

MAE 5870 - Aula 4

Regressão em Séries Temporais – domínio de tempo

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 $t = 1, \dots, n$, is being influenced by a collection of possible inputs or *independent* series, say, $z_{t1}, z_{t2}, \dots, z_{tq}$, 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*

$$x_t = \beta_0 + \beta_1 z_{t1} + \beta_2 z_{t2} + \dots + \beta_q z_{tq} + w_t, \quad (2.1)$$

where $\beta_0, \beta_1, \dots, \beta_q$ 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

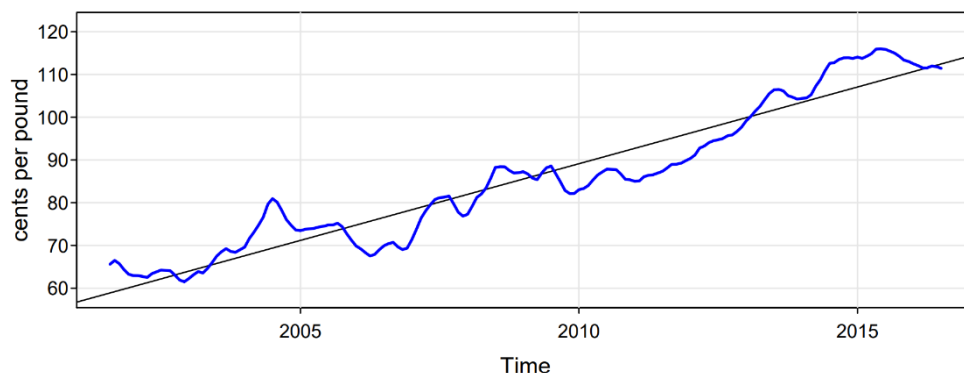


Fig. 2.1. The price of chicken: monthly whole bird spot price, Georgia docks, US cents per pound, August 2001 to July 2016, with fitted linear trend line.

