum_{t=1}^{n} z_t z_t' \right) \hat{\beta} = \sum_{t=1}^{n} z_t x_t.
\]

(2.4)

A further simplification of notation results from defining the matrix

\[ Z = (z_1, z_2, \ldots, z_n)' \]

as the \( n \times q \) matrix composed of the \( n \) samples of the input variables and the observed \( n \times 1 \) vector

\[ x = (x_1, x_2, \ldots, x_n)' \]

This identification yields

\[
(Z'Z) \hat{\beta} = Z'x
\]

(2.5)

and the solution

\[
\hat{\beta} = (Z'Z)^{-1}Z'x
\]

(2.6)

when the matrix \( Z'Z \) is of rank \( q \). The minimized residual sum of squares (2.3) has the equivalent matrix forms

\[
RSS = (x - Z \hat{\beta}')(x - Z \hat{\beta})
\]

\[
= x'x - \hat{\beta}'Z'x
\]

\[
= x'x - x'Z(Z'Z)^{-1}Z'x,
\]

(2.7)

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

If the errors \( w_t \) are normally distributed (Gaussian), \( \hat{\beta} \) is also the maximum likelihood estimator for \( \beta \) and is normally distributed with

\[
\text{cov}(\hat{\beta}) = \sigma_w^2 \left( \sum_{t=1}^{n} z_t z_t' \right)^{-1}
\]

\[
= \sigma_w^2 (Z'Z)^{-1}
\]

\[
= \sigma_w^2 C,
\]

(2.8)

where

\[
C = (Z'Z)^{-1}
\]

(2.9)

is a convenient notation for later equations. An unbiased estimator for the variance \( \sigma_w^2 \) is

\[
s_w^2 = \frac{RSS}{n - q},
\]

(2.10)
Table 2.1 Analysis of Variance for Regression

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_{t,q_1+1}, \ldots, z_{t,q}$</td>
<td>$q - q_1$</td>
<td>$SS_{reg} = RSS_1 - RSS$</td>
<td>$MS_{reg} = SS_{reg} / (q - q_1)$</td>
</tr>
<tr>
<td>Error</td>
<td>$n - q$</td>
<td>$RSS$</td>
<td>$s_w^2 = RSS / (n - q)$</td>
</tr>
<tr>
<td>Total</td>
<td>$n - q_1$</td>
<td>$RSS_1$</td>
<td></td>
</tr>
</tbody>
</table>

contrasted with the maximum likelihood estimator $\hat{\sigma}_w^2 = RSS / n$, which has the divisor $n$. Under the normal assumption, $s_w^2$ is distributed proportionally to a chi-squared random variable with $n - q$ degrees of freedom, denoted by $\chi^2_{n - q}$, and independently of $\hat{\beta}$. It follows that

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

(2.11)

has the t-distribution with $n - q$ degrees of freedom; $c_{ii}$ denotes the $i$th diagonal element of $C$, as defined in (2.9).

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 $q_1 < q$ independent variables, say, $z_{1t} = (z_{t1}, z_{t2}, \ldots, z_{tq_1})'$ is influencing the dependent variable $x_t$, so the model

$$x_t = \beta_1' z_{1t} + w_t$$

(2.12)

becomes the null hypothesis, where $\beta_1 = (\beta_1, \beta_2, \ldots, \beta_{q_1})'$ is a subset of coefficients of the original $q$ variables. We can test the reduced model (2.12) against the full model (2.2) by comparing the residual sums of squares under the two models using the F-statistic

$$F_{q - q_1, n - q} = \frac{RSS_1 - RSS}{RSS / (q - q_1)}$$

(2.13)

which has the central $F$-distribution with $q - q_1$ and $n - q$ degrees of freedom when (2.12) is the correct model. 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 (2.13) as the ratio of the two mean squares, obtaining

$$F_{q - q_1, n - q} = \frac{MS_{reg}}{s_w^2}.$$
A special case of interest is $q_1 = 1$ and $z_{1t} = 1$, so the model in (2.12) becomes

$$x_t = \beta_1 + w_t,$$

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

$$R^2_{xz} = \frac{RSS_0 - RSS}{RSS_0},$$

where the residual sum of squares under the reduced model

$$RSS_0 = \sum_{t=1}^{n} (x_t - \bar{x})^2,$$

in this case is just the sum of squared deviations from the mean $\bar{x}$. The measure $R^2_{xz}$ is also the squared multiple correlation between $x_t$ and the variables $z_{12}, z_{13}, \ldots, z_{1q}$.

