Statistics for Biology and Health

Adrian G. Barnett · Annette J. Dobson

Analysing Seasonal Health Data

Springer

Statistics for Biology and Health

Series
M. Gail
K. Krickeberg
J. Samet
A. Tsiatis
W. Wong

For further volumes:
http://www.springer.com/series/2848

Adrian G. Barnett · Annette J. Dobson

Analysing Seasonal Health Data

Springer

Dr. Adrian G. Barnett
Queensland University of Technology
Institute Health and Biomedical Innovation
and School of Public Health
60 Musk Avenue
Kelvin Grove QLD 4059
Australia
a.barnett@qut.edu.au

Prof. Annette J. Dobson
University of Queensland
School of Population Health
Herston Road
Herston QLD 4006
Australia
a.dobson@sph.uq.edu.au

ISSN 1431-8776
ISBN 978-3-642-10747-4 e-ISBN 978-3-642-10748-1
DOI 10.1007/978-3-642-10748-1
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009942900

© Springer-Verlag Berlin Heidelberg 2010
This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permissions for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.
The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To Hope, Mum and Dad

Preface

Seasonality in disease was first recognised by Hippocrates (460–370 BC) who said, "All diseases occur at all seasons of the year, but certain of them are more apt to occur and be exacerbated at certain seasons." We first encountered seasonality when examining time series of cardiovascular disease. We found a strong seasonal pattern with increases in winter and decreases in summer. Rather oddly, we found that warmer climates showed the biggest seasonal changes. This odd pattern was explained by studies which found that populations in colder climates had better insulated homes and wore more clothes in cold weather. Recent studies that improved home insulation found an improvement in residents' general health and reductions in their blood pressure.

By investigating seasonal patterns in disease it is possible to generate hypotheses about aetiology. The changes in the seasons cause changes in many environmental and social variables. These changes are repeated year after year and create a natural experiment for studying links between seasonal exposure and disease.

This book details a wide variety of methods for investigating seasonal patterns in disease. We use a range of health examples based on data collected daily, weekly and monthly, and using Binomial, Poisson and Normal distributions. Chapter 1 introduces the statistical methods that we will build on in later chapters. In Chap. 2, we define a "season" and show some methods for investigating and modelling a seasonal pattern. Chapter 3 is concerned with cosinor models, which are easy to apply but have some important limitations. In Chap. 4, we show a number of different methods for decomposing data to a trend, season(s) and noise. Seasonality is not always the focus of the study; in Chap. 5, we show a number of methods designed to control seasonality when it is an important confounder. In Chap. 6, we demonstrate the analysis of seasonal patterns that are clustered in time or space.

We hope this book will be accessible to non-statistical researchers as well as statisticians. To aid its implementation we have created an R library "season" that contains most of the functions and data.

The methods shown in this book are all based on the Gregorian calendar running from January to December, but any of the calendar-based methods could be equally applied to the Islamic calendar.

Acknowledgements

Thanks to Professors Alun Evans and Frank Kee at the Centre of Excellence for Public Health, Queen's University, Belfast, for providing a quiet space for writing.

This book would not have been possible without access to such interesting data. Our thanks go to: Queensland Health for the data on still births; Professor John McGrath, University of Queensland, for the schizophrenia data; Professor Elizabeth Eakin and colleagues, University of Queensland, for the exercise data; Dr Dan Mullany, The Prince Charles Hospital, Brisbane, for the long-term survival data; Ms Hope Becker for the footballers data; the UK Office of National Statistics for the UK births data; and the Department of Biostatistics at the Johns Hopkins Bloomberg School of Public Health and the Health Effects Institute, for making the National Morbidity and Mortality Air Pollution Study data publicly available.

Thanks to Dr Peter Baker for help with creating the R library. Thanks also to Dr John Fraser for organising the celebrations.

Computational resources and services used in this work were provided by the High Performance Computer and Research Support Unit, Queensland University of Technology, Brisbane.

Brisbane, *Adrian Barnett*
September 2009 *Annette Dobson*

Contents

Acronyms

ACF	Autocovariance *or* autocorrelation function
AFL	Australian Football League
AIC	Akaike information criterion
BMI	Body mass index
CAR	Conditional autoregression
CI	Confidence interval *or* credible interval
CVD	Cardiovascular disease
DIC	Deviance information criterion
EPL	English Premier League
ERR	Exposure–risk relationship
GAM	Generalized additive model
GLM	Generalized linear model
GLMM	Generalized linear mixed model
MCMC	Markov chain Monte Carlo
MVN	Multivariate normal
NMMAPS	National morbidity and mortality air pollution study
ppb	Parts per billion
RR	Rate ratio
RSS	Residual sum of squares
SEIFA	Socio-economic indexes for areas
SOI	Southern oscillation index
STL	Seasonal-trend decomposition procedure based on loess
A	Amplitude
P	Phase
s	Season
x	Independent variable
y	Dependent variable
δ	Month
ε	Residuals or noise
μ	Trend or mean
ρ	Correlation
σ	Standard deviation
ω	Frequency

Chapter 1
Introduction

1.1 Example Data Sets

This section describes the example data sets that we will use to demonstrate methods of analysing seasonal data. The examples aim to cover a range of health outcomes and measurement scales. The diet and exercise example uses *continuous* body mass index data that may have a Normal distribution. The cardiovascular disease data are *counts* that may have a Poisson distribution. The stillbirth data are *binary* and will have a Binomial distribution. The cardiovascular disease, schizophrenia and flu data sets are *time series*, as the results are measured at successive and equally spaced times. The exercise data are from an *intervention study*, and the times of observations depended on when people joined the study.

1.1.1 Cardiovascular Disease Deaths

Figure 1.1 shows the monthly counts of cardiovascular disease (CVD) deaths in people aged $\geq$75 in Los Angeles for the years 1987–2001 (14 years of data, 168 months). The data are from the National Morbidity and Mortality Air Pollution Study (NMMAPS) study [75]. There is a large peak in cardiovascular deaths every winter and a dip in summer. There is also a general decline in the average number of deaths from the start of the study to around 1992. Additionally there are also smaller peaks in deaths in some summers.

The data are counts and so we should consider using methods that assume the response has a Poisson distribution. However, the mean number of deaths is very large, and so the Normal approximation to the Poisson distribution may well apply, even though the data have a positive skew because of the large winter peak in deaths.

The data are arranged with one row per month (per year). The first three rows and last row of data are shown in Table 1.1. The variable "pop" is the population size which was only estimated in 2000 and so is the same for every row. "tmpd" is the mean monthly temperature in degrees Fahrenheit. "cvd" is the monthly total number of CVD deaths. "yrmon" is the fraction of time given by year + (month − 1)/12.

A.G. Barnett and A.J. Dobson, *Analysing Seasonal Health Data*, Statistics for Biology and Health, DOI 10.1007/978-3-642-10748-1_1,
© Springer-Verlag Berlin Heidelberg 2010

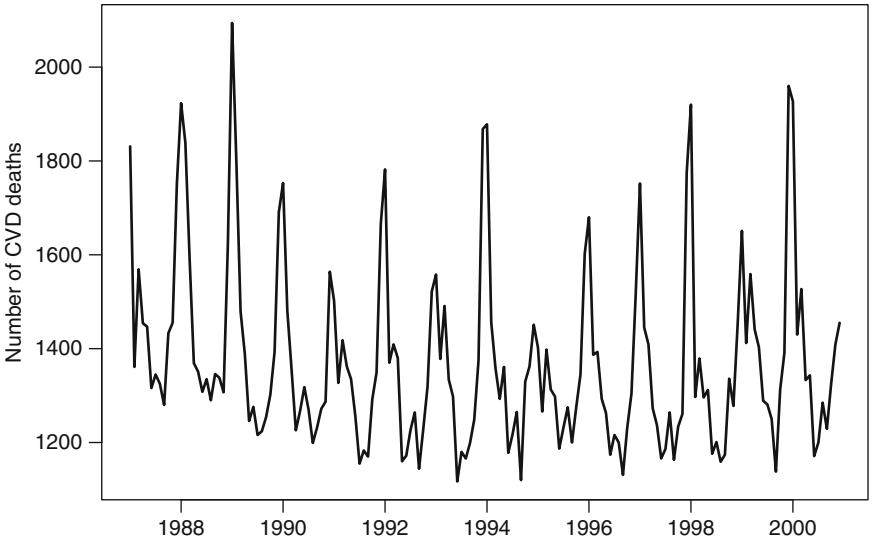


Fig. 1.1 Monthly counts of cardiovascular disease deaths in people aged ≥75 in Los Angeles for the years 1987–2000

Table 1.1 First three rows and last row of data from the cardiovascular disease study

Year	Month	yrmon	pop	cvd	tmpd
1987	1	1987.000	429,474	1,831	54.75
1987	2	1987.083	429,474	1,361	57.98
1987	3	1987.167	429,474	1,569	58.97
⋮	⋮	⋮	⋮	⋮	⋮
2000	12	2000.917	429,474	1,455	58.03

This is useful for plotting the data on the correct time scale, for example using the R command:

```
> plot(CVD$yrmon,CVD$cvd)
```

The CVD data are also available as daily counts and daily temperatures using the `NMMAPSlite` package in R [66]. In this case the time series is 5,114 days long.

1.1.2 Schizophrenia

The time series shown in the left-hand panel of Fig. 1.2 shows the number of people with schizophrenia in Australia by their month and year of birth from 1930 to 1971. Schizophrenia is a serious mental disorder which significantly reduces quality of life and shortens life expectancy. We use the broad diagnostic criteria based on a

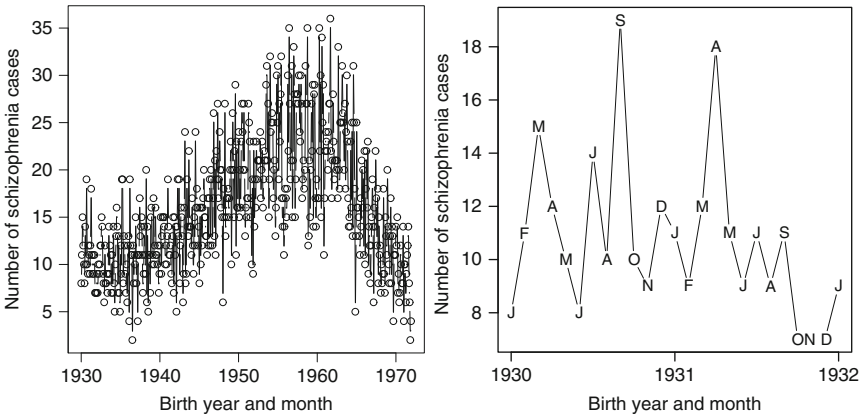


Fig. 1.2 Number of schizophrenia cases in Australia by date of birth from 1930 to 1971 (*left panel*), January 1930 to January 1932 (*right panel*)

broader definition of the number of symptoms needed for a sufferer to be classed as schizophrenic.

The monthly number of births for schizophrenics is shown in Fig. 1.2. The dominant feature of these data is the long-term trend, which rose from the late 1930s to 1960s, and then had a sharp decline from the 1960s onwards.

It is difficult to see any seasonal pattern in these data. The right-hand panel of Fig. 1.2 shows the data for the first two years. In this panel each month's starting letter is used to label the points. There is no clear seasonal pattern in this small section of data.

The schizophrenia data have a similar format to the cardiovascular disease data, as they both deal with monthly counts of disease, and are recorded as one month per row. The schizophrenia data cover a much longer time period, and an increase in the population in Australia from 1940 to 1960 is largely responsible for the increased trend. Data on the total number of births per month are also available and can be used as an *offset* (Sect. 1.4.6).

The schizophrenia data set also contains the southern oscillation index (SOI). The SOI is a weather variable that measures the monthly fluctuations in the air pressure difference between Tahiti and Darwin, Australia. Positive values for the SOI are usually accompanied by an increase in rainfall over Australia and hence a decrease in sunshine. Sunshine is the key producer of vitamin D and insufficient maternal vitamin D has been associated with an increased risk of schizophrenia [22].

There are 42 years of data in the study, giving 504 monthly observations. However, the number of people with schizophrenia born in January 1960 is missing. We keep a record in the data for January 1960, but set the number of births to missing. This ensures that the observations remain equally spaced.

1.1.3 Influenza

The flu is an infectious disease that flourishes in cold temperatures. Most people with the flu suffer pains, headache and coughing. In frailer people the symptoms and consequences can be more serious and can lead to death.

Figure 1.3 shows the weekly number of influenza cases in two flu seasons using data from the United States. The number of flu cases is monitored weekly by the Centers for Disease Control and Prevention [15]. Week 1 is the first week of January. The number of cases is monitored from week 40 (October) in one year to week 20 (May) in the next year. This period should capture the flu season. In 2006 the monitoring continued until week 25, as the number of cases was still reasonably large in week 20.

The data are arranged with one row per week. The first three rows and last row of the data are shown in Table 1.2. The data contain the weekly counts of four different types of influenza: B, A (unsubtyped), A (H1) and A (H3).

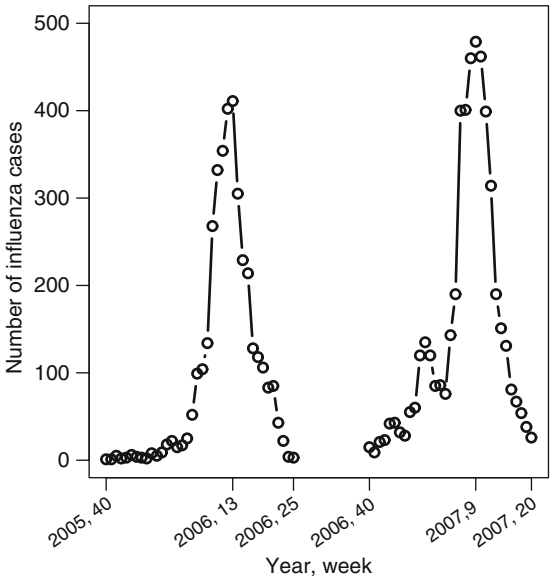


Fig. 1.3 Weekly number of positive influenza type B samples from the Centers for Disease Control and Prevention, United States, for the 2005/2006 and 2006/2007 flu seasons

Table 1.2 First three rows and last row of data from the influenza study

Year	Week	B	A_un	A_H1	A_H3
2005	40	1	4	0	6
2005	41	1	4	0	6
2005	42	5	6	0	12
⋮	⋮	⋮	⋮	⋮	⋮
2007	20	26	39	3	23

1.1.4 Exercise

Keeping physically active reduces the risks of chronic disease such as diabetes and hypertension. Levels of physical activity may depend on season, as many activities (e.g., walking) are done outdoors. Figure 1.4 shows the total walking time in the last week (in minutes) against date for 40 randomly selected subjects. We only plot 40 subjects because the plot becomes too busy if the data from all subjects are used. The data contain repeated results from the same subjects over time, and this design is known as a *longitudinal* study. It is difficult to see any seasonal pattern in the data, partly because walking time is strongly skewed, with lots of zeros.

The data are from a randomised controlled trial of a physical activity intervention in Logan, Queensland [31]. Subjects were recruited into the trial as they became available and so the dates of responses are not equally spaced. Subjects were eligible for the trial if they were not meeting Australian guidelines for adequate physical activity. Data were collected at an initial recruitment visit, and at two follow-up visits 4 and 12 months later. However, as Fig. 1.4 shows, not all subjects completed the follow-up. In total there were 434 subjects and 1,152 responses, giving an average of 2.7 responses per subject.

The data are arranged in longitudinal format with one row per visit. The first six rows of data are shown in Table 1.3. "NA" means missing in R, so the third response for both these subjects is missing as these subjects *dropped-out* from the study. The dates are in the format of day/month/year.

BMI was only measured at baseline. Walking time per week in minutes was measured at every follow-up. Walking time is therefore a *time-dependent* variable, whereas BMI is a *time-independent* variable.

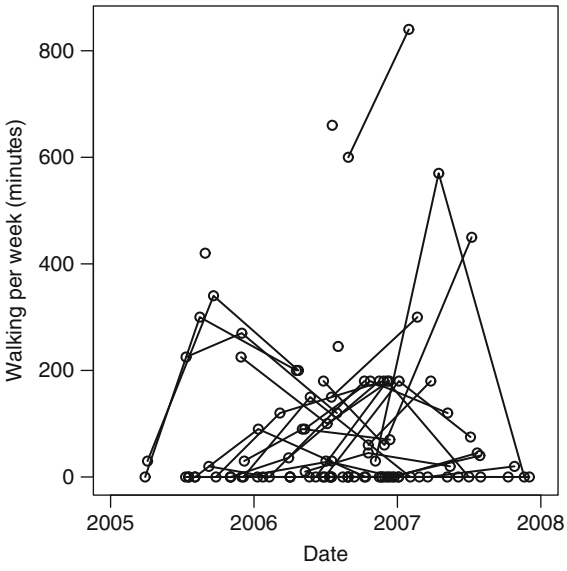


Fig. 1.4 Walking time in the last week (minutes) against date for 40 randomly selected subjects from the exercise study. Results from the same subjects are joined

Table 1.3 First six rows of data from the exercise study

ID	Visit	Date	BMI	Walking
1	1	5/4/2005	26.0	0
1	2	18/10/2005	26.0	60
1	3	NA	NA	NA
2	1	1/4/2005	22.5	0
2	2	22/8/2005	22.5	115
2	3	NA	NA	NA
⋮	⋮	⋮	⋮	⋮

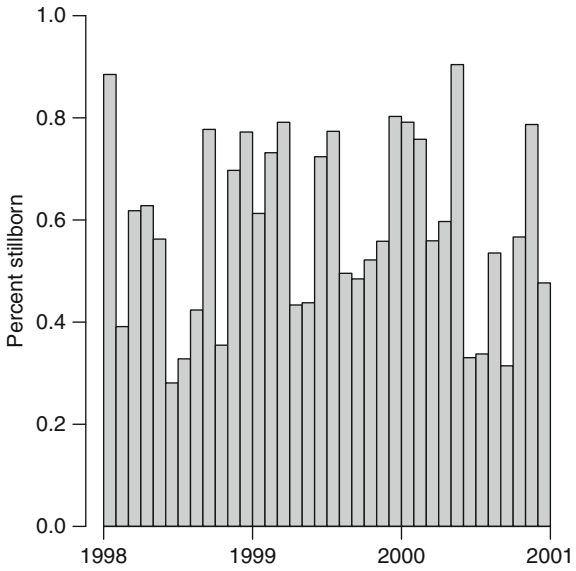


Fig. 1.5 Monthly percentages of stillbirth in Queensland, 1998–2000 ($n = 60, 110$)

The data from this study are irregularly spaced, as the responses were recorded on the dates when the interviews were conducted. This is in contrast to the CVD time series data, which were collected at regular monthly intervals. The results are also *unbalanced* as some subjects gave only one response, whereas others gave two or three. The data would only have been balanced if every subject gave the same number of responses.

1.1.5 Stillbirths

Figure 1.5 shows the monthly percentages of stillbirth from Queensland by month for the years 1998–2000. Although we have plotted the monthly rates the data are in the form of individual births, with dates of birth. To preserve anonymity the dates of birth have been scrambled by randomly re-ordering the day of birth. The first three rows of the data are shown in Table 1.4.

Table 1.4 First three rows of data from the stillbirth study

Dateb	Gestation	SEIFA	Stillborn
01/01/1998	39	4	0
01/01/1998	40	1	0
01/01/1998	40	2	0
⋮	⋮	⋮	⋮

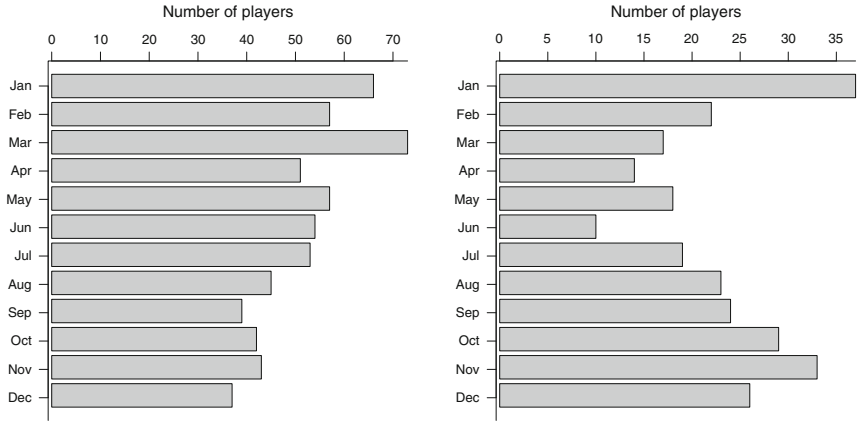


Fig. 1.6 Frequencies of months of birth of Australian Football League (AFL; *left panel*) and English Premier League (EPL; *right panel*) players

The main dependent variable is binary: stillborn (yes $= 1$ / no $= 0$). Incidence of stillbirths during this period was very low, there were 352 stillbirths out of 60,110 births in total (0.6%). This is a large data set, as it is only possible to look for seasonality in such rare conditions with a large number of subjects. The data also contains information on the length of gestation (in weeks), and an area level measure of socio-economic status called the SEIFA score (scaled to quintiles) [84].

It is difficult to spot a clear seasonal pattern in Fig. 1.5, but the rates are generally higher in summer and lower in winter (southern hemisphere data). This is not the best way to present these data, and we show an alternative method in Sect. 2.2.3.

1.1.6 Footballers

Figure 1.6 shows the distribution of players' months of births for the Australian Football League (AFL) and English Premier League. The AFL data is for the 2009 season (617 players), and the EPL data for the 2008/2009 season (272 players). Foreign born players were excluded in both data sets. Players' birthdays do not appear to be evenly distributed throughout the year in either sport. The total number of births by month in Australia and the UK is also available and we can use this information to calculate the expected number of births in each month.

1.2 Time Series Methods

In this section we give some time series methods that will be useful for, and expanded in, later chapters. We only give a brief introduction to time series. Three good introductory books on time series are by Chatfield [16], Diggle [25] and Fuller [38].

We restrict ourselves to equally spaced time series, with an index for time $t = 1, 2, \ldots, n$, and use y_t to refer to the single observed value of the series $\mathbf{y}$ at time t. We refer to an entire time series using bold font $\mathbf{y}$. For the schizophrenia data (Fig. 1.2) the first four observations of the time series are $\mathbf{y} = 8, 11, 15, 12, \ldots$. These observations are referred to using the lower case letter $\mathbf{y}$ whereas unknown or modelled values are referred to using the upper case letter. For example, Y_{505} is the predicted number of people with schizophrenia born in January 1972 (which is one month after the study ended).

We assume that any time series can be split into two parts: *signal* and *noise*,

$$Y_t = X_t + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{1.1}$$

where n is the total length of the time series or *sample size*. In this equation X_t represents the signal part of the series, and the Greek letter ε_t represents the noise (also referred to as *residual* or *error*). Time series in health always contain some noise, and common causes of this noise are measurement error or random variability. In general, we want to remove the noise so that we can focus on the signal (if there is any), as the signal will tell us what patterns exist in the series, whereas the noise should be unpredictable and without any interesting patterns. We should perform *residuals checks* to verify the approximate randomness of the residuals (Sect. 1.4.3).

The *signal-to-noise* ratio gives the relative amount of signal to noise. The lower the signal-to-noise ratio the harder it will be to extract the signal. Looking again at the plot of the schizophrenia data (Fig. 1.2) we can expect a low signal-to-noise ratio as the number of people with schizophrenia is thankfully small. In contrast the cardiovascular data (Fig. 1.1) have a strong seasonal signal partly because of the much larger number of cases. There is some variation from year to year, and we discuss whether this is signal or noise later in Sect. 2.2.2.

The signal part of the time series in (1.1) can be split into different kinds of signal, for example,

$$X_t = Z_t + S_t, \qquad t = 1, \ldots, n,$$

where Z_t captures the long-term trend in X_t, and S_t captures the seasonality. We discuss *decomposing* a time series into separate signals in Chap. 4.

Bold letters $\mathbf{X}$, $\boldsymbol{\varepsilon}$, $\mathbf{Z}$ and $\mathbf{S}$ are used to denote the entire series of values of X_t, ε_t, Z_t and S_t.

1.2.1 Autocovariance and Autocorrelation

In this section we explain autocovariance and autocorrelation, which are useful statistics for describing time series. These statistics use the sample mean and variance, so we start by explaining these most fundamental statistics. Also, before using any statistical tests for the autocovariance or autocorrelation, we briefly outline the principles of hypothesis tests and statistical inference.

1.2.1.1 Sample Mean, Variance and Median

For an observed time series we define the *sample mean* as

$$\overline{y} = \frac{1}{n}\sum_{t=1}^{n} y_t.$$

The mean is a useful measure of the centre of a time series. A useful measure of the spread of the data is the *standard deviation*. For an observed time series we define the sample standard deviation as

$$s = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \overline{y})^2}. \tag{1.2}$$

If the time series data have a Normal distribution, then a 95% *tolerance interval* is defined as

$$\overline{y} \pm 1.96s,$$

where 1.96 is the value from a standard Normal distribution such that $P(-1.96 < Z < 1.96) = 0.95$, where $P()$ means probability. A 95% tolerance interval gives the limits within which 95% of the data would be contained if the observations could be repeated.

