Chapter 2

Time Series Regression and Exploratory Data Analysis

In this chapter we introduce classical multiple linear regression in a time series context, model selection, exploratory data analysis for preprocessing nonstationary time series (for example trend removal), the concept of differencing and the backshift operator, variance stabilization, and nonparametric smoothing of time series.

2.1 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*

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \cdots + \beta_q z_{tq} + w_t, \tag{2.1}$$

where $\beta_0, \beta_1, \ldots, \beta_q$ 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. For time series regression, it is rarely the case that the noise is white, and we will need to eventually relax that assumption. 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 monthly price (per pound) of a chicken in the US from mid-2001 to mid-2016 (180 months), say x_t, shown in Figure 2.1. There is an obvious upward trend in the series, and we might use simple linear regression to estimate that trend by fitting the model

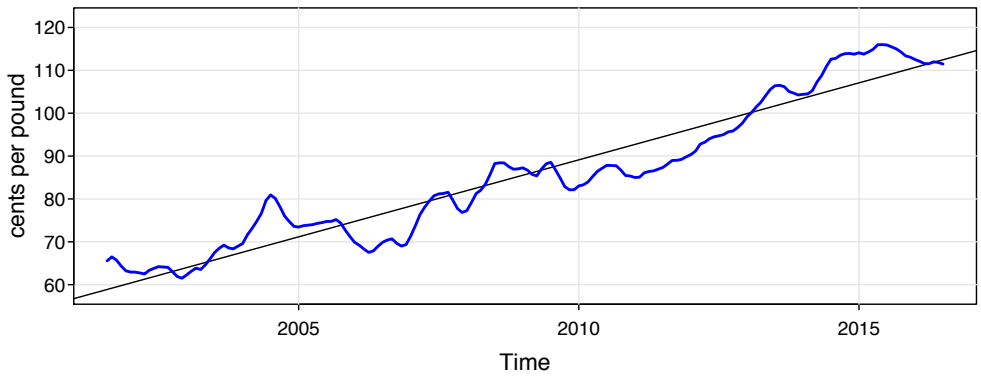

Fig. 2.1. *The price of chicken: monthly whole bird spot price, Georgia docks, US cents per pound, August 2001 to July 2016, with fitted linear trend line.*

$$x_t = \beta_0 + \beta_1 z_t + w_t, \quad z_t = 2001\tfrac{7}{12}, 2001\tfrac{8}{12}, \ldots, 2016\tfrac{6}{12}.$$

This is in the form of the regression model (2.1) with $q = 1$. Note that we are making the assumption that the errors, w_t, are an iid normal sequence, which may not be true; the problem of autocorrelated errors is discussed in detail in Chapter 3.

In ordinary least squares (OLS), we minimize the error sum of squares

$$Q = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} (x_t - [\beta_0 + \beta_1 z_t])^2$$

with respect to β_i for $i = 0, 1$. In this case we can use simple calculus to evaluate $\partial Q/\partial \beta_i = 0$ for $i = 0, 1$, to obtain two equations to solve for the βs. The OLS estimates of the coefficients are explicit and given by

$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(z_t - \bar{z})}{\sum_{t=1}^{n}(z_t - \bar{z})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{x} - \hat{\beta}_1 \bar{z},$$

where $\bar{x} = \sum_t x_t/n$ and $\bar{z} = \sum_t z_t/n$ are the respective sample means.

Using R, we obtained the estimated slope coefficient of $\hat{\beta}_1 = 3.59$ (with a standard error of .08) yielding a significant estimated increase of about 3.6 cents per year. Finally, Figure 2.1 shows the data with the estimated trend line superimposed. R code with partial output:

```
summary(fit <- lm(chicken~time(chicken), na.action=NULL))
                Estimate  Std.Error t.value
  (Intercept)   -7131.02     162.41   -43.9
  time(chicken)     3.59       0.08    44.4
  --
  Residual standard error: 4.7 on 178 degrees of freedom
plot(chicken, ylab="cents per pound")
abline(fit)           # add the fitted line
```

The multiple linear regression model described by (2.1) can be conveniently written in a more general notation by defining the column vectors $\mathbf{z}_t = (1, z_{t1}, z_{t2}, \ldots, z_{tq})'$

and $\beta = (\beta_0, \beta_1, \ldots, \beta_q)'$, where $'$ denotes transpose, so (2.1) can be written in the alternate form

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_q z_{tq} + w_t = \beta' z_t + w_t. \tag{2.2}$$

where $w_t \sim$ iid $\mathrm{N}(0, \sigma_w^2)$. As in the previous example, OLS estimation finds the coefficient vector β that minimizes the error sum of squares

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

with respect to $\beta_0, \beta_1, \ldots, \beta_q$. This minimization can be accomplished by differentiating (2.3) with respect to the vector β or by using the properties of projections. Either way, the solution must satisfy $\sum_{t=1}^{n} (x_t - \hat{\beta}' z_t) z_t' = 0$. This procedure gives the *normal equations*

$$\left(\sum_{t=1}^{n} z_t z_t' \right) \hat{\beta} = \sum_{t=1}^{n} z_t x_t. \tag{2.4}$$

If $\sum_{t=1}^{n} z_t z_t'$ is non-singular, the least squares estimate of β is

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

The minimized error sum of squares (2.3), denoted SSE, can be written as

$$SSE = \sum_{t=1}^{n} (x_t - \hat{\beta}' z_t)^2. \tag{2.5}$$

The ordinary least squares estimators are unbiased, i.e., $\mathrm{E}(\hat{\beta}) = \beta$, and have the smallest variance within the class of linear unbiased estimators.

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

$$\mathrm{cov}(\hat{\beta}) = \sigma_w^2 C, \tag{2.6}$$

where

$$C = \left(\sum_{t=1}^{n} z_t z_t' \right)^{-1} \tag{2.7}$$

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

$$s_w^2 = MSE = \frac{SSE}{n - (q+1)}, \tag{2.8}$$

where MSE denotes the *mean squared error*. Under the normal assumption,

$$\mathrm{t} = \frac{(\hat{\beta}_i - \beta_i)}{s_w \sqrt{c_{ii}}} \tag{2.9}$$

Table 2.1. *Analysis of Variance for Regression*

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

has the t-distribution with $n-(q+1)$ degrees of freedom; c_{ii} denotes the i-th diagonal element of C, as defined in (2.7). This result is often used for individual tests of the null hypothesis $H_0: \beta_i = 0$ for $i = 1, \ldots, q$.

Various competing models are often 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, $z_{t,1:r} = \{z_{t1}, z_{t2}, \ldots, z_{tr}\}$ is influencing the dependent variable x_t. The reduced model is

$$x_t = \beta_0 + \beta_1 z_{t1} + \cdots + \beta_r z_{tr} + w_t \tag{2.10}$$

where $\beta_1, \beta_2, \ldots, \beta_r$ are a subset of coefficients of the original q variables.

The null hypothesis in this case is $H_0: \beta_{r+1} = \cdots = \beta_q = 0$. We can test the reduced model (2.10) against the full model (2.2) by comparing the error sums of squares under the two models using the F-statistic

$$F = \frac{(SSE_r - SSE)/(q-r)}{SSE/(n-q-1)} = \frac{MSR}{MSE}, \tag{2.11}$$

where SSE_r is the error sum of squares under the reduced model (2.10). Note that $SSE_r \geq SSE$ because the full model has more parameters. If $H_0: \beta_{r+1} = \cdots = \beta_q = 0$ is true, then $SSE_r \approx SSE$ because the estimates of those βs will be close to 0. Hence, we do not believe H_0 if $SSR = SSE_r - SSE$ is big. Under the null hypothesis, (2.11) has a central F-distribution with $q-r$ and $n-q-1$ degrees of freedom when (2.10) is the correct model.

These results are 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 (SSR). The null hypothesis is rejected at level α if $F > F^{q-r}_{n-q-1}(\alpha)$, the $1-\alpha$ percentile of the F distribution with $q-r$ numerator and $n-q-1$ denominator degrees of freedom.

A special case of interest is the null hypothesis $H_0: \beta_1 = \cdots = \beta_q = 0$. In this case $r = 0$, and the model in (2.10) becomes

$$x_t = \beta_0 + w_t \,.$$

We may measure the proportion of variation accounted for by all the variables using

$$R^2 = \frac{SSE_0 - SSE}{SSE_0}, \tag{2.12}$$

where the residual sum of squares under the reduced model is

$$SSE_0 = \sum_{t=1}^{n} (x_t - \bar{x})^2. \tag{2.13}$$

In this case SSE_0 is the sum of squared deviations from the mean $\bar{x}$ and is otherwise known as the adjusted total sum of squares. The measure R^2 is called the *coefficient of determination*.

The techniques discussed in the previous paragraph can be used to test various models against one another using the F test given in (2.11). 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

$$\hat{\sigma}_k^2 = \frac{SSE(k)}{n}, \tag{2.14}$$

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.[2.1]

Definition 2.1 Akaike's Information Criterion (AIC)

$$\text{AIC} = \log\, \hat{\sigma}_k^2 + \frac{n + 2k}{n}, \tag{2.15}$$

where $\hat{\sigma}_k^2$ is given by (2.14) 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 $\hat{\sigma}_k^2$ 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 (2.15) 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 Problem 2.4 and Problem 2.5). The corrected form is defined as follows.