The techniques discussed in the previous paragraph can be used to test various models against one another using the $F$ test given in (2.13), (2.14), 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 regression model with $k$ coefficients and denote the maximum likelihood estimator for the variance as

$$\hat{\sigma}^2_k = \frac{RSS_k}{n},$$

where $RSS_k$ 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

**Definition 2.1 Akaike’s Information Criterion (AIC)**

$$\text{AIC} = \ln \hat{\sigma}^2_k + \frac{n + 2k}{n},$$

where $\hat{\sigma}^2_k$ is given by (2.17) 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 $\hat{\sigma}^2_k$ would be a reasonable objective, except that it decreases monotonically as $k$ increases. Therefore, we ought penalize the error variance by a term proportional to the number of parameters. The choice for the penalty term given by (2.18) is not the only one, and a considerable
literature is available advocating different penalty terms. A corrected 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 Problems 2.4 and 2.5). The corrected form is defined as

**Definition 2.2 AIC, Bias Corrected (AICc)**

\[
\text{AICc} = \ln \hat{\sigma}^2_k + \frac{n + k}{n - k - 2},
\]

where \(\hat{\sigma}^2_k\) is given by (2.17), \(k\) 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

**Definition 2.3 Schwarz’s Information Criterion (SIC)**

\[
\text{SIC} = \ln \hat{\sigma}^2_k + \frac{k \ln n}{n},
\]

using the same notation as in Definition 2.2.

SIC is also called the Bayesian Information Criterion (BIC) (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 SIC 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 \(s^2_w\), 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 daily 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

\[
M_t = \beta_0 + \beta_1 t + w_t
\]

(2.21)

\[
M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T) + w_t
\]

(2.22)

\[
M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T) + \beta_3 (T_t - T)^2 + w_t
\]

(2.23)

\[
M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T) + \beta_3 (T_t - T)^2 + \beta_4 P_t + w_t
\]

(2.24)

where we adjust temperature for its mean, $T = 74.6$, to avoid scaling problems. It is clear that (2.21) is a trend only model, (2.22) is linear temperature, (2.23) is curvilinear temperature and (2.24) is curvilinear temperature and pollution. We summarize some the statistics given for this particular case in Table 2.2. The values of $R^2$ were computed by
We note that each model does substantially better than the one before it and that the model including both temperature, temperature squared and particulates does the best, accounting for some 60% of the variability noting that $RSS_0 = 50,687$ using (2.16).

Table 2.2 Summary Statistics for Mortality Models

<table>
<thead>
<tr>
<th>Model</th>
<th>RSS (2.3)</th>
<th>$s^2_\epsilon$ (2.10)</th>
<th>$R^2$ (2.15)</th>
<th>AICc (2.19)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2.21)</td>
<td>40,020</td>
<td>79.09</td>
<td>.21</td>
<td>5.38</td>
</tr>
<tr>
<td>(2.22)</td>
<td>31,413</td>
<td>62.20</td>
<td>.38</td>
<td>5.14</td>
</tr>
<tr>
<td>(2.23)</td>
<td>27,985</td>
<td>55.52</td>
<td>.45</td>
<td>5.03</td>
</tr>
<tr>
<td>(2.24)</td>
<td>20,509</td>
<td>40.77</td>
<td>.60</td>
<td>4.72</td>
</tr>
</tbody>
</table>
and with the best value for AICc. Note that one can compare any two models using the residual sums of squares and (2.13). Hence, a model with only trend could be compared to the full model using \( q = 5, q_1 = 2, n = 508 \), so

\[
F_{3,503} = \frac{(40,020 - 20,509)}{20,509} \frac{503}{3} = 160,
\]

which exceeds \( F_{3,\infty}(0.001) = 5.42 \). We obtain the best prediction model,

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

for mortality, where the standard errors, computed from (2.8)-(2.10), 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 \( \hat{w}_t = \hat{M}_t - \hat{M}_t \) for autocorrelation, but we defer this question to the section on correlated least squares, in which the incorporation of time correlation changes the estimated standard errors.