For the cardiovascular disease data (Sect. 1.1.1) the mean number of deaths per month is 1,373.2 and the standard deviation is 194.5. This gives a 95% tolerance interval of 992.0 to 1,754.4. The mean and tolerance interval are plotted in the second panel of Fig. 1.7. In this example the lower limit of the tolerance interval is outside the observed range of the data. This is because the distribution of the data is not Normally distributed but is positively skewed, as shown by the histogram.

Two better summary statistics to describe the centre and spread of these data are the *median* and *inter-quartile range*, respectively. The interquartile range goes from the first quartile to the third quartile. The first quartile is the point that divides the data into the lower quarter and upper three-quarters (the median is the second quartile). The interquartile range is a useful measure of spread as it contains the central 50% of the data. The first quartile of the data is also known as the 25th *percentile*. For the number of cardiovascular deaths the median is 1,325 and the second and third quartiles are 1,248 and 1,435, respectively.

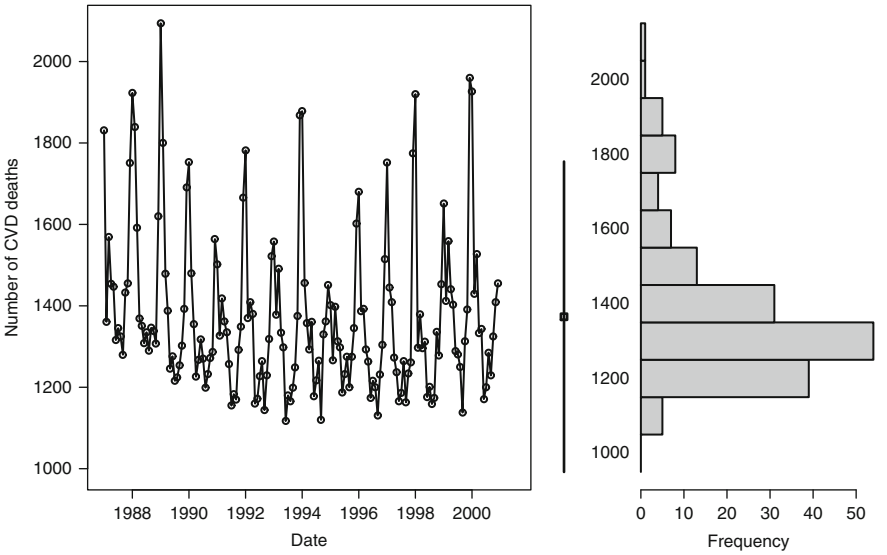


Fig. 1.7 A time series (*left panel*) and histogram (*right panel*) of the monthly counts of cardiovascular disease deaths in people aged ≥75 in Los Angeles for the years 1987–2000. Mean (*square*) and 95% tolerance interval (*vertical line*) in the *centre panel*

The mean, median and standard deviation are all *marginal* statistics, meaning they describe the distribution of the data in one dimension (that is, collapsed over time). Importantly, they do not describe how the series changes over time. Similarly the histogram in Fig. 1.7 describes the marginal distribution of the data. To illustrate how the mean, median and histogram do not capture changes in time, we could rearrange the order of the series (e.g., swapping 1988 and 1989) but would obtain the same mean, median and histogram.

1.2.1.2 Hypotheses Testing

A statistical hypothesis test is used to make formal decisions based on observed data. For example, we might be interested in the mean of a random variable X. The null and alternative hypothesis could be:

- H_0: the mean of X is zero ($E(X) = 0$).
- H_1: the mean of X is not zero ($E(X) \neq 0$).

H_0 is the *null hypothesis* and H_1 is the *alternative hypothesis*. We gather some data $(x_1, x_2, \ldots, x_n)$, and then calculate a *test-statistic* and associated *p-value*. The p-value is the probability of observing more extreme data assuming that the null hypothesis is true. So smaller p-values indicate less support for the null hypothesis and greater *statistical significance*.

Table 1.5 Four possible outcomes depending on which hypothesis is true and whether the null hypothesis is accepted

	H_0 accepted	H_0 rejected
H_0 is true	Correct decision	Type I error
H_1 is true	Type II error	Correct decision

Standard practice is often to reject the null hypothesis if the p-value is less than 0.05. In other words, if there is a 1 in 20 chance or less of observing more extreme data when the null hypothesis is true. For simplicity we say that H_0 is "accepted" if it is not rejected, although this is not strictly correct – it is really just not rejected but still retained as a possibility.

As soon as we make a decision about which hypothesis to accept, there is a chance that we have made the wrong decision, as shown in Table 1.5.

A Type I error is also known as a *false positive*, and a type II error as a *false negative*. The probability of a Type I error is equal to the level of statistical significance (usually arbitrarily set at 0.05). Although this probability is only meaningful when considering the long-term frequency of Type I errors, as for an individual test we will have either made completely the wrong or completely right decision.

In the epidemiological literature there are many examples of an over-reliance or abuse of p-values [81]. In this book we use p-values to judge the likelihood that the null hypothesis is true, and avoid the arbitrary splitting of results into statistically significant and statistically non-significant. We do this by also considering *clinical significance*, when we take in account if the observed result is likely to have noticeable implications in practice. Judging this significance is aided by using, where possible, confidence intervals as well as (or instead of) the p-value.

In later chapters we use Bayesian methods, whose principles and inferences depend less on p-values.

1.2.1.3 Autocovariance

A statistic that describes the pattern of a series over time is the *autocovariance*. "Auto" here meaning from the same series, and "covariance" meaning how it varies together. The sample autocovariance between observations k times apart is given by

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(y_t - \overline{y})(y_{t+k} - \overline{y}), \qquad k = 0, 1, \ldots, n-1, \tag{1.3}$$

where $\overline{y}$ is the sample mean. We calculate this statistic for increasing values of $k = 0, 1, 2, \ldots, m$, up to some limit m, and then plot the statistic against k. The value k is referred to as the *lag* as it describes the time difference between two observations (and m is the maximum lag). This plot is sometimes called the correlogram [16]. The plot will highlight how observations vary together and is useful for detecting

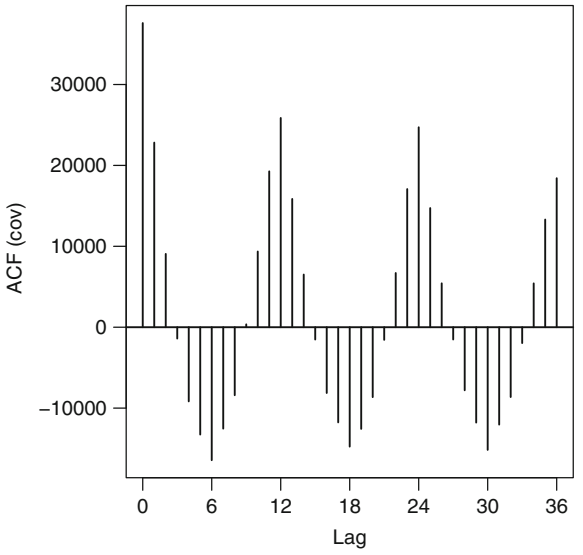


Fig. 1.8 Autocovariance of the cardiovascular disease data for lags $k = 0, \ldots, 36$

periodic patterns in the series. Note that

$$c_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})^2,$$

is the sample variance s^2; that is, the square of the sample standard deviation s in (1.2), except that the denominator is n rather than $n - 1$.

We have plotted the autocovariance for the cardiovascular disease data in Fig. 1.8. This plot was created using the R command:

```
> acf(CVD$cvd,type="covariance",demean=TRUE,
  lag.max=36)
```

The option `demean` ensures that the sample mean ($\overline{y}$) is subtracted from each observation as in our definition (1.3). The option `lag.max=36` shows the covariance for observations up to 3 years (36 months) apart (the value m defined above).

The autocovariance in Fig. 1.8 shows steady rises and falls that mirrors the rises and falls in the original data (Fig. 1.1). The largest autocovariance is at lag zero (which is – approximately – the variance). The next largest is 12 months later, which indicates a positive covariance between the months in neighbouring years (e.g., January 1987 and January 1988). This covariance persists for the same months in non-neighbouring years, as shown by the peaks at $k = 24$ and 36, but the covariance has become smaller with the widening time gap. Results that are six months apart (i.e., in the opposite season) have a negative covariance, reflecting the fact that the highest peaks in the data are followed six months later by the lowest points.

1.2.1.4 Autocorrelation

The autocorrelation is the sample autocovariance (1.3) divided by the autocovariance at lag zero (c_0^2). The equation to estimate the autocorrelation is

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \overline{y})(y_{t+k} - \overline{y})}{\sum_{t=1}^{n}(y_t - \overline{y})^2},$$

which gives us the correlation between observations k distances apart. Confusingly both the autocovariance and autocorrelation are referred to as the ACF.

The autocorrelation is often more useful than the autocovariance because it is on a standardised scale, with a range from -1 for observations that are perfectly negatively correlated to 1 for observations that are perfectly positively correlated. An autocorrelation of 0 indicates no correlation between the observations at the two time points.

We can calculate and plot the autocorrelation using the R command:

```
> acf(CVD$cvd,type="correlation",demean=TRUE)
```

This command produces the plot shown in Fig. 1.9. Data from the same time point (lag 0) are perfectly correlated ($r_0 = 1$). Data one observation apart (lag 1) are strongly positively correlated, with a correlation around 0.6. With increasing distance the observations become less correlated until at six observations apart the results are negatively correlated. The patterns in Figs. 1.8 and 1.9 are identical, only the scales on the y-axes are different.

Another advantage to plotting the autocorrelation instead of the autocovariance, is that we can add horizontal lines that examine the statistical significance of the

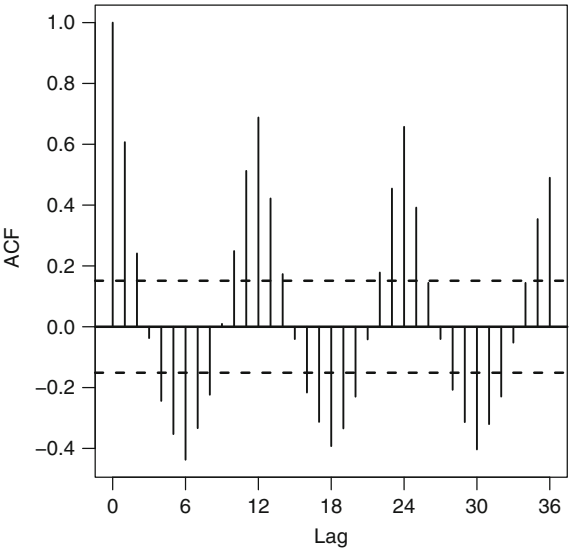


Fig. 1.9 Autocorrelation of the cardiovascular disease data for lags $k = 0, \ldots, 36$

autocorrelations. These horizontal lines correspond to the largest autocorrelation we would expect to see if the series were completely random (and so had no autocorrelation). These limits are shown in Fig. 1.9 at ± 0.15 (explained below), and clearly many of the autocorrelations for this series exceed these limits, which strongly indicates that the series is not random.

If the series was completely random then we would expect the autocorrelations to be zero for $k > 0$, but because of random noise they would never be perfectly zero. The limits are derived from the fact that a completely random series has the approximate variance [16]

$$\text{var}(r_k) \simeq \frac{1}{n}.$$

So approximate 95% confidence limits for the autocorrelation can be constructed using $0 \pm 1.96 \times \sqrt{1/n}$. We recommend that these limits are used as a visual guide to examine autocorrelation, and not to formally reject the null hypothesis that the series is random if any of the points are beyond the limits (where $0 < k \leq m$). Firstly, multiple hypothesis testing is involved (for each value of the lag k) and for large values of m (the maximum lag) the chance of making a type I error is high. Second, the autocorrelation describes just one aspect of the randomness of a series, and it may be that complex (non-random) patterns exist which cannot be detected by the autocorrelation [10].

It is worth noting that the number of observations available for calculating the autocorrelation and autocovariance decreases for larger values of k. So larger values of c_k and r_k are based on less observed data and so should be interpreted with more caution than values at shorter lags.

Patterns in the autocorrelation plot may not always be as clear as for the cardiovascular series. Figure 1.10 shows the autocorrelation for the first 10 years of the schizophrenia data, again for observations up to 3 years apart. In this plot only two values are outside the limits expected if the series was random, at $k = 4$ and 15, and both are only just outside the limits of ± 0.18. There is also no clear cyclic pattern in the plot, which indicates a lack of seasonal pattern in the data.

1.3 Fourier Series

The Fourier representation of a time series is a key approach for modelling seasonality. It uses the trigonometric functions, cosine and sine, so we first introduce these important functions. We refer to cosine and sine functions as *sinusoidal*.

1.3.1 Cosine and Sine Functions

Figure 1.11 shows the cosine and sine functions over the range 0 to 4π. In this range both functions have moved through two complete *cycles* (or *periods*). The

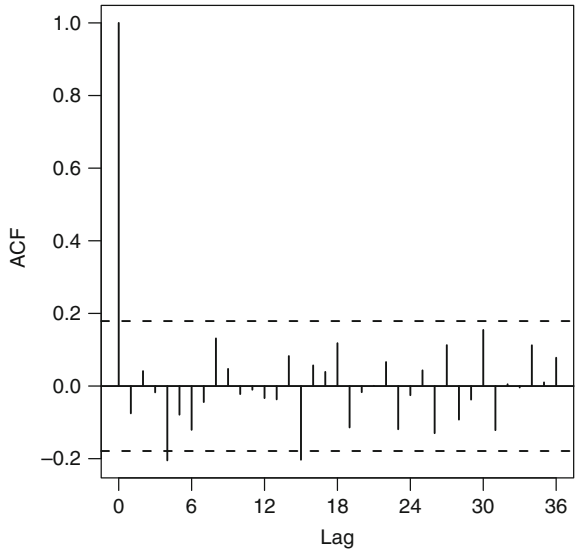


Fig. 1.10 Autocorrelation of the first 10 years of the schizophrenia data for lags $k = 0, \ldots, 36$

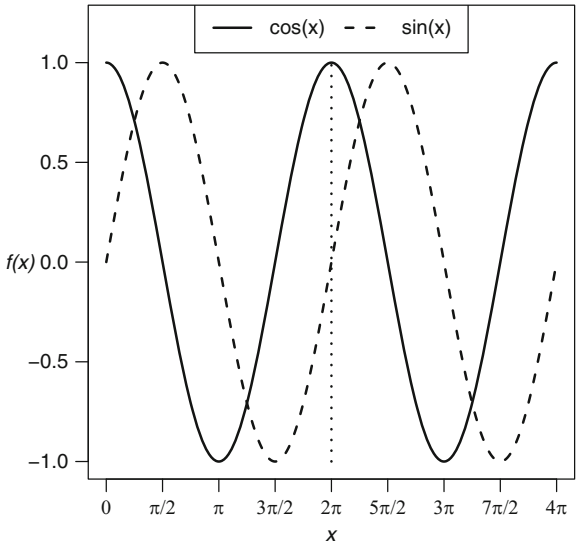


Fig. 1.11 Cosine and sine functions over the range 0 to 4π. *Dotted vertical line* at 2π

cosine function started at a height of 1, $\cos(0) = 1$, then dipped to a value of -1 at π, $\cos(\pi) = -1$, and returned to 1 at 2π, $\cos(2\pi) = 1$. The sine function followed the same pattern, but it started at zero. In fact $\sin(a + \pi/2) = \cos(a)$ for any value of a. So the sine curve is equal to a cosine curve that has been shifted to the right by $\pi/2$.

The steady rise-and-fall of the cosine and sine functions makes them ideal for modelling seasonality. Another useful property is that this rise-and-fall pattern is repeated. So the patterns from 2π to 4π are the same as those in 0 to 2π.

The repeating property of the cosine and sine functions means that we only need to consider times from $(0, 2\pi]$. The usual curved parenthesis ((), means "greater than", and the square parenthesis (]), means "less than or equal to". Because of the repeating property, for any value of a and any integer value of k we have

$$\cos(a + 2\pi k) = \cos(a),$$
$$\sin(a + 2\pi k) = \sin(a).$$

The value 2π is a key constant because it is the circumference of circle with radius 1. Figure 1.12 shows a cosine and sine curve over a time period of 0 to 2π and the corresponding points on a circle. The top panel shows a time of $\pi/4$. The cosine curve started on the y-axis at $\cos(0) = 1$ at time 0, and has moved down to $\cos(\pi/4) = 0.707$ by $\pi/4$. The sine curve started at $\sin(0) = 0$ at time 0, and has moved up to $\sin(\pi/4) = 0.707$ by $\pi/4$. The accompanying circle starts at 3 o'clock and moves in an anti-clockwise direction. The point on the circle is defined by a right-angled triangle with an adjacent (horizontal) side of length of $\cos(\pi/4) = 0.707$ and opposite (vertical) side of length of $\sin(\pi/4) = 0.707$. The second time is at $\pi/2$, where the cosine function is zero, and the adjacent length of the triangle is zero (hence it appears as a vertical line). The third time is at π, where the cosine function is at its lowest value and the circle is at the opposite point from where it started. The final plot shows the completion of the curves and circle at a time of 2π.

Together, the cosine and sine functions can represent any point on the curve and the circle. They are called *trigonometric functions*, as trigonometry is the branch of mathematics that deals with triangles (and other shapes).

Figure 1.11 shows that the cosine and sine curves both have two turning points (at their maxima and minima). Apart from these points they are either increasing or decreasing. The rate of change is given by the first derivative which is

$$\frac{\mathrm{d}}{\mathrm{d}x}\cos x = -\sin x, \qquad \frac{\mathrm{d}}{\mathrm{d}x}\sin x = \cos x.$$

So the rate of change in $\cos(x)$ is given by $\sin(x)$ and vice versa. So when $\sin(x) = 0$, $\cos(x)$ is at one of its two turning points (Fig. 1.11). When $\sin(x) = -1$, $\cos(x)$ is at its fastest rate of increase. So the biggest changes occur half-way between the highest and lowest points, and the slowest changes occur close to the highest and lowest points.

We can illustrate this aspect of a sinusoidal pattern using astronomical data. Figure 1.13 shows the day length in New York for 2009, defined as the difference (in hours) between sunrise and sunset. The longest day is 21 June, the summer solstice. The difference in day length between 20 June and 21 June is just 1 second. The vernal equinox is 20 March. The difference in day length between 20 March and 21 March is 164 seconds, the biggest increase in the year. The day length does not perfectly match a sinusoidal function because the Earth's orbit is not perfectly circular (instead it is elliptical).

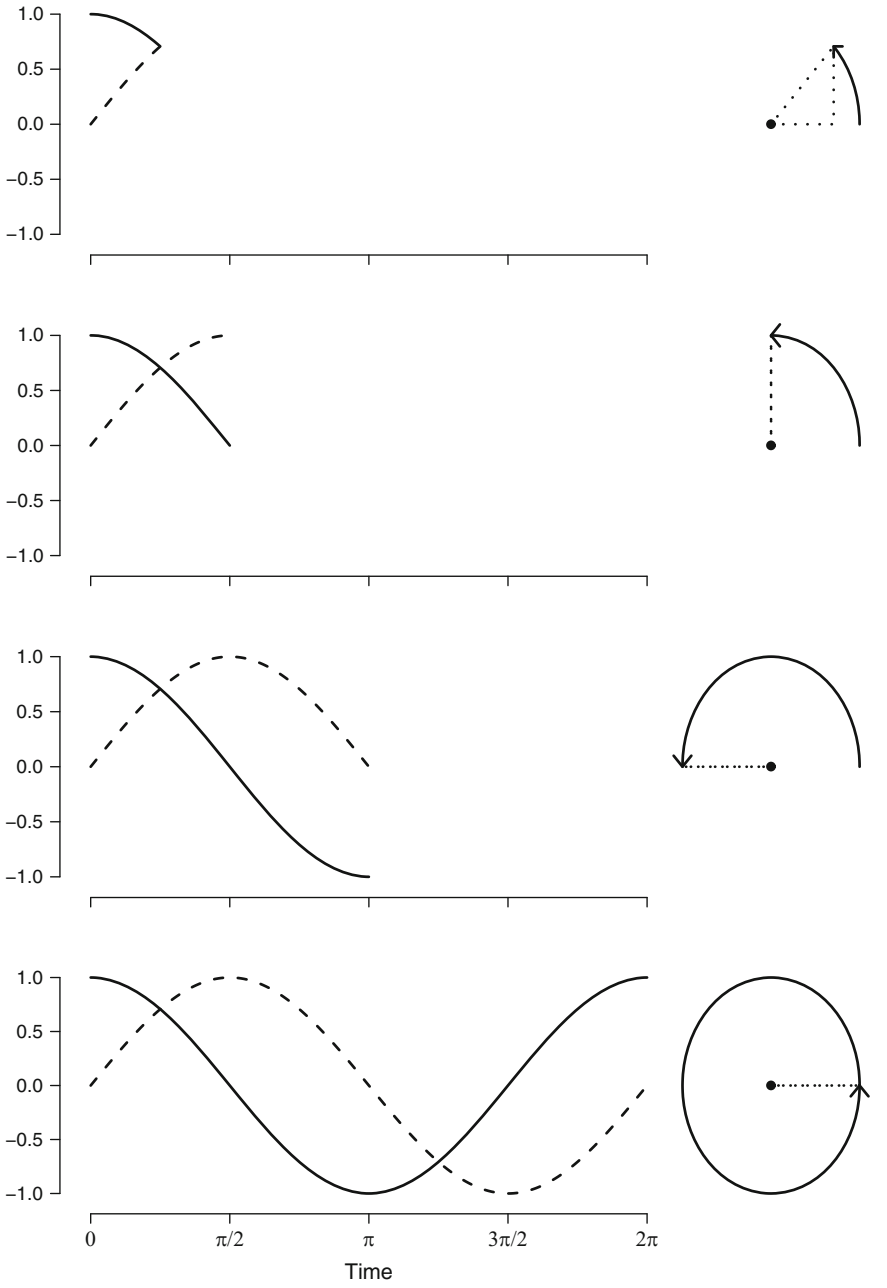


Fig. 1.12 Cosine and sine curves at four times and the corresponding points on a circle with centre at co-ordinates (0,0). The height of the *solid cosine curve* is equal to the length of the adjacent (horizontal) side of the triangle (*dotted line*). The height of the *dashed sine curve* is equal to the length of the opposite (vertical) side of the triangle (*dotted line*)

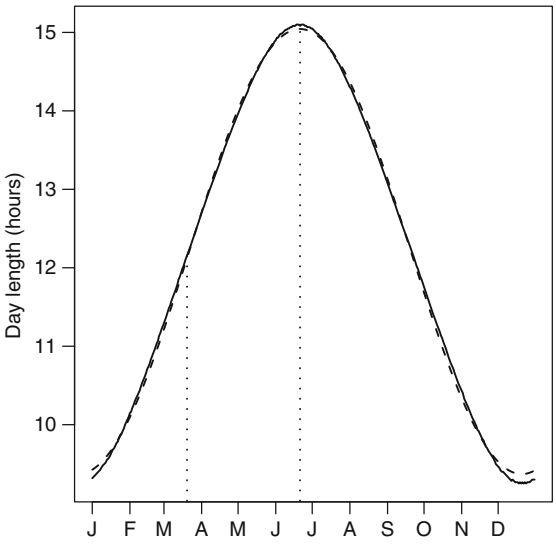


Fig. 1.13 Day length in New York for 2009 (*solid line*) and fitted sinusoid (*dashed line*). The *dotted vertical lines* show the vernal equinox (20 March) and summer solstice (21 June)

1.3.2 Fourier Series

To create a Fourier series we first modify a cosine function using the equation

$$Y_t = A\cos(\omega t - P). \tag{1.4}$$

A is the *amplitude* (or height) of the sinusoid, whilst ω is the *frequency* (cycle length) and P the *phase* (peak location). We use a minus sign in front of the phase so that it moves in a clockwise direction around the circle (Fig. 1.12).

Figure 1.14 illustrates the use of the formula (1.4), showing a sinusoidal pattern for two combinations of amplitude, frequency and phase. The solid line sinusoid has an amplitude of 1, which represents its maximum value. The dashed line has an amplitude twice the size. The solid line has a frequency of 1, which means it completes a cycle in the range 0 to 2π. The dashed line has a frequency of 2, which means it completes two cycles in the range 0 to 2π.

The phase of the solid sinusoid is $\pi/2$ and this is where it peaks. The phase of the dashed sinusoid is also $\pi/2$ but it peaks at $\pi/4$ and again at $5\pi/4$ because the phase is relative to the frequency. If we define the start as the point at which the sinusoids are zero, then the phase for both sinusoids occurs one-quarter of the way through their cycle. To calculate the point of the peak in time we use $P^* = P/\omega$. You can experiment with varying the amplitude, frequency and phase using the R function `sinusoid` from our "season" library.