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

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 β s. 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
```

Outro exemplo: Global temperature data

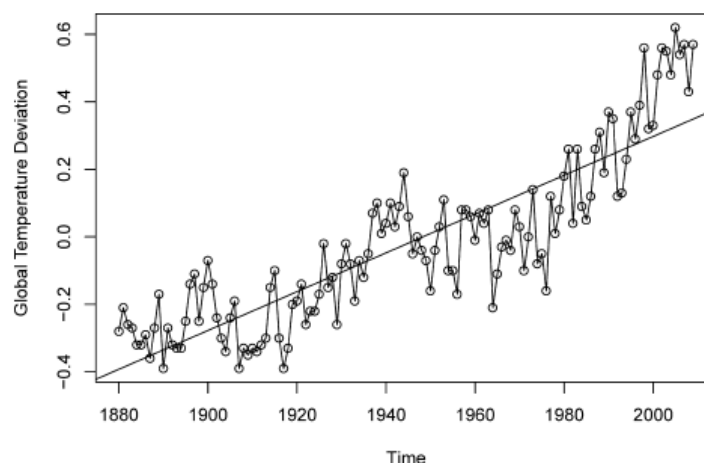


Fig. 2.1. Global temperature deviations shown in Figure 1.2 with fitted linear trend line.

$$\hat{x}_t = -11.2 + .006t.$$

The multiple linear regression model described by (2.1) can be conveniently written in a more general notation by defining the column vectors $\mathbf{z}_t = (1, z_{t1}, z_{t2}, \dots, z_{tq})'$ and $\beta = (\beta_0, \beta_1, \dots, \beta_q)'$, where ' denotes transpose, so (2.1) can be written in the alternate form

$$x_t = \beta_0 + \beta_1 z_{t1} + \dots + \beta_q z_{tq} + w_t = \beta' z_t + w_t. \quad (2.2)$$

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

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

with respect to $\beta_0, \beta_1, \dots, \beta_q$. 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*

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

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

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

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

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

where

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

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

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

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

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

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 $i = 1, \dots, q$.

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, $z_{t,1:r} = \{z_{t1}, z_{t2}, \dots, z_{tr}\}$ is influencing the dependent variable x_t . The reduced model is

$$x_t = \beta_0 + \beta_1 z_{t1} + \dots + \beta_r z_{tr} + w_t \quad (2.10)$$

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

The null hypothesis in this case is $H_0: \beta_{r+1} = \dots = \beta_q = 0$. 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

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

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 $H_0: \beta_{r+1} = \dots = \beta_q = 0$ is true, then $SSE_r \approx SSE$ because the estimates of those β s 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.

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	$MSE = SSE/(n - q - 1)$	

A special case of interest is the null hypothesis $H_0: \beta_1 = \dots = \beta_q = 0$. 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

$$R^2 = \frac{SSE_0 - SSE}{SSE_0}, \quad (2.12)$$

where the residual sum of squares under the reduced model is

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

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*.

Critério de seleção

Suppose we consider a normal regression model with k coefficients and denote the maximum likelihood estimator for the variance as

$$\hat{\sigma}_k^2 = \frac{SSE_k}{n}, \quad (2.18)$$

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

Definition 2.1 Akaike's Information Criterion (AIC)

$$\text{AIC} = \log \hat{\sigma}_k^2 + \frac{n + 2k}{n}, \quad (2.19)$$

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

Definition 2.2 AIC, Bias Corrected (AICc)

$$\text{AICc} = \log \hat{\sigma}_k^2 + \frac{n+k}{n-k-2}, \quad (2.20)$$

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

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

Definition 2.3 Bayesian Information Criterion (BIC)

$$\text{BIC} = \log \hat{\sigma}_k^2 + \frac{k \log n}{n}, \quad (2.21)$$

using the same notation as in Definition 2.2.

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

Example 2.2 Pollution, Temperature and Mortality

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

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