To display the scatterplot matrix, perform the final regression and compute AIC in R, use the following commands:

\[
\begin{align*}
> & \text{mort} = \text{scan("/mydata/cmort.dat")} \\
> & \text{temp} = \text{scan("/mydata/temp.dat")} \\
> & \text{part} = \text{scan("/mydata/part.dat")} \\
> & \text{temp} = \text{temp - mean(temp)} \\
> & \text{temp2} = \text{temp}^2 \\
> & \text{t} = 1:\text{length(mort)} \\
> & \text{fit} = \text{lm(mort}\sim t + \text{temp} + \text{temp2} + \text{part}) \\
> & \text{summary(fit)} \ # \text{Results} \\
> & \text{AIC(fit)/508} \ # \text{R gives n*AIC} \\
> & \text{pairs(cbind(mort, temp, part))} \ # \text{scatterplot matrix}
\end{align*}
\]

### 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

\[ x_t = \mu_t + y_t \]  

(2.25)

where \( x_t \) are the observations, \( \mu_t \) denotes the trend, and \( y_t \) is a stationary process. Quite often, strong trend, \( \mu_t \), will obscure the behavior of the stationary process, \( y_t \), as we shall see in numerous examples in Chapter 3. 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 \( \hat{\mu}_t \), and then work with the residuals

\[ \hat{y}_t = x_t - \hat{\mu}_t. \]  

(2.26)

Consider the following example.

**Example 2.3 Detrending Global Temperature**

Here we suppose the model is of the form of (2.25),

\[ x_t = \mu_t + y_t, \]

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\(^1\)

\(^1\)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.
and found
\[ \hat{\mu}_t = -12.186 + .006 t. \]

Figure 2.1 shows the data with the estimated trend line superimposed. To obtain the detrended series we simply subtract \( \hat{\mu}_t \) from the observations, \( x_t \), to obtain the detrended series
\[ \hat{y}_t = x_t + 12.186 - .006 t. \]

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).

To detrend in R, assuming the data are in `gtemp`:
```r
> x = gtemp[45:142]  # use only 1900 to 1997
> t = 1900:1997
> fit = lm(x~t)       # detrended series in fit$resid
> plot(t, fit$resid, type="o", ylab="detrended gtemp")
```

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 2.3), we might model trend as a stochastic component using the random walk with drift model,
\[ \mu_t = \delta + \mu_{t-1} + w_t, \] (2.27)
Regression and Exploratory Data Analysis

\[ w_t \text{ is white noise and is independent of } y_t. \] If the appropriate model is (2.25), then differencing the data, \( x_t \), yields a stationary process; that is,

\[
x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \delta + w_t + y_t - y_{t-1}. \tag{2.28}
\]

We leave it as an exercise (Problem 2.7) to show (2.28) is stationary.\(^2\)

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 (2.28). 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

\(^2\)The key to establishing the stationarity of these types of processes is to recall that if \( U = \sum_{j=1}^{m} a_j X_j \) and \( V = \sum_{k=1}^{r} b_k Y_k \) are linear combinations of random variables \( \{X_j\} \) and \( \{Y_k\} \), respectively, then \( \text{cov}(U, V) = \sum_{j=1}^{m} \sum_{k=1}^{r} a_j b_k \text{cov}(X_j, Y_k) \).
differencing may be more appropriate. Differencing is also a viable tool if the trend is fixed, as in Example 2.3. That is, e.g., if $\mu_t = \beta_1 + \beta_2 t$ in the model (2.25), 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

$$\nabla x_t = x_t - x_{t-1}. \quad (2.29)$$

As we have seen, the first difference eliminates a linear trend. A second difference, that is, the difference of (2.29), can eliminate a quadratic trend, and so on. In order to define higher differences, we need a variation in notation that we use, for the first time here, and often in our discussion of ARIMA models in Chapter 3.