The sinusoidal representation (1.4) has many advantages for modelling seasonality. It is a very *parsimonious* equation, as with just three parameters (A, ω and P) it is possible to describe a large variety of seasonal patterns. The sinusoid is symmetric about the horizontal axis and so the area of the curve above zero is equal to

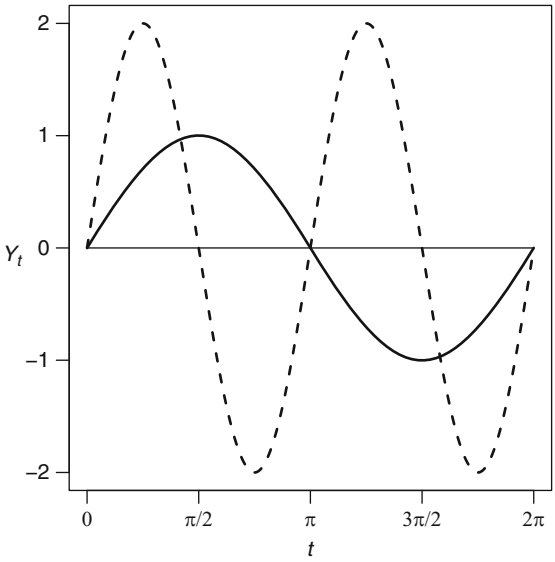


Fig. 1.14 Two sinusoids using the formula (1.4). The *solid sinusoid* has an {amplitude, frequency, phase} of $\{1, 1, \pi/2\}$, the *dashed sinusoid* $\{2, 2, \pi/2\}$. The *horizontal reference line* is at zero

the area of the curve below zero. The sinusoid is also symmetric about the vertical axis (using a point of symmetry at the phase), so the increase in the seasonal pattern is mirrored by the decrease. This property may be too restrictive for seasonal health conditions that do not have a steady rise-and-fall, and we investigate relaxing this feature in later sections.

It is possible to combine multiple sinusoids using the formula

$$Y_t = \sum_{j=1}^{m} A_j \cos(\omega_j t - P_j). \tag{1.5}$$

This is the finite version of the Fourier series. For our purposes this combination would be useful if there were multiple seasonal patterns in the data. Figure 1.15 shows the result of combining the two sinusoids from Fig. 1.14. Notice how the resulting time series looks like a sinusoid with a frequency of 2 combined with a downward trend.

Figure 1.16 shows the result of combining five sinusoids. Notice how the resulting time series looks very noisy, even though it is purely *deterministic*, meaning that there is no noise.

1.3.3 Periodogram

The periodogram is a useful graphical statistic for uncovering the important frequencies in a time series (i.e., the values of ω_j in (1.5)). The principle is that if we

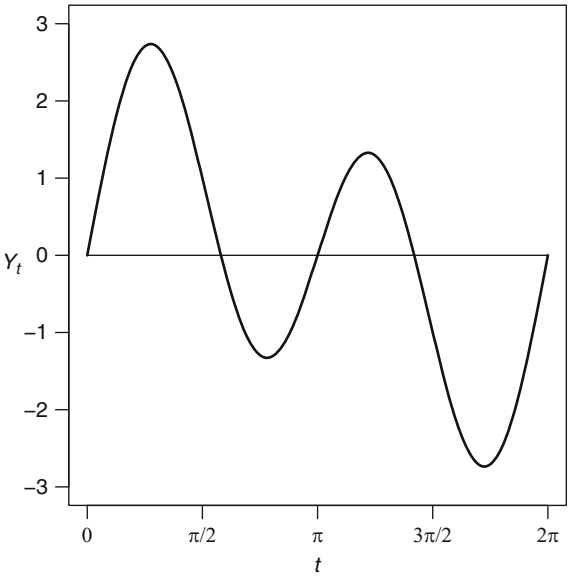


Fig. 1.15 Plot of the sum of the two sinusoids from Fig. 1.14. *Horizontal reference line* at zero

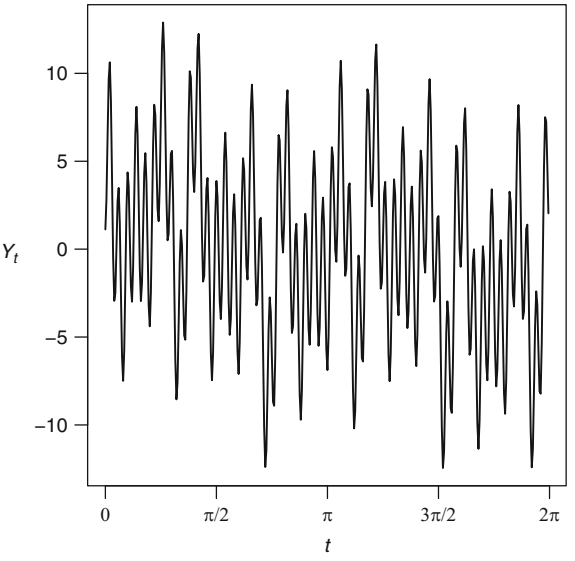


Fig. 1.16 Plot of the sum of five sinusoids with {amplitude, frequency, phase} of $\{1, 1, \pi/2\}$, $\{2, 2, \pi/2\}$, $\{3, 10, \pi/4\}$, $\{4, 15, 0\}$, $\{5, 50, \pi\}$

make m in (1.5) very large, and cover a range of frequencies, then we can closely represent any real time series using this equation. Figure 1.16 demonstrated how we can create a comparatively complex time series using only five sinusoids.

The frequencies examined by the periodogram range from the highest possible to the lowest possible. As an example, if we had monthly data collected over two years, then the highest possible frequency would be a sinusoid that changed from its lowest to highest point from one month to the next month (i.e., 2-month cycle), and the

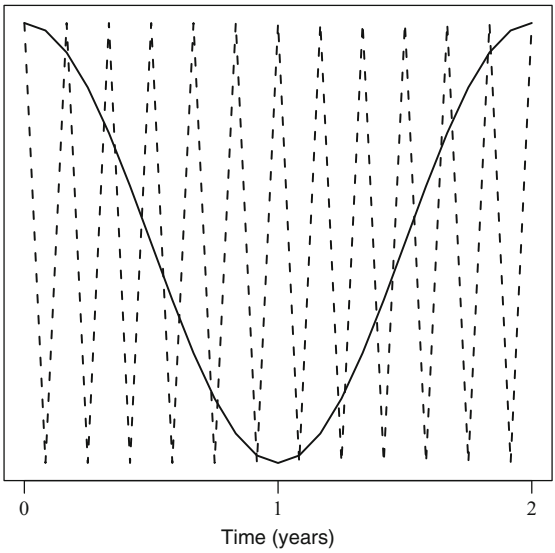


Fig. 1.17 Plot of the highest (*dashed*) and lowest (*solid*) observable frequencies for monthly data collected over two years

lowest possible observable frequency would complete a cycle over two years. This example is shown in Fig. 1.17. If there is a seasonal pattern that changes with a very high frequency, say from week-to-week, then we would not be able to observe this change using monthly data. We could only solve this problem by collecting more detailed data (i.e., sampling at least every week). Similarly, if there is a seasonal pattern that has a very long frequency (say every 4 years) then we would not be able to observe it in these data, and would need to collect data for a longer time.

The frequencies examined by the periodogram are called the *Fourier frequencies* defined as

$$\omega_j = \frac{2\pi j}{n}, \qquad j = 1, \ldots, n/2. \tag{1.6}$$

When $j = 1$, then $\omega_j = 2\pi/n$, which is the highest observable frequency. At the other end of the scale when $j = n/2$, then $\omega_j = \pi$, which is the lowest observable frequency.

We can express the Fourier frequencies as the number of time points needed for a complete cycle using $c_j = 2\pi/\omega_j$; or alternatively as the number of cycles per unit of time using $f_j = \omega_j/2\pi$. For the example in Fig. 1.17, we have the highest frequency of $c_1 = 2$ months (one cycle every two months), or $f_1 = 0.5$ (half a cycle per month); and the lowest frequency of $c_{12} = 24$ months, or $f_{12} = 1/24$ cycles per month.

The formula for the periodogram arises from expanded version of (1.5) (see [16] for details) and is

$$I(\omega_j) = \frac{2}{n}(\hat{C}_j^2 + \hat{S}_j^2), \qquad j = 0, \ldots, n/2, \tag{1.7}$$

where

$$\hat{C}_j = 2\sum_{t=1}^{n} y_t \cos(\omega_j t)/n,$$
$$\hat{S}_j = 2\sum_{t=1}^{n} y_t \sin(\omega_j t)/n,$$

where $\mathbf{y}$ are the observed data and $\boldsymbol{\omega}$ are the Fourier frequencies. The periodogram $I(\omega_j)$ is always positive, as it is the sum of two squared values, but it will be larger at frequencies that are strongly represented in the data. So we can give a graphical representation of the important frequencies in the data by plotting $I(\omega_j)$ against ω_j.

We can calculate and plot the periodogram using the R commands:

```
> spec<-spec.pgram(CVD$cvd,demean=TRUE)
> plot.spec(spec,ci=-1,main="",xlab="Frequency",
  ylab="Periodogram",log="no")
```

These commands produce the plot shown in Fig. 1.18. The frequency range on the x-axis goes from 0 to 0.5 (not 0 to 2π), as it is based on a modified version of the Fourier frequencies (1.6) with $\omega_j^* = \omega_j/2\pi$. So the number of times points for a complete cycle are given by $c_j^* = 1/\omega_j^*$. The phrases "low frequency" and "high frequency" used here are synonymous with those used in radio broadcasting.

There is a clear peak in the periodogram at a frequency of 0.0833, which gives a cycle of $1/0.0833 = 12$ months. So the periodogram has done a good job of finding the annual signal in the data. The second largest peak in the periodogram

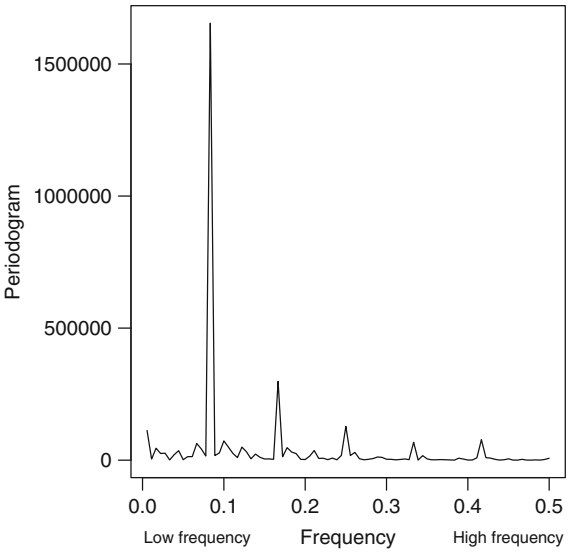


Fig. 1.18 Periodogram of the cardiovascular disease data

is at a frequency of 0.1667, which gives a cycle of $1/0.1667 = 6$ months. This corresponds to the peaks in deaths in summer and winter.

The formula for the periodogram (1.7) started at a Fourier frequency of $\omega_j = 0$. This zero frequency corresponds to an infinite cycle and gives $\hat{C}_0 = \overline{y}$ and $\hat{S}_0 = 0$. So the periodogram at $\omega_j = 0$ is proportional to the mean. As we discussed in Sec 1.2.1, the mean is a marginal statistic and is not important for judging periodicity. Using the `demean=TRUE` option in R means that the series mean is zero and hence $I(0) = 0$. If we did not subtract the mean then the scale of the periodogram plot could be difficult to judge if $I(0)$ was very large.

Some properties of the periodogram worth noting are:

- $I(\omega_j)$ indicates the contribution of ω_j to the total variance in the series, and the area under the periodogram is equal to the total series variance.
- The periodogram contains the same information as the autocovariance function (Sect. 1.2.1). The periodogram is a *frequency domain* statistic; the x-axis in Fig. 1.18 is units of frequency. The autocovariance function is a *time domain* statistic; the x-axis in Fig. 1.8 is in units of time.
- The periodogram is not *consistent*, which means its variance does not decrease with increasing sample size. For many statistics, such as the sample mean, we get a more accurate estimate as the sample size increases. To overcome this problem it is common to smooth the periodogram. The smoothed estimate is called the *spectrum*.
- A periodogram of a series without any signal will be roughly flat with no large peaks. This is the derivation of the term "white noise" as white light is made up of all frequencies of light (or colours) equally, and has a flat colour spectrum.

Figure 1.19 shows the periodogram of the artificial data from Fig. 1.16. Apart from the four clear peaks the rest of the results are approximately zero. The largest periodogram value is at a frequency of 0.125. The series was generated using 400 observations, so to transform this frequency to a cycle we use $400 \times 0.125 = 50$, which corresponds to the frequency of the largest amplitude. The second largest periodogram value is at a frequency of $400 \times 0.0375 = 15$, and the third largest at $400 \times 0.025 = 10$. The fourth largest value is at a low frequency of 0.005 or two cycles (dashed line in Fig. 1.14), but there is no large value for the longest cycle (solid line in Fig. 1.14). It is not surprising that the periodogram has missed this frequency as it occurs only once in the data.

1.3.4 Cumulative Periodogram

A useful re-expression of the periodogram is the cumulative periodogram [85, p. 395], defined as

$$C(k) = \frac{\sum_{j=1}^{k} I(\omega_j)}{\sum_{j=1}^{n/2} I(\omega_j)}, \qquad k = 1, \ldots, n/2.$$

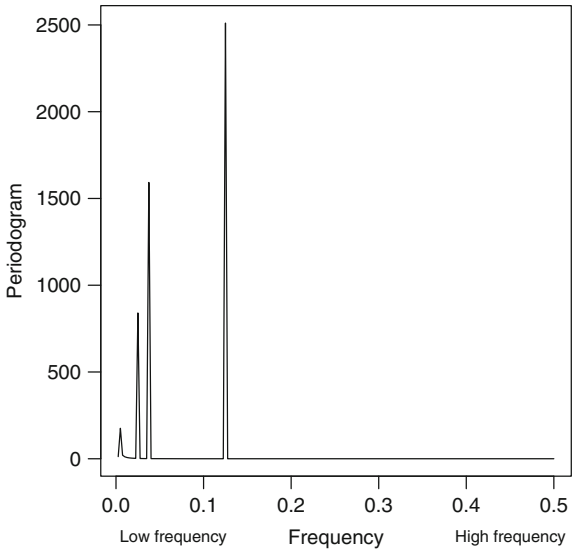


Fig. 1.19 Periodogram of the artificial data from Fig. 1.16

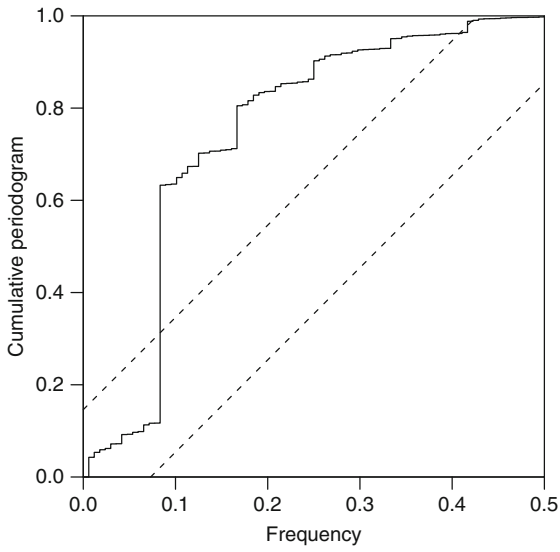


Fig. 1.20 Cumulative periodogram for the CVD data (*solid line*) and limits for the null hypothesis that the series is purely noise (*dashed lines*)

For a completely uncorrelated series $C(k)$ should increase steadily from 0 to 1. Limits for this increase, based on the null hypothesis that the series is purely noise, were developed by Bartlett [11]. We can plot the cumulative periodogram and these limits for the CVD data using the R command:

```
> cpgram(CVD$cvd)
```

The resulting plot in Fig. 1.20 shows the cumulative periodogram crossing the upper limit at a frequency of 0.0833. There is another smaller vertical step at a frequency of 0.1667.

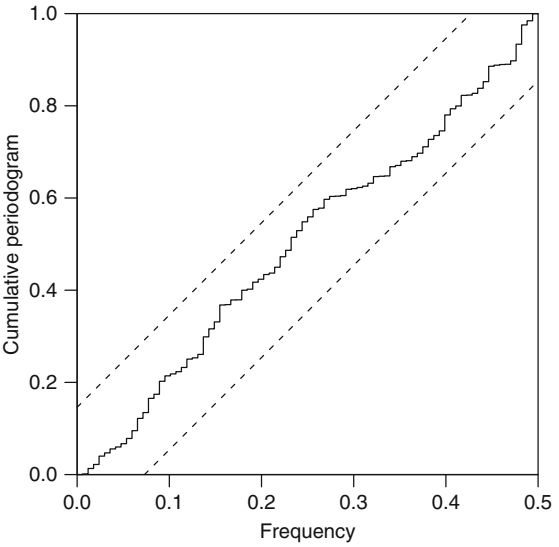


Fig. 1.21 Cumulative periodogram for independently generated data (*solid line*) and limits for the null hypothesis that the series is purely noise (*dashed lines*)

As a contrast we show the cumulative periodogram for a time series of the same length ($n = 168$), but generated using random or independent data with no signal. We used the R commands:

```
> set.seed(0)
> independent<-rnorm(168)
> cpgram(independent)
```

We used the `set.seed` command to fix a point for the random number generation, which means that re-running this code will exactly reproduce the time series, and hence the cumulative periodogram shown in Fig. 1.21. The cumulative periodogram has remained well within the upper and lower boundaries.

1.4 Regression Methods

Regression is a commonly used statistical method for explaining the association between a *dependent variable* and *independent variable(s)*. Dependent variables are also called *response variables*, and independent variables are also called *explanatory variables* or sometimes *predictor variables*. In this section we give a brief introduction to some important aspects of regression.

We define a simple linear regression equation, with one independent variable X_t, as

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{1.8}$$

where β_0 is the *intercept* and β_1 is the *slope*, which measures the average change in Y_t for a unit change in X_t and ε_t describes the random variation or *noise* in Y_t.

We can save space by representing regression models in *matrix* notation. A matrix is a rectangular collection of variables. For example a 2×3 matrix is

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 2 & 0 \end{bmatrix}.$$

The *transpose* of this matrix is denoted $\mathbf{M}^T$ and transposes the columns into columns, e.g.,

$$\mathbf{M}^T = \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ -1 & 0 \end{bmatrix}.$$

In matrix notation, the multiple linear regression equation, (1.8), is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is the $n \times p$ *design matrix*, and $\boldsymbol{\beta}^T = \beta_0, \ldots, \beta_p$ are the parameters, of which there are $p + 1$.

1.4.1 Scatter Plot

Before applying a regression model it is always useful to first plot the data in a *scatter plot*. This is a plot of the dependent variable on the y-axis and an independent variable on the x-axis. For multiple independent variables we would create multiple plots. If there is only a weak association between the dependent and independent variables, or if this association is complex, then these plots may be difficult to interpret. However, these plots are generally useful for:

- Spotting unusually large or small values in the dependent or independent variables.
- Showing the association between the dependent and independent variables, and whether this association is linear (or perhaps curvilinear), and also whether it is "noisy".

For the cardiovascular disease data (Sect. 1.1.1) we are interested in the explanatory variable mean monthly temperature. A scatter plot of the number of deaths against temperature is shown in Fig. 1.22, and shows a clear association between temperature and death. The most dominant pattern is a strong negative relationship with fewer deaths at warmer temperatures. There is a steeper rise in deaths below 60°F, which suggests a *non-linear* association between temperature and death. There is also a greater variance in deaths at lower temperatures than at higher temperatures, which is known as *heteroscedasticity*.

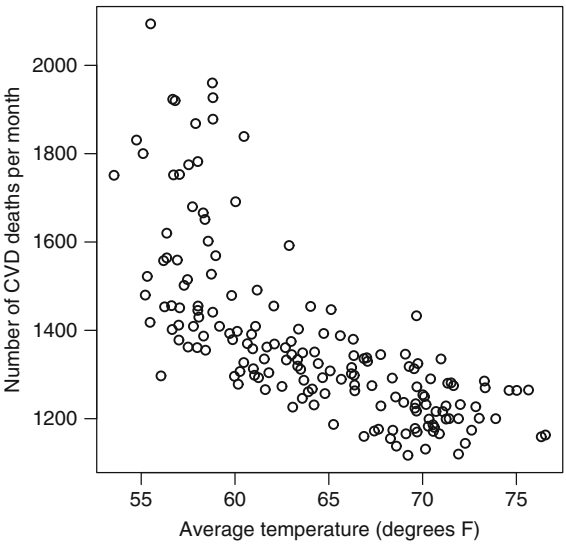


Fig. 1.22 Scatter plot of the monthly number of CVD deaths against the monthly average temperature

1.4.2 Linear Regression

A linear regression model assumes that the association between the independent variable(s) and response is linear, and that the residual error is constant (homoscedastic). The scatter plot for the cardiovascular data (Fig. 1.22) indicates that both these assumptions will likely be violated, but we will fit a linear model for illustrative purposes. We can fit a linear regression model using the R commands

```
> model<-lm(cvd~tmpd,data=CVD)
> summary(model)
```

which gives the following output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2933.246    118.681   24.71   <2e-16 ***
tmpd         -24.296      1.841  -13.20   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 136.3 on 166 degrees of freedom
Multiple R-squared: 0.512,      Adjusted R-squared: 0.509
F-statistic: 174.2 on 1 and 166 DF,  p-value: < 2.2e-16
```

The estimated coefficient for temperature is the slope of the regression line, and tells us that for every 1°F increase in temperature the average number of deaths decreases by 24.3. This change in deaths is strongly statistically significant with a p-value less than `2e-16`, which is in *scientific notation* and so means 2×10^{-16} which is 0.0000000000000002 (2 after 15 zeros). We also refer to the regression coefficients as regression *parameter estimates*.

The F-statistic gives a test of the fit of the regression model, which in this case is equivalent to testing the effect of temperature. If we had multiple independent variables, then the F-statistic would test their overall value.

1.4.2.1 R-Squared

A useful statistic in the above output from R is *R-squared* (also labelled "R^2"). It is the proportion of variability in the dependent variable explained by the independent variable(s). For this example, 50.9% of the variability in the monthly counts of cardiovascular deaths is explained by mean monthly temperature. We use the adjusted version of the R-squared, as this gives a slightly more conservative value by adjusting for the number of independent variables in the model.

The formula for the multiple (unadjusted) R-squared is

$$R^2 = 1 - \frac{SS_{error}}{SS_{total}},$$

where SS stands for *sum of squares* which is a key measure of variability. The R-squared divides the variability due to error divided by the total variability in the data, to give the proportion of unexplained variability. One minus this value is the proportion of explained variability. We can multiply the proportion by 100 to give a percentage. Adjusted R^2 takes into account the number of parameters ($\beta_0, \beta_1, \ldots$) estimated for the model.

For the above example the R command `anova(model)` gives the error sum of squares of 3,082,569 and total sum of squares of 6,316,605, which gives $R^2 = 0.512$.

1.4.2.2 Centring and Scaling

The intercept in the previous output is the estimated mean number of daily CVD deaths at 0°F, and is 2,375 in this example. We can give the intercept a more meaningful value by working with a centred version of temperature. For example, to base the intercept on 50°F we would use the commands

```
> CVD$tmpd.c<-CVD$tmpd-50
> model<-lm(cvd~tmpd.c,data=CVD)
```

which gives the following output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1718.442     28.196   60.95   <2e-16 ***
tmpd.c       -24.296      1.841  -13.20   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 136.3 on 166 degrees of freedom
Multiple R-squared: 0.512,      Adjusted R-squared: 0.509
F-statistic: 174.2 on 1 and 166 DF,  p-value: < 2.2e-16
```

Note the slope for temperature has not changed, but the intercept is now a mean of 1,718 deaths when the temperature is 50°F. This value can be roughly seen on the scatter plot (Fig. 1.22). We do not interpret the p-value for the intercept, as it tests the hypothesis that the intercept is significantly different from zero (at 50°F), but we know that the number of deaths will be positive. Note also that the R-squared has not changed, as we are fitting the same model, but are presenting the results on a more interpretable scale.