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

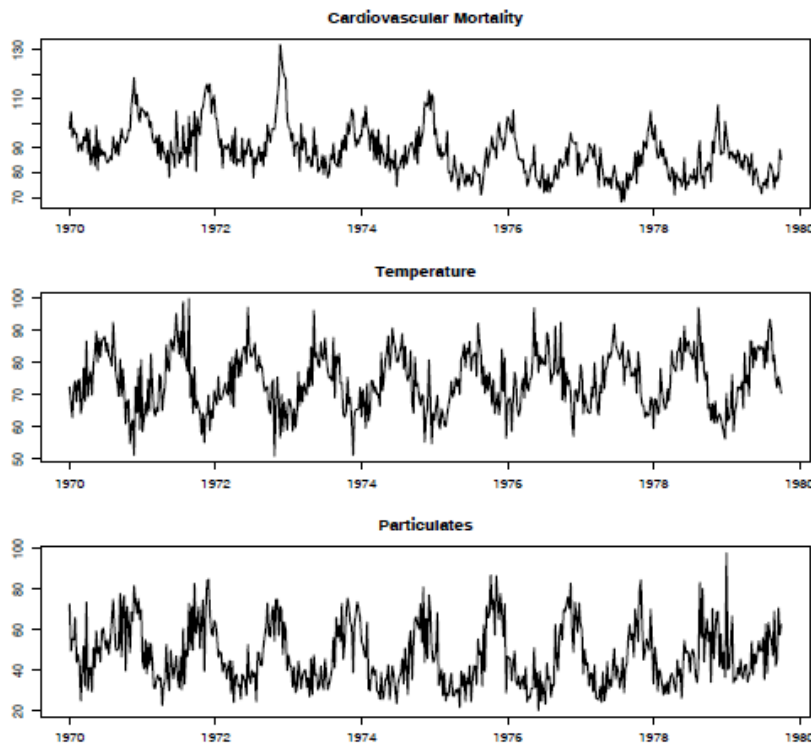


Fig. 2.2. Average weekly cardiovascular mortality (top), temperature (middle) and particulate pollution (bottom) in Los Angeles County. There are 508 six-day smoothed averages obtained by filtering daily values over the 10 year period 1970-1979.

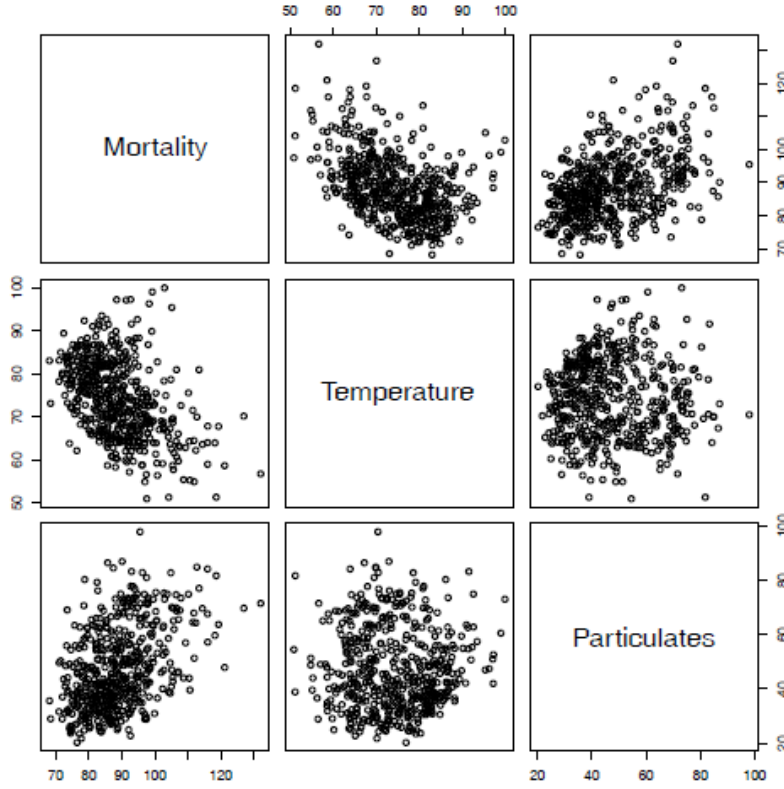


Fig. 2.3. Scatterplot matrix showing plausible relations between mortality, temperature, and pollution.

$$M_t = \beta_1 + \beta_2 t + w_t \quad (2.22)$$

$$M_t = \beta_1 + \beta_2 t + \beta_3(T_t - T.) + w_t \quad (2.23)$$

$$M_t = \beta_1 + \beta_2 t + \beta_3(T_t - T.) + \beta_4(T_t - T.)^2 + w_t \quad (2.24)$$

$$M_t = \beta_1 + \beta_2 t + \beta_3(T_t - T.) + \beta_4(T_t - T.)^2 + \beta_5 P_t + w_t \quad (2.25)$$

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

Table 2.2. Summary Statistics for Mortality Models

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

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 (2.14). Hence, a model with only trend could be compared to the full model using $q = 5, r = 2, n = 508$, so

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

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

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

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

Example 2.3 Regression With Lagged Variables

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

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

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

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

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

Definition 1.16 *The estimators for the cross-covariance function, $\gamma_{xy}(h)$, as given in (1.28) and the cross-correlation, $\rho_{xy}(h)$, in (1.11) are given, respectively, by the **sample cross-covariance function***

$$\widehat{\gamma}_{xy}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}), \quad (1.40)$$

where $\widehat{\gamma}_{xy}(-h) = \widehat{\gamma}_{yx}(h)$ determines the function for negative lags, and the **sample cross-correlation function**

$$\widehat{\rho}_{xy}(h) = \frac{\widehat{\gamma}_{xy}(h)}{\sqrt{\widehat{\gamma}_x(0)\widehat{\gamma}_y(0)}}. \quad (1.41)$$

The sample cross-correlation function can be examined graphically as a function of lag h to search for leading or lagging relations in the data using the property mentioned in Example 1.24 for the theoretical cross-covariance function. Because $-1 \leq \widehat{\rho}_{xy}(h) \leq 1$ the practical importance of peaks can be assessed by comparing their magnitudes with their theoretical maximum values. Furthermore, for x_t and y_t independent linear processes of the form (1.31), we have the following property.

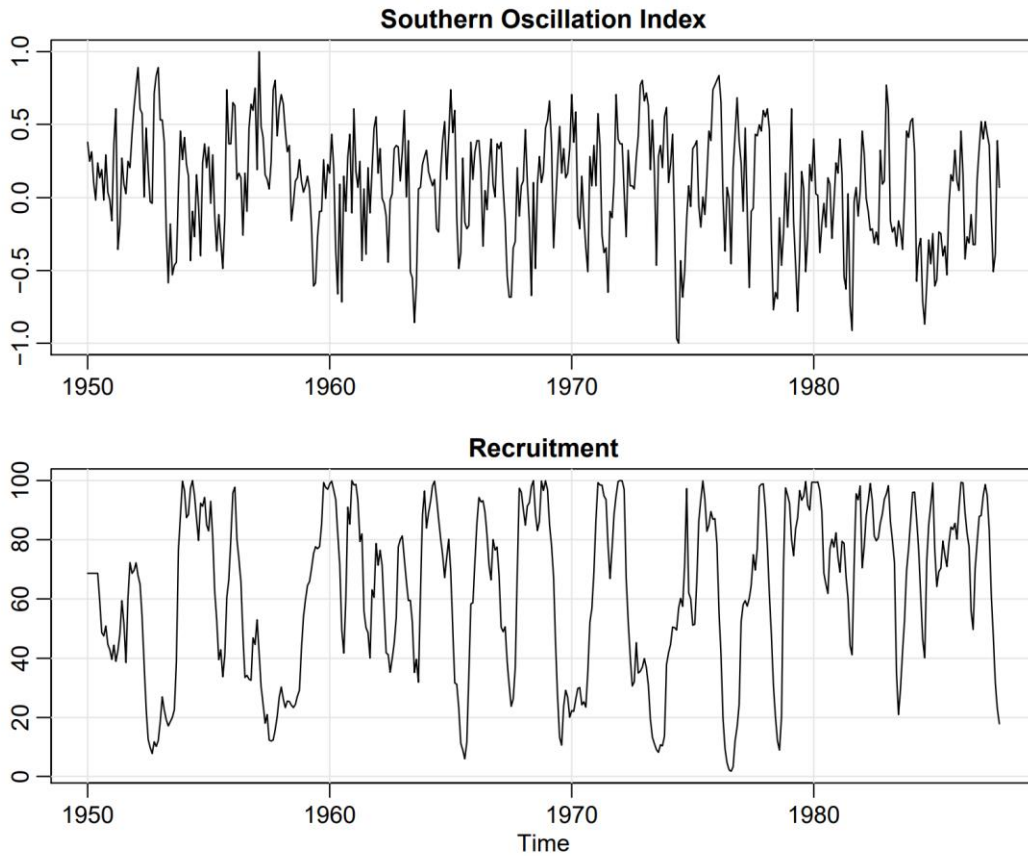


Fig. 1.5. Monthly SOI and Recruitment (estimated new fish), 1950-1987.

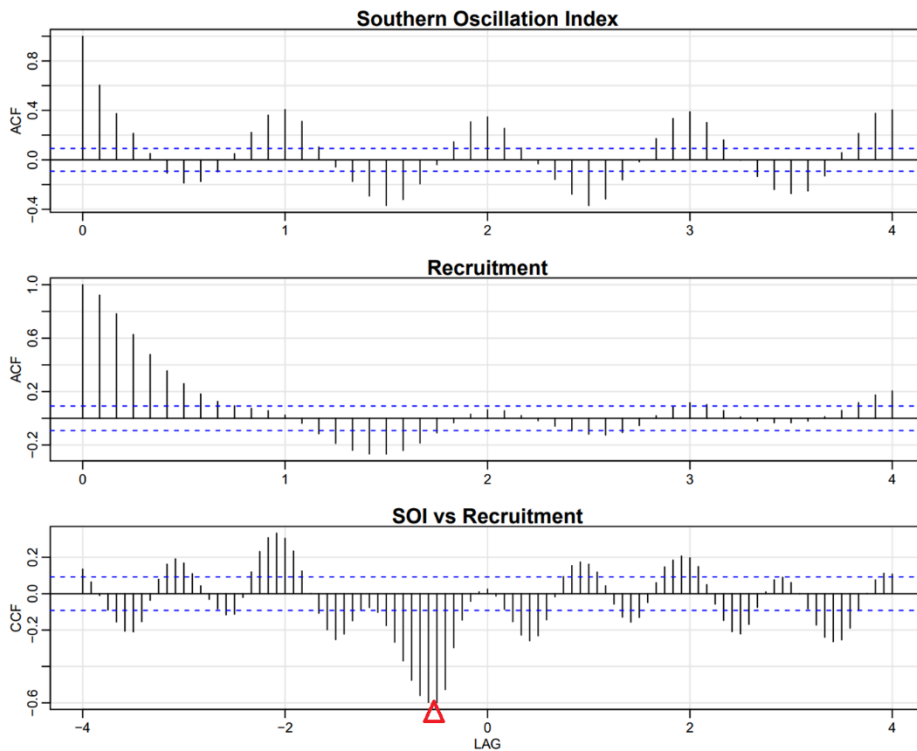


Fig. 1.16. Sample ACFs of the SOI series (top) and of the Recruitment series (middle), and the sample CCF of the two series (bottom); negative lags indicate SOI leads Recruitment. The lag axes are in terms of seasons (12 months).

Exploratory Data Analysis

In general, it is necessary for time series data to **be stationary**, so averaging lagged products over time 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**

Often, this is not the case, and we will mention some methods for playing down **the effects of nonstationarity** so the stationary properties of the series may be studied.

Examples:

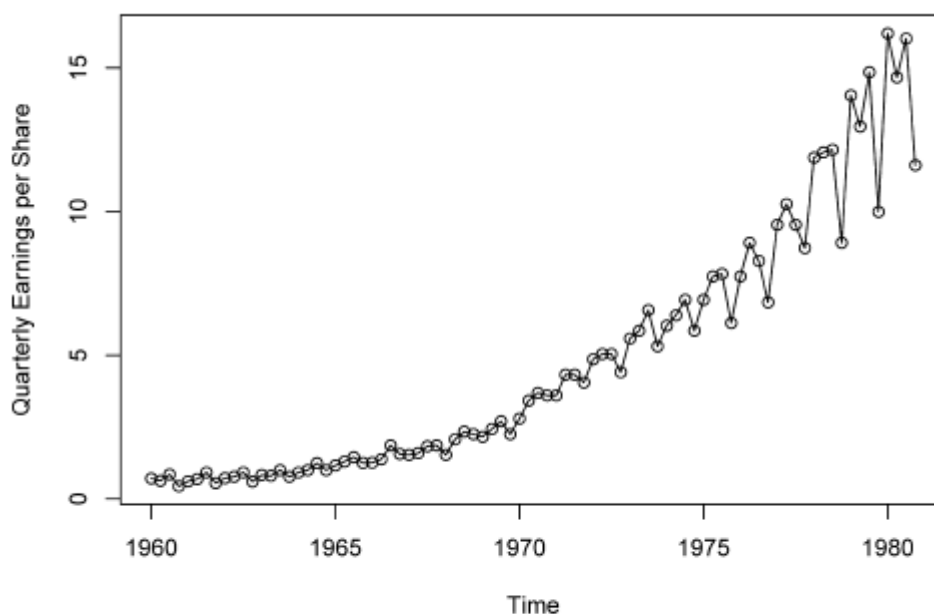


Fig. 1.1. Johnson & Johnson quarterly earnings per share, 84 quarters, 1960-I to 1980-IV.

Example 2.4 Detrending Global Temperature

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

$$x_t = \mu_t + y_t,$$