Definition 2.2 AIC, Bias Corrected (AICc)

$$\text{AICc} = \log\, \hat{\sigma}_k^2 + \frac{n + k}{n - k - 2}, \tag{2.16}$$

[2.1] 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 (2.15). 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.

where $\hat{\sigma}_k^2$ is given by (2.14), 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)

$$\text{BIC} = \log\ \hat{\sigma}_k^2 + \frac{k \log n}{n}, \tag{2.17}$$

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. Notice that the penalty term in BIC is much larger than in AIC, consequently, BIC tends to choose smaller models. 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

$$M_t = \beta_0 + \beta_1 t + w_t \tag{2.18}$$

$$M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + w_t \tag{2.19}$$

$$M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + \beta_3 (T_t - T.)^2 + w_t \tag{2.20}$$

$$M_t = \beta_0 + \beta_1 t + \beta_2 (T_t - T.) + \beta_3 (T_t - T.)^2 + \beta_4 P_t + w_t \tag{2.21}$$

where we adjust temperature for its mean, $T. = 74.26$, to avoid collinearity problems. It is clear that (2.18) is a trend only model, (2.19) is linear temperature, (2.20)

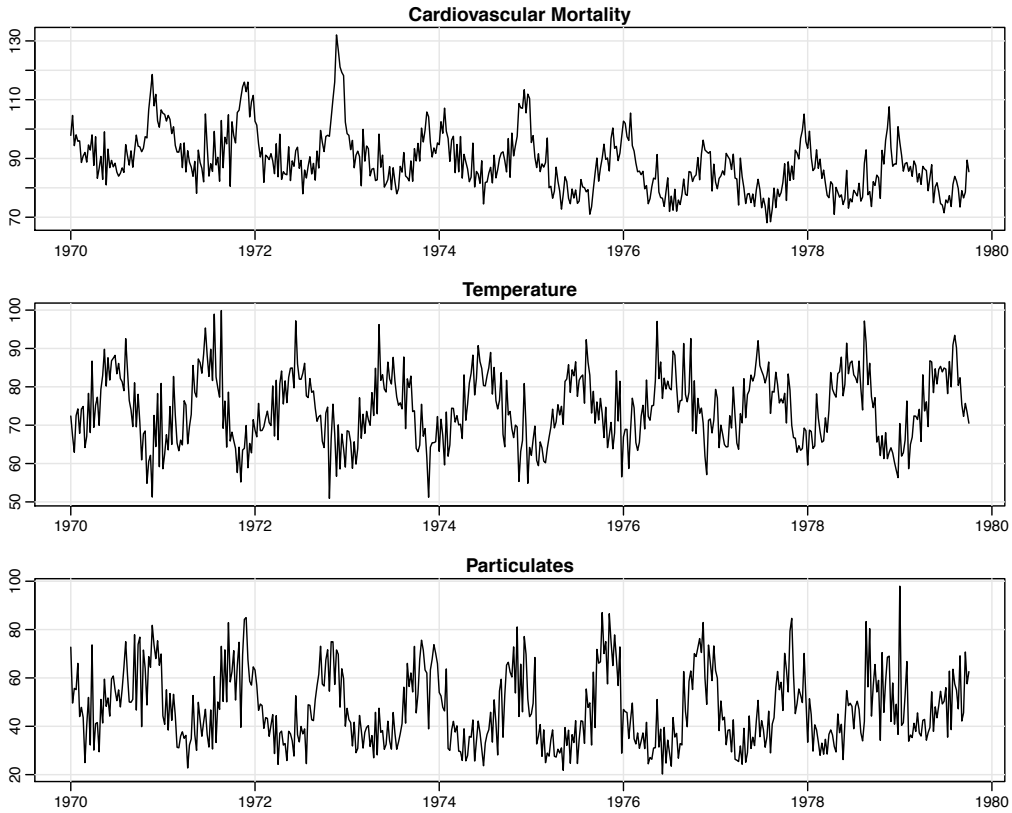

Fig. 2.2. *Average weekly cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.*

Table 2.2. *Summary Statistics for Mortality Models*

Model	k	SSE	df	MSE	R^2	AIC	BIC
(2.18)	2	40,020	506	79.0	.21	5.38	5.40
(2.19)	3	31,413	505	62.2	.38	5.14	5.17
(2.20)	4	27,985	504	55.5	.45	5.03	5.07
(2.21)	5	20,508	503	40.8	.60	4.72	4.77

is curvilinear temperature and (2.21) is curvilinear temperature and pollution. We summarize some of the statistics given for this particular case in Table 2.2.

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 are nearly the same). Note that one can compare any two models using the residual sums of squares and (2.11). Hence, a model with only trend could be compared to the full model, $H_0: \beta_2 = \beta_3 = \beta_4 = 0$, using $q = 4, r = 1, n = 508$, and

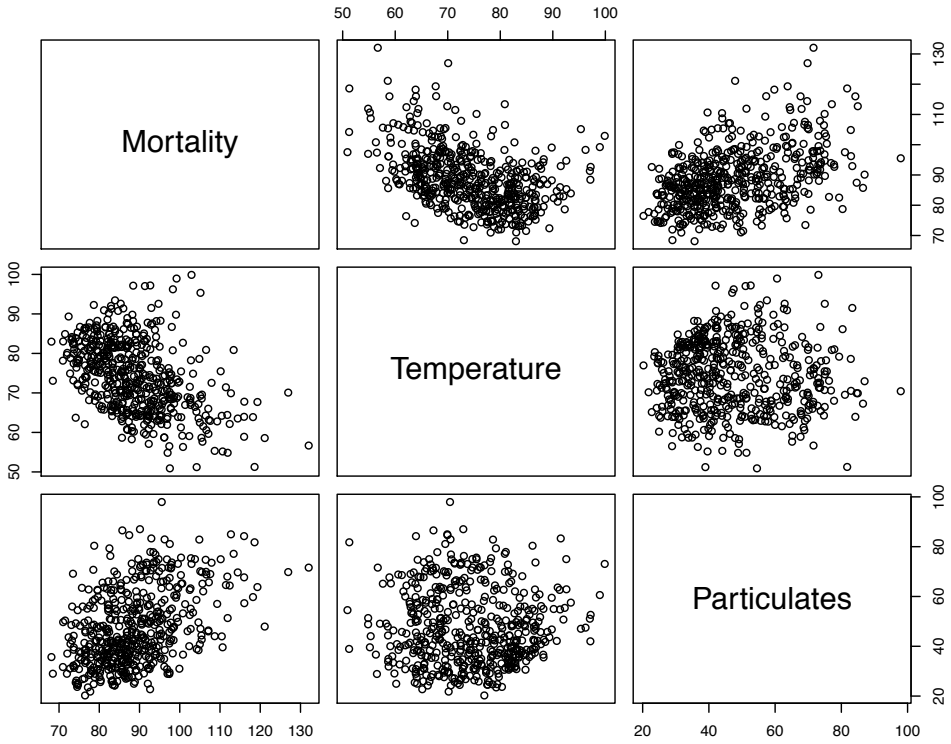

Fig. 2.3. *Scatterplot matrix showing relations between mortality, temperature, and pollution.*

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

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

$$\hat{M}_t = 2831.5 - 1.396_{(.10)}t - .472_{(.032)}(T_t - 74.26) + .023_{(.003)}(T_t - 74.26)^2 + .255_{(.019)}P_t,$$

for mortality, where the standard errors, computed from (2.6)–(2.8), 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 $\hat{w}_t = M_t - \hat{M}_t$ for autocorrelation (of which there is a substantial amount), but we defer this question to Section 3.8 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 (2.21), and compute the corresponding values of AIC, AICc and BIC.[2.2] Finally, the use of `na.action` in `lm()` is to retain the time series attributes for the residuals and fitted values.

[2.2] 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 (2.15) and (2.17) from the R output. It is more difficult to obtain AICc.

```
par(mfrow=c(3,1))  # plot the data
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
ts.plot(cmort,tempr,part, col=1:3)  # all on same plot (not shown)
dev.new()
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 2.10. The following is a simple example of lagged regression.

Example 2.3 Regression With Lagged Variables

In Example 1.28, 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 2.8 and Example 2.9), consider the following regression,

$$R_t = \beta_0 + \beta_1 S_{t-6} + w_t, \tag{2.22}$$

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

$$\hat{R}_t = 65.79 - 44.28_{(2.78)} S_{t-6} \tag{2.23}$$

with $\hat{\sigma}_w = 22.5$ 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 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(fit1 <- lm(rec~soiL6, data=fish, na.action=NULL))
```

The headache of aligning the lagged series can be avoided by using the R package `dynlm`, which must be downloaded and installed.

```
library(dynlm)
summary(fit2 <- dynlm(rec~ L(soi,6)))
```

We note that `fit2` is similar to the `fit1` object, but the time series attributes are retained without any additional commands.