As well as centring it can also be useful to *scale* the effect of an independent variable. For example, a 1°F decrease in temperature may be too small to have any clinical meaning, so we could look at a 10°F decrease (centred on an intercept of 50°F) using the following code

```
> CVD$tmpd.s<-(CVD$tmpd-50)/10
> model<-lm(cvd~tmpd.s,data=CVD)
```

which gives the following output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1718.44      28.20   60.95   <2e-16 ***
tmpd.s       -242.96      18.41  -13.20   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 136.3 on 166 degrees of freedom
Multiple R-squared: 0.512,      Adjusted R-squared: 0.509
F-statistic: 174.2 on 1 and 166 DF,  p-value: < 2.2e-16
```

The average increase in deaths for a 10°F decrease in temperature is 243. The intercept and R-squared are the same as in the previous model. We could have obtained the same scaled slope by multiplying the slope for temperature from the previous output (−24.296) by 10. However, in general we prefer to scale explanatory variables in advance because: (1) it encourages a priori thought about the scale of the explanatory variables, (2) in more complex models we will examine the covariance between parameter estimates, and these estimates cannot simply be re-scaled using multiplication after fitting the model, (3) scaling independent variables is often necessary when fitting Bayesian models in the WinBUGS software (Sect. 1.6).

1.4.3 Residual Checking

The *residuals* are the difference between the observed and *fitted* values and we use them to check a number of important assumptions about the regression model. The fitted values are generated from the regression equation. So using the last regression equation we would calculate the fitted values using the equation

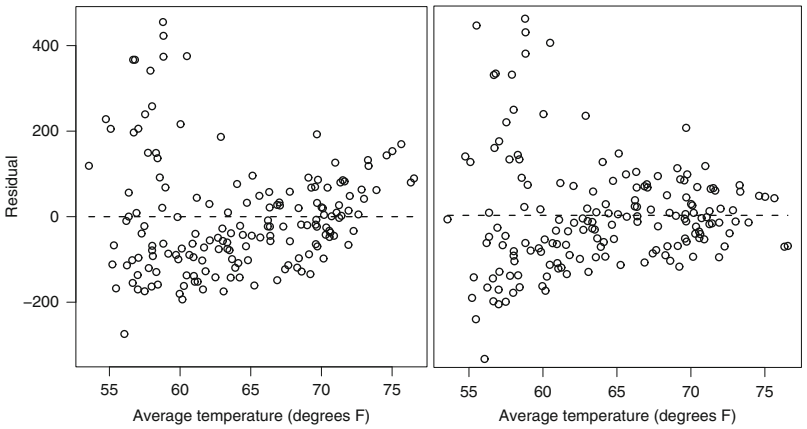


Fig. 1.23 Residuals of the linear model (*left panel*) and quadratic model (*right panel*) for the effect of temperature in predicting CVD deaths. *Dashed horizontal reference line* at zero corresponds to a perfect fit

$$\widehat{\text{Deaths}} = 1,718.44 - 242.96 \times (\text{Temperature} - 50)/10,$$

where the "hat" symbol ($\widehat{\ }$) indicates an estimate. The residuals are the difference between the observed and fitted values (Deaths$-\widehat{\text{Deaths}}$). We can calculate the fitted values and residuals using the R commands:

```
> model<-lm(cvd~tmpd,data=CVD)
> fit<-fitted(model)
> res<-resid(model)
```

If the model is a good fit to the data then there should be no signal remaining in the residuals. So a scatter plot of the residuals against any independent variable, or the dependent variable, should have no pattern. A plot of the residuals from the linear model against mean temperature in the left panel of Fig. 1.23 shows a *quadratic* pattern (U-shaped pattern). The model systematically underestimates the number of deaths at low and high temperatures. The horizontal reference line at zero indicates a perfect fit.

We can fit a quadratic model to the number of deaths by adding a squared value to temperature to the linear model. We use the R commands:

```
> CVD$tmpd2<-CVD$tmpd^2
> model2<-lm(cvd~tmpd+tmpd2,data=CVD)
> summary(model2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8904.889   1390.225   6.405 1.50e-09 ***
tmpd        -210.745     43.298  -4.867 2.63e-06 ***
tmpd2          1.444      0.335   4.310 2.80e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 129.6 on 165 degrees of freedom
Multiple R-squared: 0.5614,     Adjusted R-squared: 0.556
F-statistic: 105.6 on 2 and 165 DF,  p-value: < 2.2e-16
```

The residuals from this model are shown in the right panel of Fig. 1.23. The quadratic (squared) term has improved the fit somewhat as the residuals are more evenly scattered around the horizontal line, but there is still noticeable heteroscedasticity as the spread of the residuals becomes smaller at high temperatures. The assumption that the residual variance is homoscedastic will be broken here, as our predictions at low temperatures will have more error than our predictions at high temperatures.

The adjusted R^2 for the quadratic model is 55.6%, indicating a modest improvement in fit compared with the linear model (adjusted $R^2 = 50.9\%$).

It is useful to first standardise the residuals before creating the scatter plots so that relatively large residuals can more easily be spotted. The *studentized residuals* are defined as

$$\widehat{\varepsilon}_t^s = \frac{\widehat{\varepsilon}_t}{\widehat{\sigma}_\varepsilon \sqrt{1 - h_{tt}}}, \qquad t = 1, \ldots, n,$$

where h_{tt} is the diagonal of the so-called "*hat*" matrix and σ_ε is the standard deviation of the residuals. Using the Normality assumption, any residuals outside ± 2 may be considered somewhat unusual.

1.4.3.1 Independence of Residuals Over Time

The scatter plots in Fig. 1.23 were useful for examining any association between the residuals and the independent variable average temperature. For time series data it is also important to test independence over time. We can do this using two previous defined statistics, the autocorrelation (Sect. 1.2.1) and the periodogram (Sect. 1.3.3).

We can explore whether the independent variable temperature has completely explained the seasonal pattern in the data by using the autocorrelation function on the residuals.

```
> acf(res,type='correlation',lag.max=36)
```

We do not need the `demean=TRUE` option as by definition the residuals have a zero mean. The left-hand panel of Fig. 1.24 shows the autocorrelation of the residuals from the linear model. We can compare this plot to the autocorrelation using the number of deaths (Fig. 1.9). The autocorrelations for the residuals are much smaller than those using the death numbers, but there are still a number of significant values. The moderate positive correlations at lags 12, 24 and 36 indicate that some part of the annual seasonal pattern has not been captured by the linear model. The autocorrelation at lag 1 may indicate some trend, which we discuss later (Chap. 4).

The cumulative periodogram is another useful check of the independence of the residuals over time. The right-hand panel of Fig. 1.24 shows the cumulative periodogram for the residuals from the linear model. We can compare this plot to the cumulative periodogram using the number of deaths (Fig. 1.20). The cumulative

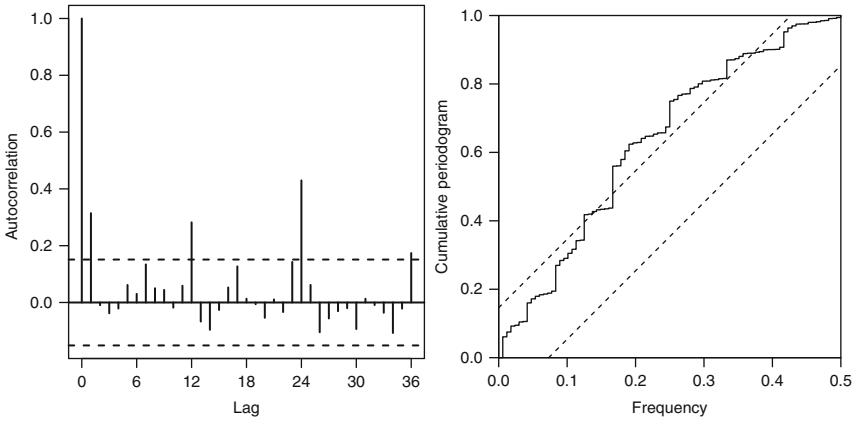


Fig. 1.24 Autocorrelation (*left panel*) and cumulative periodogram (*right panel*) of the residuals of the linear model for temperature predicting CVD deaths

periodogram for the residuals indicates that there remains some pattern over time. The cumulative periodogram first crosses the upper test limit at a frequency of 0.125 (cycle of 8 months) and again at 0.167 (cycle of 6 months).

The test limits of the ACF and cumulative periodogram have different *type I error* probabilities (see Sect. 1.2). The test limits from the cumulative periodogram have an overall type I error of 5%, whereas those from the autocorrelation function have an individual type I error of 5%. In general, we recommend the cumulative periodogram for formal significance testing, but a thorough investigation of the residuals will involve looking at both plots rather than relying on a test.

1.4.3.2 Distributional Assumptions

We can also use the residuals to check distributional assumptions. The left panel of Fig. 1.25 shows the histogram of the residuals from the linear model, and the right panel using the quadratic model. There is a positive *skew* to the distribution, which is slightly lessened for the quadratic model. Although inferences from simple regression models are based on the assumption that the residuals are Normally distributed, in practice this assumption is often not important because for large sample size the *central limit theorem* ensures that the fitted values will (approximately) have a Normal distribution. We still recommend plotting a histogram of the residuals though, as it is useful for spotting:

- A skewed distribution, which may indicate that a *data transformation* such as the log-transform would be useful
- Large outliers, which should be investigated further and may indicate a problem with the data collection (e.g., erroneous data) or that an independent variable is missing from the model (e.g., all the outliers have a similar set of independent variables)

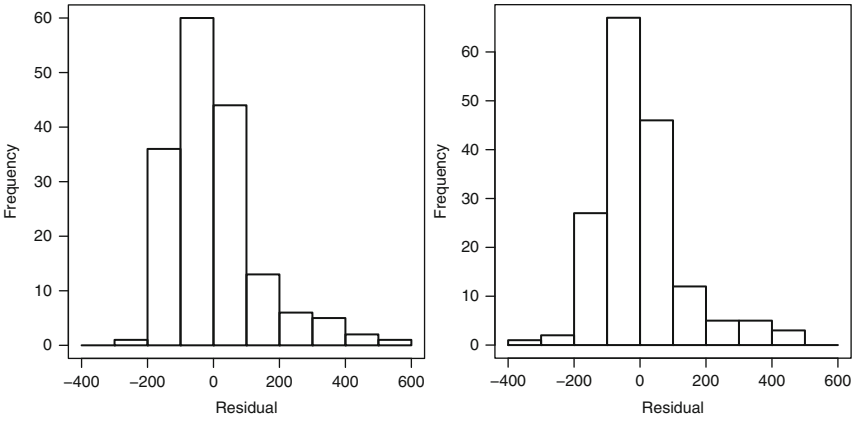


Fig. 1.25 Histogram of the residuals of the linear model (*left panel*) and quadratic model (*right panel*) for the effect of temperature in predicting CVD deaths

1.4.4 Influential Observations

Another useful check of the fit of a regression model is to look for *influential observations*. These are observations that have a large influence on the regression line (either its intercept or slope). We can check the size of the influence by leaving out each observation in turn and measuring the change in the regression line.

For the linear regression equation, (1.8), the statistic *delta-beta* (or *df-beta*) is defined by

$$\Delta_t \hat{\beta}_k = \hat{\beta}_k - \hat{\beta}_{k(t)},$$

where $\hat{\beta}_{k(t)}$ is the estimate of β_k obtained when the tth observation is omitted from the data. $\hat{\beta}_k$ is the estimate of β_k obtained using all the observations. The symbol Δ is commonly used to indicate difference. A plot of $\Delta_t \hat{\beta}_k$ against t can then be used to identify influential observations.

Once an influential observation (or observations) is detected, the first step is to determine whether it might be a measurement error, transcription error or some other mistake. It should only be removed from the data (or truncated to a more reasonable value) if there is a solid scientific basis (e.g., impossible value). Otherwise the influential observation(s) should be retained, and regression models including and excluding them reported.

For the exercise data, Fig. 1.26 shows a scatter plot of body mass index against age. There are two very large BMIs that could have a large influence on a regression model.

A linear regression model of body mass index against age gives a regression slope of -1.82 for a 10-year increase in age, with a t-value of -6.88 and p-value < 0.001. So older subjects have a lower BMI on average. Figure 1.27 shows the df-beta's for the linear regression slope plotted against the subjects' ID number (Table 1.3). The

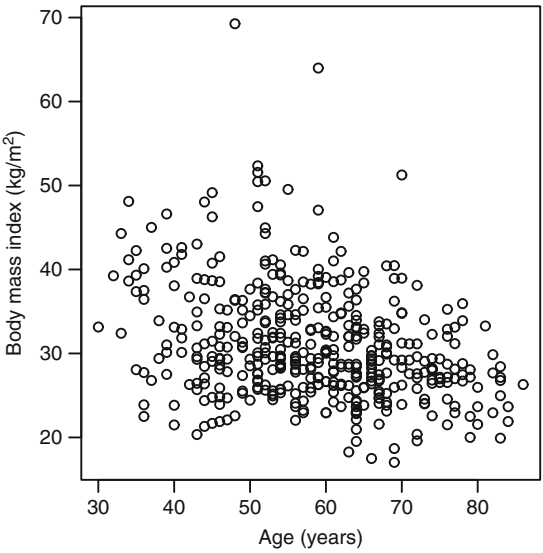


Fig. 1.26 Body mass index (kg/m^2) at baseline against age for all 434 subjects from the exercise data

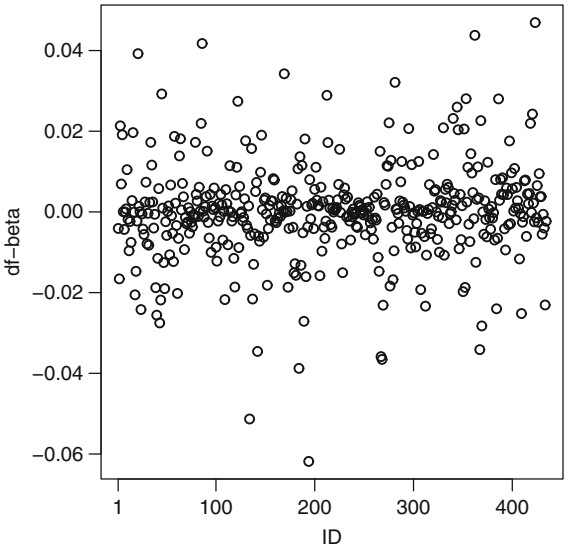


Fig. 1.27 Influence of leaving out observations (df-beta statistic) on the effect of the linear slope between age and BMI

df-beta's are calculated in R using the command `dfbeta(model)`. The largest absolute change in the regression slope came from leaving out a subject with a BMI over 60. Removing this subject reduces the regression slope to a shallower -1.76, so the association between age and BMI is weaker. The df-beta is $-1.82 - (-1.76) = -0.06$, which is quite small. Also, the remaining points are reasonably close to zero, so there are no overtly influential observations for the regression slope in this model.

1.4.5 Generalized Linear Model

For previous regression examples in this chapter we have assumed that the dependent variable had a Normal distribution. For dependent variables with a non-Normal distribution we can use much of the same theory by generalising the model [26]. The generalisation to non-Normal data is achieved using a *link function*. The equation below extends the simple linear regression equation (1.8) to a *generalized linear model*,

$$g(Y_t) = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_p X_{pt}, \qquad t = 1, \ldots, n, \tag{1.9}$$

where $g()$ is the link function. This regression model has $p + 1$ regression parameters ($\boldsymbol{\beta}$) including the intercept (β_0). The distribution of the dependent variable is modelled through the link function rather than as a Normally distributed random error term on the right-hand side of the equation, as in (1.8).

Table 1.6 shows the link functions that we will need, and gives an example of when the link function would be appropriate.

The logit link is appropriate for Binomial data and is defined as

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right),$$

where p is the probability of "success", for example the probability of a stillbirth. The logit link transforms p, which is *bounded* between 0 and 1, to an unbounded value. The parameter estimates from the model are on the logit scale. We can obtain the *odds ratio* of "success" for a change in an independent variable by exponentiating the parameter estimates in (1.9); the odds ratio corresponding to parameter β_j is $\exp(\beta_j)$. A parameter estimate $\hat{\beta}_j = 0$ corresponds to no change in the probability of success, which is an odds ratio of 1.

The log link is appropriate for count data and is simply the log transformation of the observed counts. Count data often have a natural positive skew, and the log transformation helps to reduce this skew and make the distribution appear more Normal. Count data may have a *Poisson* distribution. For data where the mean count is greater than 15, then the Normal approximation to the Poisson distribution may be adequate, and we can fit a model without a link function.

When using Poisson regression the regression parameters are estimated on a log scale. We can obtain the *rate ratio* for a change in an independent variable by exponentiating the parameter estimates in (1.9); rate ratio RR = $\exp(\beta_j)$. It is also possible to express the result as a *percentage change* using the formula

Table 1.6 Link functions for Normal, Binomial and Poisson dependent variables

Link	Distribution (type)	Seasonal example	Section
None (identity)	Normal (continuous)	Body mass index	1.1.4
Logit	Binomial (dichotomous)	Stillbirth	1.1.5
Log	Poisson (counts)	Schizophrenia births	1.1.2

$$\Delta\% = 100 \times (\text{RR} - 1).$$

A rate ratio of 1 indicates no change in mean counts.

1.4.5.1 Poisson Regression Example

We fit a Poisson regression model to the monthly birth counts for schizophrenia cases. We used a single continuous explanatory variable of the southern oscillation index (SOI). Because the SOI has quite a wide range we first scaled it by 10 (Sect. 1.4.2). The R commands are:

```
> schz$SOI.10<-schz$SOI/10
> model<-glm(SczBroad~SOI.10,data=schz,family=
  poisson(link='log'))
> summary(model)
```

This model has the parameter estimates

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.77231    0.01120  247.55   <2e-16 ***
SOI.10       0.01210    0.01249    0.97    0.332
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

If we exponentiate the parameter estimate for the SOI we get a rate ratio of $\exp(0.01210) = 1.012$ for every 10 unit increase in the SOI, but this increase is not statistically significant (p-value = 0.332). The results of Poisson regression are on a *multiplicative scale*, whereas those using Normal regression are on an *additive scale* (Sect. 1.4.2). The estimated percentage change in cases for a 10 unit increase in the SOI is $100 \times (1.012 - 1) = 1.2\%$.

An important consideration when using Poisson regression is whether the data closely follow a Poisson distribution. When data follow a Poisson distribution their mean is equal to their variance. It is possible for data to have a roughly Poisson distribution, but with a variance that is greater than the mean. In this case the data is *over-dispersed* [26]. Conversely, and more rarely, if the variance is less than the mean then the data is *under-dispersed*.

If we ignore over-dispersion then the standard errors of our parameter estimates will be too small (because we are missing the extra variance due to over-dispersion). The output below shows the R code and parameter estimates for the schizophrenia cases after accounting for over-dispersion. We have used the "gam" function from the "mgcv" library [88]. The `scale=-1` option accounts for any over-dispersion.

```
> library(mgcv)
> model<-gam(SczBroad~SOI.10,data=schz,family=
  poisson(link='log'),scale=-1)
> summary(model)
```

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.77231    0.01887 146.911   <2e-16 ***
SOI.10       0.01210    0.02104   0.575    0.565
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The mean estimates have not changed, but the standard errors (for both the intercept and effect of SOI) have increased by a factor of 1.7.

1.4.5.2 Logistic Regression Example

We can fit a logistic regression model to the stillbirth data using the R commands:

```
> model<-glm(stillborn~as.factor(seifa),
  data=stillbirth,family=binomial(link = "logit"))
```

As an example we have used the single explanatory variable "seifa" (SEIFA score) which is an area-level measure of socio-economic disadvantage (higher scores indicate less disadvantage). In this data it was an *ordinal* variable on an integer scale of 1–5. We used the `as.factor` command to fit this variable as a nominal *categorical* explanatory variable, rather than as a *continuous* variable. This command gives the following parameter estimates

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.17325    0.10942 -47.280   <2e-16 ***
as.factor(seifa)2  0.02166    0.15134   0.143    0.886
as.factor(seifa)3  0.08272    0.16716   0.495    0.621
as.factor(seifa)4  0.01218    0.17210   0.071    0.944
as.factor(seifa)5  0.10818    0.17304   0.625    0.532
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

As we fitted SEIFA as a categorical variable we have five estimates which are all compared with the reference category of a SEIFA score of 1 (the most disadvantaged group). The odds of a stillbirth for a woman with a SEIFA score of 2 are $\exp(0.02166) = 1.022$ times the odds for a woman with a SEIFA score of 1. This is a small increase in odds which is not statistically significant (p-value = 0.886). There are no statistically significant differences in odds between SEIFA scores of 2–5 compared with a SEIFA score of 1.

The alternative to fitting SEIFA as a categorical independent variable is to fit it as a ordinal variable, which assumes that increasing SEIFA score is associated with a linear change in risk. The advantage of this method is that we would only need one parameter (for the slope) as opposed to four parameters (one for each category other than the reference category). So it is a more *parsimonious* model. In this case, however, the parameter estimates based on a categorical variable do not show any clear linear trend, so using a single parameter for the slope is probably not justified.

1.4.6 Offsets

Offsets are used to adjust regression models to account for a *denominator*. For example, the number of schizophrenia cases with birth dates in a particular month (Fig. 1.2) is likely to depend on the total number of births in that month. Similarly, if we counted the area of a patient's skin covered by eczema, this would depend on the patient's total skin area.

For descriptive purposes we calculate the rate of schizophrenia per 1,000 births,

$$\text{rate} = 1{,}000 \times (\text{schizophrenia births} \,/\, \text{total births}).$$

This rate may be clinically meaningful, and it makes it easier to compare rates of schizophrenia in Australia and other countries. The R command to calculate this rate is

```
> schz$rate<-1000*(schz$SczBroad/schz$NBirths)
```

Figure 1.28 shows the rate of schizophrenia births over time. This figure should be compared to the "raw" numbers in Fig. 1.2.

The two figures tell very different stories, and after adjusting for the number of births the risk of schizophrenia now looks fairly constant from 1930 to 1960, followed by a decline from 1960 to 1971. Ignoring the change in population birth rates would have caused us to wrongly conclude that the risk of schizophrenia had increased from 1930 to 1960.

To include an offset in a regression model we would use Poisson regression with a log link. This is because we are interested in the rate of events, so our regression equation in words is

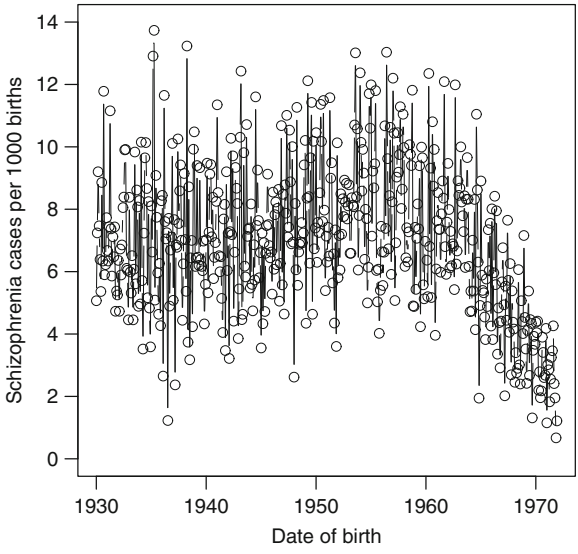


Fig. 1.28 Rates of schizophrenia cases per 1,000 births in Australia from 1930 to 1971

$$\log(\text{counts}/\text{denominator}) = \text{explanatory variables}$$

but as $\log(a/b) = \log(a) - \log(b)$ we can re-arrange this equation to

$$\log(\text{counts}) = \text{explanatory variables} + \log(\text{denominator}).$$

So we need to log-transform the denominator when we use it as an offset.

We can fit an offset for the population size (and account for over-dispersion) using the R commands:

```
> model<-gam(SczBroad~SOI,data=schz,family=
  poisson(link='log'),offset=log(NBirths),scale=-1)
```

This command gives the parameter estimates

```
Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.98359    0.01637 -304.52   <2e-16 ***
SOI         -0.01508    0.01818   -0.83    0.407
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

A 10 unit increase in the SOI changes the rate of schizophrenia by a factor of RR = $\exp(-0.01508) = 0.985$, but this decrease is not statistically significant (p-value = 0.407). The intercept is on a scale of "per birth". We do not usually interpret the parameter estimate for the intercept (or its p-value).

1.4.7 Akaike Information Criterion

A useful statistic for model selection is the *Akaike information criterion* (AIC) [1]. In words the equation is

$$\text{AIC} = -2 \times \text{log-likelihood} + 2 \times \text{number of parameters}.$$

The lower the AIC the better model. The *log-likelihood* is a measure of model fit, with smaller values of $-2\times$ log-likelihood indicating a better fit to the data. The number of parameters is a measure of model complexity. The equation aims to penalize the addition of new independent variables by adding twice the number of parameters to the model fit. This is important because the addition of *any* variable will improve the log-likelihood, even if it is not associated with the dependent variable. The AIC therefore aims to find variables that meaningfully improve the model fit.