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

$$B x_t = x_{t-1}$$

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

$$B^k x_t = x_{t-k}. \quad (2.30)$$

It is clear that we may then rewrite (2.29) as

$$\nabla x_t = (1 - B)x_t, \quad (2.31)$$

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

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t$$

$$= x_t - 2x_{t-1} + x_{t-2}$$

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**

$$\nabla^d = (1 - B)^d, \quad (2.32)$$

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 (2.29) 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 2.4 Differencing Global Temperature

The first difference of the global temperature series, also shown in Figure 2.4, 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 may be white noise, which implies that the global temperature series is a random walk. Finally, notice that removing trend by detrending (i.e., regression techniques) produces different results than removing trend by differencing.

Continuing from Example 2.3, to difference and plot the data in R:

```r
> x = gtemp[44:142]  # start at 1899
> plot(1900:1997, diff(x), type="o", xlab="year")
```

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 (2.32) to fractional powers \(-0.5 < d < 0.5\), which still define stationary processes. Granger and Joyeux (1980) and Hosking (1981) introduced long memory time series, which corresponds to the case when \(0 < d < 0.5\). 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

\[
y_t = \ln x_t,
\]

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

\[
y_t = \begin{cases} 
  (x_t^\lambda - 1)/\lambda, & \lambda \neq 0 \\
  \ln x_t, & \lambda = 0.
\end{cases}
\]

Methods for choosing the power \(\lambda\) 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.

Example 2.5 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
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 (2.31) 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.

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, $x_{t-h}$. 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 $x_t$ on the vertical axis plotted against $x_{t-h}$ on the horizontal axis for the SOI $x_t$. Similarly, we might want to look at values of one series $y_t$ plotted against another series at various...
2.3: Exploratory Data Analysis

Figure 2.8 Scatterplot matrix of the Recruitment series, \( y_t \), on the vertical axis plotted against the SOI series, \( x_{t-h} \), on the horizontal axis at lags \( h = 0, 1, \ldots, 8 \).

Example 2.6 Scatterplot Matrices, SOI, and Recruitment Series

Consider the possibility of looking for nonlinear functional relations at lags in the SOI series, \( x_{t-h} \), for \( h = 0, 1, 2, \ldots \), and the Recruitment series, \( y_t \). Noting first the top panel in Figure 2.7, we see strong posi-
tive and linear relations at lags \( h = 1, 2, 11, 12 \), that is, between \( x_t \) and \( x_{t-1}, x_{t-2}, x_{t-11}, x_{t-12} \), 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. Figure 2.8 shows linearity in relating Recruitment, \( y_t \), with the SOI series at \( x_{t-5}, x_{t-6}, x_{t-7}, x_{t-8} \), 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, and vice versa. Some possible nonlinear behavior shows as the relation tends to flatten out at both extremes, indicating a logistic type transformation may be useful.

To reproduce Figure 2.7 in R assuming the data are in \( \text{soi} \) and \( \text{rec} \) as before:

```r
> lag.plot(soi, lags=12, layout=c(3,4), diag=F)
```

Reproducing Figure 2.8 in R is not as easy, but here is how the figure was generated:

```r
> soi=ts(soi) # make the series
> rec=ts(rec) # time series objects
> par(mfrow=c(3,3), mar=c(2.5, 4, 4, 1)) # set up plot area
> for(h in 0:8){ # loop through lags 0-8
> plot(lag(soi,-h),rec, main=paste("soi(t-",h,")"), ylab="rec(t)",xlab="")
> }
```

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 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 2.7 Using Regression to Discover a Signal in Noise**

Recall, in Example 1.12 we generated \( n = 500 \) observations from the model

\[
x_t = A \cos(2\pi \omega t + \phi) + w_t,
\]

(2.35)

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. At this point we assume the frequency of oscillation \( \omega = 1/50 \) is known, but \( A \) and \( \phi \) are unknown parameters. In
2.3: Exploratory Data Analysis

Figure 2.9 Data generated by (2.35) [dashed line] with the fitted [solid] line, (2.37), superimposed.

This case the parameters appear in (2.35) in a nonlinear way, so we use a trigonometric identity and write

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

where \( \beta_1 = A \cos(\phi) \) and \( \beta_2 = -A \sin(\phi) \). Now the model (2.35) can be written in the usual linear regression form given by (no intercept term is needed here)

\[ x_t = \beta_1 \cos(2\pi t/50) + \beta_2 \sin(2\pi t/50) + w_t. \]  

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

\[ \hat{x}_t = -0.84_{(0.32)} \cos(2\pi t/50) - 1.99_{(0.32)} \sin(2\pi t/50) \]  

with \( \hat{\sigma}_w = 5.08 \), where the values in parentheses are the standard errors. We note the actual values of the coefficients for this example are \( \beta_1 = 2 \cos(0.6\pi) = -0.62 \) and \( \beta_2 = -2 \sin(0.6\pi) = -1.90 \). 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 (2.35) with the fitted line, (2.37), superimposed.