2.2 Exploratory Data Analysis

In general, it is necessary for time series data to be stationary so that 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 the 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

$$x_t = \mu_t + y_t \tag{2.24}$$

where x_t are the observations, μ_t denotes the trend, and y_t is a stationary process. Quite often, strong trend 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 $\hat{\mu}_t$, and then work with the residuals

$$\hat{y}_t = x_t - \hat{\mu}_t. \tag{2.25}$$

Example 2.4 Detrending Chicken Prices

Here we suppose the model is of the form of (2.24),

$$x_t = \mu_t + y_t,$$

where, as we suggested in the analysis of the chicken price data presented in Example 2.1, a straight line might be useful for detrending the data; i.e.,

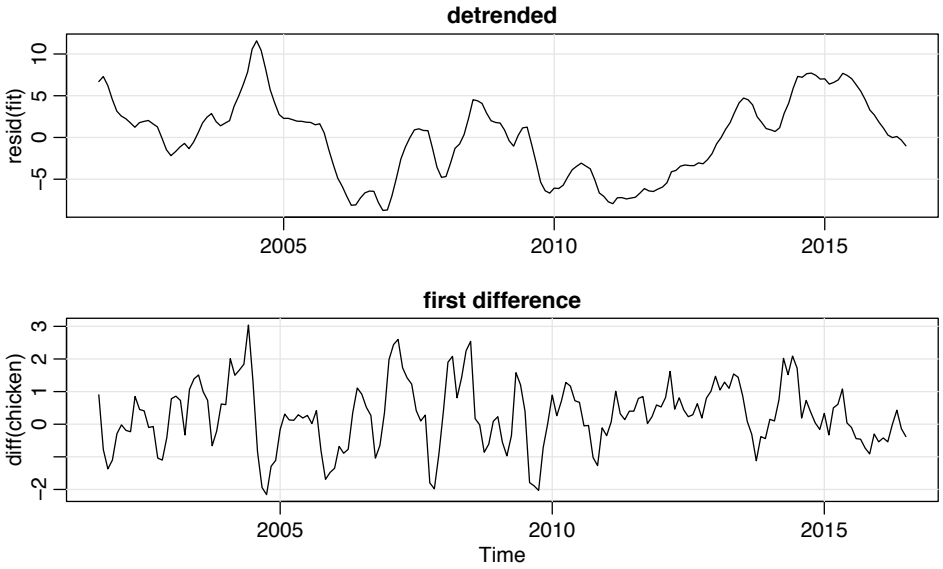

Fig. 2.4. *Detrended (top) and differenced (bottom) chicken price series. The original data are shown in Figure 2.1.*

$$\mu_t = \beta_0 + \beta_1 t.$$

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

$$\hat{\mu}_t = -7131 + 3.59\, t$$

where we are using t instead of z_t for time. 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[2.3]

$$\hat{y}_t = x_t + 7131 - 3.59\, t.$$

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).

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,

$$\mu_t = \delta + \mu_{t-1} + w_t, \tag{2.26}$$

where w_t is white noise and is independent of y_t. If the appropriate model is (2.24), then *differencing* the data, x_t, yields a stationary process; that is,

[2.3] 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 with regard to efficiency.

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) \tag{2.27}$$
$$= \delta + w_t + y_t - y_{t-1}.$$

It is easy to show $z_t = y_t - y_{t-1}$ is stationary using Property 1.1. That is, because y_t is stationary,

$$\gamma_z(h) = \text{cov}(z_{t+h}, z_t) = \text{cov}(y_{t+h} - y_{t+h-1}, y_t - y_{t-1})$$
$$= 2\gamma_y(h) - \gamma_y(h+1) - \gamma_y(h-1)$$

is independent of time; we leave it as an exercise (Problem 2.7) to show that $x_t - x_{t-1}$ in (2.27) 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 (2.27). 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 $\mu_t = \beta_0 + \beta_1 t$ in the model (2.24), differencing the data produces stationarity (see Problem 2.6):

$$x_t - x_{t-1} = (\mu_t + y_t) - (\mu_{t-1} + y_{t-1}) = \beta_1 + y_t - y_{t-1}.$$

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

$$\nabla x_t = x_t - x_{t-1}. \tag{2.28}$$

As we have seen, the first difference eliminates a linear trend. A second difference, that is, the difference of (2.28), 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,*

$$B^k x_t = x_{t-k}. \tag{2.29}$$

The idea of an inverse operator can also be given if we require $B^{-1}B = 1$, so that

$$x_t = B^{-1}Bx_t = B^{-1}x_{t-1}.$$

That is, B^{-1} is the *forward-shift operator*. In addition, it is clear that we may rewrite (2.28) as

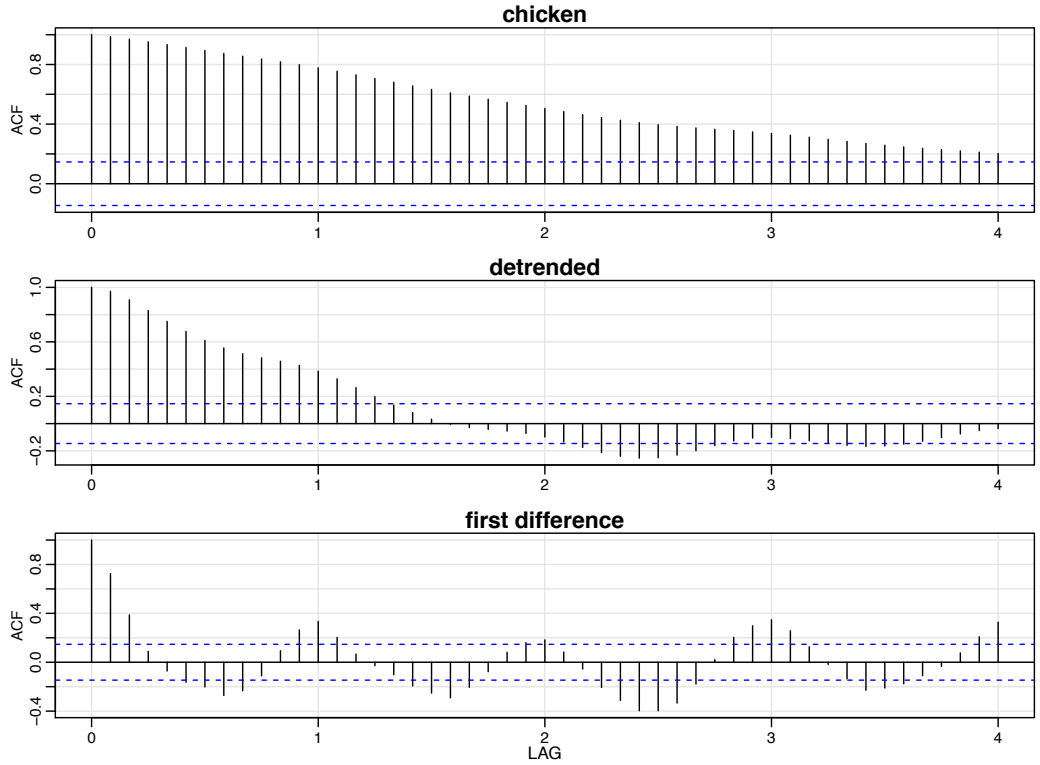

Fig. 2.5. *Sample ACFs of chicken prices (top), and of the detrended (middle) and the differenced (bottom) series. Compare the top plot with the sample ACF of a straight line:* `acf(1:100)`.

$$\nabla x_t = (1 - B)x_t, \tag{2.30}$$

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

$$\nabla^2 x_t = (1 - B)^2 x_t = (1 - 2B + B^2)x_t = x_t - 2x_{t-1} + x_{t-2} \tag{2.31}$$

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*

$$\nabla^d = (1 - B)^d, \tag{2.32}$$

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 (2.28) 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 4. 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 Chicken Prices

The first difference of the chicken prices 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 (five-year) cycle we observe in the detrended series. The ACF of this series is also shown in Figure 2.5. In this case, the differenced series exhibits an annual cycle that was obscured in the original or detrended data.

The R code to reproduce Figure 2.4 and Figure 2.5 is as follows.

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

Example 2.6 Differencing Global Temperature

The global temperature series shown in Figure 1.2 appears to behave more as a random walk than a trend stationary series. Hence, rather than detrend the data, it would be more appropriate to use differencing to coerce it into stationarity. The detreded data are shown in Figure 2.6 along with the corresponding sample ACF. 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 .008, or an increase of about one degree centigrade per 100 years.

The R code to reproduce Figure 2.4 and Figure 2.5 is as follows.

```
par(mfrow=c(2,1))
plot(diff(globtemp), type="o")
 mean(diff(globtemp))  # drift estimate = .008
acf(diff(gtemp), 48)
```

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 $-.5 < d < .5$, 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 Section 5.1.