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

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

In that example, we estimated the trend using ordinary least squares³ and found

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

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

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

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

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

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

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,

$$\mu_t = \delta + \mu_{t-1} + w_t, \quad (2.30)$$

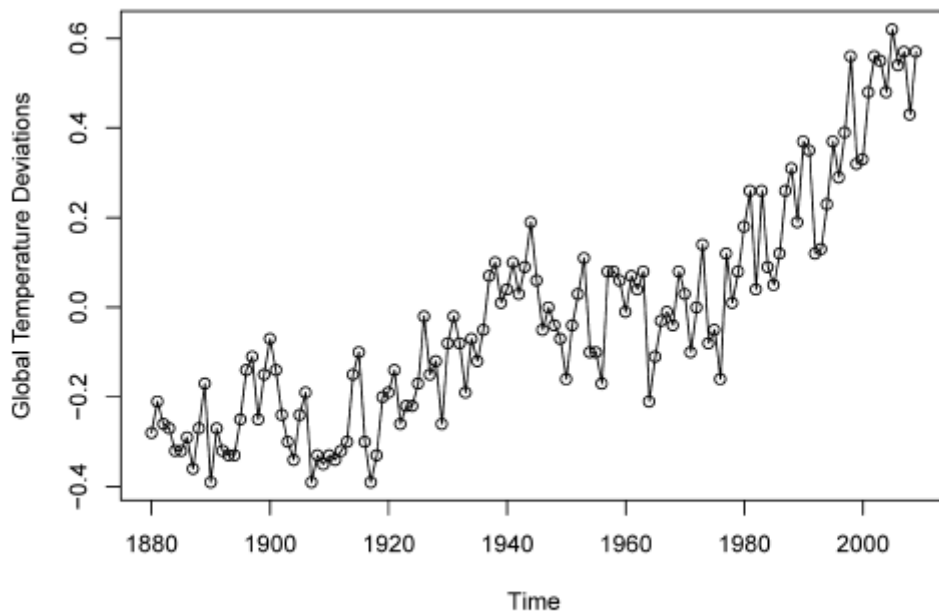


Fig. 1.2. Yearly average global temperature deviations (1880–2009) in degrees centigrade.

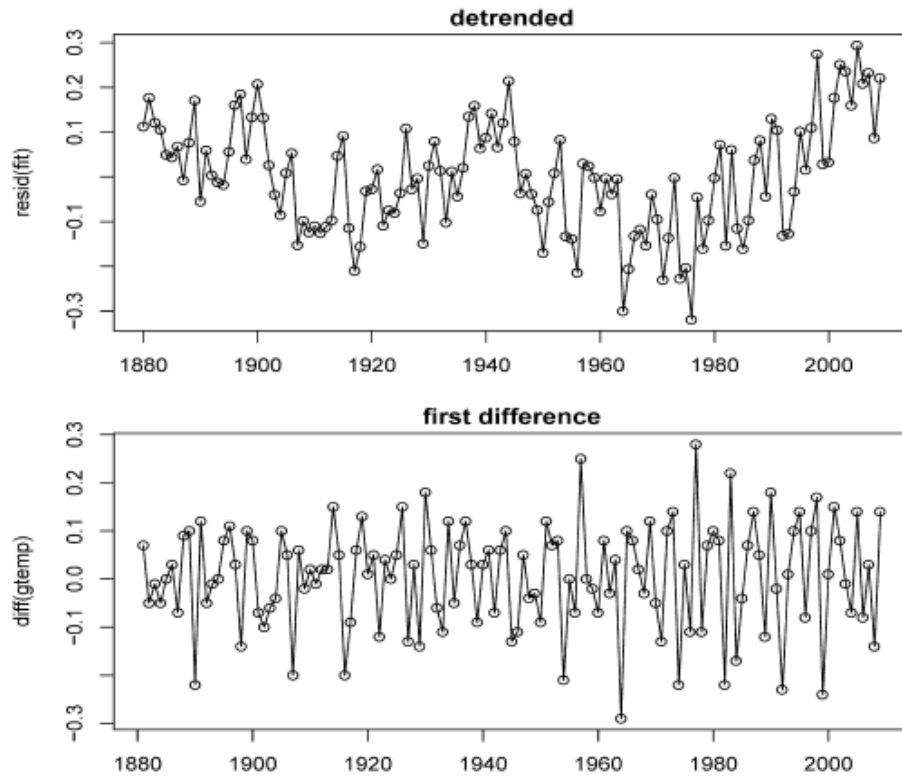


Fig. 2.4. Detrended (top) and differenced (bottom) global temperature series. The original data are shown in Figures 1.2 and 2.1.

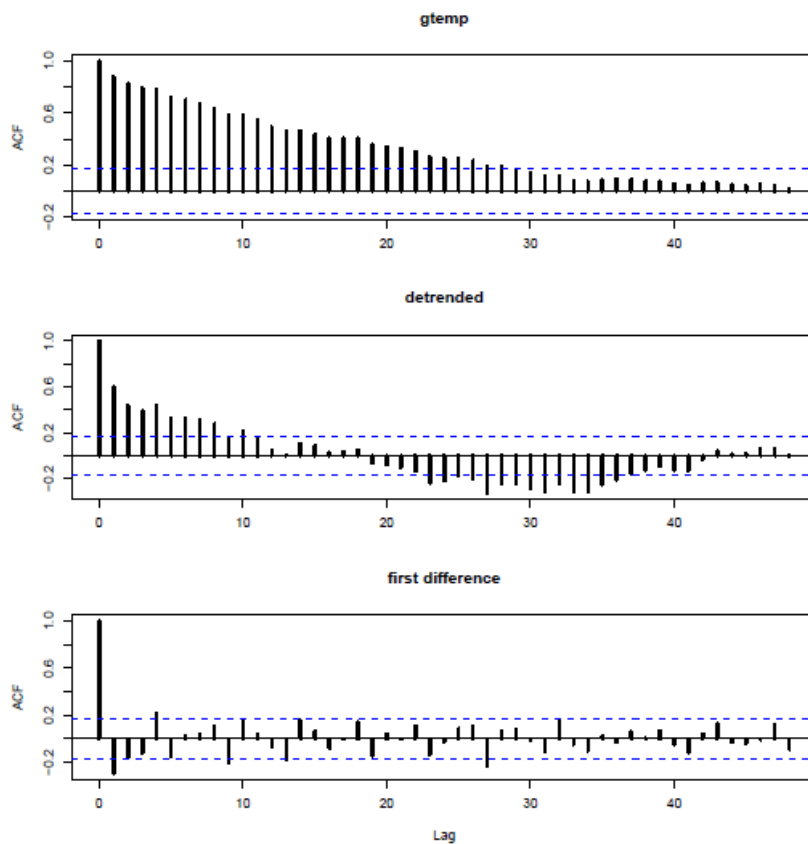


Fig. 2.5. Sample ACFs of the global temperature (top), and of the detrended (middle) and the differenced (bottom) series.

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

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

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

$$B^kx_t = x_{t-k}. \quad (2.33)$$

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

$$\nabla x_t = (1 - B)x_t, \quad (2.34)$$

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

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

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

Definition 2.5 Differences of order d are defined as

$$\nabla^d = (1 - B)^d, \quad (2.35)$$

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

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

$$y_t = \log x_t, \quad (2.36)$$

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

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

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

Example 2.6 Paleoclimatic Glacial Varves

Melting glaciers deposit yearly layers of sand and silt during the spring melting seasons, which can be reconstructed yearly over a period ranging from the time deglaciation began in New England (about 12,600 years ago) to the time it ended (about 6,000 years ago). Such sedimentary deposits, called varves, can be used as proxies for paleoclimatic parameters, such as temperature, because, in a warm year, more sand and silt are deposited from the receding glacier. Figure 2.6 shows the thicknesses of the yearly varves collected from one location in Massachusetts for 634 years, beginning 11,834 years ago. For further information, see Shumway and Verosub (1992). Because the variation in thicknesses increases in proportion to the amount deposited, a logarithmic transformation could remove the nonstationarity observable in the variance as a function of time. Figure 2.6 shows the original and transformed varves, and it is clear that this improvement has occurred.

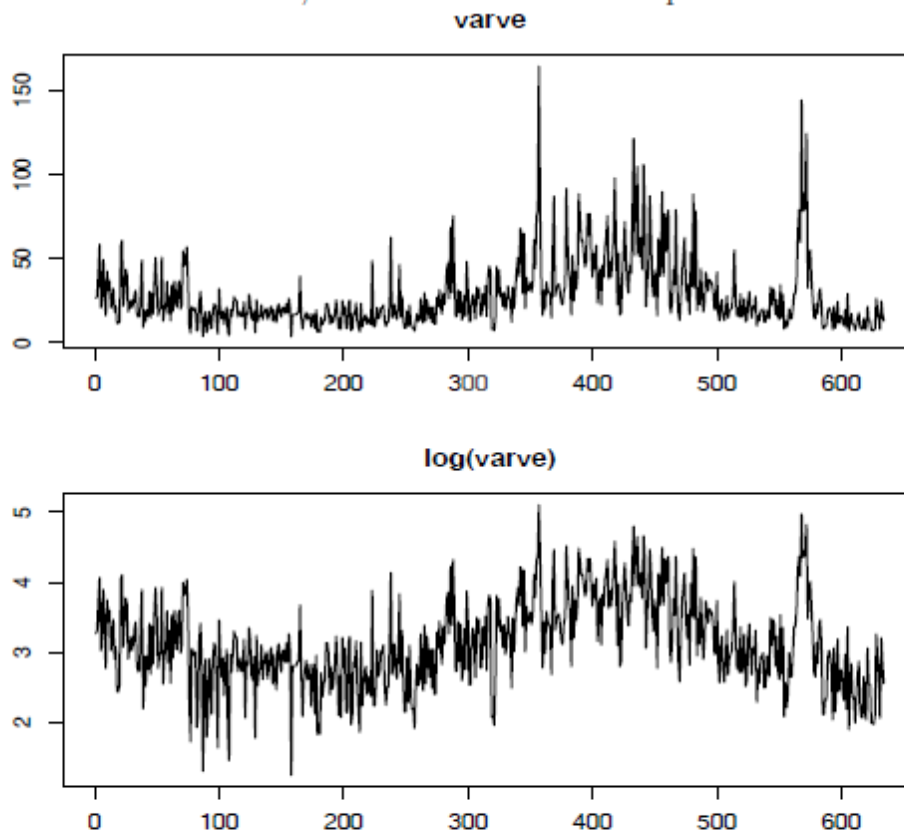


Fig. 2.6. Glacial varve thicknesses (top) from Massachusetts for $n = 634$ years compared with log transformed thicknesses (bottom).

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 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.

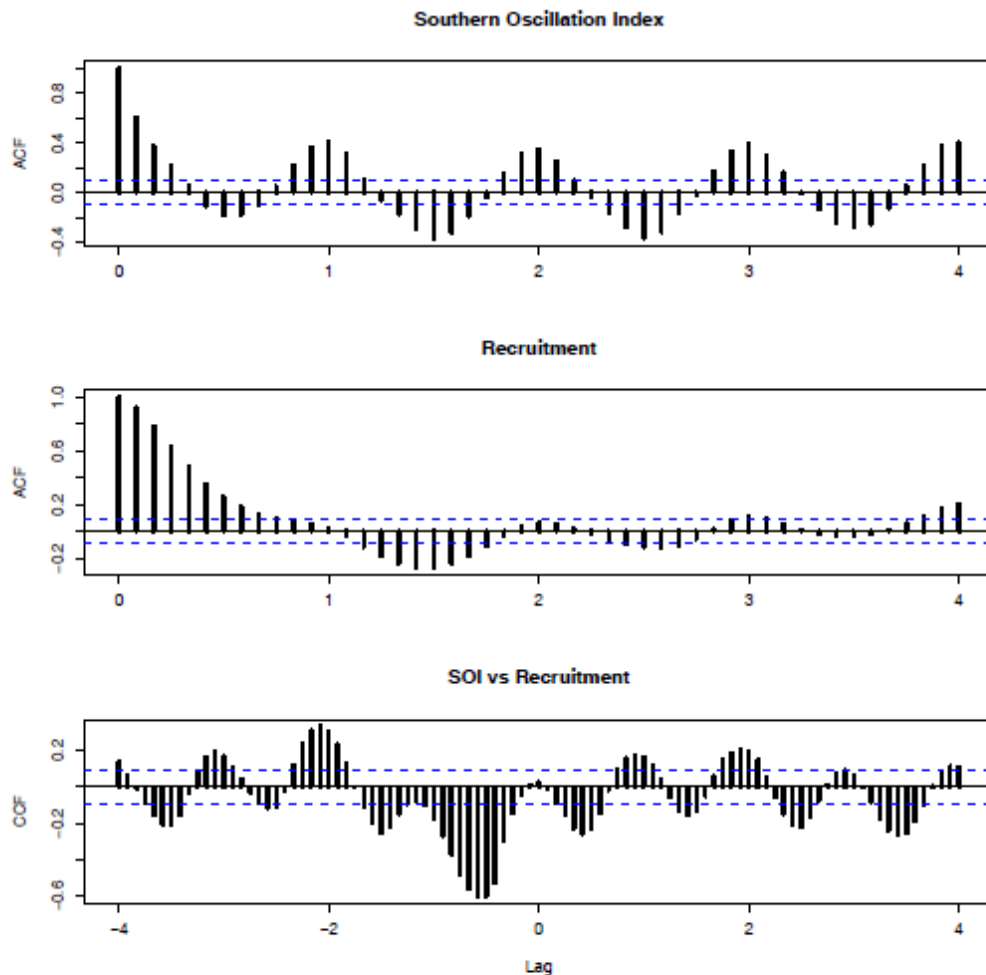


Fig. 1.14. Sample ACFs of the SOI series (top) and of the Recruitment series (middle), and the sample CCF of the two series (bottom); negative lags indicate SOI leads Recruitment. The lag axes are in terms of seasons (12 months).

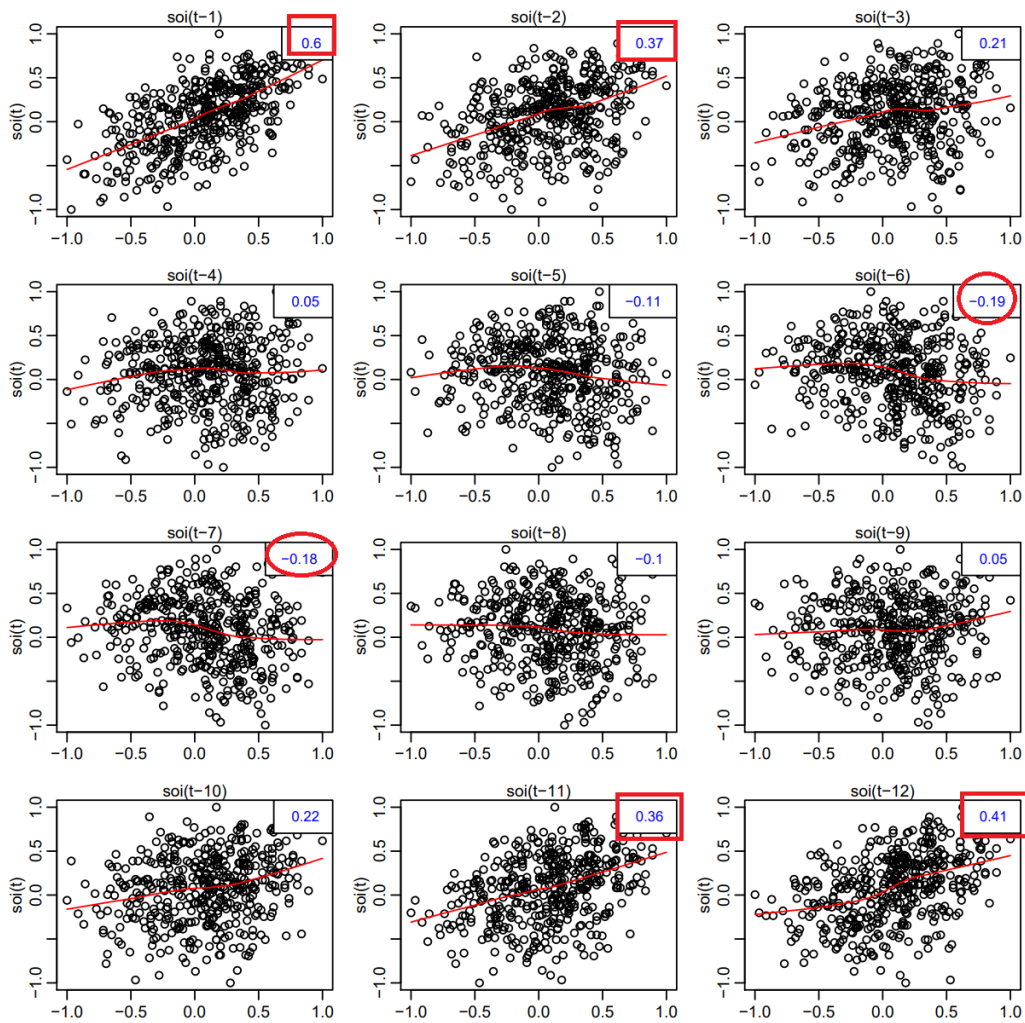


Fig. 2.8. Scatterplot matrix relating current SOI values, S_t , to past SOI values, S_{t-h} , at lags $h = 1, 2, \dots, 12$. The values in the upper right corner are the sample autocorrelations and the lines are a lowess fit.

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
```