We can use the AIC to choose between two alternative models by taking the difference in AIC. The size of this difference then describes the improvement (or worsening) in model fit. Cut-offs for the difference in AIC have been suggested [13,

Table 1.7 Difference in AIC between two alternative models and some suggested interpretations [13]

ΔAIC	Interpretation
0–2	No difference in models, the simpler model is preferred
4–7	Model with smaller AIC is probably better
>10	Model with smaller AIC is definitely better

p. 70], which seem to work well in our experience. Assuming that we have subtracted the larger AIC from the smaller AIC, these cut-offs are shown in Table 1.7. These cut-offs are only a guideline and we do not recommend using them as absolute rules. Note also that there are gaps in the cut-offs 2–4 and 7–10, in these regions the interpretation is less clear cut.

As well as using the AIC, the best model should also be selected based on the results of residual checks and discussions about the inference. In some cases there may not be a single best model, in which case it is valid to present the results from a number of plausible alternatives.

As an example of using the AIC for model selection, consider the model fitted to the schizophrenia data in Sect. 1.4.6. The R commands below fit models with and without the SOI and then calculates the difference in AIC.

```
> model.0<-gam(SczBroad~1,data=schz,family=poisson(link=
  'log'),offset=log(NBirths),scale=-1)
> model.1<-gam(SczBroad~SOI.10,data=schz,family=poisson(link=
  'log'),offset=log(NBirths),scale=-1)
> aic.diff<-model.0$aic-model.1$aic
> aic.diff
[1] -0.530384
```

The AIC for a model without the SOI as an independent variable is 3,415.7. The AIC with the SOI is 3,416.2, a difference of 0.5. This small difference indicates that we should favour the simpler model without the SOI. Selecting the simpler of two models with a similar fit is an example of the principle of *parsimony*.

1.4.8 Non-linear Regression Using Splines

The relationship between CVD death and temperature shown in Fig. 1.22 appears to be non-linear, as the scatter plot has a curved shape. To account for this we previously added a non-linear quadratic effect for temperature (Sect. 1.4.3). However, a quadratic term is quite restricted in its capability to describe possible non-linear shapes. For example, it is symmetric about its point of inflexion. A more flexible class of non-linear shapes are called *splines*.

There are many different types splines (e.g., natural splines, thin-plate splines), see [88] for details. To demonstrate the major features of a spline we use a *regression spline* defined as follows

$$s(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \sum_{k=1}^{K} \beta_k^\star (t - \kappa_k)_+^2, \qquad t = 1, \ldots, n, \tag{1.10}$$

where $\boldsymbol{\kappa}$ are a set of *knots*, $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\star$ are sets of unknown parameters. There are $K + 3$ regression parameters and K knots. The function $(.)_+$ is used to "switch on" parts of the model, and is defined as

$$(x)_+ = \begin{cases} 0, & x \leq 0, \\ x, & x > 0. \end{cases}$$

The first part of the spline (involving the terms $\boldsymbol{\beta}$) is a parametric function (quadratic in this case). The second part of the spline (involving the terms $\boldsymbol{\beta}^\star$) is non-parametric. This spline is said to have a *quadratic basis* because it uses the squared value $(t - \kappa_k)_+^2$.

The splines that we use later in this section for modelling seasonality are often more complex than (1.10). This is because much research has been done on the best form of the spline basis, knot placement and parameters [73, 88]. However, the key principles of splines can be illustrated using (1.10).

We illustrate the use of a spline in Fig. 1.29. The spline was created using the following equation with $K = 2$ knots,

$$s(t) = t^2 + 0.2(t - \kappa_1)_+^2 - 1.8(t - \kappa_2)_+^2.$$

The spline is shown as a solid line, the parametric (quadratic) part of the spline is shown as a dashed line. To the left of the first knot (κ_1) the spline is purely quadratic because $t \leq \kappa_1$ and so none of the non-parametric estimates are "switched on". After the first knot the spline is greater than the quadratic curve as $\beta_1^\star = 0.2$ is positive. After the second knot (κ_2) the spline turns back towards, and then crosses, the quadratic curve as $\beta_1^\star + \beta_2^\star = -1.6$ is negative.

The spline has a more complex non-linear pattern than the quadratic function. The shape of the spline between any two knots is quadratic (U-shaped) as (1.10) has a quadratic basis. These shapes are joined because the spline uses a cumulative sum of the non-parametric parts, and because it uses time since the previous knot $(t - \kappa_k)$ in the model.

The spline shown in Fig. 1.29 can only change shape after passing a knot, so the positioning of the knots is a key factor for determining the shape of the spline. Usually the knots, $\boldsymbol{\kappa}$, are equally spaced over time (or over the range of an independent variable). The two knots in Fig. 1.29 are equally spaced along the x-axis.

If $\beta_k^\star = 0$ for all values of k, then the non-parametric part of the spline will be zero and the spline will be completely parametric. Conversely if the values of $\boldsymbol{\beta}^\star$ are large, then the spline will depart greatly from the parametric curve and so be very flexible. The flexibility of the line therefore depends upon the size of the non-parametric part of the spline. A *penalized spline* applies a weight (or penalty) to the non-parametric part of the spline [73]. We do not go into the details of these

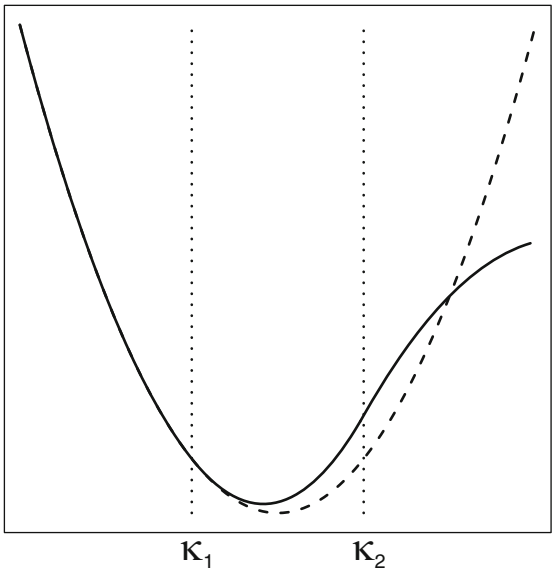


Fig. 1.29 Simple example of a regression spline (*solid line*) with a quadratic basis and two knots

methods here, but instead focus on the practical application of using splines for modelling seasonality.

1.4.8.1 Example of a Non-linear Spline

We show an example of splines by examining the non-linear association between temperature and CVD death in Fig. 1.30. The spline was fitted using (1.10) with four equally spaced knots. The result shows a slightly non-linear curve, as the association becomes weaker at higher temperatures.

1.5 Box Plots

A box plot is a graphical method for illustrating the association between a continuous variable and a categorical variable. The box extends from the first to the third quartile of the continuous variable. The box therefore contains the central 50% of the data and its length is the inter-quartile range (IQR; Sect. 1.2.1). The second quartile (or median) is marked as a line within the box. The "whiskers" extend from the top and bottom of the box to the maximum and minimum, respectively. However, these maximum and minimum exclude outliers, which are defined as those observations which are more than 1.5× IQR from the top or bottom of the box. Values outside this range are flagged as outliers.

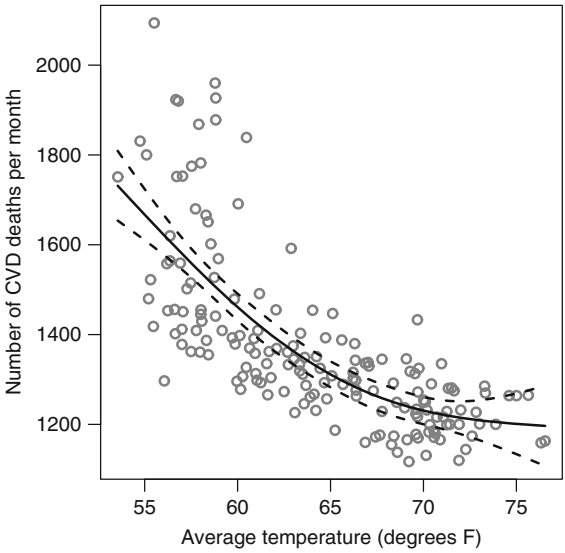


Fig. 1.30 Scatter plot of the monthly number of CVD deaths against the monthly average temperature (*grey dots*) and a fitted non-linear spline (*solid line*) and 95% confidence intervals (*dashed lines*)

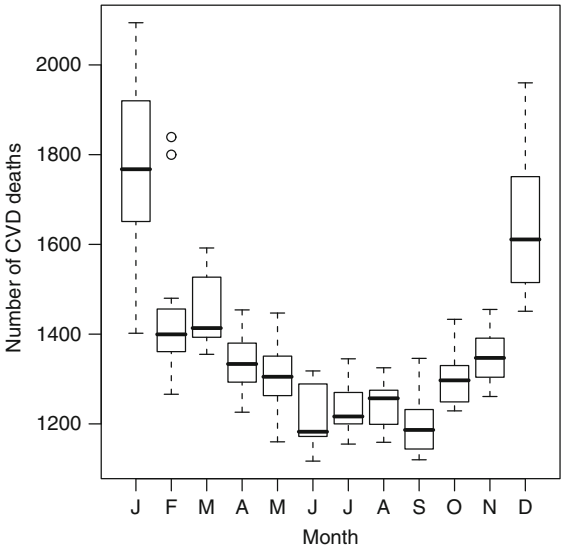


Fig. 1.31 Box plot of the number of cardiovascular deaths against month

We can obtain a box plot of the number of CVD deaths (continuous) against month (categorical) using the R commands:

```
> boxplot(cvd~month,data=CVD)
```

This command produces the plot show in Fig. 1.31. The seasonal pattern in deaths is quite clear, as is the larger variability in deaths in January and December (when the mean is also highest). There are two large outliers in February (more than 1,800 deaths) flagged using circles. The distribution in most months is symmetric,

although February looks somewhat negatively skewed due to its longer lower whisker, and December looks positively skewed due to its longer upper whisker. This plot has not adjusted for the different number of days between months, so the results for February may be unfairly low. We show how to adjust for this in the next chapter (Sect. 2.2.1).

1.6 Bayesian Statistics

Bayesian statistics is an alternative statistical paradigm to the more commonly used *frequentist* statistics. It has some advantages over standard frequentist theory that we will use in later chapters.

In this section we give a very brief overview. Details of Bayesian analysis is available in the books by Berger [12] or Gelman et al. [40]. A basic introduction is also available in the book by Dobson and Barnett [26].

At the heart of Bayesian statistics is the equation

$$P(\boldsymbol{\theta}|\mathbf{y}) \propto P(\mathbf{y}|\boldsymbol{\theta})P(\boldsymbol{\theta}), \tag{1.11}$$

where $\propto$ means "proportional to", $\boldsymbol{\theta}$ is an unknown parameter(s), $\mathbf{y}$ is the observed data, P indicates probability and the vertical bar $|$ is read as "given". The unknown parameter might be a linear regression slope, or an odds ratio (for example). The equation combines the likelihood, $P(\mathbf{y}|\boldsymbol{\theta})$, with a prior, $P(\boldsymbol{\theta})$, to give a posterior, $P(\boldsymbol{\theta}|\mathbf{y})$. The posterior is our updated view of $\boldsymbol{\theta}$, formed by combining our prior view with the latest data. If we have no prior opinion then all values of $\boldsymbol{\theta}$ are considered equally likely so (1.11) becomes

$$P(\boldsymbol{\theta}|\mathbf{y}) \propto P(\mathbf{y}|\boldsymbol{\theta}).$$

So by using an *uninformative prior* our inference about $\boldsymbol{\theta}$ will be based solely on the likelihood. The likelihood is the key function in frequentist methodology, so Bayesian and frequentist approaches will often give the same inference.

Some distinct advantages of a Bayesian approach over a frequentist approach are:

- A Bayesian approach views data as fixed and parameters as random (a frequentist approach takes the opposite view). Most scientists find it more intuitive to think of the collected data as being fixed. A frequentist approach requires us to think about alternative data that might have been collected if the experiment could be repeated.
- A Bayesian approach gives p-values and confidence intervals (called *credible intervals* or posterior intervals) that make more intuitive sense. A 95% credible interval contains the true parameter value with 95% probability. A frequentist 95% confidence interval would contain the true parameter on 95% of occasions *if* the data could be collected again.

- Priors can be used to specify assumptions about the model, whilst remaining non-informative. We give an example of this type of prior in Sect. 4.4.1.
- For some statistical models it is impossible to write down the exact likelihood, hence it is impossible to find estimates for $\boldsymbol{\theta}$ using a frequentist approach. Bayesian statistics is able to empirically estimate the posterior by using a numerical technique called *Markov chain Monte Carlo*. Hence Bayesian statistics can be used to fit models that cannot be fitted using standard theory.

1.6.1 Markov Chain Monte Carlo Estimation

Using standard statistical theory, we can estimate the model parameters by maximising the likelihood, $L(\mathbf{y}|\boldsymbol{\theta})$. The maximum value of $L()$ is found using an iterative algorithm that progressively updates the values of $\boldsymbol{\theta}$, until no better values can be found (known as *convergence*). This method relies on knowing the likelihood function (or a function proportional to the likelihood), which is not possible for all models (especially more complex models).

Markov chain Monte Carlo (MCMC) estimation deals with complex likelihoods by breaking the problem into smaller steps. Like maximum likelihood estimation it progressively updates the values of $\boldsymbol{\theta}$ forming a chain of estimates. At each step a new value for θ is chosen probabilistically, importantly the probabilities are weighted towards more likely estimates of θ. A Markov chain does not stop when it reaches the most likely value of θ, but instead continues to sample values around the maximum in order to estimate the complete distribution of θ. Based on the distribution we can then create statistics that describe aspects of θ, such as the mean or fifth percentile. To start the chain we need to give *initial values* for $\boldsymbol{\theta}$. It is often worth thinking carefully about these initial values, as if they are very poor the chain can find it hard to get started.

We show an example MCMC chain in Fig. 1.32, together with the *deviance* which is $-2\times$ log-likelihood, so a lower deviance indicates a better fit. The model is a linear regression of the number of CVD deaths against temperature (Fig. 1.22), which has an intercept (θ_0) and slope parameter (θ_1). We used initial values of zero for both the intercept and slope. From this initial value the chain for the slope $\hat{\theta}_1$ in Fig. 1.32 has progressed towards a range of values from around -0.3 to -0.5. The decline in $\hat{\theta}_1$ is associated with a similar decline in the deviance, showing that the chain has successfully progressed towards a better estimate.

The estimates of $\hat{\theta}$ from iterations 1 to around 50 are comparatively poor, and we would not want to use them to make inferences about θ. So we would discard these iterations as a *burn-in* as the chain is recovering from the poor initial estimate. Judging the correct length of the burn-in can be difficult for some chains, and some may need to be run for a very long time ($>$10,000 iterations) before they can be said to have converged. We can often improve the speed at which a chain converge by centring the data (Sect. 1.4.2).

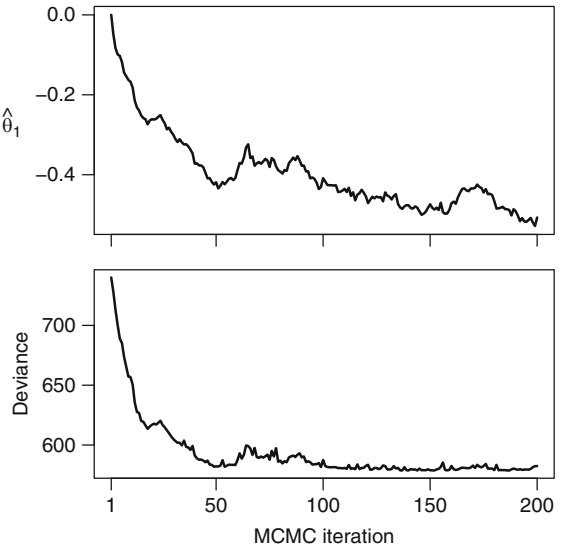


Fig. 1.32 Estimated slope $\hat{\theta}_1$ and deviance for 200 MCMC iterations after an initial value of $\hat{\theta}_1 = 0$

Table 1.8 Difference in DIC between two alternative models and some suggested interpretations [79]

ΔDIC	Interpretation
0–5	No difference in models, the simpler model is preferred
5–10	Model with smaller DIC is probably better
>10	Model with smaller DIC is definitely better

For a more detailed description of MCMC methods, see [40, Chaps. 10–11] or [42].

1.6.2 Deviance Information Criterion

A useful model selection statistic when using Bayesian inference is the *Deviance Information Criterion* (DIC) [80]. The DIC is the Bayesian equivalent of the Akaike Information Criterion (Sec 1.4.7), and like the AIC, it is a trade-off between model fit and complexity. The formula for the DIC is

$$\text{DIC} = D(\mathbf{y}|\overline{\boldsymbol{\theta}}) + 2p_D,$$

where $D(\mathbf{y}|\overline{\boldsymbol{\theta}})$ is the deviance evaluated at the parameter means ($\overline{\boldsymbol{\theta}}$), and p_D is the effective number of parameters. The deviance is an estimate of model fit and the number of parameters a penalty for complexity. The DIC uses the *effective* number of parameters, which is not necessarily an integer and can be thought of as the amount of "information" needed to fit the model.

The cut-offs in Table 1.8 for the difference in DIC have been suggested [79]. We stress that these are guidelines and selecting the best model should be based on other factors, such as residual checks. It is also important to check the inference and interpretation of the best model. If two models have a similar DIC, but have very different inferences (e.g., one shows a strong positive association between exposure and disease, whilst the other shows a strong negative association), then it is probably best to present the results of both models, or find another reason for selecting one model over another (e.g., a more random residual pattern).

Chapter 2
Introduction to Seasonality

In this chapter we define what we mean by a season and show some methods for investigating and modelling a seasonal pattern.

Three excellent, but more technical, books on seasonality are by Mardia [59], Fisher [35], and Ghysels and Osborn [41]. We also recommend Chap. 8 of West and Harrison [87]. An excellent book on the history and construction of the Christian calendar is Duncan [28].

2.1 What is a Season?

The word "season" has many different meanings. It can be used for:

- Events that happen annually at fixed dates (such as the festive season).
- Events that tend to happen at particular times but with less certainty (such as the flu season).
- Dividing the year into distinct climatic periods. Typically the seasonal divisions are spring, summer, autumn and winter, although in tropical climates the only important differences are between the wet and dry seasons.

In this book we are interested in all three of these types of seasons.

A seasonal pattern is not always regular in timing or in its shape. For example, the flu season is a period when the number of cases of flu reaches an epidemic, but the timing and length of this season varies from year-to-year. This variation can be seen in Fig. 1.3. The peak number of cases in the 2005/2006 season happened in week 13. In the 2006/2007 season there was a larger peak at a slightly early time (week 9). The change in the timing reflects the dynamic nature of a flu epidemic.

Seasonal patterns can have many different shapes. Figure 1.1 shows a repeating seasonal pattern in the number of CVD deaths that appears reasonably sinusoidal, with a steady increase followed by a decrease in deaths. The seasonal pattern in flu is less sinusoidal, as there are many weeks with low numbers, followed by a sharp increase and equally sharp decline. More irregular seasonal shapes are also possible. A good example is the festive season, which is a regular seasonal pattern in exposure that has been linked to spikes in cardiovascular disease [68].

A.G. Barnett and A.J. Dobson, *Analysing Seasonal Health Data*, Statistics for Biology and Health, DOI 10.1007/978-3-642-10748-1_2,
© Springer-Verlag Berlin Heidelberg 2010

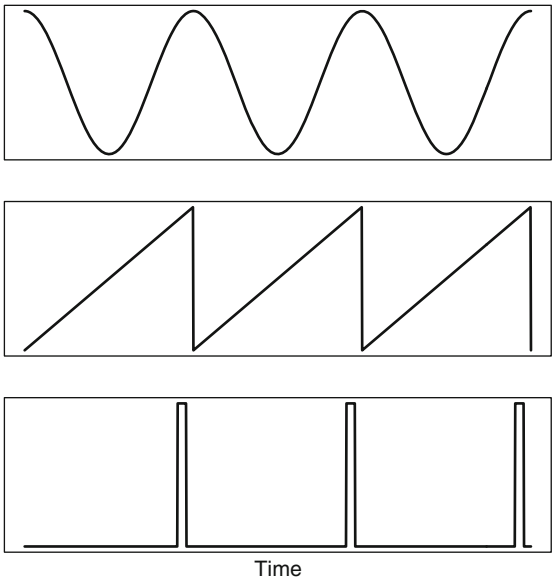


Fig. 2.1 Three seasonal patterns: sinusoidal (*top*), sawtooth (*middle*), spiked (*bottom*)

Figure 2.1 shows three very different patterns, which can all be classed as seasonal. In each case the seasonal pattern is repeated three times. The top panel shows a very smooth *sinusoidal* seasonal pattern. The middle panel shows a *sawtooth* seasonal pattern with a steady linear rise over time, followed by a short, sharp fall. The lower panel shows a *spiked* seasonal pattern with a short high period and long low period.

To capture the great variety of seasonal patterns we use the broad definition below:

Box 2.1: Definition of a season

A season is a pattern in a health outcome or exposure that increases and then decreases with some regularity.

This definition allows the size, timing and shape of the seasonal pattern to change over time. Also, this definition is not restricted to annual seasonal patterns, and can be applied to seasonal patterns that happen during a week, a month, or any other period of time. This means that the cycle of the seasonal pattern can be any length, from milliseconds to millennia.

2.1.1 Seasonality and Health

Many health conditions and exposures (or risk factors) have some seasonal element. Table 2.1 lists a selection of seasonal health conditions and exposures. An important distinction is whether a health condition is seasonal in its severity or

Table 2.1 Some examples of seasonal health conditions and exposures. Those health conditions in italics are seasonal by birth, those in normal font are seasonal by incidence, those underlined are seasonal exposures

Air pollution [77]	Asthma [50]	*Birth defects* [39]
Birth weight [54]	Blood pressure [9]	Cardiovascular disease [8]
Cholera [34]	Depression [51]	Diabetes [44]
Kawasaki syndrome [14]	Lead [89]	Malaria [3]
Pesticide [57]	Pulmonary disease [18]	*Schizophrenia* [22]
Semen quality [77]	SIDS [69]	Suicide [64]

SIDS sudden infant death syndrome

incidence (e.g., cardiovascular disease, Sect. 1.1.1) or seasonal *by birth* (e.g., stillbirth, Sect. 1.1.5). In Table 2.1 those conditions with a seasonal pattern by birth are shown in italics. As well as these health conditions, mortality is also strongly seasonal [70].

Seasonal patterns can also occur in other aspects of health. For example, survival times for colon cancer are longer for subjects diagnosed in summer and autumn [63].

Seasonality in disease has been a concern for some time. Hippocrates in around 400 BC said, "All diseases occur at all seasons of the year, but certain of them are more apt to occur and be exacerbated at certain seasons." A seasonal pattern in pulmonary disease was reported in 1853 [18], and in suicide in 1886 [64]. More recently there has been concern that seasons will become more intense due to global warming, particularly episodes of extreme heat [62]. What consequences this will have for human health, and how soon any changes will happen, are not yet known.

Seasonal changes provide a natural experiment for uncovering the aetiology of disease, by creating periods of high and low incidence. For annual seasonal patterns, every year of data gives us the opportunity to examine the association between the seasonal incidence and candidate exposures, which should also have some seasonal pattern. Finding a seasonal pattern in a disease can also act as a hypotheses generator, and as a trigger for further investigation.

2.1.1.1 Environmental Seasonal Exposures

The two major changes caused by the astronomical year are that the days are longer and hotter in summer, and shorter and colder in winter. These changes may be risk factors for disease in themselves (e.g., increased sunlight exposure as a cause of skin cancer), but the longer and hotter days also cause a cascade of changes in other environmental variables. For example, increasing sunlight leads to increases in plant growth, which causes a large seasonal increase in pollen in many places. High levels of pollen are known to trigger hay fever and possibly asthma.

An example of the environmental changes in winter is shown in Fig. 2.2. In many places around the world, rates of CVD have been shown to increase in winter [7]. It is still not certain what the cause (or causes) of this increase are. As shown in the figure the onset of winter leads to a drop in temperature. This may be the cause of

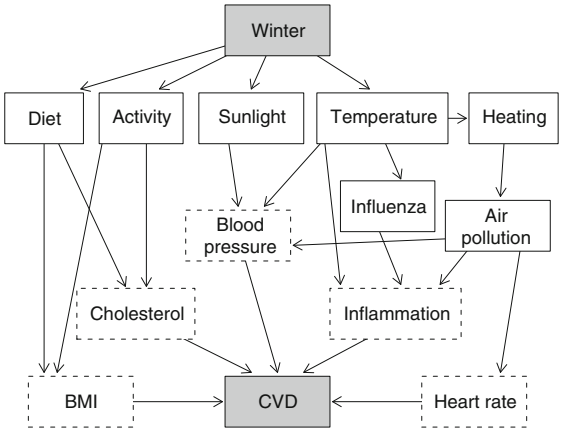


Fig. 2.2 The possible pathways between winter and the increase cardiovascular disease. Seasonal exposures are shown in *boxes with solid lines*, and risk factors in *boxes with dotted lines*. Based on Scragg [76]

the winter increase in CVD as colder temperatures lead to higher blood pressures. However, colder temperatures also mean that people use more heaters, which leads to an increased exposure to air pollution. Exposure to air pollution has been linked to an increased heart rate, which could also explain the winter increase in CVD.