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

$$y_t = \log x_t, \tag{2.33}$$

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

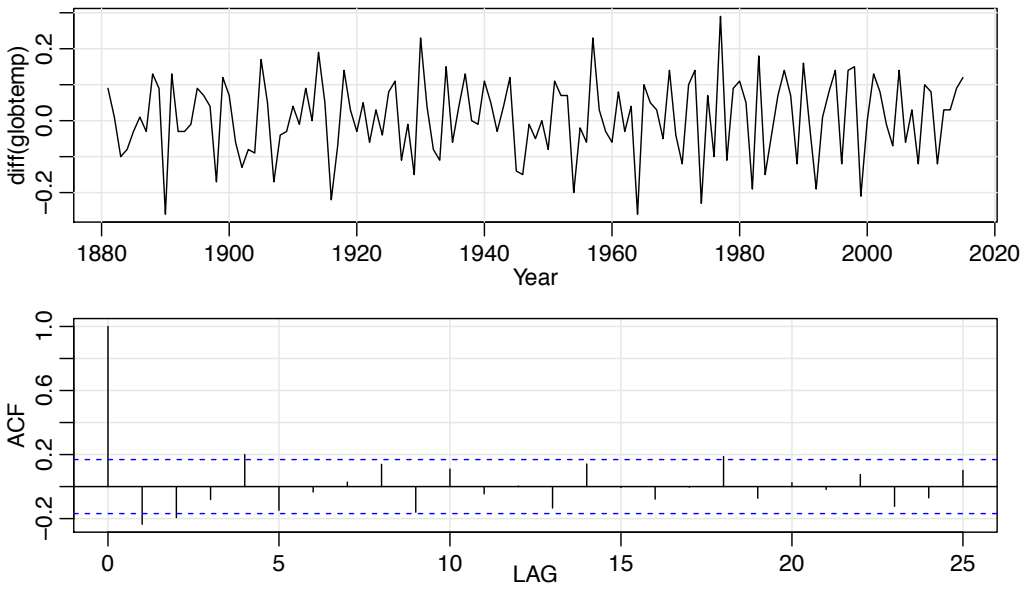

Fig. 2.6. *Differenced global temperature series and its sample ACF.*

Box–Cox family of the form

$$y_t = \begin{cases} (x_t^\lambda - 1)/\lambda & \lambda \neq 0, \\ \log x_t & \lambda = 0. \end{cases} \tag{2.34}$$

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.

Example 2.7 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 2.7 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 2.7 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 (2.30) are also computed in Problem 2.8, and we note that the first differences have

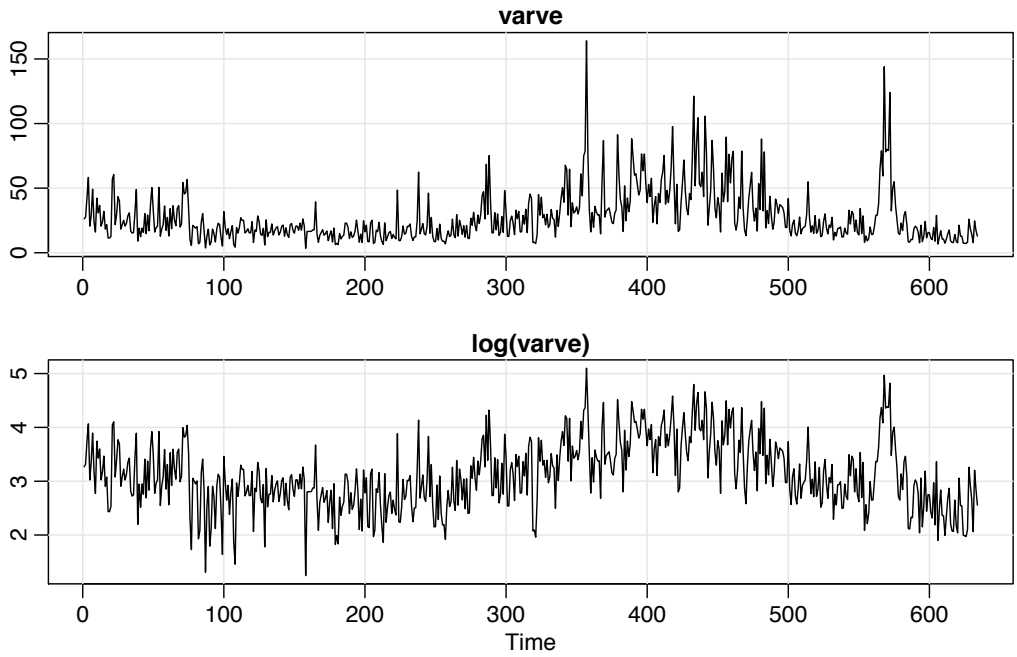

Fig. 2.7. *Glacial varve thicknesses (top) from Massachusetts for* $n = 634$ *years compared with* log *transformed thicknesses (bottom).*

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 2.7 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 2.8 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 2.8, 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

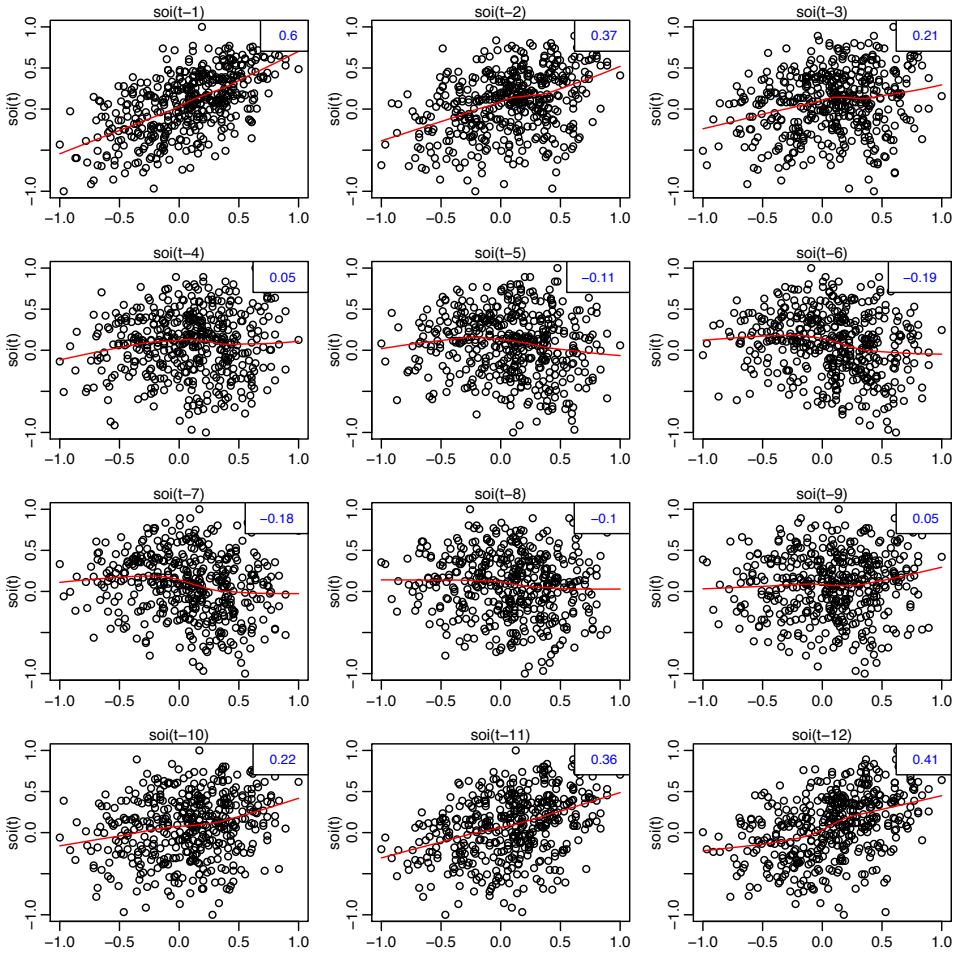

Fig. 2.8. *Scatterplot matrix relating current SOI values, S_t, to past SOI values, S_{t-h}, at lags $h = 1, 2, \ldots, 12$. The values in the upper right corner are the sample autocorrelations and the lines are a lowess fit.*

discover any nonlinearities. We discuss smoothing in the next section, but for now, think of lowess as a robust method for fitting local regression.

In Figure 2.8, 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 1.16.