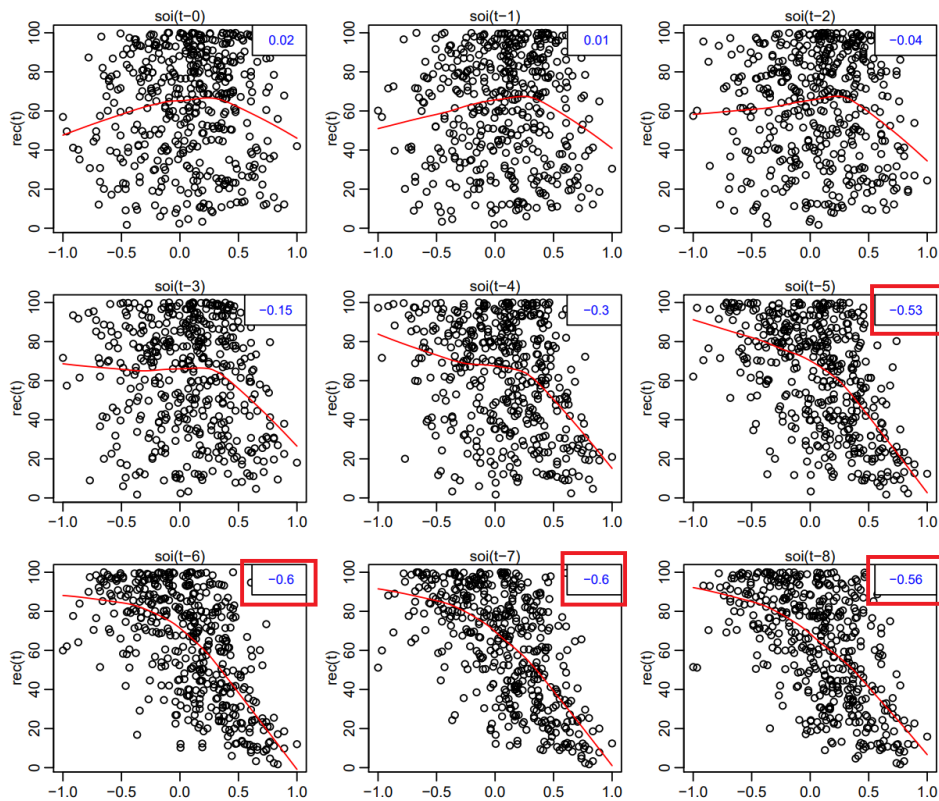


Fig. 2.9. Scatterplot matrix of the Recruitment series, R_t , on the vertical axis plotted against the SOI series, S_{t-h} , on the horizontal axis at lags $h = 0, 1, \dots, 8$. The values in the upper right corner are the sample cross-correlations and the lines are a loess fit.