Figure 2.2 highlights the inter-related nature of seasonal exposures. This can create significant epidemiological challenges, because exposures that change together will be *confounded*. This means we will not be able to determine statistically which exposure is the most likely cause of the disease. For this reason we need to exploit any differences in seasonal exposures in order to find the strongest links with disease. Using the example in Fig. 2.2, the influenza season may vary from year-to-year in intensity and timing. Its seasonal pattern may differ from the seasonal pattern in the coldest temperatures. We can then examine which pattern has the closest match with the seasonal pattern in disease. In later sections (Sect. 4.7) we discuss ways of exploiting this variability in exposure in order to find the strongest link with disease.

2.1.1.2 Social Seasonal Exposures

Social seasonal exposures are those due to regular events based on the calendar. Important changes include seasonal work (e.g., in agriculture), the school and academic year, and public holidays. An example of the health effects of a social seasonal exposure is the increase in CVD at Christmas and New Year's in the US [68]. The extra eating and drinking at this time may be the cause, or the fact that many people delay going to see the doctor during holidays. In contrast, suicide rates decrease in the periods before Christmas and other major holidays [48].

An interesting social example is the increase in accidents when the clocks go back at the start of summer, and decrease in accidents when they go forward at the end of summer, possibly caused by the loss and gain of an hour's sleep [19]. This is a social seasonal exposure caused by our reaction to astronomical changes.

Another example is an increase in asthma risk 2–4 weeks after children return to school [56].

Some social seasonal exposures, such as Christmas, occur on fixed dates, whereas others, such as Easter or the start of the school year, have changing dates from year-to-year. For statistical purposes some variation in the date is useful because social seasonal exposures can often be confounded with environmental exposures. For example, Christmas always occurs just after midwinter in the northern hemisphere, and is therefore confounded with some of the year's coldest temperatures. We show some methods for separating these two exposures in Sect. 5.3. Interestingly the Islamic calendar is lunar based, and therefore important events (such as Ramadan) occur in different meteorological conditions from year-to-year.

2.2 Descriptive Seasonal Statistics and Plots

For any epidemiological problem, seasonal or not, it is important to investigate the data before fitting a statistical model. In this section we outline some descriptive statistics for seasonal data, and some methods for representing seasonal data graphically. First we look at the small but important adjustment that needs to be considered when dealing with monthly count data.

2.2.1 Adjusting Monthly Counts

When examining seasonal patterns in count data we should always remember the denominator, which is the size of the *at-risk population*. This can change depending on the size of the population (e.g., the number of people in the city) and the length of time considered. When using monthly counts of disease the duration of exposure changes because of the unequal numbers of days in the months. This means we would expect larger counts in January (31 days) than February (28 or 29 days). We should adjust for this difference when plotting or regressing monthly count data.

For plotting monthly counts we can standardise to a common month length. This common length could be 30 days, or even $365.25/12 = 30.44$ days (the average length of a month). Figure 2.3 shows the mean rate of CVD deaths in each month using the cardiovascular disease data (Sect. 1.1.1) before and after adjusting the results to an average month length. The results are also adjusted for population size, and the y-axis shows the number of deaths per 100,000 people.

The adjusted rate of CVD was created using the formula

$$\text{rate}_t = \frac{\text{count}_t \times 100{,}000 \times 365.25/12}{\text{population}_t \times \text{days per month}_t}, \qquad t = 1, \ldots, n.$$

The numerator contains the size of the population and month length that we want to standardise to. The denominator contains the changes in the population and month length over time. The variable "days per month" is either 28, 29, 30 or 31.

The adjusted rates can be calculated using the following R code

```
> adjmean<-monthmean(data=CVD,pop=pop/100000,
  resp=cvd,adjmonth='average')
```

The population is the variable `pop` in this example. The option `adjmonth='average'` means the results are given for an average month length of 30.44 days.

The means based on an adjusted rate of events in Fig. 2.3 follow a smoother curve than the means based on the unadjusted counts. Particularly noticeable is the large increase in the rate for February after adjusting for its shorter length.

When using a generalized linear model we can account for the unequal number of days using an *offset* (Sect. 1.4.6). As an example, to add an offset using the cardiovascular disease data we use the R code

```
> monthglm(formula=cvd~-1,offsetpop=pop/100000,
  offsetmonth=TRUE,data=CVD,family=poisson())
```

The `-1` option in the `formula` means that the model is fitted without an intercept. The code has two offsets, one to adjust for the unequal number of days in the months (`offsetmonth`), and one to adjust for the population size at the time (`offsetpop`). The model was fitted assuming that the counts of deaths from cardiovascular disease follow a Poisson distribution. We give more details on regression modelling later (Sect. 2.3).

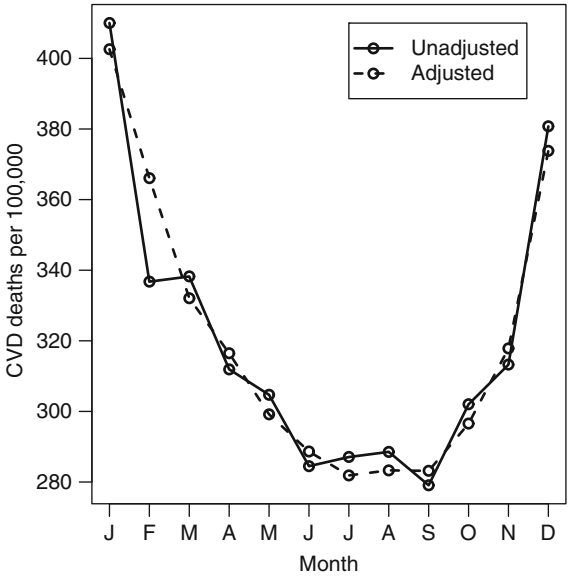


Fig. 2.3 Monthly mean rate of CVD before and after adjustment for the unequal numbers of days in the months

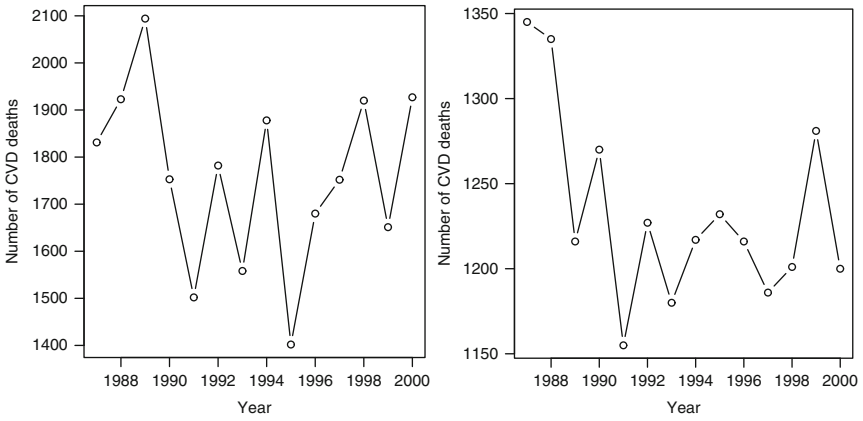


Fig. 2.4 Numbers of CVD deaths over time for the years 1987–2000 for January (*left panel*) and July (*right panel*). The scales on the y-axes are different in the *left and right panels*

2.2.2 Data Reduction

Data reduction is one of the simplest methods for investigating seasonality. It involves ignoring or transforming a selection of the data. Sometimes this may be the most appropriate approach, as the seasonal pattern may only occur in a specific part of the year.

As an example of data reduction we show the rates of CVD death over time for January and July in Fig. 2.4. January has the highest rates in every year, whilst July has some of the lowest. It can be useful to consider just the worst and best levels of a seasonal pattern. The rates in 1987 to 1990 seem generally higher, but this may be due to long-term trend and a gradual reduction in CVD death with time. The rates for January are quite erratic from year-to-year. The rates for July are reasonably stable after 1990 except for 1999 which had a comparatively high number of deaths. In the same year the number of deaths in January was comparatively low. A possible explanation of this pattern is that temperature-related CVD is thought to affect the most vulnerable people in the population. A mild winter with fewer cold-related deaths would mean that there would be more vulnerable people in the following summer, who are at greater risk of heat-related deaths [72]. This may also help to explain the roughly alternating pattern in the January peak, as a winter with many deaths means fewer vulnerable people in the following winter (whereas a mild winter with a small number of deaths increases the vulnerable "pool" in the following winter).

Figure 2.5 shows the results for all 12 months. It is important to note that the scales on the y-axes are not equal. This is actually data rearrangement, not data reduction, as all 168 observations are used and the information is the same as in Fig. 1.1.

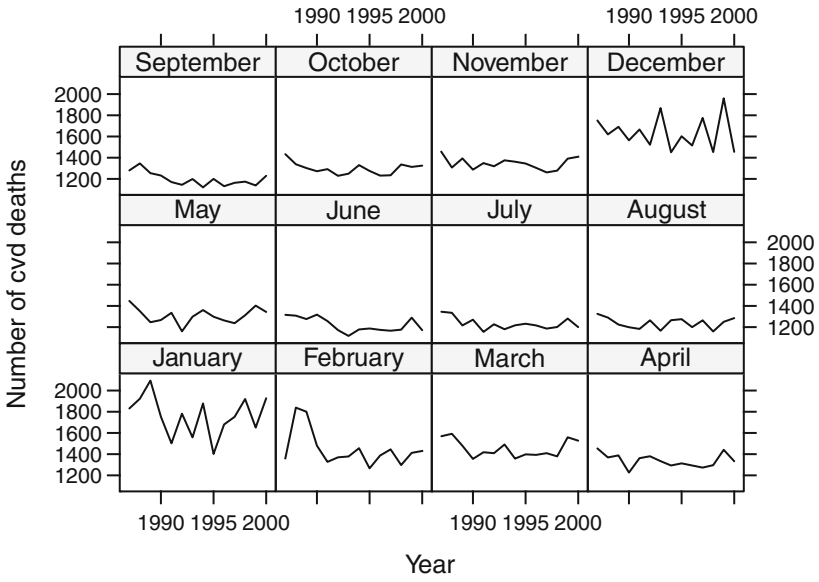


Fig. 2.5 Counts of CVD deaths over time for the years 1987–2000 by month.

The R code to create this plot is

```
> CVD$month <- factor(CVD$month, labels=month.name)
> xyplot(cvd~year|month,data=CVD,type="l")
```

Another useful form of data reduction for seasonal patterns is to look at the range of responses in a year. As an example, we subtract the total number of births in May from the number in the following November using data from the UK (Sect. 1.1.6). The differences are shown in Fig. 2.6. May has 31 days, whereas November only has 30, hence we first standardised the number of births in May to 30 days by multiplying the observed number by $\frac{30}{31}$ (Sect. 2.2.1).

The plot shows a declining difference from 1947 to 2003, indicating a decline in the seasonal pattern. There are a number of possible reasons. One possible explanation is the increased trend for children to be born outside of marriage [20, 46]. As marriages are more popular in the summer months of June to August, this leads to a peak in births approximately nine months later in March to May (birth rates were also higher in March and April, although we only used the data for May). During the war (1939–1946) the number of births was reduced and most social norms (such as marriage times) were disrupted. A gradual improvement in nutrition after the war would have reduced the seasonality in female fertility, which peaks in summer during periods of poor nutrition. Also, after 1960 there was a large increase in both contraceptive pill use and abortions. These changes gave women more control over when they had children.

It is worth noting that by subtracting the counts of births from the same year we have removed the strong secular trend in births. This de-trending makes it easier to focus on the seasonal aspects of birth numbers.

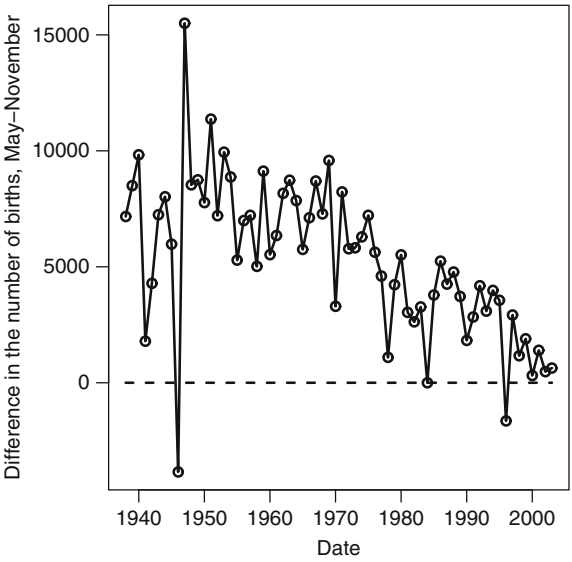


Fig. 2.6 Difference in the total numbers of births between May and the following November in the UK for the years 1938–2003. Births in May adjusted to a 30-day month length. *Dotted horizontal reference line* at an equal number of births in May and November

Using this technique for detecting a long-term trend in seasonality relies on the seasonal pattern being well characterised using dates for just two months. Ideally these two months would be chosen to capture the highest and lowest levels.

To illustrate the effect we show two sinusoidal disease patterns in Fig. 2.7 together with the difference between the rates in January and July. The seasonal pattern in one disease has remained constant over time, whereas the seasonal variation for the other has decreased (although the overall mean has remained the same at a rate of 1). The reduction in the seasonal pattern is clear in this example using either plot. The data reduction plot may help where the results are noisy. The example also illustrates an important drawback of data reduction, as we have reduced the sample size from 48 (4 years of monthly data) to just 4.

2.2.2.1 Grouping Data into the Four Seasons

A common method of data reduction when analysing seasonal data is to group the data into the four seasons. These seasons may be defined in a number of ways, and we give three definitions in Table 2.2. These definitions are for the Northern hemisphere; for the Southern hemisphere the seasons are simply reversed. The *solstices* are the two extremes in the year when the Earth is most tilted toward (or away) from the sun (depending on the hemisphere). They are therefore the longest and shortest days. The *equinoxes* are the two points in the year when the length of day and night are roughly equal. These four important events do not always occur on the same day every year. For example, the summer solstice is not always on 21 June, and can sometimes be on 20 June. This movement is caused by the imbalance between the

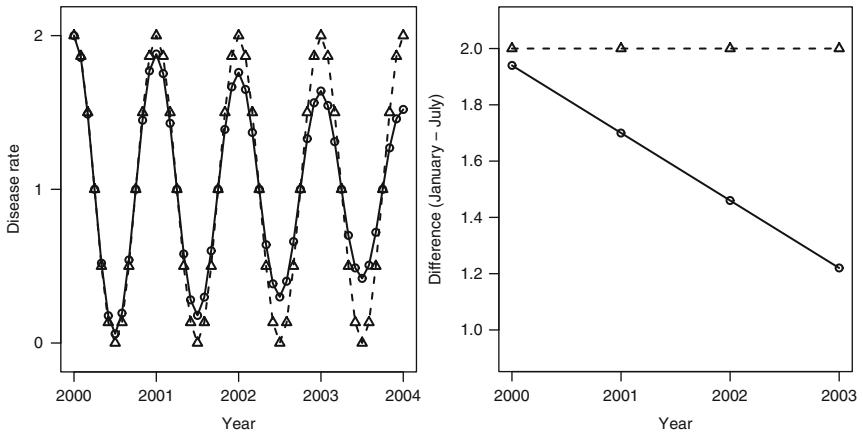


Fig. 2.7 Plot of two sinusoidal disease patterns (*left panel*) and the difference between the rates in January and July (*right panel*)

Table 2.2 Three alternative definitions of the four seasons in the Northern hemisphere. The table shows the range of dates and the number of days in each season

	Meteorological
Spring	Mar, Apr, May (92 days)
Summer	Jun, Jul, Aug (92 days)
Autumn	Sep, Oct, Nov (91 days)
Winter	Dec, Jan, Feb (90/91 days)
	Astronomical
Spring	20 Mar (vernal equinox) to 20 Jun (93 days)
Summer	21 Jun (summer solstice) to 21 Sep (93 days)
Autumn	22 Sep (autumnal equinox) to 20 Dec (90 days)
Winter	21 Dec (winter solstice) to 19 Mar (89/90 days)
	Mid-solstice
Spring	5 Feb to 6 May (91/92 days)
Summer	7 May to 5 Aug (91 days)
Autumn	6 Aug to 5 Nov (92 days)
Winter	6 Nov to 4 Feb (91 days)

calendar year (which is 365 or 366 days long) and the astronomical year (which is approximately 365.25 days long) [28].

The Meteorological definition uses whole months to create four seasons. It is called the Meteorological definition as it roughly breaks the year into the coldest and hottest quarters. The Astronomical definition is based on changing the season at the start of every solstice or equinox. The Mid-solstice definition has the solstices and equinoxes at the centres of the seasons.

Figure 2.8 shows the three definitions over an annual time line (for the Northern hemisphere). The Mid-Solstice seasons start earliest in the calendar year, followed by the Meteorological and Astronomical.

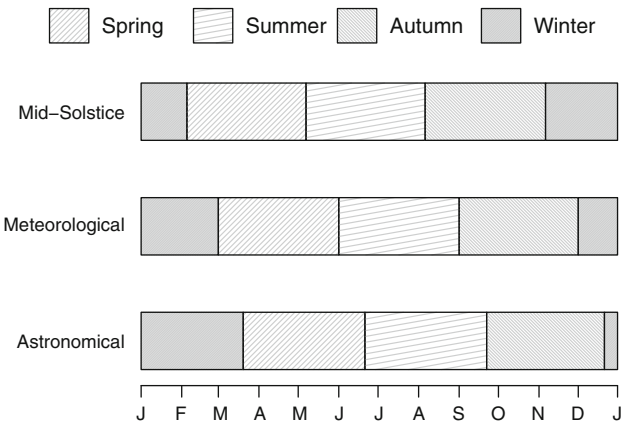


Fig. 2.8 Three definitions of the four seasons for the Northern hemisphere

For any of the three definitions there is a small difference in the number of days in each season (Table 2.2). This could cause a small bias when using count data as the denominator is not consistent. It is possible to remove this bias using an offset, as described in Sect. 2.2.1.

The best seasonal definition to use will depend upon the data. The Mid-Solstice seasons most evenly break the year in terms of day length, but average temperatures often lag behind day length, hence the warmest and coolest periods are better captured by the Meteorological definition. The Astronomical definition is, arguably, the most familiar to the public, although the most commonly used popular seasons vary between countries. In tropical and sub-tropical regions there are only two important seasons, wet and dry, which have different dates. In some cases the data may not have been collected using day of the month, but collected using only month. In this case it is only possible to use the Meteorological definition.

Reducing the data to seasons can make results easier to communicate. For example, if we wanted to tabulate average rates of CVD, using months would require 12 rows (or columns depending on the layout), whereas using seasons requires only 4. This space saving becomes particularly important if results are presented for multiple groups (e.g., by gender). Reporting the data by seasons also avoids the confusion of January being a winter month in the northern hemisphere and a summer month in the southern.

A disadvantage of using defined seasons is the arbitrary differences it creates. For example, using the Astronomical definition the 20 March and 20 June are considered the same, whereas the 19 and 20 March are not. These arbitrary changes ignore more plausible smooth changes in a seasonal disease. To illustrate the problems this can cause let us assume that we are examining a disease with a seasonal pattern that is sinusoidal. Figure 2.9 shows two sinusoidal disease patterns, one with a phase on 5 February and another with a phase on 20 March. The dashed line shows the mean estimates based on using the astronomical seasons. Using the four seasons

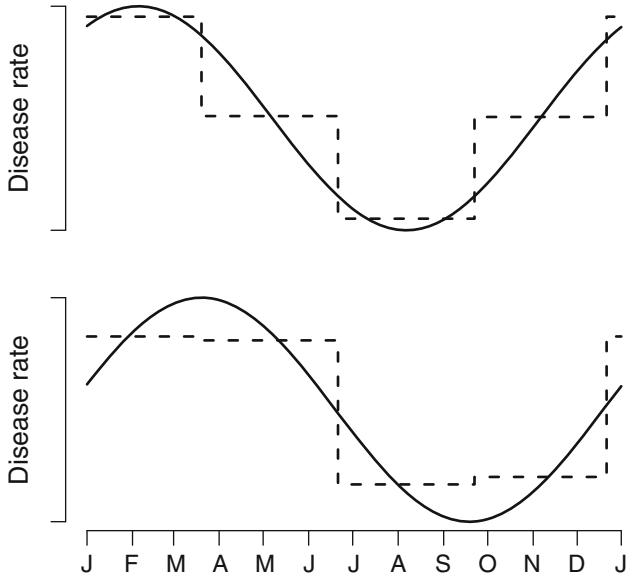


Fig. 2.9 Sinusoidal seasonal disease pattern (*solid line*) estimated using the four Astronomical seasons (*dashed line*). Sinusoid with a phase of 5 February (*top panel*) and 20 March (*bottom panel*)

represents smooth change in "steps". The closeness of these steps to the sinusoid depends on the phase. With a phase on the cusp of winter and spring (20 March) the estimates show only two disease rates: high in winter and spring, and low in summer and autumn.

Care needs to be taken when interpreting estimates based on four seasons, as they represent the *average rates* in each season. As shown in Fig. 2.9 these averages may not capture the amplitude of the seasonal variation in disease, which ranges from the highest to the lowest point of the sinusoid. Conversely, when we calculate seasonal estimates based on sinusoids (Chap. 3) we need to remember that this maximum difference only occurs for a short period of time.

Figure 2.10 shows the same sinusoidal changes in disease as in Fig. 2.9, this time estimated using months instead of seasons. Again the estimates are in steps, but the overall seasonal pattern is somewhat smoother. Also the largest difference in disease rates is captured regardless of the phase.

All of the problems highlighted above only occur if the true seasonal pattern in disease is sinusoidal, or follows some other gradually changing seasonal pattern. Step-like seasonal patterns are possible when an exposure changes abruptly, as may happen with a social seasonal exposure (Sect. 2.1.1). In this case a categorical seasonal estimate might be preferable to a smooth seasonal estimate.

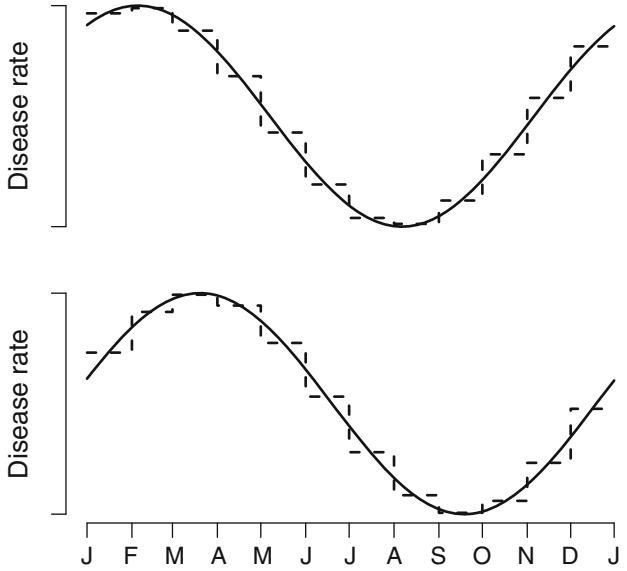


Fig. 2.10 Sinusoidal season (*solid line*) estimated using months (*dashed line*). Sinusoid with a phase of 5 February (*top panel*) and 20 March (*bottom panel*)

2.2.3 Circular Plot

So far we have summarised seasonal patterns using linear plots, with month on the x-axis and the seasonal summary on the y-axis (e.g., Fig. 1.31). A disadvantage of this type of plot is that the results from December and January are not side-by-side. An alternative is a *circular plot* and an example using the mean rates of CVD is shown in Fig. 2.11. These plots are also known as *rose diagrams*. The distance from the centre to edge of the segment (or petal) is proportional to the size of the effect (e.g., the mean or rate).

The mean rates in Fig. 2.11 have been adjusted for the unequal number of days in the month (Sect. 2.2.1). January had the highest mean rate of 403 deaths per 100,000 people. The plot shows the steady decline from January to July, and then the steady rise back again.

The R code to produce this plot is

```
> adjmean<-monthmean(data=CVD,resp=cvd,pop=pop/
  100000,adj='average')
> plotCircular(radii1=adjmean$mean,dp=1,labels=month.abb,
  scale=0.7)
```

The first line of code adjusts the counts to account for the unequal number of days and standardises to a population size of 100,000.