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 2.9 shows the lagged scatterplot of the Recruitment series R_t on the

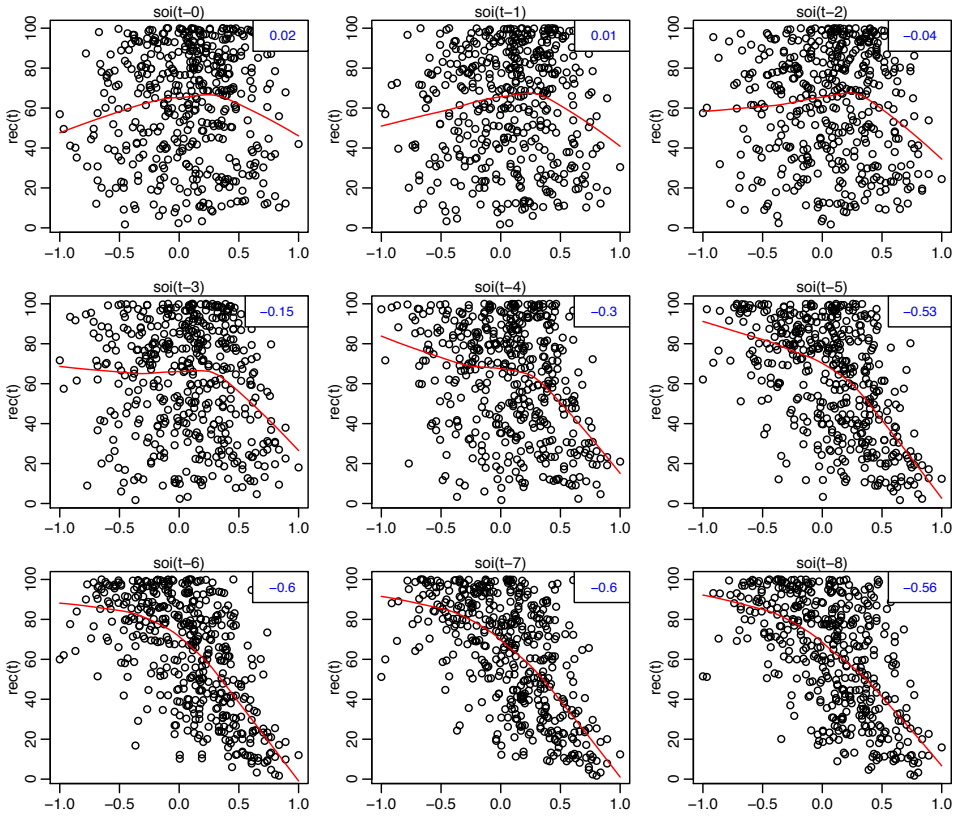

Fig. 2.9. *Scatterplot matrix of the Recruitment series, R_t, on the vertical axis plotted against the SOI series, S_{t-h}, on the horizontal axis at lags $h = 0, 1, \ldots, 8$. The values in the upper right corner are the sample cross-correlations and the lines are a lowess fit.*

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 2.9 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 nonlinearity observed in the scatterplots (with the help of the superimposed lowess fits) indicates 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 2.8 and Figure 2.9 may be reproduced using the following scripts provided with `astsa`:

```
lag1.plot(soi, 12)      # Figure 2.8
lag2.plot(soi, rec, 8)  # Figure 2.9
```

Example 2.9 Regression with Lagged Variables (cont)

In Example 2.3 we regressed Recruitment on lagged SOI,

$$R_t = \beta_0 + \beta_1 S_{t-6} + w_t.$$

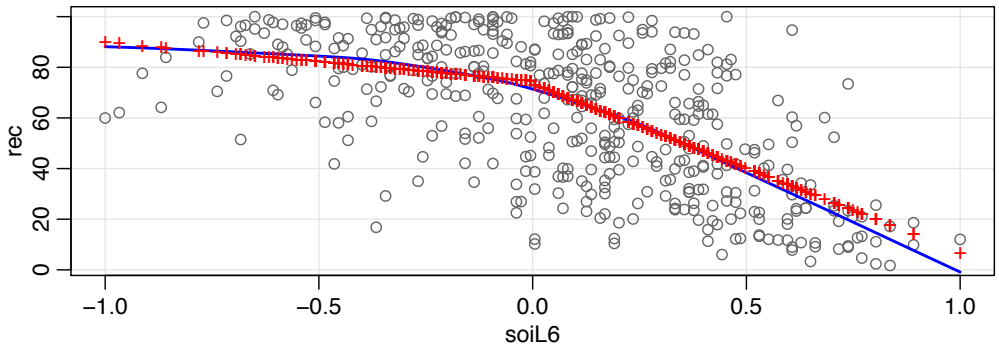

Fig. 2.10. *Display for Example 2.9: Plot of Recruitment (R_t) vs SOI lagged 6 months (S_{t-6}) with the fitted values of the regression as points (+) and a lowess fit (—).*

However, in Example 2.8, we saw that the relationship is nonlinear and different when SOI is positive or negative. In this case, we may consider adding a dummy variable to account for this change. In particular, we fit the model

$$R_t = \beta_0 + \beta_1 S_{t-6} + \beta_2 D_{t-6} + \beta_3 D_{t-6} S_{t-6} + w_t,$$

where D_t is a dummy variable that is 0 if $S_t < 0$ and 1 otherwise. This means that

$$R_t = \begin{cases} \beta_0 + \beta_1 S_{t-6} + w_t & \text{if } S_{t-6} < 0, \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) S_{t-6} + w_t & \text{if } S_{t-6} \geq 0. \end{cases}$$

The result of the fit is given in the R code below. Figure 2.10 shows R_t vs S_{t-6} with the fitted values of the regression and a lowess fit superimposed. The piecewise regression fit is similar to the lowess fit, but we note that the residuals are not white noise (see the code below). This is followed up in Example 3.45.

```
dummy = ifelse(soi<0, 0, 1)
fish = ts.intersect(rec, soiL6=lag(soi,-6), dL6=lag(dummy,-6), dframe=TRUE)
summary(fit <- lm(rec~ soiL6*dL6, data=fish, na.action=NULL))
  Coefficients:
              Estimate  Std.Error  t.value
  (Intercept)   74.479      2.865   25.998
  soiL6        -15.358      7.401   -2.075
  dL6           -1.139      3.711   -0.307
  soiL6:dL6    -51.244      9.523   -5.381
  ---
  Residual standard error: 21.84 on 443 degrees of freedom
  Multiple R-squared:  0.4024
  F-statistic: 99.43 on 3 and 443 DF
attach(fish)
plot(soiL6, rec)
lines(lowess(soiL6, rec), col=4, lwd=2)
points(soiL6, fitted(fit), pch='+', col=2)
plot(resid(fit))  # not shown ...
acf(resid(fit))   # ... but obviously not noise
```

As a final exploratory tool, we discuss assessing periodic behavior in time series data using regression analysis. 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. 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, which obscures the slower El Niño cycle.

Example 2.10 Using Regression to Discover a Signal in Noise

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

$$x_t = A\cos(2\pi\omega t + \phi) + w_t, \tag{2.35}$$

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 ϕ are unknown parameters. In this case the parameters appear in (2.35) in a nonlinear way, so we use a trigonometric identity[2.4] 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 (2.35) can be written in the usual linear regression form given by (no intercept term is needed here)

$$x_t = \beta_1\cos(2\pi t/50) + \beta_2\sin(2\pi t/50) + w_t. \tag{2.36}$$

Using linear regression, we find $\hat{\beta}_1 = -.74_{(.33)}$, $\hat{\beta}_2 = -1.99_{(.33)}$ with $\hat{\sigma}_w = 5.18$; 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 $\beta_2 = -2\sin(.6\pi) = -1.90$. It is clear that we are able to detect the signal in the noise using regression, even though the signal-to-noise ratio is small. Figure 2.11 shows data generated by (2.35) with the fitted line superimposed.

To reproduce the analysis and Figure 2.11 in R, use the following:

```
set.seed(90210)           # 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
  Coefficients:
     Estimate Std. Error t value
  z1  -0.7442     0.3274  -2.273
  z2  -1.9949     0.3274  -6.093
  Residual standard error: 5.177 on 498 degrees of freedom
par(mfrow=c(2,1))
plot.ts(x)
plot.ts(x, col=8, ylab=expression(hat(x)))
lines(fitted(fit), col=2)
```

We will discuss this and related approaches in more detail in Chapter 4.

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

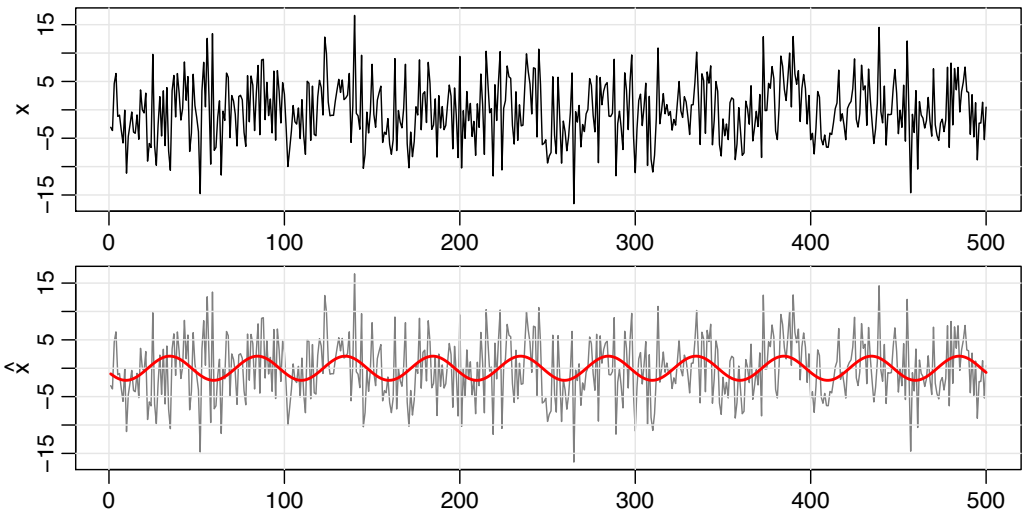

Fig. 2.11. *Data generated by (2.35) [top] and the fitted line superimposed on the data [bottom].*

2.3 Smoothing in the Time Series Context

In Section 1.2, we introduced the concept of filtering or 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

$$m_t = \sum_{j=-k}^{k} a_j x_{t-j}, \tag{2.37}$$

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

Example 2.11 Moving Average Smoother

For example, Figure 2.12 shows the monthly SOI series discussed in Example 1.5 smoothed using (2.37) with weights $a_0 = a_{\pm 1} = \cdots = a_{\pm 5} = 1/12$, and $a_{\pm 6} = 1/24$; $k = 6$. This particular method removes (filters out) the obvious annual temperature cycle and helps emphasize the El Niño cycle. To reproduce Figure 2.12 in R:

```
wgts = c(.5, rep(1,11), .5)/12
soif = filter(soi, sides=2, filter=wgts)
plot(soi)
lines(soif, lwd=2, col=4)
par(fig = c(.65, 1, .65, 1), new = TRUE)  # the insert
nwgts = c(rep(0,20), wgts, rep(0,20))
plot(nwgts, type="l", ylim = c(-.02,.1), xaxt='n', yaxt='n', ann=FALSE)
```

Although the moving average smoother does a good job in highlighting the El Niño effect, it might be considered too choppy. We can obtain a smoother fit using the normal distribution for the weights, instead of boxcar-type weights of (2.37).