Example 2.9 Regression with Lagged Variables (cont)

In [Example 2.3](#) we regressed Recruitment on lagged SOI,

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

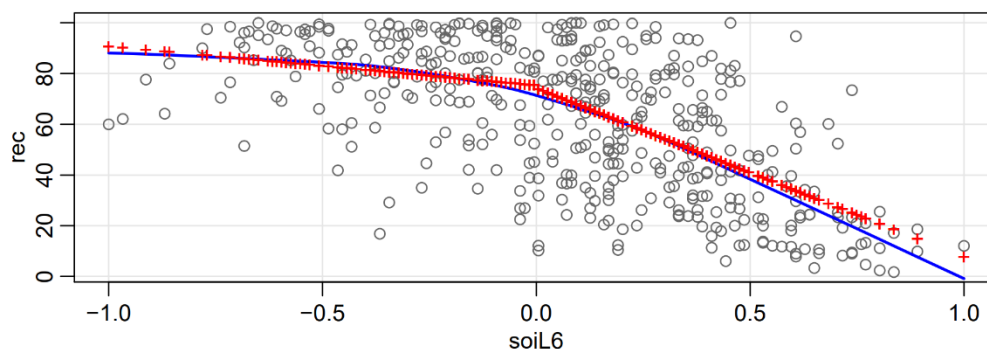


Fig. 2.10. Display for [Example 2.9](#): Plot of Recruitment (R_t) vs SOI lagged 6 months (S_{t-6}) with the fitted values of the regression as points (+) and a loess fit (—).