Example 2.8 Using the Periodogram to Discover a Signal in Noise

The analysis in Example 2.7 may seem like cheating because we assumed we knew the value of the frequency parameter \( \omega \). If we do not know \( \omega \),
we could try to fit the model (2.35) using nonlinear regression with $\omega$ as a parameter. Another method is to try various values of $\omega$ in a systematic way. Using the regression results of §2.2 (also, see Problem 4.10), we can show the estimated regression coefficients in Example 2.7 take on the special form\footnote{In the notation of §2.2, the estimates are $\hat{\beta}_1 = \frac{\sum_{t=1}^{n} x_t \cos(2\pi t/50)}{\sum_{t=1}^{n} \cos^2(2\pi t/50)}$ and $\hat{\beta}_2 = \frac{\sum_{t=1}^{n} x_t \sin(2\pi t/50)}{\sum_{t=1}^{n} \sin^2(2\pi t/50)}$. Here, $z_t = \cos(2\pi t/50)$ or $z_t = \sin(2\pi t/50)$.} given by

\begin{align*}
\hat{\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); \\
\hat{\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).
\end{align*}

This suggests looking at all possible regression parameter estimates, say

\begin{align*}
\hat{\beta}_1(j/n) &= \frac{2}{n} \sum_{t=1}^{n} x_t \cos(2\pi t j/n); \\
\hat{\beta}_2(j/n) &= \frac{2}{n} \sum_{t=1}^{n} x_t \sin(2\pi t j/n),
\end{align*}

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

For this particular example, the values calculated in (2.38) and (2.39) are $\hat{\beta}_1(10/500)$ and $\hat{\beta}_2(10/500)$. By doing this, we have regressed a series, $x_t$, of 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 $\hat{\beta}_1(j/n)$ and $\hat{\beta}_2(j/n)$, for each $j$, are essentially measuring the correlation of the data with a sinusoid oscillating at $j$ cycles in $n$ time points.\footnote{In the notation of §2.2, the regression coefficients (2.40) and (2.41) are of the form $\sum_{t} x_t z_t / \sum_{t} z_t^2$ whereas sample correlations are of the form $\sum_{t} x_t z_t / (\sum_{t} x_t^2 \sum_{t} z_t^2)^{1/2}$.} Hence, an appropriate measure of the presence of a frequency of oscillation of $j$ cycles in $n$ time points in the data would be

\begin{equation}
P(j/n) = \hat{\beta}_1^2(j/n) + \hat{\beta}_2^2(j/n),
\end{equation}

which is basically a measure of squared correlation. The quantity (2.42) 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 (2.35), and it
2.3: Exploratory Data Analysis

Figure 2.10 The scaled periodogram, (2.42), of the 500 observations generated by (2.35). The data are displayed in Figures 1.11 and 2.9.

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

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

(2.43)

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 composite integer. There is no error in (2.43) 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

$$d(j/n) = n^{-1/2} \sum_{t=1}^{n} x_t \exp(-2\pi itj/n),$$

(2.44)

and values $j/n$ are called the Fourier or fundamental frequencies. Because of a large number of redundancies in the calculation, (2.44) may be computed quickly using the fast Fourier transform (FFT), which is available in many computing packages such as Matlab, S-PLUS and R. We note that

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

(2.45)

$^5 e^{-i\alpha} = \cos(\alpha) - i \sin(\alpha)$ and if $z = a - ib$, then $|z|^2 = z\overline{z} = (a - ib)(a + ib) = a^2 + b^2$. 
and it is this quantity that is called the periodogram; we will write

\[ I(j/n) = |d(j/n)|^2. \]

So, we may calculate the scaled periodogram, (2.42), using the periodogram as

\[ P(j/n) = \frac{4}{n} I(j/n). \]  \hspace{1cm} (2.46)