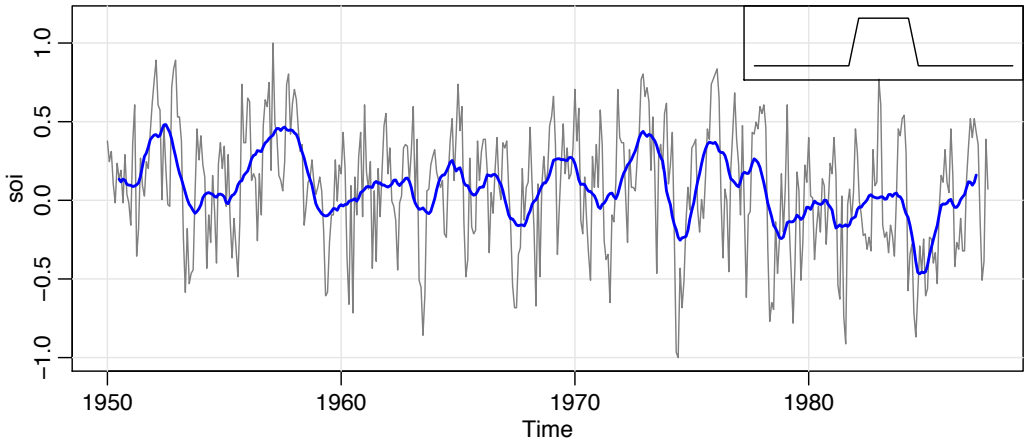

Fig. 2.12. *Moving average smoother of SOI. The insert shows the shape of the moving average ("boxcar") kernel [not drawn to scale] described in* (2.39).

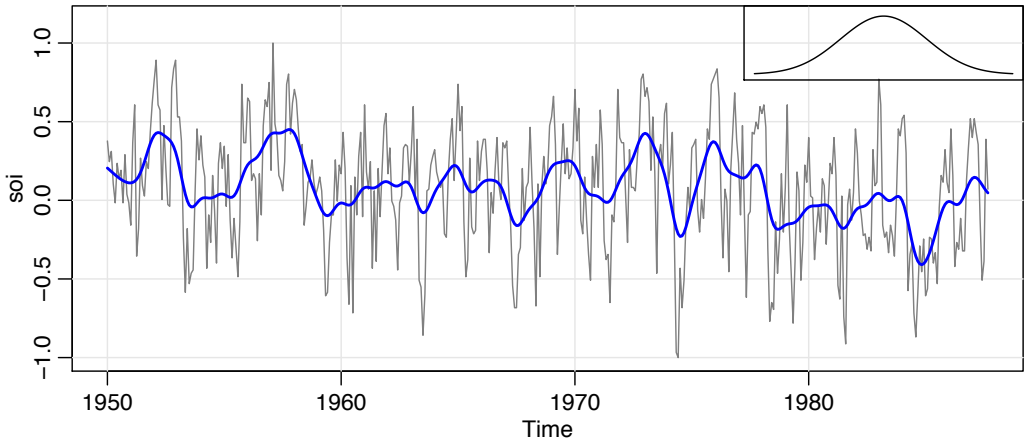

Fig. 2.13. *Kernel smoother of SOI. The insert shows the shape of the normal kernel [not drawn to scale].*

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 SOI series, where m_t is now

$$m_t = \sum_{i=1}^{n} w_i(t) x_i, \tag{2.38}$$

where

$$w_i(t) = K\left(\tfrac{t-i}{b}\right) \Big/ \textstyle\sum_{j=1}^{n} K\left(\tfrac{t-j}{b}\right) \tag{2.39}$$

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). In this example, and typically, the normal kernel, $K(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$, is used.

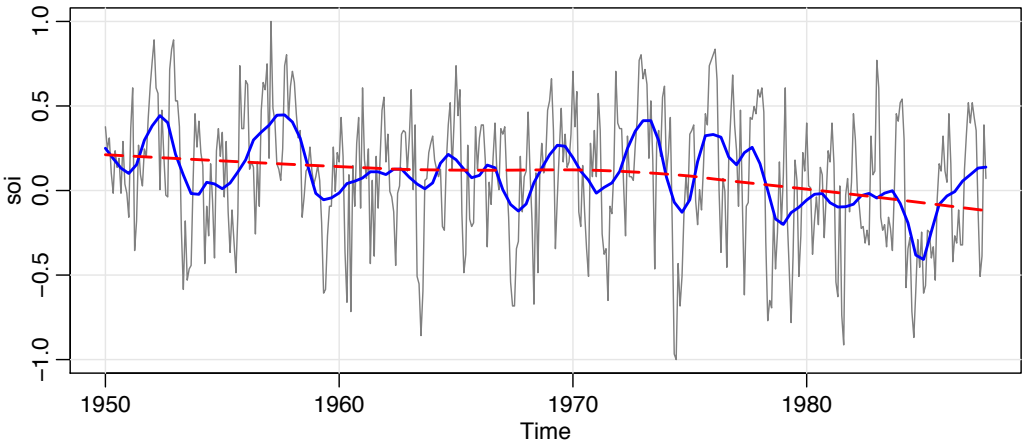

Fig. 2.14. *Locally weighted scatterplot smoothers (`lowess`) of the SOI series.*

To implement this in R, use the `ksmooth` function where a bandwidth can be chosen. The wider the bandwidth, b, the smoother the result. From the R `ksmooth` help file: The kernels are scaled so that their quartiles (viewed as probability densities) are at $\pm$ 0.25$*$bandwidth. For the standard normal distribution, the quartiles are $\pm.674$. In our case, we are smoothing over time, which is of the form $t/12$ for the SOI time series. In Figure 2.13, we used the value of $b = 1$ to correspond to approximately smoothing a little over one year. Figure 2.13 can be reproduced in R as follows.

```
plot(soi)
lines(ksmooth(time(soi), soi, "normal", bandwidth=1), lwd=2, col=4)
par(fig = c(.65, 1, .65, 1), new = TRUE)  # the insert
gauss = function(x) { 1/sqrt(2*pi) * exp(-(x^2)/2) }
x = seq(from = -3, to = 3, by = 0.001)
plot(x, gauss(x), type ="l", ylim=c(-.02,.45), xaxt='n', yaxt='n', ann=FALSE)
```

Example 2.13 Lowess

Another approach to smoothing a time plot is nearest neighbor regression. The technique is based on k-nearest neighbors regression, wherein one uses only the data $\{x_{t-k/2}, \ldots, x_t, \ldots, x_{t+k/2}\}$ to predict x_t via regression, and then sets $m_t = \hat{x}_t$.

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 SOI using the R 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 values m_t. The larger the fraction of nearest neighbors included, the smoother the fit will be. In Figure 2.14, one smoother uses 5% of the data to obtain an estimate of the El Niño cycle of the data.

In addition, a (negative) trend in SOI would indicate the long-term warming of the Pacific Ocean. To investigate this, we used lowess with the default smoother span of `f=2/3` of the data. Figure 2.14 can be reproduced in R as follows.

```
plot(soi)
lines(lowess(soi, f=.05), lwd=2, col=4)  # El Nino cycle
```

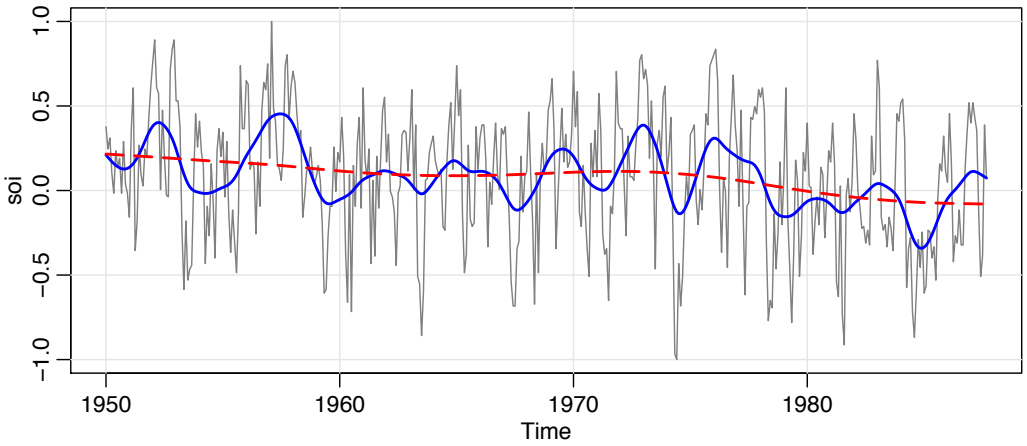

Fig. 2.15. *Smoothing splines fit to the SOI series.*

```
lines(lowess(soi), lty=2, lwd=2, col=2)  # trend (with default span)
```

Example 2.14 Smoothing Splines

An obvious way to smooth data would be to fit a polynomial regression in terms of time. For example, a cubic polynomial would have $x_t = m_t + w_t$ where

$$m_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3.$$

We could then fit m_t via ordinary least squares.