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
```

Example 2.8 Using Regression to Discover a Signal in Noise

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

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

where $\omega = 1/50$, $A = 2$, $\phi = .6\pi$, and $\sigma_w = 5$; the data are shown on the bottom panel of Figure 1.11 on page 16. At this point we assume the frequency of oscillation $\omega = 1/50$ is known, but A and ϕ are unknown parameters. In this case the parameters appear in (2.38) in a nonlinear way, so we use a trigonometric identity⁴ and write

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

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

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

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

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

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

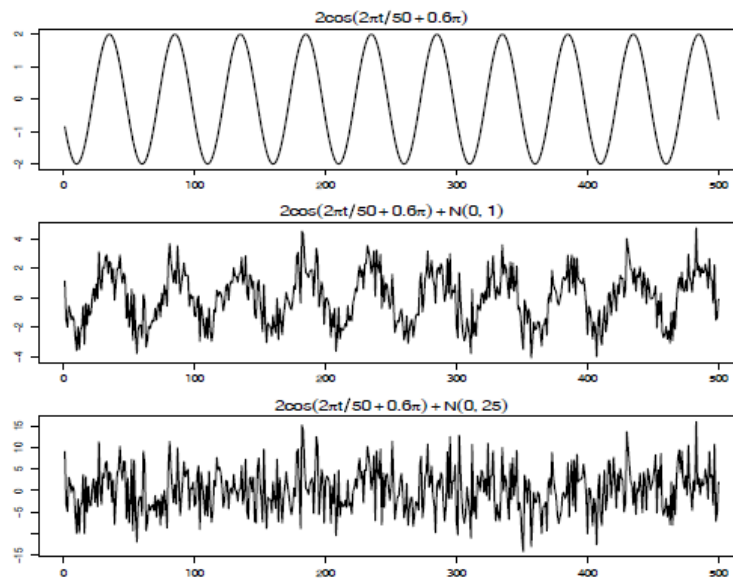


Fig. 1.11. Cosine wave with period 50 points (top panel) compared with the cosine wave contaminated with additive white Gaussian noise, $\sigma_w = 1$ (middle panel) and $\sigma_w = 5$ (bottom panel); see (1.5).

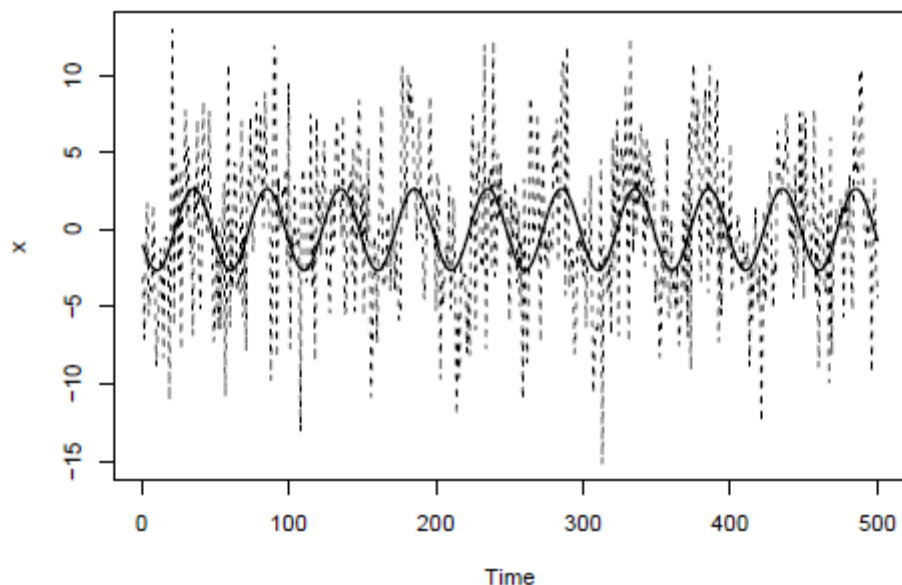


Fig. 2.9. Data generated by (2.38) [dashed line] with the fitted [solid] line, (2.40), superimposed.

Example 2.9 Using the Periodogram to Discover a Signal in Noise

The analysis in Example 2.8 may seem like cheating because we assumed we knew the value of the frequency parameter ω . If we do not know ω , we could try to fit the model (2.38) using nonlinear regression with ω as a parameter. Another method is to try various values of ω in a systematic way. Using the regression results of §2.2, we can show the estimated regression coefficients in Example 2.8 take on the special form given by

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

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

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

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

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

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

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

For this particular example, the values calculated in (2.41) and (2.42) 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 $\beta_1(j/n)$ and $\beta_2(j/n)$, for each j , are essentially measuring the correlation of the data with a sinusoid oscillating at j cycles in n time points.⁶ Hence, an appropriate measure of the presence of a frequency of oscillation of j cycles in n time points in the data would be

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

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

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

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

to obtain the values of $\beta_1(j/n)$ and $\beta_2(j/n)$ [with $\beta_2(0) = \beta_2(1/2) = 0$] because they can be computed quickly if n (assumed even here) is a highly composite integer. There is no error in (2.46) because there are n observations and n parameters; the regression fit will be perfect.

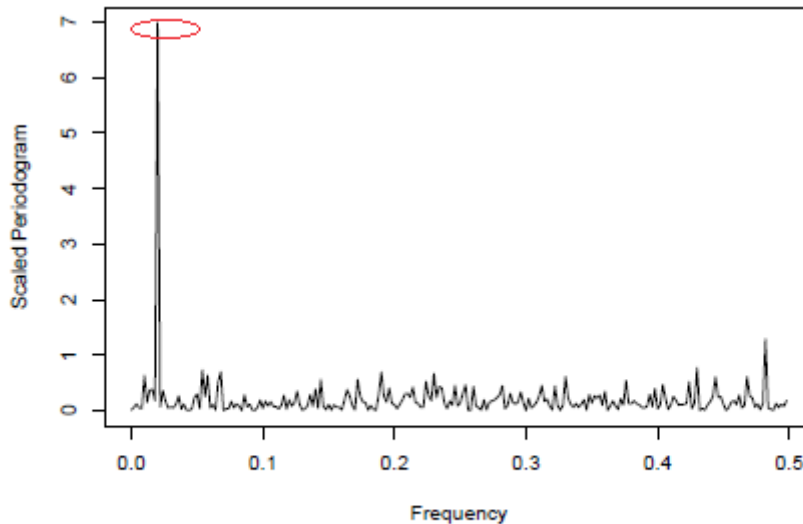


Fig. 2.10. The scaled periodogram, (2.45), of the 500 observations generated by (2.38); the data are displayed in Figures 1.11 and 2.9.

Smoothing in the Time Series Context

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

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

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

Example 2.10 Moving Average Smoother

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

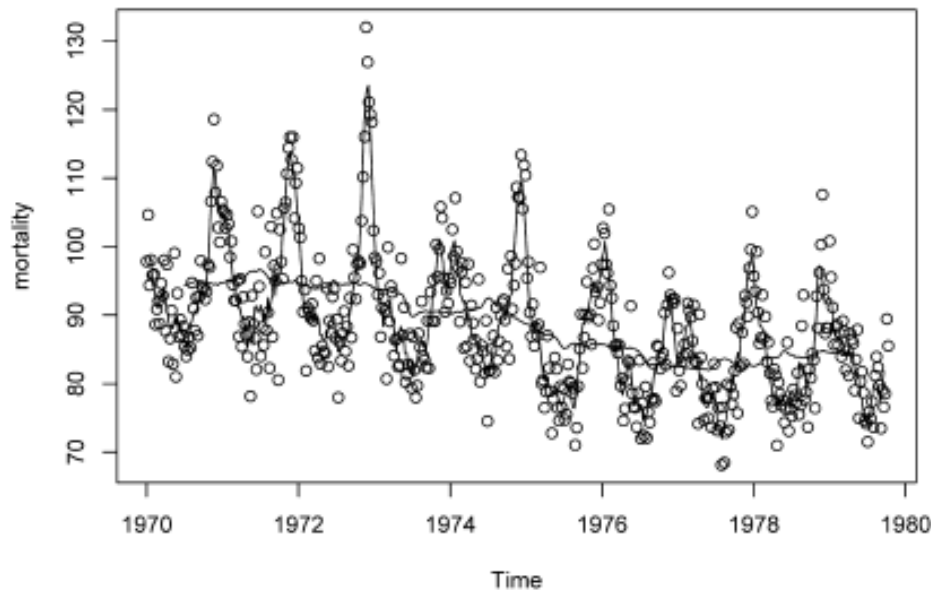


Fig. 2.11. The weekly cardiovascular mortality series discussed in Example 2.2 smoothed using a five-week moving average and a 53-week moving average.

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

$$x_t = f_t + y_t, \quad (2.51)$$

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

$$f_t = \beta_0 + \beta_1 t + \cdots + \beta_p t^p. \quad (2.52)$$

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

$$f_t = \alpha_0 + \alpha_1 \cos(2\pi\omega_1 t) + \beta_1 \sin(2\pi\omega_1 t) \\ + \cdots + \alpha_p \cos(2\pi\omega_p t) + \beta_p \sin(2\pi\omega_p t), \quad (2.53)$$

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

Example 2.11 Polynomial and Periodic Regression Smoothers

Figure 2.12 shows the weekly mortality series with an estimated (via ordinary least squares) cubic smoother

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

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.

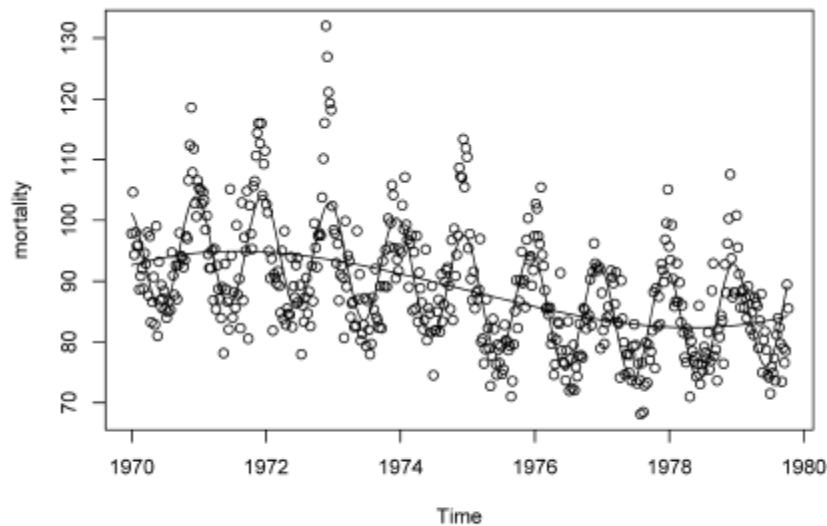


Fig. 2.12. The weekly cardiovascular mortality series with a cubic trend and cubic trend plus periodic regression.

Example 2.12 Kernel Smoothing

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

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

where

$$w_i(t) = K\left(\frac{t-t_i}{b}\right) / \sum_{j=1}^n K\left(\frac{t-t_j}{b}\right). \quad (2.55)$$

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

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

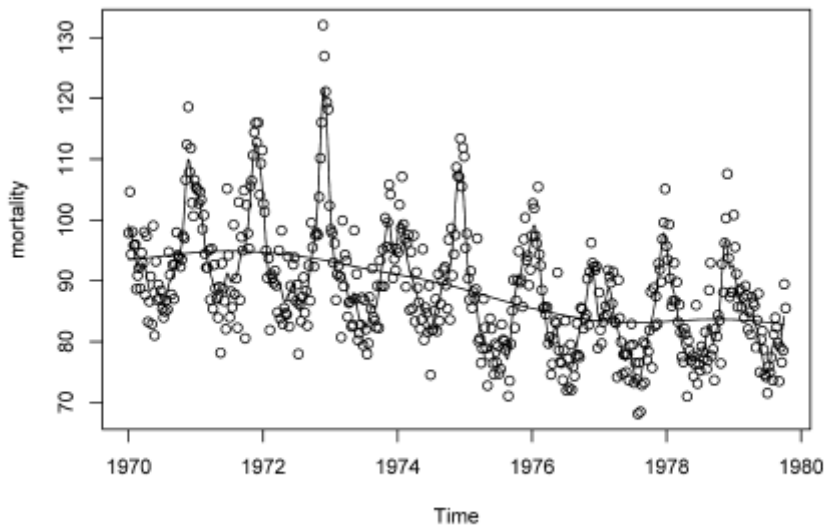


Fig. 2.13. Kernel smoothers of the mortality data.

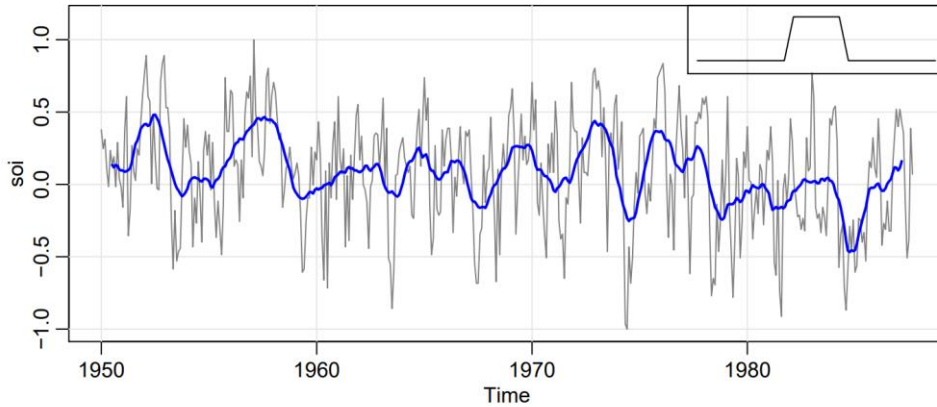


Fig. 2.12. Moving average smoother of SOI. The insert shows the shape of the moving average (“boxcar”) kernel [not drawn to scale] described in (2.39).

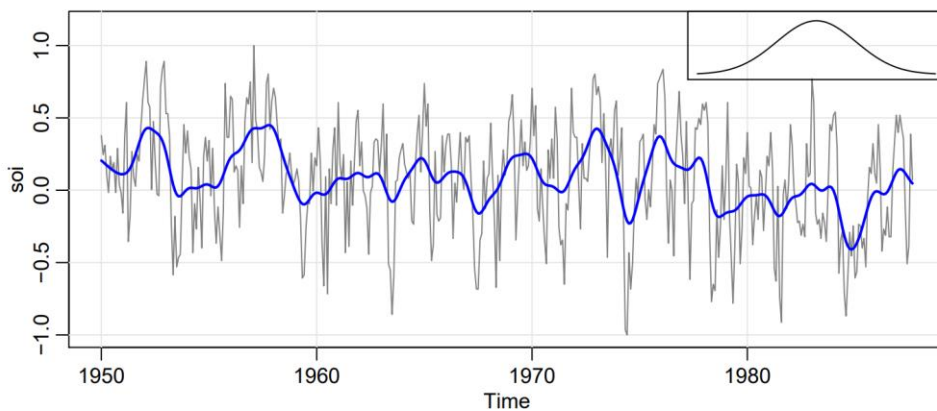


Fig. 2.13. Kernel smoother of SOI. The insert shows the shape of the normal kernel [not drawn to scale].

Example 2.13 Lowess and Nearest Neighbor Regression

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

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

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

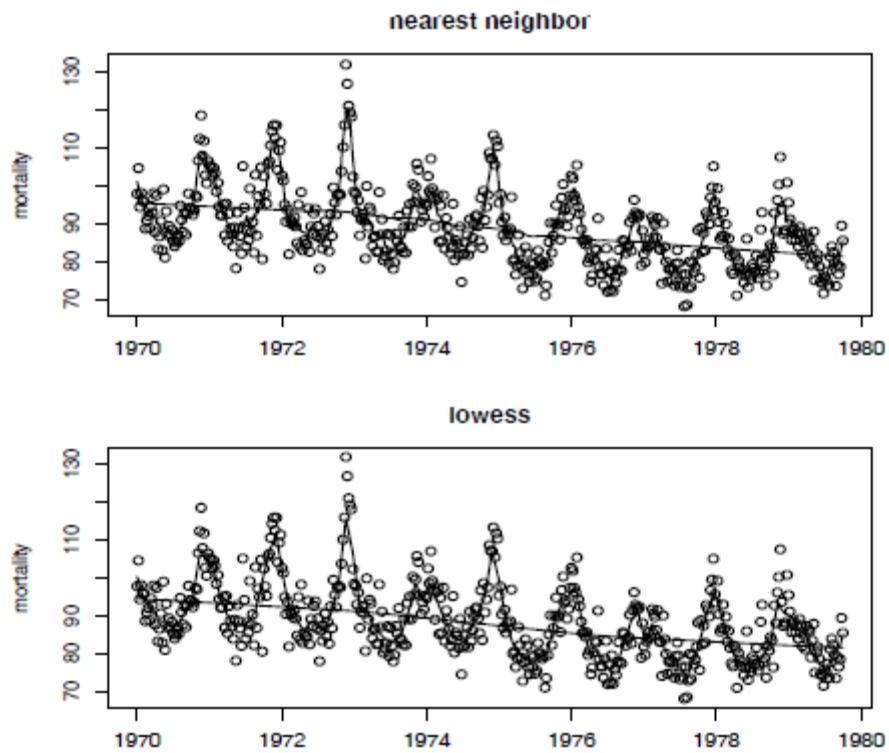
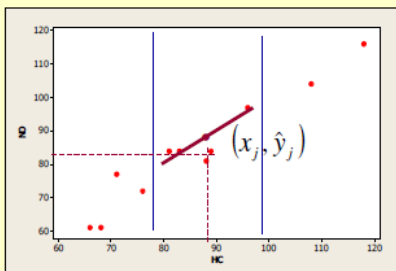


Fig. 2.14. Nearest neighbor (`supsmu`) and locally weighted regression (`lowess`) smoothers of the mortality data.

Suavização - Lowess

Lowess: Locally weighted regression scatter plot smoothing



Para obter (x_j, \hat{y}_j) :

- Abrir uma faixa vertical centrada em (x_j, y_j) , contendo $q = [fn]$ pontos ($0 < f < 1$). Quando maior o valor de f , mais suave será o ajustamento. S-Plus

$$1/3 \leq f \leq 2/3$$

- Definir pesos para os pontos vizinhos de (x_j, y_j)

$$h(u) = \begin{cases} (1 - |u|^3)^3 & \text{se } |u| < 1 \\ 0 & \text{cc} \end{cases} \Rightarrow \text{o peso atribuído a } (x_k, y_k) \text{ é } h(x_k) = h\left(\frac{x_j - x_k}{d_j}\right)$$

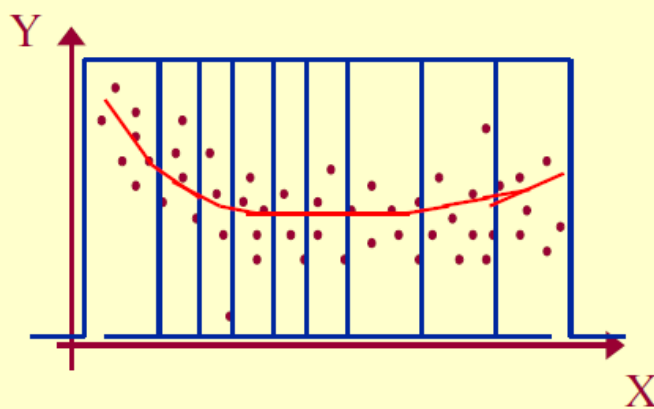
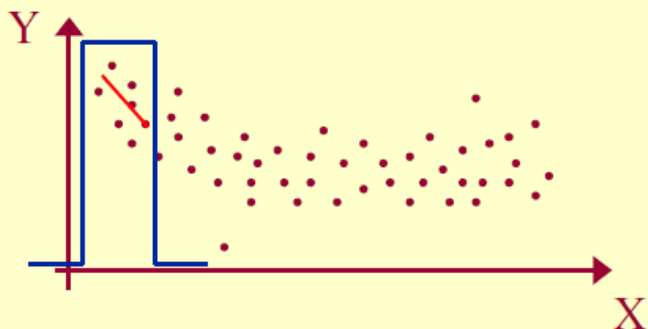
- Ajustamos uma reta aos q pontos (M.Q.P.)

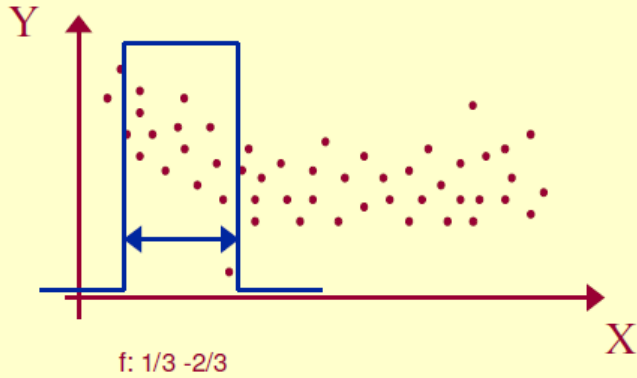
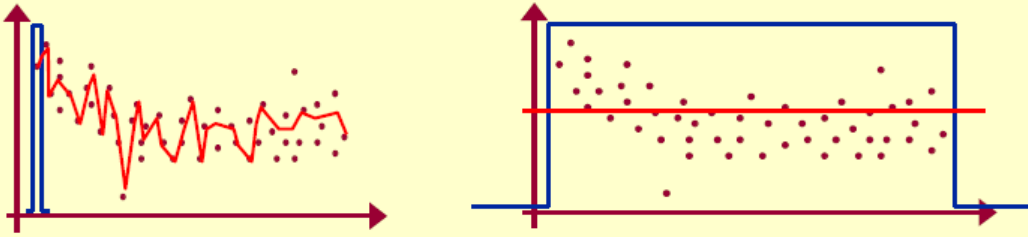
↑ distância ao vizinho mais afastado

$$\hat{y}_j = \hat{\alpha} + \hat{\beta} x_j; \quad \sum_{k=1}^n h(x_k) (y_k - \alpha - \beta x_k)^2$$

↑ resíduo: ↓ peso

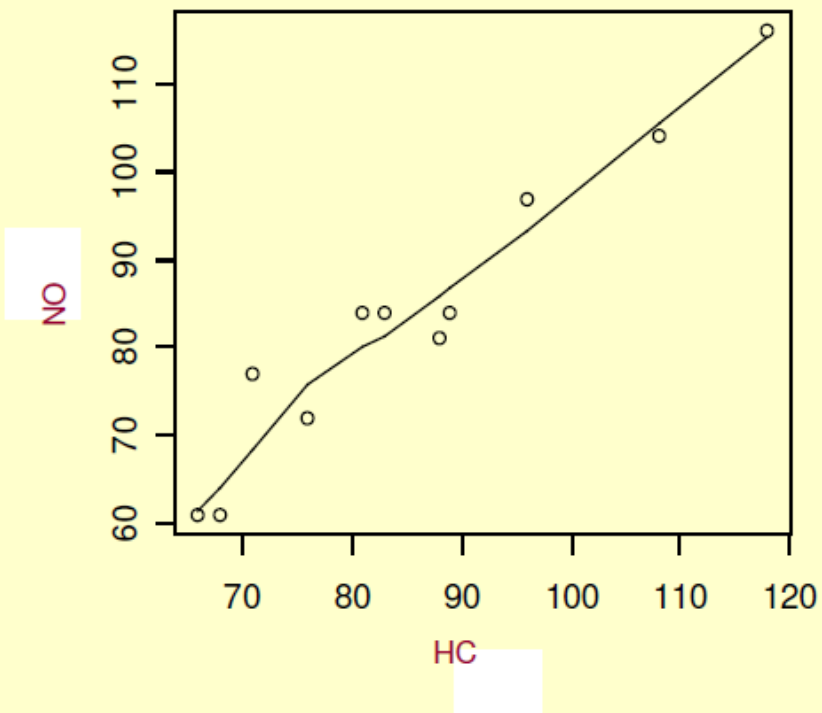
57





Suavização - Lowess

Dados de Poluente