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

A figure similar to Figure 2.10 can be created in R using the following commands\(^6\):

```r
> t = 1:500
> x = 2*cos(2*pi*t/50 + .6*pi) + rnorm(500,0,5)
> I = abs(fft(x)/sqrt(500))^2  # the periodogram
> P = (4/500)*I                # the scaled periodogram
> f = 0:250/500
> plot(f, P[1:251], type="l", xlab="frequency", ylab=" ")
> abline(v=seq(0,.5,.02), lty="dotted")
```

**Example 2.9 The Periodogram as a Matchmaker**

Another way of understanding the results of the previous example is to consider the problem of matching the data with sinusoids oscillating at various frequency. For example, Figure 2.11 shows \( n = 100 \) observations (as a solid line) generated by the model

\[ x_t = \cos(2\pi t [2/100]) + \omega_t, \]  \hspace{1cm} (2.47)

where \( \omega_t \) is Gaussian white noise with \( \sigma_\omega = 1 \). Superimposed on \( x_t \) are cosines oscillating at frequency 1/100, 2/100, and 3/100 (shown as dashed lines). Also included in the figure are correlations of \( x_t \) with the particular cosine, \( \cos(2\pi tj/100) \), for \( j = 1, 2, 3 \). Note that the data match up well with the cosine oscillating at 2 cycles every 100 points (with a correlation of .57), whereas the data do not match up well with the other two cosines. For example, in the top panel of Figure 2.11, there is a decreasing trend in the data until observation 25, and then the data start an increasing trend to observation 50, whereas the cosine making one cycle (1/100) continues to decrease until observation 50.

---

\(^6\)Different packages scale the FFT differently; consult the documentation. R calculates (2.44) without scaling by \( n^{-1/2} \).
2.4: Smoothing

2.4.1: 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 trend and seasonal components. In particular, if \( x_t \) represents the observations, then

\[
m_t = \sum_{j=-k}^{k} a_j x_{t-j},
\]

(2.48)

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 2.12 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 =

Figure 2.11 Data generated by (2.47) represented as a solid line with cosines oscillating at various frequencies superimposed (dashed lines). The correlation indicates the degree to which the two series line up.
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 $1/(2k + 1)$.

To reproduce Figure 2.12 in R assuming the mortality series is in `mort`:

```r
> t = 1:length(mort)
> ma5 = filter(mort, sides=2, rep(1,5)/5)
> ma53 = filter(mort, sides=2, rep(1,53)/53)
> plot(t, mort, xlab="week", ylab="mortality")
> lines(ma5)
> lines(ma53)
```

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

$$ x_t = f_t + y_t, $$

(2.49)

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 (2.48), as an estimator of $f_t$. An obvious choice for $f_t$ in (2.49) is polynomial regression

$$ f_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p. $$

(2.50)

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

$$ f_t = a_0 + a_1 \cos(2\pi \omega_1 t) + a_1 \sin(2\pi \omega_1 t) + \cdots + a_p \cos(2\pi \omega_p t) + a_p \sin(2\pi \omega_p t), $$

(2.51)

Sometimes, the end weights, $a_{-k}$ and $a_k$ are set equal to half the value of the other weights.
where $\omega_1, \ldots, \omega_p$ are distinct, specified frequencies. In addition, one might consider combining (2.50) and (2.51). These smoothers can be applied using classical linear regression.

Example 2.11 Polynomial and Periodic Regression Smoothers

Figure 2.13 shows the weekly mortality series with an estimated (via ordinary least squares) cubic smoother

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

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

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

superimposed to emphasize trend and seasonality.

The R commands for this example are:
```
> t = 1:length(mort)
> t2 = t^2
> t3 = t^3
> c = cos(2*pi*t/52)
> s = sin(2*pi*t/52)
> fit1 = lm(mort~t + t^2 + t^3)
> fit2 = lm(mort~t + t^2 + t^3 + c + s)
> plot(t, mort)
> lines(fit1$fit)
> lines(fit2$fit)
```
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 2.14 shows kernel smoothing of the mortality series, where \(f_t\) in (2.49) is estimated by

\[
\hat{f}_t = \sum_{i=1}^{n} w_t(i)x_t, \tag{2.52}
\]

where

\[
w_t(i) = K\left(\frac{t-i}{b}\right) / \sum_{j=1}^{n} K\left(\frac{t-j}{b}\right). \tag{2.53}
\]

This estimator is called the Naradaya–Watson estimator (Watson, 1966). In (2.53), \(K(\cdot)\) is a kernel function; 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 2.14, the values of \(b\) for this example were \(b = 10\) (roughly weighted

![Figure 2.14 Kernel smoothers of the mortality data.](image-url)
monthly averages; that is, \( b/2 \) is the inner quartile range of the kernel) for the seasonal component, and \( b = 104 \) (roughly weighted yearly averages) for the trend component.