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 polynomial regression, typically the order is 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

$$\sum_{t=1}^{n} [x_t - m_t]^2 + \lambda \int \left(m_t''\right)^2 dt, \tag{2.40}$$

where m_t is a cubic spline with a knot at each t and primes denote differentiation. The degree of smoothness is controlled by $\lambda > 0$.

Think of taking a long drive where m_t is the position of your car at time t. In this case, m_t'' is instantaneous acceleration/deceleration, and $\int (m_t'')^2 dt$ is a measure of the total amount of acceleration and deceleration on your trip. A smooth drive would be one where a constant velocity, is maintained (i.e., $m_t'' = 0$). A choppy ride would be when the driver is constantly accelerating and decelerating, such as beginning drivers tend to do.

If $\lambda = 0$, we don't care how choppy the ride is, and this leads to $m_t = x_t$, the data, which are not smooth. If $\lambda = \infty$, we insist on no acceleration or deceleration ($m_t'' = 0$); in this case, our drive must be at constant velocity, $m_t = c + vt$, and

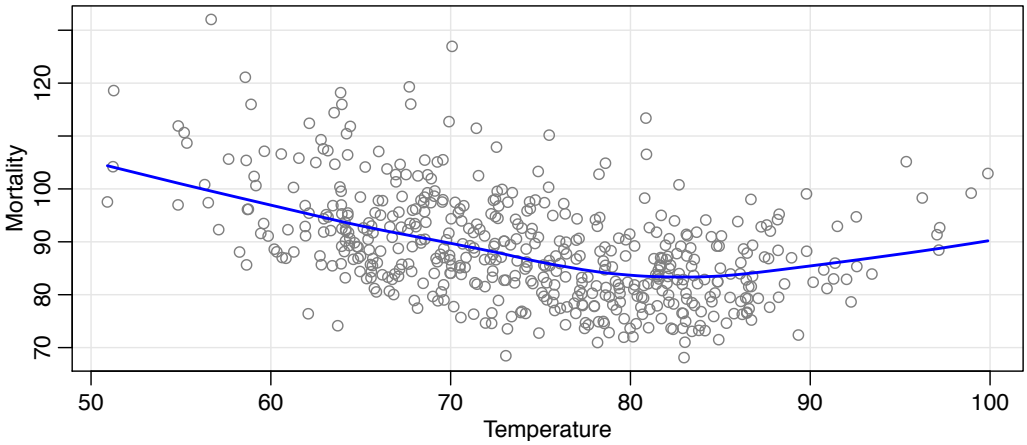

Fig. 2.16. *Smooth of mortality as a function of temperature using lowess.*

consequently very smooth. Thus, λ is seen as a trade-off between linear regression (completely smooth) and the data itself (no smoothness). The larger the value of λ, the smoother the fit.

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 SOI series using `spar=.5` to emphasize the El Niño cycle, and `spar=1` to emphasize the trend. The figure can be reproduced in R as follows.

```
plot(soi)
lines(smooth.spline(time(soi), soi, spar=.5), lwd=2, col=4)
lines(smooth.spline(time(soi), soi, spar= 1), lty=2, lwd=2, col=2)
```

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. We have already seen this idea used in Example 2.8 when we used lowess to visualize the nonlinear relationship between Recruitment and SOI at various lags. 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 show a scatterplot of mortality, M_t, and temperature, T_t, along with M_t smoothed as a function of T_t using lowess. Note that 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 83° F.

Figure 2.16 can be reproduced in R as follows using the defaults.

```
plot(tempr, cmort, xlab="Temperature", ylab="Mortality")
lines(lowess(tempr, cmort))
```

Problems

Section 2.1

2.1 A Structural Model For the Johnson & Johnson data, say y_t, shown in Figure 1.1, let $x_t = \log(y_t)$. In this problem, we are going to fit a special type of structural model, $x_t = T_t + S_t + N_t$ where T_t is a trend component, S_t is a seasonal component, and N_t is noise. In our case, time t is in quarters (1960.00, 1960.25, . . .) so one unit of time is a year.

(a) Fit the regression model

$$x_t = \underbrace{\beta t}_{\text{trend}} + \underbrace{\alpha_1 Q_1(t) + \alpha_2 Q_2(t) + \alpha_3 Q_3(t) + \alpha_4 Q_4(t)}_{\text{seasonal}} + \underbrace{w_t}_{\text{noise}}$$

where $Q_i(t) = 1$ if time t corresponds to quarter $i = 1, 2, 3, 4$, and zero otherwise. The $Q_i(t)$'s are called indicator variables. We will assume for now that w_t is a Gaussian white noise sequence. *Hint:* Detailed code is given in Code R.4, the last example of Section R.4.

(b) If the model is correct, what is the estimated average annual increase in the logged earnings per share?

(c) If the model is correct, does the average logged earnings rate increase or decrease from the third quarter to the fourth quarter? And, by what percentage does it increase or decrease?

(d) What happens if you include an intercept term in the model in (a)? Explain why there was a problem.

(e) Graph the data, x_t, and superimpose the fitted values, say $\hat{x}_t$, on the graph. Examine the residuals, $x_t - \hat{x}_t$, and state your conclusions. Does it appear that the model fits the data well (do the residuals look white)?

2.2 For the mortality data examined in Example 2.2:

(a) Add another component to the regression in (2.21) that accounts for the particulate count four weeks prior; that is, add P_{t-4} to the regression in (2.21). State your conclusion.

(b) Draw a scatterplot matrix of M_t, T_t, P_t and P_{t-4} and then calculate the pairwise correlations between the series. Compare the relationship between M_t and P_t versus M_t and P_{t-4}.

2.3 In this problem, we explore the difference between a random walk and a trend stationary process.

(a) Generate *four* series that are random walk with drift, (1.4), of length $n = 100$ with $\delta = .01$ and $\sigma_w = 1$. Call the data x_t for $t = 1, \ldots, 100$. Fit the regression $x_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .01\, t$) and the fitted line, $\hat{x}_t = \hat{\beta}\, t$, on the same graph. *Hint:* The following R code may be useful.

```
par(mfrow=c(2,2), mar=c(2.5,2.5,0,0)+.5, mgp=c(1.6,.6,0))  # set up
 for (i in 1:4){
  x = ts(cumsum(rnorm(100,.01,1)))        # data
  regx = lm(x~0+time(x), na.action=NULL)  # regression
  plot(x, ylab='Random Walk w Drift')     # plots
   abline(a=0, b=.01, col=2, lty=2)       # true mean (red - dashed)
   abline(regx, col=4)                    # fitted line (blue - solid)
}
```

(b) Generate *four* series of length $n = 100$ that are linear trend plus noise, say $y_t = .01\,t + w_t$, where t and w_t are as in part (a). Fit the regression $y_t = \beta t + w_t$ using least squares. Plot the data, the true mean function (i.e., $\mu_t = .01\,t$) and the fitted line, $\hat{y}_t = \hat{\beta}\,t$, on the same graph.

(c) Comment (what did you learn from this assignment).