It is also possible to show weekly data using a circular plot. Figure 2.12 shows the number of CVD deaths by day of the week. We used the total counts of CVD rather

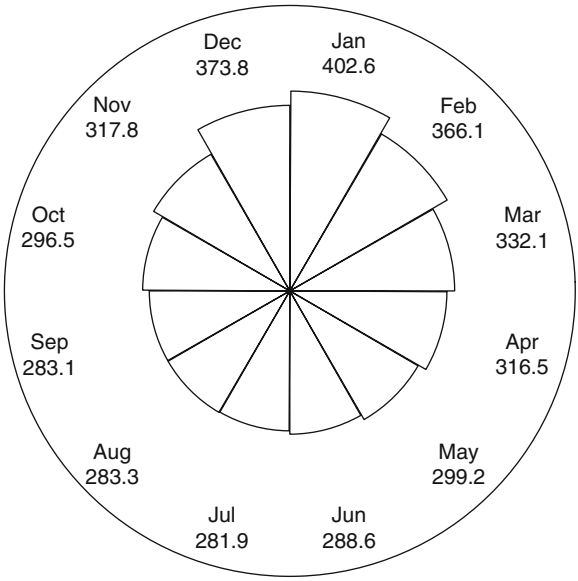


Fig. 2.11 Circular plot of the adjusted mean monthly rates of CVD per 100,000

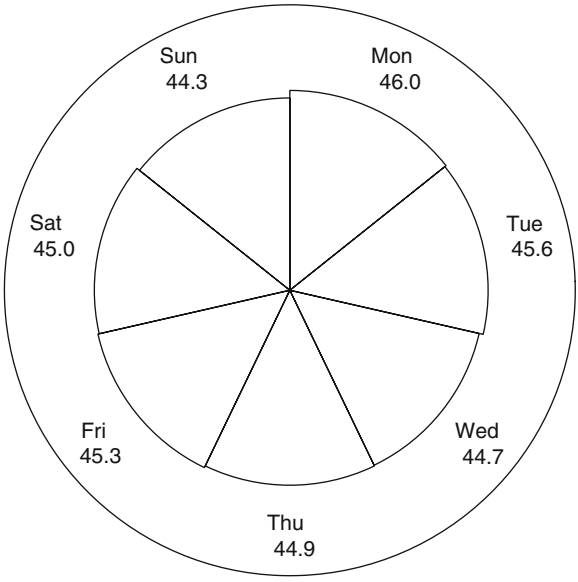


Fig. 2.12 Circular plot of the average numbers of CVD deaths by day of the week

than an adjusted rate because each of the seven days was almost equally represented (Monday to Wednesday each occurred 730 times and the other days 731). The mean counts are similar, although there was a slight increase on Mondays (46.0 deaths) compared to other days.

We can also create a circular plot for Binomial data and we use the stillbirth data as an example (Sect. 1.1.5). Using the rate of stillbirths, defined as the number of stillbirths divided by the total number ofbirths, automatically adjusts for the

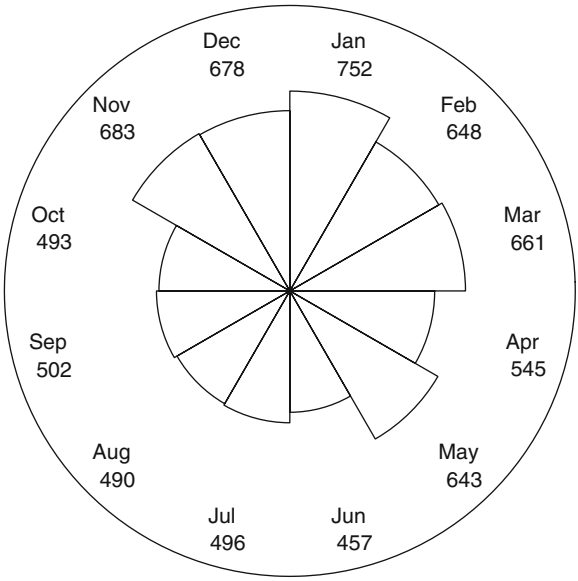


Fig. 2.13 Circular plot of the rates of stillbirth per 100,000 births, 1998–2000, Queensland

unequal lengths of the months. Also, the total number of births has its own seasonal pattern, and this is also accounted for by using the number of births as a denominator. Figure 2.13 shows the rate of stillbirths per 100,000 births. The rate of stillbirths was highest in January and lowest in June. The rates appear to have a step-like seasonal pattern with low rates in June to October (southern hemisphere winter) and high rates in November to May.

It is also possible to add an estimate of uncertainty to the circular plot using "spokes" that radiate from the centre of the circular axes. Figure 2.14 shows the mean baseline BMI from the exercise study according to the month of joining the study (Sect. 1.1.4). The segments shown the mean BMI and the spokes show the standard error of the mean. The highest mean BMI was 36.5 kg/m^2 in February. This was also the month with the greatest amount of uncertainty, as there were only three subjects with baseline data in February.

As a last illustration of the use of circular plots, we show what the three seasonal patterns in Fig. 2.1 look like as circular plots in Fig. 2.15. These are idealised representations of a seasonal pattern, without any noise. The figures based on real data earlier in the section are not always as smooth.

2.2.4 Smooth Plot of Season

A circular plot summarises the overall seasonal pattern, but it is also useful to look at the seasonal pattern from year-to-year. This can be achieved by fitting a spline for time as part of a regression model (Sect. 1.4.8). The spline then represents the mean value over time. A spline that is flexible enough will include any seasonal variation.

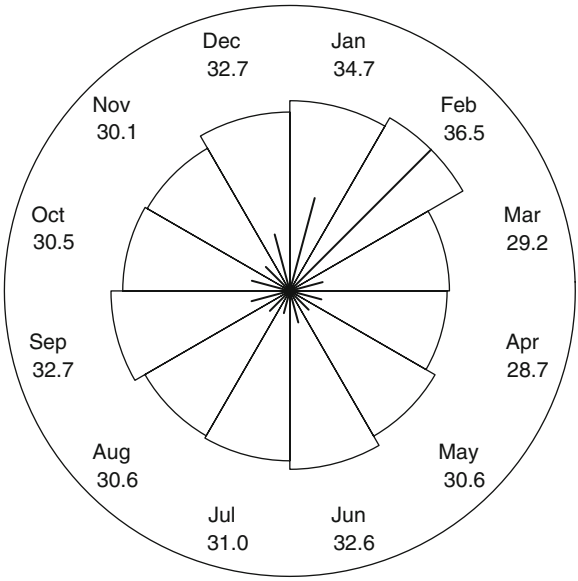


Fig. 2.14 Circular plot of mean BMI (kg/m^2) and standard error of the mean using spokes

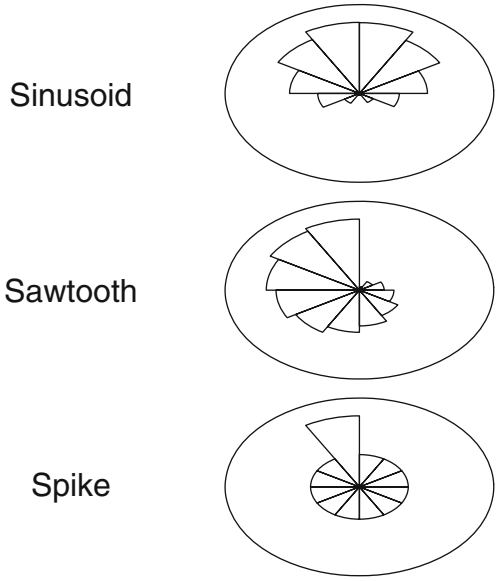


Fig. 2.15 Circular plots of three seasonal patterns shown as a linear plot in Fig. 2.1

As an example of the use of a spline we examine the exercise data (Sect. 1.1.4). This is a particularly good example because it is difficult to visualise any seasonal pattern in the raw data as it is very skewed, and there is a large variation between subjects.

Figure 2.16 shows a spline for the walking time per week. The spline represents the mean walking time, and the dashed lines are the 95% confidence interval of the

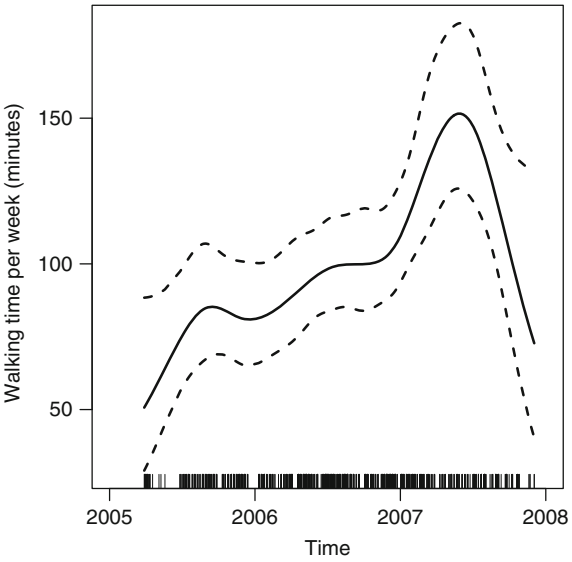


Fig. 2.16 Smooth estimate of mean walking time per week using a penalized spline. Mean estimate (*solid line*) and 95% confidence interval (*dashed line*)

mean. For this graphical summary we have ignored the longitudinal design. There is a clearly increasing trend in walking time, which may be due to the encouragement the participants received during the course of the trial. There is also a seasonal pattern, with increases in the mean walking time in July or August (two of the coolest months in Queensland). It is possible that overall levels of walking decline in summer, when it is often too hot to go outside for long periods during the day.

The smoothed mean was estimated using a penalized cubic regression spline with seven degrees of freedom. We discuss choosing the degrees of freedom in Sect. 5.2. We assumed that walking time had an over-dispersed Poisson distribution. The R code to produce the plot is

```
> s.model<-gam(walking~s(as.numeric(date),k=8,bs='cr'),
  family=poisson(),data=exercise,scale=-1)
> plot(s.model)
```

This uses the `mgcv` R library [88].

The times of the observations are shown in a *rug plot* along the x-axis of Fig. 2.16 (using the R command `rug`). This is particularly useful for unequally spaced data, as it highlights time periods that have few responses, and hence where we should be less certain about the mean response. This uncertainty is also reflected in the width of the confidence intervals. In Fig. 2.16 the data are spread reasonably evenly over time.

2.3 Modelling Monthly Data

In this section we describe regression modelling methods that use month to capture annual seasonal patterns. In words the regression equation is

$$\text{disease} = \text{month} + \text{other independent variables.}$$

In this section month will be fitted as a categorical independent variable, and we explore simple *fixed effects* as well as more complex *random effects*. By using *generalized linear models* (GLMs) we are able to model data that can be represented by Normal, Poisson or Binomial distributions (Sect. 1.4.5).

We have already plotted summary statistics by month (e.g., Figs. 1.31, 2.3 and 2.11), and we have seen how months can give a reasonably close approximation to an annual sinusoidal seasonal pattern (Fig. 2.10).

For now we ignore any long-term trend, as we deal with this important issue later in Chap. 4.

2.3.1 Month as a Fixed Effect

We start by fitting month as a fixed categorical effect. Each month is treated as a separate category with a fixed parameter. We define $m(t)$ to be the month at time t (or observation t) so $m(t) \in \{1, 2, \ldots, 12\}$. An example of $m(t)$ can be seen for balanced data in Table 1.1, and we could extract month from the date for the unbalanced data in Table 1.3. We always assume that the months are numbered in order, so that 1 = January, 2 = February, etc.

The GLM regression equation including month can be written as

$$g(Y_t) = \delta_{m(t)} + \sum_{k=1}^{p} \beta_k X_{t,k}, \qquad t = 1, \ldots, n, \tag{2.1}$$

where $\delta_{m(t)}$ is the effect of month. There are p other independent variables ($\beta_1, \ldots, \beta_p$), which would typically include an intercept by setting the first column of the design matrix to 1 ($x_{t,1} = 1$ for all t). If we use an intercept then we need to create a reference month by fixing one of the δ's to be zero. Without this reference month the model would be over-parameterised.

An example GLM fit is shown in Table 2.3 using the CVD data. For this model we assumed that the number of CVD deaths followed a Poisson distribution and we adjusted for the unequal number of days in the month using an offset (Sec 2.2.1).

The R code to produce the estimates is

```
> monthglm(formula=cvd~1,data=CVD,family=quasipoisson(),
  offsetpop=pop/100000,offsetmonth=TRUE)
```

Using the option `family=quasipoisson` accounts for any over-dispersion (Sect. 1.4.5).

As we assumed a Poisson distribution the results are in terms of rate ratios (RRs) (Sect. 1.4.5). The rates are relative to January (the reference month), so for example, the rate of CVD deaths in February is 0.90 times the rate in January (with a 95%

Table 2.3 Estimates of the mean rate ratios (RRs) and 95% CIs of CVD deaths using an over-dispersed Poisson model with month as a fixed categorical independent variable and January as a reference month

Month	RR	95% CI	Z-value	p-Value
January	1	–		
February	0.90	0.86, 0.95	−4.22	<0.001
March	0.82	0.79, 0.87	−7.73	<0.001
April	0.79	0.75, 0.83	−9.44	<0.001
May	0.74	0.71, 0.78	−11.57	<0.001
June	0.72	0.68, 0.75	−12.71	<0.001
July	0.70	0.67, 0.74	−13.66	<0.001
August	0.70	0.67, 0.74	−13.48	<0.001
September	0.70	0.67, 0.74	−13.37	<0.001
October	0.74	0.70, 0.77	−11.89	<0.001
November	0.79	0.75, 0.83	−9.29	<0.001
December	0.93	0.89, 0.97	−3.07	<0.001

CI of 0.88, 0.92). A test statistic is given comparing the rate in each month with January. In this case the rates in every month are statistically significantly lower than the rate in January. Comparisons of other months are possible (e.g., testing the hypothesis that the rate in June is equal to the rate in July).

Rather than focus on the results of paired tests we prefer to look for an overall seasonal pattern. The mean rate ratios are plotted in Fig. 2.17 together with the 95% confidence intervals. The seasonal pattern is remarkably smooth, with a strong U-shaped pattern. An alternative plot is shown in the right-hand panel of Fig. 2.17 which plots the mean death rates per 100,000 people. The estimated seasonal pattern is exactly the same, only the scale is different, and the results for January now have a confidence interval. These estimates were made by fitting the model without an intercept, using the R code `formula=cvd~-1`.

Using a GLM it is possible to apply the same model, (2.1), to data with a variety of distributions. The stillbirth data (Sect. 1.1.5) are binary and so we can use a logit link. However, as the probability of stillbirth is small we prefer to use the complementary log-log link, as this is often better for modelling very small or very large probabilities [26, Chap. 8]. However, for this example the results are almost identical using the logit link.

The R code to produce the estimates is

```
> monthglm(formula=stillborn~1,data=stillbirth,
  family=binomial(link="cloglog"))
```

The estimates are plotted in Fig. 2.18. The odds ratios for stillbirth are lower from July to October and higher from November to March. These ORs follow the same pattern as the rates of stillbirth per 100,000 (Fig. 2.13). The 95% confidence interval for each month includes 1, indicating no significant difference in any month compared with January.

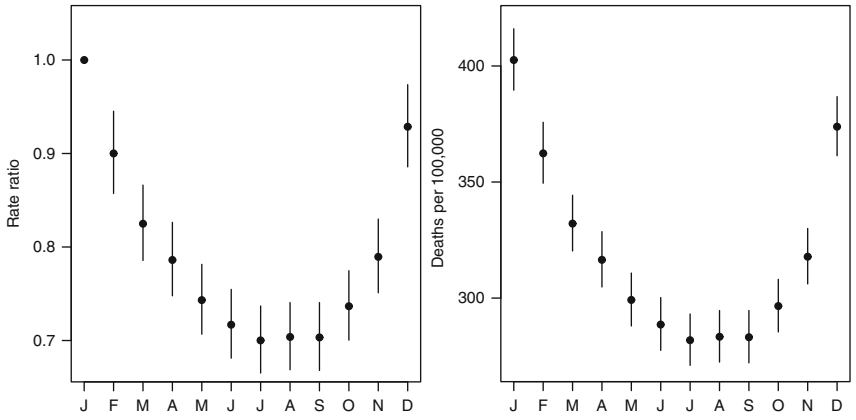
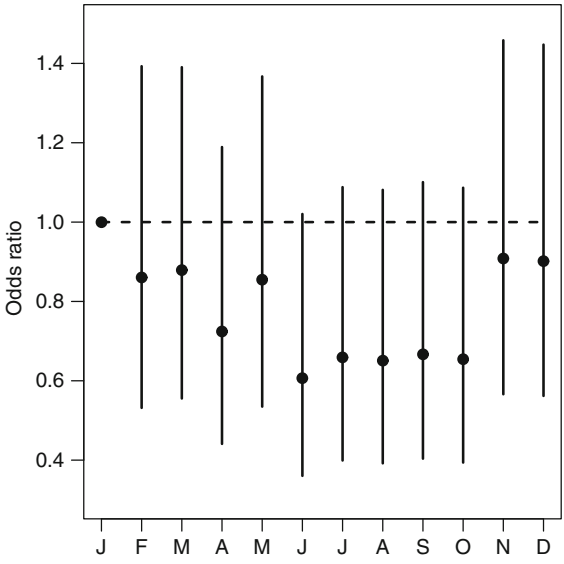


Fig. 2.17 Monthly mean estimates of CVD (and 95% confidence intervals) using an over-dispersed Poisson model. The *left panel* shows the rate ratios fitted with January as the reference month. The *right panel* shows the death rates per 100,000, fitted without an intercept

Fig. 2.18 Monthly mean estimates of the odds ratios of stillbirth (and 95% confidence intervals) using January as the reference month



Rather than testing individual months, we can test the overall value of adding month by comparing the AIC of the model including month to the AIC from a model with only an intercept (Sect. 1.4.7). The AIC for the model with month is 4,335.3, the model without month has a smaller AIC of 4,322.7, a difference of 12.6. The model with month used 11 extra parameters, which is a reasonably large increase in complexity. This large difference in the AIC indicates that adding month has not sufficiently improved the fit of the model, and the simpler model without month should be preferred. Two possibilities are suggested: (1) there is no seasonal pattern

in stillbirth, (2) the seasonal pattern in stillbirth is simpler than a model assuming different rates in each month. The circular plot of the rates (Fig. 2.13) does seem to suggest a simpler two-level seasonal pattern.

2.3.2 Month as a Random Effect

In the above examples month was fitted as a *fixed effect*, meaning that each month was given a fixed parameter value. We can instead fit month as a *random effect*, meaning that we allow the estimates to follow some distribution. The distribution most often used for random effects is the Normal distribution. The random effects in (2.1) would be

$$\delta_j \sim \mathrm{N}(0, \sigma_\delta^2), \qquad j = 1, \ldots, 12.$$

For this parameterisation we do not use a reference month. Instead we fit an intercept which is the overall average and the random effects represent the difference from the overall average.

The monthly estimates are constrained to be Normally distributed, and the spread of this Normal distribution is controlled by the variance parameter (σ_δ^2), a larger variance means a larger spread in the monthly estimates. This translates to a greater difference in the highest and lowest seasonal estimates. We use a subscript "δ" for the variance to emphasise that it describes the spread of the $\boldsymbol{\delta}$ estimates. An important feature of the Normal distribution is that it is symmetric, so generally the monthly estimates will be equally spread above and below the overall mean.

The fixed effect used 11 parameters for month, whereas the random model uses 13: one for each month (δ_j) plus the variance (σ_δ^2). This is a relatively small increase in complexity, and is usually offset by some borrowing of strength due to the use of the Normal distribution.

To fit this model in R we prefer to use the Bayesian library "BRugs" [83]. Because of this the results are presented using credible intervals rather than confidence intervals (Sect. 1.6).

The monthly estimates using a random effect are shown in Fig. 2.19. As in Fig. 2.17 there is a clear U-shaped pattern. The estimates for April and November are close to the average rate. There are four months above the average rate, and six months below, indicating a positive skew in the seasonal pattern.

2.3.3 Month as a Correlated Random Effect

A disadvantage of the previous model is that is assumes complete independence between months; δ_i and δ_j are uncorrelated if $i \neq j$. For many seasonal patterns this is unlikely to be true, as neighbouring months are likely to be positively correlated. We can introduce a correlation between months using a technique called *conditional*

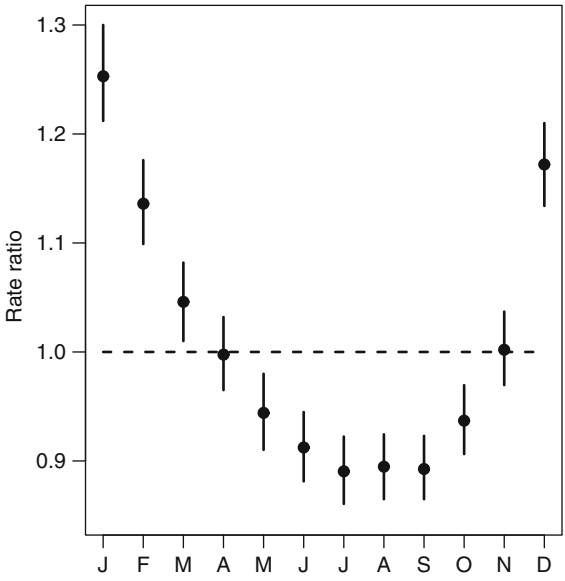


Fig. 2.19 Monthly estimates of the rate ratios of CVD (and 95% credible intervals) using a random effect for month. The *horizontal reference line* is at a rate ratio of 1

autoregression (CAR). This method has been widely used for modelling *spatial correlation* between neighbouring areas (e.g., counties or countries) [67].

We first assume the month effects have a *Multivariate Normal* (MVN) distribution

$$\boldsymbol{\delta} \sim \text{MVN}(\mathbf{0}, \mathbf{V}_\delta),$$

where $\mathbf{0}$ is a vector of zeros of length 12, and $\mathbf{V}_\delta$ is a 12×12 variance–covariance matrix. This means that the estimates for each δ are no longer independent but can depend upon other δ's.

To make every neighbouring month correlated we specify the 12×12 variance–covariance matrix as

$$\mathbf{V}_\delta = \sigma_\delta^2 \begin{bmatrix} 1 & \rho & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho \\ \rho & 1 & \rho & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \rho & 1 & \rho & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \rho & 1 & \rho & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \rho & 1 & \rho & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho & 1 & \rho & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \rho & 1 & \rho & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \rho & 1 & \rho & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho & 1 & \rho & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho & 1 & \rho & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho & 1 & \rho \\ \rho & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho & 1 \end{bmatrix}. \tag{2.2}$$

The matrix has 12 rows and 12 columns to represent the 12 months. This formulation splits the variance–covariance matrix into the overall variance σ_δ^2 and a correlation matrix. The correlation between neighbouring months is $0 \leq \rho \leq 1$, the correlation between results from the same month is 1, and the correlation between results more than one month apart is zero. The matrix is symmetric as the lower-left triangle is the reverse of the upper-right triangle. This means that the correlation between months i and j is equal to the correlation between months j and i. An important feature of the matrix is the ρ that appears in the lower-left and upper-right corner, which creates a link between January and December.

In the previous section we used an uncorrelated random effect, which is equivalent to specifying the above variance–covariance matrix with $\rho = 0$.

The matrix $\mathbf{V}_\delta$ is a *sparse matrix* as it contains many zeros, which means it cannot be readily inverted. However, inverting the variance–covariance matrix is essential for most estimation techniques (e.g., least squares). This means that the above model cannot be fitted using standard estimation techniques. Instead we can use conditional autoregression (CAR) to break the large matrix into smaller parts.

The CAR model fits the matrix (2.2) by assuming that the estimate for each month conditional on its neighbouring months has a Normal distribution

$$\delta_j | \delta_{\underline{j}} \sim \mathrm{N}\left(\overline{\delta}_j, \frac{v}{n_j}\right), \qquad j = 1, \ldots, 12,$$

where $\delta_{\underline{j}}$ denotes the estimates from the months that neighbour month j, and n_j is the number of neighbouring months, which is always 2 when using matrix (2.2). The parameter v describes the variance, which is constant across months. The mean of each month is defined using its two neighbours

$$\overline{\delta}_j = \begin{cases} \left(\overline{\delta}_2 + \overline{\delta}_{12}\right)/2, & j = 1, \\ \left(\overline{\delta}_{j-1} + \overline{\delta}_{j+1}\right)/2, & j = 2, \ldots, 11, \\ \left(\overline{\delta}_1 + \overline{\delta}_{11}\right)/2, & j = 12. \end{cases}$$

In words, the above equation simply states that the mean for each month will be the means of the neighbouring months. This will smooth the estimates by producing a more gradual change between months.

A graphical representation of the matrix (2.2) is given in Fig. 2.20. Each month has two neighbours. The important feature is the join between December and January, which creates a circular, rather than linear, model.