Figure 2.14 can be reproduced in R (or S-PLUS) as follows; we assume \( t \) and \( \text{mort} \) are available from the previous example:

```r
> plot(t, mort)
> lines(ksmooth(t, mort, "normal", bandwidth=5))
> lines(ksmooth(t, mort, "normal", bandwidth=104))
```

**Example 2.13 Nearest Neighbor and Locally Weighted 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 \( \hat{f}_t \). For example, Figure 2.15 shows cardiovascular mortality and the nearest neighbor method using the R (or S-PLUS) smoother \texttt{supsmu}. We used \( k = n/2 \) to estimate the trend and \( k = n/100 \) to estimate the seasonal component. In general, \texttt{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 2.15 shows smoothing of mortality using the R or S-PLUS function \texttt{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 \( \hat{f}_t \) will be. In Figure 2.15, 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 2.15 can be reproduced in R or S-PLUS as follows (assuming \( t \) and \( \text{mort} \) are available from the previous example):

```r
> par(mfrow=c(2,1))
> plot(t, mort, main="nearest neighbor")
> lines(supsmu(t, mort, span=.5))
> lines(supsmu(t, mort, span=.01))
> plot(t, mort, main="lowess")
> lines(lowess(t, mort, .02))
> lines(lowess(t, mort, 2/3))
```
Figure 2.15 Nearest neighbor (supsmu) and locally weighted least squares (lowess) smoothers of the mortality data.

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 (2.50); 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

\[
\sum_{i=1}^{n} (x_i - f_i)^2 + \lambda \int (f_i^{''})^2 \, dt,
\]

(2.54)

where \( f_i \) is a cubic spline with a knot at each \( t \). The degree of smoothness is controlled by \( \lambda > 0 \). Figure 2.16 shows smoothing splines on mortality using \( \lambda = 10^{-7} \) for the seasonal component, and \( \lambda = 0.1 \) for the trend.
2.4: Smoothing

Figure 2.16 Smoothing splines fit to the mortality data.

Figure 2.16 can be reproduced in R or S-PLUS as follows (assuming \( t \) and \( mort \) are available from the previous example):

\[
\begin{align*}
&> \text{plot}(t, \text{mort}) \\
&> \text{lines} \left( \text{smooth.spline}(t, \text{mort}, \text{spar}=0.0000001) \right) \\
&> \text{lines} \left( \text{smooth.spline}(t, \text{mort}, \text{spar}=0.1) \right)
\end{align*}
\]

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 2.17 shows scatterplots of mortality, \( M_t \), and temperature, \( T_t \), along with \( M_t \) is smoothed as a function of \( T_t \) using lowess and using smoothing splines. In both cases, mortality increases at extreme 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 2.17 can be reproduced in R or S-PLUS as follows (assuming \( mort \) and \( temp \) contain the mortality and temperature data):

\[
\begin{align*}
&> \text{par(mfrow=c(2,1))} \\
&> \text{plot}(\text{temp, mort, main="lowess"}) \\
&> \text{lines(lowess(temp,mort))} \\
&> \text{plot}(\text{temp, mort, main="smoothing splines"}) \\
&> \text{lines(smooth.spline(temp,mort))}
\end{align*}
\]
As a final word of caution, the methods mentioned above do not particularly take into account the fact that the data are serially correlated, and most of the techniques mentioned have been designed for independent observations. That is, for example, the smoothers shown in Figure 2.17 are calculated under the false assumption that the pairs $(M_t, T_t)$, for $t = 1, \ldots, 508$, are iid pairs of observations. In addition, the degree of smoothness used in the previous examples were chosen arbitrarily to bring out what might be considered obvious features in the data set.