```
> plot (x,y)
```

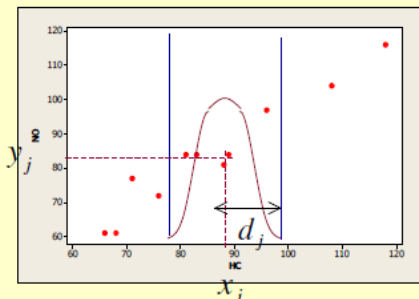
```
> lines(lowess(x,y,f=2/3))
```

Suavização – Lowess Robusto

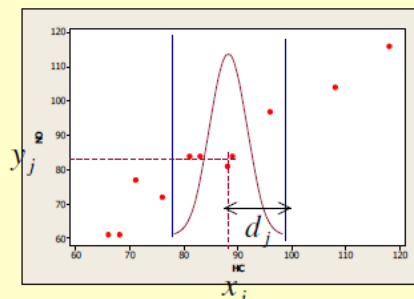
Na presença de valores atípicos \Rightarrow Lowess Robusto (Iterativo)

Atribuição de pesos robustos às observações (x_j, y_j)

$h(u)$: Tri-cúbica



$g(u)$: Bi-quadrática



- Lowess (h) \Rightarrow Lowess robusto: pesos robustos (g)

$$\hat{\varepsilon}_j = y_j - \hat{y}_j \quad Md: \text{mediana dos valores } |\hat{\varepsilon}_j|$$

$$g(u) = \begin{cases} (1 - |u|^2)^2 & \text{se } |u| < 1 \\ 0 & \text{cc} \end{cases} \Rightarrow \text{o peso atribuído a } (x_k, y_k) \text{ é } g(x_k) = g\left(\frac{\hat{\varepsilon}_k}{6Md}\right)$$

$$\hat{y}_j = \hat{\alpha} + \hat{\beta} x_j; \quad \sum_{k=1}^n g h(x_k) (y_k - \alpha - \beta x_k)^2$$

Example 2.14 Smoothing Splines

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

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

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

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

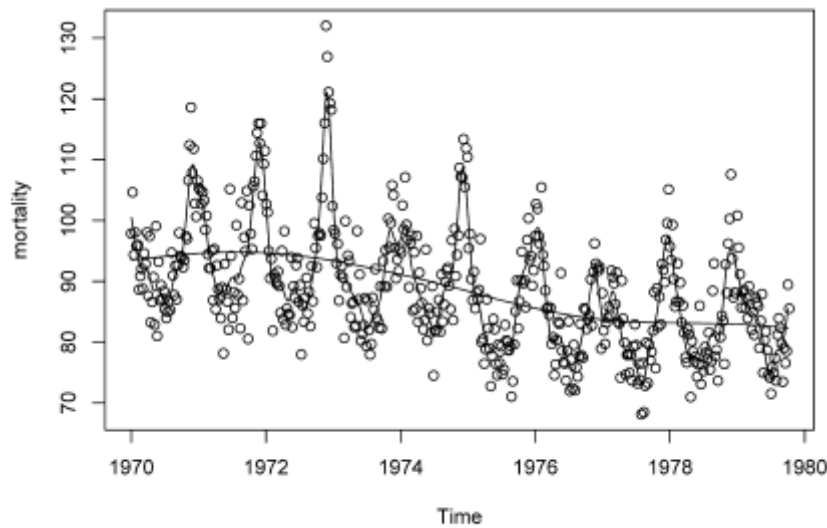


Fig. 2.15. Smoothing splines fit to the mortality data.

Example 2.15 Smoothing One Series as a Function of Another

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

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.

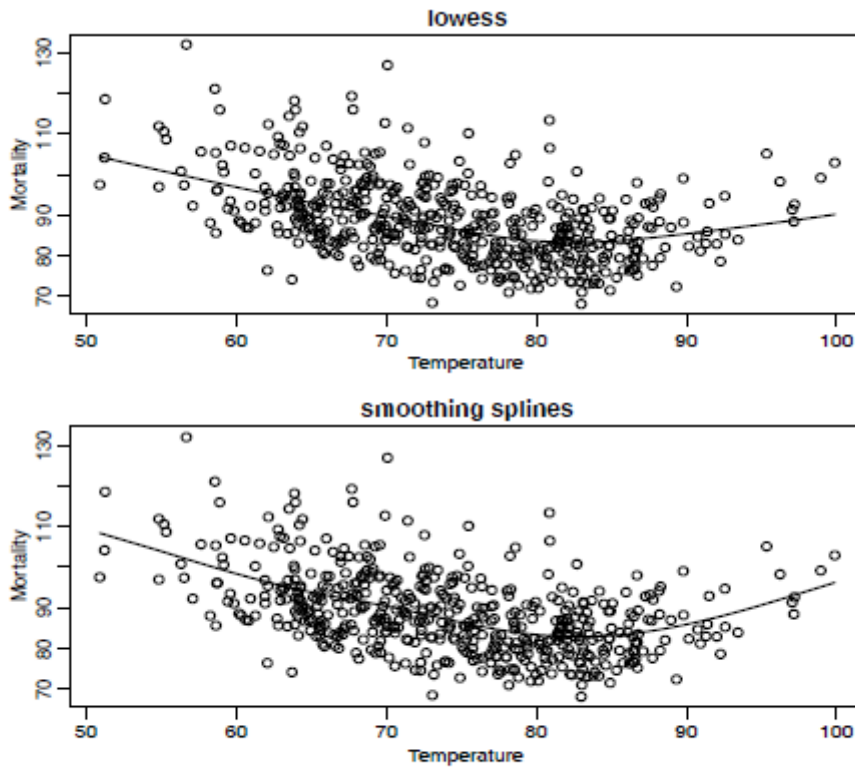


Fig. 2.16. Smoothers of mortality as a function of temperature using lowess and smoothing splines.

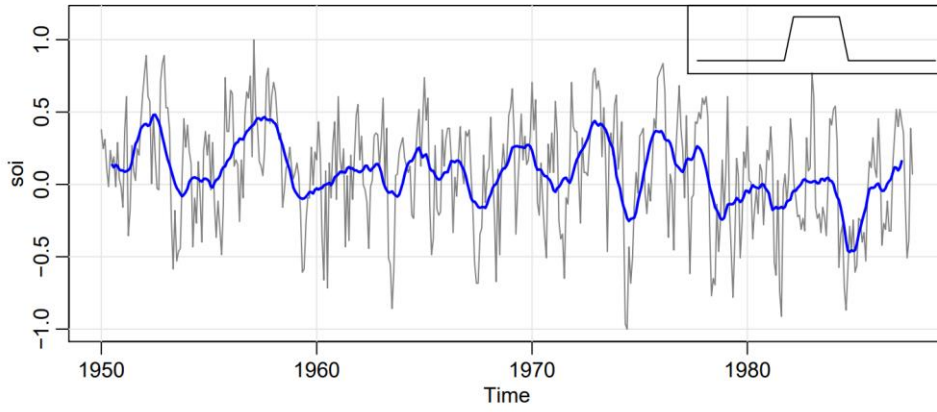


Fig. 2.12. Moving average smoother of SOI. The insert shows the shape of the moving average (“boxcar”) kernel [not drawn to scale] described in (2.39).

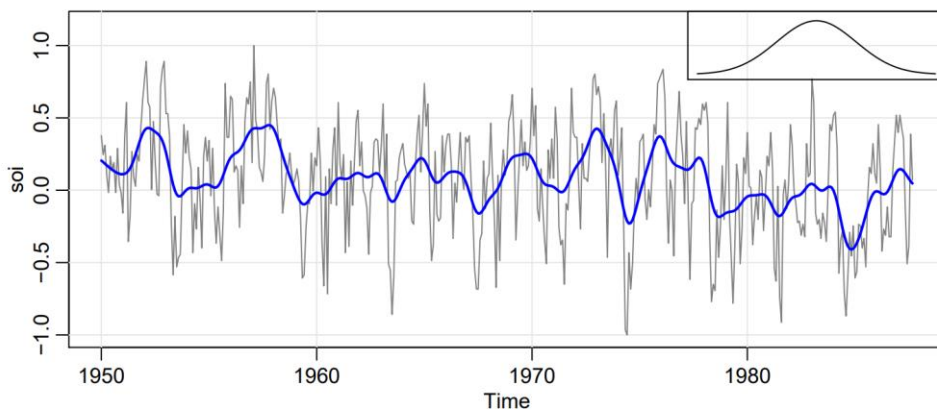


Fig. 2.13. Kernel smoother of SOI. The insert shows the shape of the normal kernel [not drawn to scale].

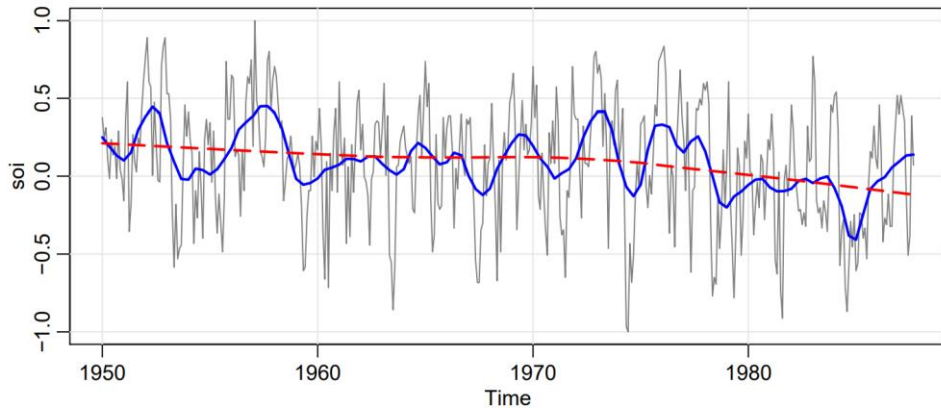


Fig. 2.14. Locally weighted scatterplot smoothers (*lowess*) of the SOI series.

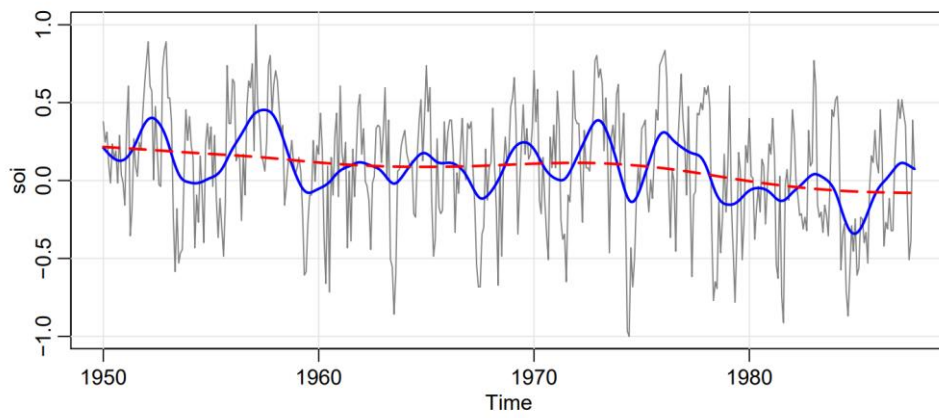


Fig. 2.15. Smoothing splines fit to the SOI series.

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

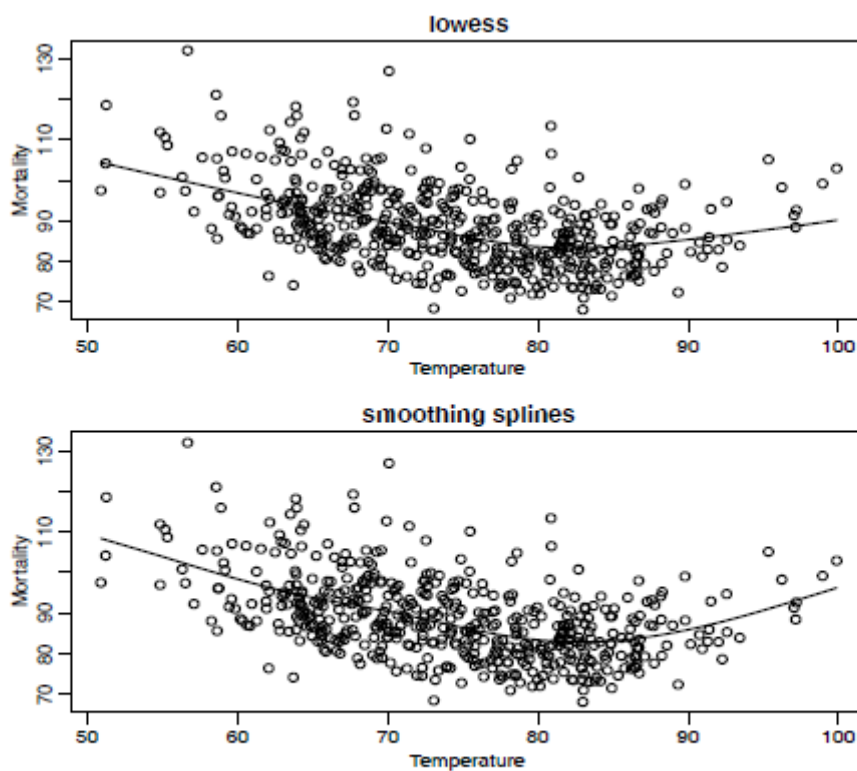


Fig. 2.16. Smoothers of mortality as a function of temperature using lowess and smoothing splines.