2.4 Kullback-Leibler Information Given the random $n \times 1$ vector y, we define the information for discriminating between two densities in the same family, indexed by a parameter θ, say $f(y;\theta_1)$ and $f(y;\theta_2)$, as

$$I(\theta_1;\theta_2) = n^{-1}\,\mathrm{E}_1 \log \frac{f(y;\theta_1)}{f(y;\theta_2)}, \tag{2.41}$$

where E_1 denotes expectation with respect to the density determined by θ_1. For the Gaussian regression model, the parameters are $\theta = (\beta', \sigma^2)'$. Show that

$$I(\theta_1;\theta_2) = \frac{1}{2}\left(\frac{\sigma_1^2}{\sigma_2^2} - \log\frac{\sigma_1^2}{\sigma_2^2} - 1\right) + \frac{1}{2}\frac{(\beta_1-\beta_2)'Z'Z(\beta_1-\beta_2)}{n\sigma_2^2}\,. \tag{2.42}$$

2.5 Model Selection Both selection criteria (2.15) and (2.16) are derived from information theoretic arguments, based on the well-known *Kullback-Leibler discrimination information* numbers (see Kullback and Leibler, 1951, Kullback, 1958). We give an argument due to Hurvich and Tsai (1989). We think of the measure (2.42) as measuring the discrepancy between the two densities, characterized by the parameter values $\theta_1' = (\beta_1', \sigma_1^2)'$ and $\theta_2' = (\beta_2', \sigma_2^2)'$. Now, if the true value of the parameter vector is θ_1, we argue that the best model would be one that minimizes the discrepancy between the theoretical value and the sample, say $I(\theta_1;\hat{\theta})$. Because θ_1 will not be known, Hurvich and Tsai (1989) considered finding an unbiased estimator for $\mathrm{E}_1[I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2)]$, where

$$I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2) = \frac{1}{2}\left(\frac{\sigma_1^2}{\hat{\sigma}^2} - \log\frac{\sigma_1^2}{\hat{\sigma}^2} - 1\right) + \frac{1}{2}\frac{(\beta_1-\hat{\beta})'Z'Z(\beta_1-\hat{\beta})}{n\hat{\sigma}^2}$$

and β is a $k \times 1$ regression vector. Show that

$$\mathrm{E}_1[I(\beta_1, \sigma_1^2; \hat{\beta}, \hat{\sigma}^2)] = \frac{1}{2}\left(-\log\sigma_1^2 + \mathrm{E}_1 \log\hat{\sigma}^2 + \frac{n+k}{n-k-2} - 1\right), \tag{2.43}$$

using the distributional properties of the regression coefficients and error variance. An unbiased estimator for $\mathrm{E}_1 \log\hat{\sigma}^2$ is $\log\hat{\sigma}^2$. Hence, we have shown that the expectation

of the above discrimination information is as claimed. As models with differing dimensions k are considered, only the second and third terms in (2.43) will vary and we only need unbiased estimators for those two terms. This gives the form of AICc quoted in (2.16) in the chapter. You will need the two distributional results

$$\frac{n\hat{\sigma}^2}{\sigma_1^2} \sim \chi^2_{n-k} \quad \text{and} \quad \frac{(\hat{\beta}-\beta_1)'Z'Z(\hat{\beta}-\beta_1)}{\sigma_1^2} \sim \chi^2_k$$

The two quantities are distributed independently as chi-squared distributions with the indicated degrees of freedom. If $x \sim \chi^2_n$, $\mathrm{E}(1/x) = 1/(n-2)$.

Section 2.2

2.6 Consider a process consisting of a linear trend with an additive noise term consisting of independent random variables w_t with zero means and variances σ_w^2, that is,

$$x_t = \beta_0 + \beta_1 t + w_t,$$

where β_0, β_1 are fixed constants.

(a) Prove x_t is nonstationary.
(b) Prove that the first difference series $\nabla x_t = x_t - x_{t-1}$ is stationary by finding its mean and autocovariance function.
(c) Repeat part (b) if w_t is replaced by a general stationary process, say y_t, with mean function μ_y and autocovariance function $\gamma_y(h)$.

2.7 Show (2.27) is stationary.

2.8 The glacial varve record plotted in Figure 2.7 exhibits some nonstationarity that can be improved by transforming to logarithms and some additional nonstationarity that can be corrected by differencing the logarithms.

(a) Argue that the glacial varves series, say x_t, exhibits heteroscedasticity by computing the sample variance over the first half and the second half of the data. Argue that the transformation $y_t = \log x_t$ stabilizes the variance over the series. Plot the histograms of x_t and y_t to see whether the approximation to normality is improved by transforming the data.
(b) Plot the series y_t. Do any time intervals, of the order 100 years, exist where one can observe behavior comparable to that observed in the global temperature records in Figure 1.2?
(c) Examine the sample ACF of y_t and comment.
(d) Compute the difference $u_t = y_t - y_{t-1}$, examine its time plot and sample ACF, and argue that differencing the logged varve data produces a reasonably stationary series. Can you think of a practical interpretation for u_t? *Hint*: Recall Footnote 1.2.

(e) Based on the sample ACF of the differenced transformed series computed in (c), argue that a generalization of the model given by Example 1.26 might be reasonable. Assume

$$u_t = \mu + w_t + \theta w_{t-1}$$

is stationary when the inputs w_t are assumed independent with mean 0 and variance σ_w^2. Show that

$$\gamma_u(h) = \begin{cases} \sigma_w^2(1+\theta^2) & \text{if } h = 0, \\ \theta\,\sigma_w^2 & \text{if } h = \pm 1, \\ 0 & \text{if } |h| > 1. \end{cases}$$

(f) Based on part (e), use $\hat{\rho}_u(1)$ and the estimate of the variance of u_t, $\hat{\gamma}_u(0)$, to derive estimates of θ and σ_w^2. This is an application of the method of moments from classical statistics, where estimators of the parameters are derived by equating sample moments to theoretical moments.

2.9 In this problem, we will explore the periodic nature of S_t, the SOI series displayed in Figure 1.5.

(a) Detrend the series by fitting a regression of S_t on time t. Is there a significant trend in the sea surface temperature? Comment.
(b) Calculate the periodogram for the detrended series obtained in part (a). Identify the frequencies of the two main peaks (with an obvious one at the frequency of one cycle every 12 months). What is the probable El Niño cycle indicated by the minor peak?

Section 2.3

2.10 Consider the two weekly time series `oil` and `gas`. The oil series is in dollars per barrel, while the gas series is in cents per gallon.

(a) Plot the data on the same graph. Which of the simulated series displayed in Section 1.2 do these series most resemble? Do you believe the series are stationary (explain your answer)?
(b) In economics, it is often the percentage change in price (termed *growth rate* or *return*), rather than the absolute price change, that is important. Argue that a transformation of the form $y_t = \nabla \log x_t$ might be applied to the data, where x_t is the oil or gas price series. *Hint*: Recall Footnote 1.2.
(c) Transform the data as described in part (b), plot the data on the same graph, look at the sample ACFs of the transformed data, and comment.
(d) Plot the CCF of the transformed data and comment The small, but significant values when `gas` leads `oil` might be considered as feedback.
(e) Exhibit scatterplots of the oil and gas growth rate series for up to three weeks of lead time of oil prices; include a nonparametric smoother in each plot and comment on the results (e.g., Are there outliers? Are the relationships linear?).

(f) There have been a number of studies questioning whether gasoline prices respond more quickly when oil prices are rising than when oil prices are falling ("asymmetry"). We will attempt to explore this question here with simple lagged regression; we will ignore some obvious problems such as outliers and autocorrelated errors, so this will not be a definitive analysis. Let G_t and O_t denote the gas and oil growth rates.

(i) Fit the regression (and comment on the results)

$$G_t = \alpha_1 + \alpha_2 I_t + \beta_1 O_t + \beta_2 O_{t-1} + w_t,$$

where $I_t = 1$ if $O_t \geq 0$ and 0 otherwise (I_t is the indicator of no growth or positive growth in oil price). *Hint:*

```
poil = diff(log(oil))
pgas = diff(log(gas))
indi = ifelse(poil < 0, 0, 1)
mess = ts.intersect(pgas, poil, poilL = lag(poil,-1), indi)
summary(fit <- lm(pgas~ poil + poilL + indi, data=mess))
```

(ii) What is the fitted model when there is negative growth in oil price at time t? What is the fitted model when there is no or positive growth in oil price? Do these results support the asymmetry hypothesis?

(iii) Analyze the residuals from the fit and comment.

2.11 Use two different smoothing techniques described in Section 2.3 to estimate the trend in the global temperature series `globtemp`. Comment.

Chapter 3

ARIMA Models

Classical regression is often insufficient for explaining all of the interesting dynamics of a time series. For example, the ACF of the residuals of the simple linear regression fit to the price of chicken data (see Example 2.4) reveals additional structure in the data that regression did not capture. Instead, the introduction of correlation that may be generated through lagged linear relations leads to proposing the *autoregressive (AR)* and *autoregressive moving average (ARMA)* models that were presented in Whittle (1951). Adding nonstationary models to the mix leads to the *autoregressive integrated moving average (ARIMA)* model popularized in the landmark work by Box and Jenkins (1970). The *Box–Jenkins method* for identifying ARIMA models is given in this chapter along with techniques for *parameter estimation* and *forecasting* for these models. A partial theoretical justification of the use of ARMA models is discussed in Section B.4.

3.1 Autoregressive Moving Average Models

The classical regression model of Chapter 2 was developed for the static case, namely, we only allow the dependent variable to be influenced by current values of the independent variables. In the time series case, it is desirable to allow the dependent variable to be influenced by the past values of the independent variables and possibly by its own past values. If the present can be plausibly modeled in terms of only the past values of the independent inputs, we have the enticing prospect that forecasting will be possible.

Introduction to Autoregressive Models

Autoregressive models are based on the idea that the current value of the series, x_t, can be explained as a function of p past values, $x_{t-1}, x_{t-2}, \ldots, x_{t-p}$, where p determines the number of steps into the past needed to forecast the current value. As a typical case, recall Example 1.10 in which data were generated using the model

$$x_t = x_{t-1} - .90x_{t-2} + w_t,$$