Fitting the correlated random effects model to the CVD data produces the monthly estimates shown in Fig. 2.21 (as in previous models, we included an estimate of the over-dispersion). In this case the estimates are almost identical to the estimates using an uncorrelated random effect. The model with uncorrelated random effects had a DIC of 1831 based on 143.7 parameters. The model with correlated random effects had a DIC of 1830 based on 143.4 parameters. This indicates a very similar fit of the two models, which is not surprising given the similarity of the

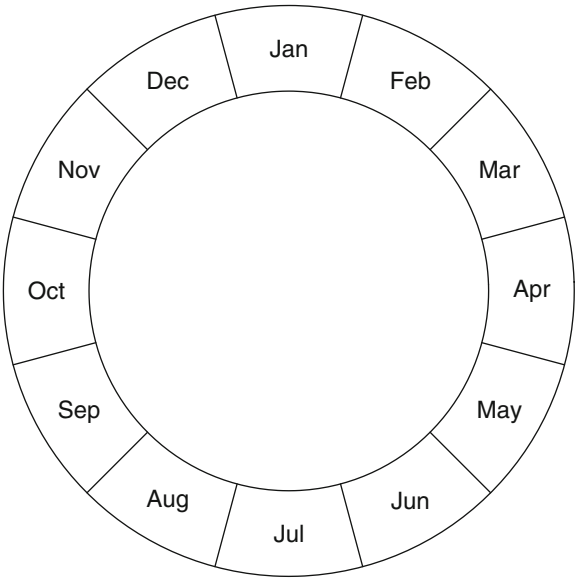


Fig. 2.20 Graphical representation of the neighbourhood for months

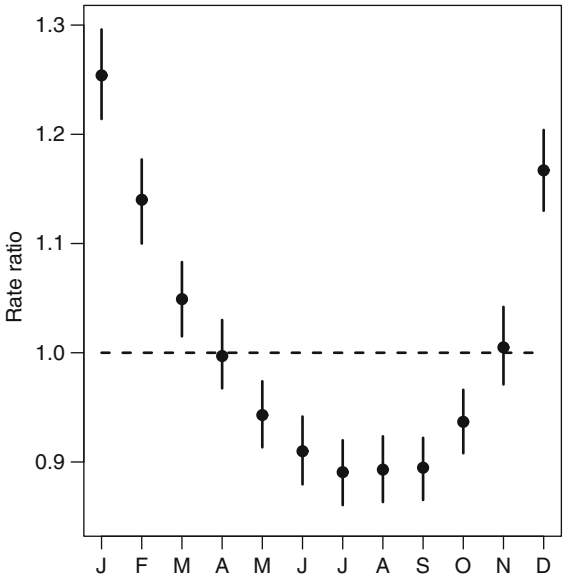


Fig. 2.21 Monthly estimates of the rate ratios of CVD deaths using correlated random effects. The *horizontal reference line* is at a rate ratio of 1

parameter estimates. In this case the monthly estimates are already very smooth, and so are not altered by allowing neighbouring months to be correlated.

The effective number of parameters for these models seems high on first inspection: 143.7 and 143.4 for the uncorrelated and correlated model, respectively. However, most of the parameters are used to model the over-dispersion. Fitting identical models without accounting for over-dispersion results in 11.9 parameters for

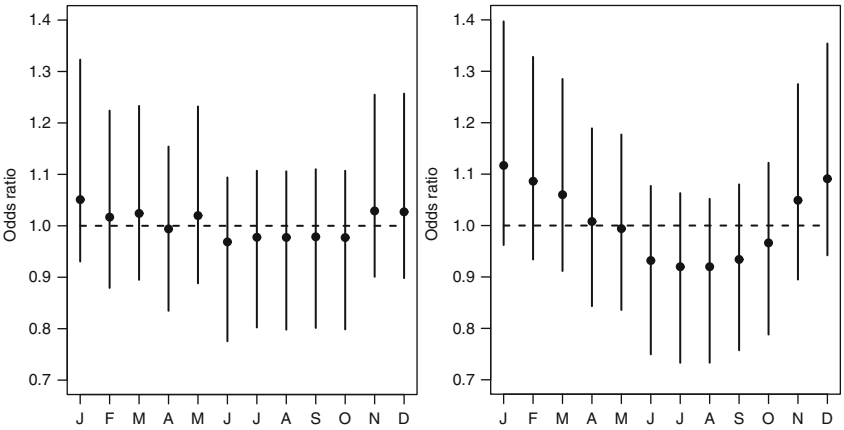


Fig. 2.22 Monthly mean estimates of odds ratios of stillbirth rates (and 95% credible intervals). Uncorrelated random effects (*left panel*), correlated random effects (*right panel*). The *horizontal reference line* is at an odds ratio of 1

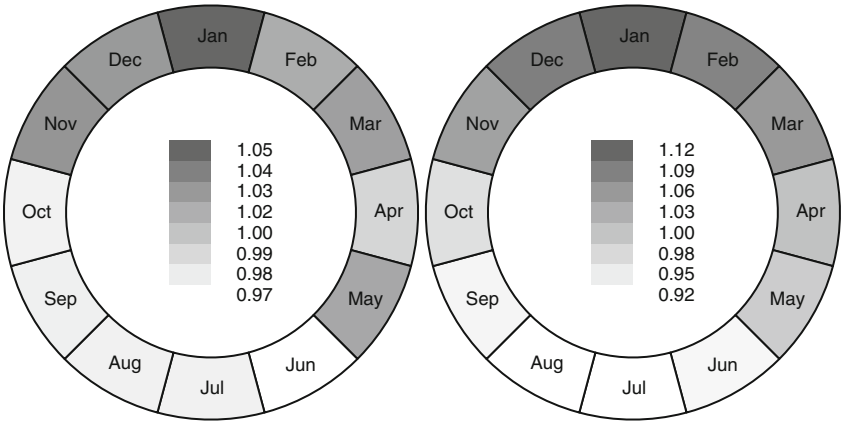


Fig. 2.23 Monthly mean odds ratios of stillbirth rates: uncorrelated random effects (*left panel*), correlated random effects (*right panel*)

the uncorrelated model and 11.7 for the correlated model. These are much closer to the expected number of regression parameters of 12 (one for each month). The slightly smaller number of effective parameters for the correlated model is due to some borrowing of strength.

An example where there is a large difference between using an uncorrelated and correlated random effect is shown in Fig. 2.22 for the seasonal pattern in stillbirth. Here the estimates based on correlated effects show a much smoother seasonal pattern.

The DIC for the uncorrelated random effects model was 183.0, using 3.6 parameters. The DIC for the correlated random effects model was 181.5, using 4.1

parameters. This indicates a small, and possibly unimportant, improvement in fit due to a small increase in the number of parameters. As a comparison, the DIC for an intercept only model was 181.2, using 1.0 parameters. This indicates that the smooth seasonal effects used 3.1 parameters, which is much less than the theoretical maximum of 12. This is due to the correlation between months, and 3.1 represents the amount of information needed to fit this seasonal pattern. However, adding this seasonal pattern did not considerably improve the overall fit (Table 1.8), and we cannot be sure that the rates of stillbirth are seasonal. More precisely, we cannot be sure that they have a smoothly varying seasonal pattern.

Figure 2.23 shows the mean monthly odds ratios plotted in a circular form as per Fig. 2.20. The smoothing has increased the estimated risks in December and February, and decreased them in September. These plots were produced by the R function `plotCircle`.

Chapter 3
Cosinor

A simple and popular method for analysing seasonality is the *cosinor*. Its popularity stems from its ease of application and interpretation [2, 60]. Also, it can be applied to time series with regular dates, or survey data with irregular dates.

In Sect. 1.3 we defined a sinusoidal equation with an amplitude and phase, Eqn. (1.4). Although this equation is relatively simple it is non-linear in time, because the phase (P) is part of the argument of the cosine function (i.e., $\cos(\omega t - P)$). This makes it difficult to estimate the amplitude and phase using standard techniques (e.g., least squares).

An identical, but linear, representation of (1.4) is

$$Y_t = c\cos(\omega_t) + s\sin(\omega_t), \qquad t = 1, \ldots, n, \tag{3.1}$$

where the amplitude is

$$A = \sqrt{c^2 + s^2}, \qquad (A \geq 0) \tag{3.2}$$

and the phase (in radians) is

$$P = \begin{cases} \arctan(s/c), & c \geq 0, \\ \arctan(s/c) + \pi, & c < 0,\ s \geq 0, \\ \arctan(s/c) - \pi, & c < 0,\ s > 0. \end{cases} \tag{3.3}$$

This model is now linear in time, and values of $\cos(\omega_t)$ and $\sin(\omega_t)$ can be added to the design matrix for a fixed frequency (ω). The model can be fitted using standard linear regression techniques. The estimates of c and s are used to estimate A and P. The values of c and s can be seen in Fig. 1.12, as c (the cosine parameter) controls the east–west point on the circle, and s (the sine parameter) controls the north–south point. So the parameters have no meaningful interpretation on their linear scale, but do when transformed into the phase and amplitude, which are the angle and radius on the circle, respectively [55].

To create ω_t we first transform time into a fraction, f_t, between zero and one. If we are interested in an annual seasonal cycle based on daily data then we

A.G. Barnett and A.J. Dobson, *Analysing Seasonal Health Data*, Statistics for Biology and Health, DOI 10.1007/978-3-642-10748-1_3,
© Springer-Verlag Berlin Heidelberg 2010

would use [55]

$$f_t = \frac{\text{day}_t - 1}{365 \text{ or } 366}, \tag{3.4}$$

where $1 \leq \text{day}_t \leq 366$ is the day in the year of the tth observation, which is divided by 366 in leap years and 365 in other years.

If we are interested in an annual seasonal cycle based on monthly data then we would use

$$f_t = \frac{\text{month}_t - 1}{12}, \tag{3.5}$$

where $1 \leq \text{month}_t \leq 12$ is the month of the tth observation. More generally, if we are interested in an seasonal pattern with k cycles per year based on monthly data then we would use

$$f_t = k\left(\frac{\text{month}_t - 1}{12}\right). \tag{3.6}$$

To transform this fraction into radians we then multiply by 2π, $\omega_t = 2\pi f_t$. We then add the columns $\cos(\omega_t)$ and $\sin(\omega_t)$ to the design matrix, which can then be used as part of a generalized linear model (Sect. 1.4.5). This means that cosinor models can be applied to Normal, Poisson or Binomial data.

To test the null hypothesis that the amplitude, A, is equal to zero, we use the following criterion: reject the null hypothesis if the p-value for the estimate of c or s is less than $\alpha/2$, where α is the required level of statistical significance (most often $\alpha = 0.05$). We divide this significance level by 2 to maintain the type I error probability at α.

The estimated phase, P, is on a scale of radians, and for interpretative purposes it is preferable to transform this to a time scale using $P^* = 365(P/2\pi) + 1$, for daily data, and $P^* = 12(P/2\pi) + 1$, for monthly data.

The cosinor is a parsimonious model, as it only requires two parameters to model a seasonal pattern. The previous models based on using month as a categorical independent variable, Sect. 2.3, required 11 parameters. However, the cosinor model assumes a sinusoidal seasonal pattern which is symmetric (Sect. 1.3.1) along the x- and y-axes. This assumption may be too restrictive for some seasonal data.

3.1 Examples

3.1.1 Cardiovascular Disease Deaths

Figure 3.1 shows the result of fitting a cosinor model to the daily CVD death data. The estimated amplitude is 7.1 deaths per day with a phase of 23 January, and this amplitude is statistically significant. The peak estimates of deaths on 23 January is 52.3 deaths per day, and the low (six months later) on 25 July is 38.0.

The sinusoid from the cosinor has not quite captured the peak in deaths in mid-winter, and a histogram of the residuals shows a strong skew. This is because the

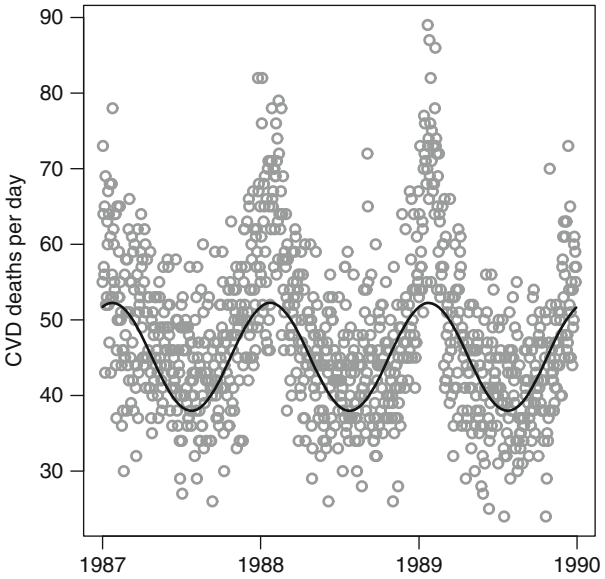


Fig. 3.1 Daily CVD death data (*grey dots*) and estimated annual seasonal pattern using a cosinor model. Results shown for the first three years only, 1987–1989

sinusoid is symmetric about the y-axis. Also, the residuals show some remaining seasonal pattern, so although the cosinor model is statistically significant, it has not fully described the seasonal pattern. Also, the figure highlights how the cosinor model is based on the assumption that the seasonal pattern is the same in every year, whereas the data indicate otherwise (Fig. 1.1).

The R code to fit the cosinor model as part of a GLM is

```
> cosinor(cvd~1,date=date,data=CVDdaily)
```

which produces the following output

```
Number of observations = 5114
GLM coefficients
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 45.110481  0.1144621 394.10845 0.000000e+00
cosw         6.617196  0.1618739  40.87872 0.000000e+00
sinw         2.684572  0.1618739  16.58435 3.305908e-60
Cosinor test
Amplitude = 7.141024
Phase: Month = January , day = 23 Low
point: Month = July , day = 25
```

In this example both the c and s parameters are strongly statistically significant (p-value < 0.001). The amplitude is 7.1 deaths per day, with a phase of 23 January. The average number of deaths is 45.1 per day.

As the model is based on a GLM we can add other independent variables. As an example, after adding mean daily temperature, the estimated seasonal amplitude is now 7.3 deaths per day, slightly larger than a model without temperature. The phase remains as 23 January.

The daily number of CVD deaths is a count variable, and although its mean is large enough for the Normal approximation to hold, it may be better to assume an over-dispersed Poisson distribution, especially as the distribution of deaths has a positive skew. The output from assuming an over-dispersed Poisson distribution is shown below.

```
Number of observations = 5114
GLM coefficients
              Estimate  Std. Error    t value     Pr(>|t|)
(Intercept) 3.80282022 0.002505911 1517.54028 0.000000e+00
cosw        0.14715057 0.003536775   41.60586 0.000000e+00
sinw        0.05969845 0.003528804   16.91747 1.636301e-62
Cosinor test
Amplitude = 7.714917 (absolute scale)
Phase: Month = January , day = 23
Low point: Month = July , day = 25
```

The phase has remained as 23 January. The amplitude is $\sqrt{0.147151^2 + 0.059699^2} = 0.158799$ on the log-scale, which gives a rate ratio of $\exp(0.158799) = 1.172$. So the rate of deaths are 1.172 times the average on 23 January. As we used the log-link, we use: $\exp(3.80282022 + 0.158799) - \exp(3.80282022) = 7.714917$, to give the amplitude in terms of the absolute number of deaths. This is reasonably similar to the amplitude of 7.1 deaths per day assuming a Normal dependent variable.

Figure 3.2 shows the result of fitting a cosinor model to the monthly CVD death data. The estimated amplitude is 215 deaths per month with a phase of 1.3 months (i.e., early January). This is based on adjusting the monthly counts to account for the unequal number of days in the month (Sect. 2.2.1).

Figure 3.3 shows the autocorrelation function and histogram of the residuals from the cosinor model shown in Fig. 3.2. The large number of strong autocorrelations indicate that a seasonal pattern still remains, and the histogram shows a positive skew.

For the monthly data the AIC for the cosinor model is 3,025.1 with three parameters. In Sect. 2.3 we fitted a model using month as a fixed factors, and this has an AIC of 2,611.8 with 12 parameters. So there is a big improvement of fit using the monthly model (AIC difference of 413.3 for nine extra parameters). These results, together with the residual checks (Fig. 3.3), indicate that the symmetric sinusoidal shape assumed by the cosinor model is too simplistic.

3.1.2 Exercise

For the exercise data (Sect. 1.1.4) we use time based on visit date (which is irregularly spaced). Body mass index (which we assume to have an Normal distribution)

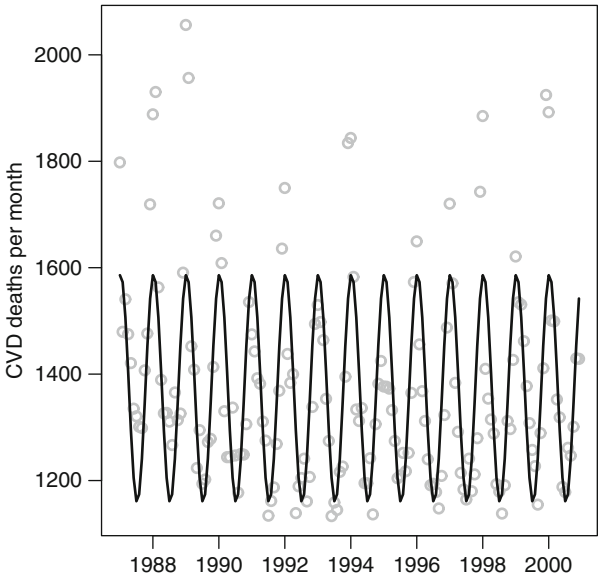


Fig. 3.2 Monthly CVD death data (*grey dots*) and estimated annual seasonal pattern using a cosinor model

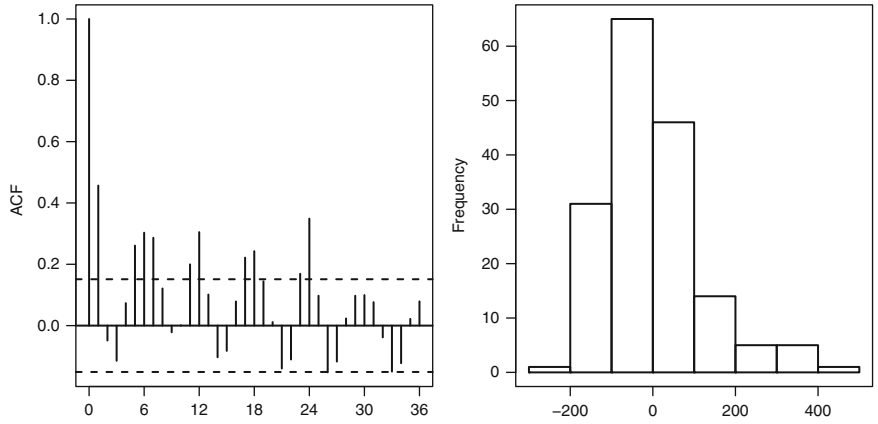


Fig. 3.3 Autocorrelation function and histogram of the residuals from the cosinor model shown in Fig. 3.2

has an estimated amplitude of 0.25 kg/m^2 with a phase of 11 August, but this amplitude is not statistically significant. Minutes of walking (which we assume to have a Poisson distribution) has an estimated amplitude of 9 minutes with a phase of 24 June, and this amplitude is statistically significant.

3.1.3 Stillbirths

For the stillbirth data (Sect. 1.1.5) we use time based on the child's date of birth and a binary dependent variable of stillborn (assuming a Binomial distribution and using a logit link). Using the cosinor model gives an amplitude of 0.0012 (on a probability scale) and a phase of 27 January, and this result is statistically significant. The estimated peak estimate in the probability of stillbirth is on 27 January is 0.0070, and the low on 29 July (six months later) is 0.0047.

3.2 Tests of Seasonality

In this section we describe two tests that are similar to the cosinor test as they view the data as circular. Both use the trigonometric cosine and sine functions to describe the peak point on the circle using an (x, y) co-ordinate (Sect. 1.3.1). This peak point is the seasonal phase. Both tests are designed to detect a seasonal pattern that is:

- Annual in frequency
- Sinusoidal in shape
- Unimodal in shape (i.e., occurs once every year)
- The same in each year (known as *stationary*)

Bimodal seasonal patterns can be detected by doubling the number of cycles, Eqn. (3.6). The tests can also be adapted to detect seasonal patterns at other frequencies (e.g., weekly).

Test of Seasonality Based on Mean Distance

This test uses individual dates (e.g., birth dates), or results grouped into months. The key statistics are the average distance along the x- and y-axes, given by

$$\overline{c} = \frac{1}{n}\sum_{t=1}^{n}\cos(\omega_t), \qquad t = 1, \ldots, n,$$

$$\overline{s} = \frac{1}{n}\sum_{t=1}^{n}\sin(\omega_t), \qquad t = 1, \ldots, n,$$

where n is the total number of dates, $\omega_t = 2\pi f_t$ and f_t is the fraction of time as described in (3.4) for individual dates (e.g., birth dates) or (3.5) for monthly data.

Figure 3.4 shows circular plots (Sect. 2.2.3) for three possibilities based on monthly data (assuming we have adjusted for the uneven number of days in the months). The top-left panel shows the axes for $\overline{c}$ and $\overline{s}$, where $(\overline{c}, \overline{s}) = (0, 0)$ is the centre of the circle. In the top-right panel the results in every month are equal, meaning the plot is perfectly circular, so the mean direction in a both a East–West ($\overline{c}$) and North–South ($\overline{s}$) direction is zero. In the bottom-left panel there is an increase in

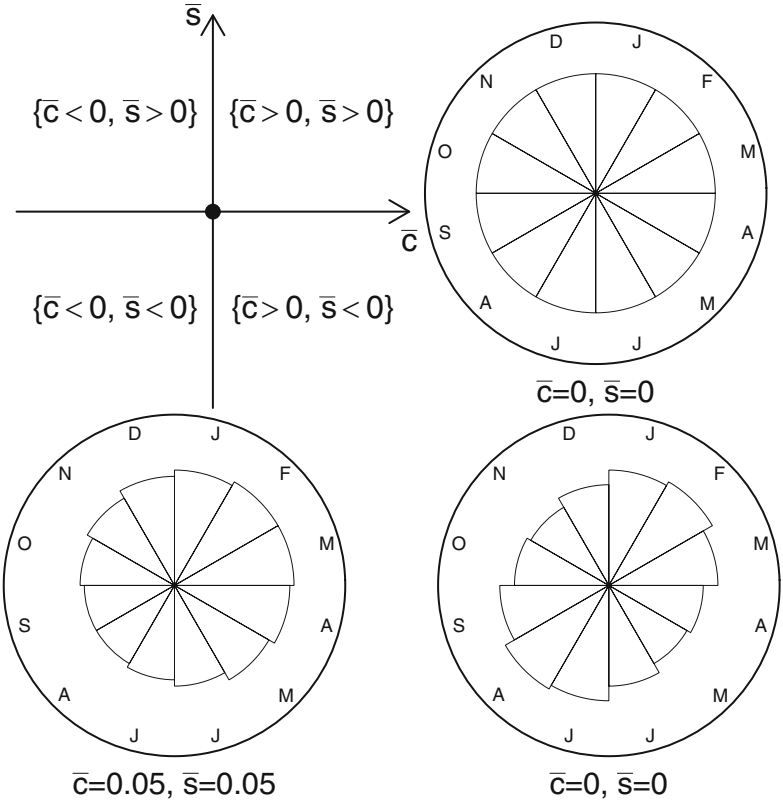


Fig. 3.4 Circular plots of monthly data and values of $\overline{c}$ and $\overline{s}$ under the null hypothesis of no seasonality (*top-right*), the alternative hypothesis of a unimodal sinusoidal seasonal pattern (*bottom-left*), and a bimodal sinusoidal seasonal pattern (*bottom-right*). The *top-left panel* shows the $\overline{c}$ and $\overline{s}$ axes with a centre $(0, 0)$ at the middle of the circle

February and March (around the 45° angle of the circle). In this case the values of $\overline{c}$ and $\overline{s}$ are both positive. In the bottom-right panel there is an increase in February which is mirrored by an increase six months later in August (bimodal seasonal pattern). In this case the mean directions are zero as the positive values of $\cos(\omega_t)$ and $\sin(\omega_t)$ are cancelled by the negative values. So in this example the values of $\overline{c}$ and $\overline{s}$ are insensitive to the bimodal seasonal pattern.

The test statistic is

$$Z = n\overline{A}^2,$$

where

$$\overline{A} = \sqrt{\overline{c}^2 + \overline{s}^2},$$

is the mean amplitude on a proportional scale ($0 \leq \overline{A} \leq 1$). When $\overline{A} = 0$ there is no seasonal pattern and the dates are spread equally throughout the year. When $\overline{A} = 1$ all dates are the same, which would be an extreme version of a spiked seasonal pattern (Fig. 2.15).

Heuristically the values of $\{\overline{c}, \overline{s}\}$ can be thought of as determining the direction of a compass needle. If there is no seasonal pattern then the needle could point in all directions equally, and a circular plot of the data should appear perfectly circular. If there is a seasonal pattern then the needle will point to the phase, and the strength of the attraction is proportional to $\overline{A}$.

The p-value for the null hypothesis that $\overline{A} = 0$ is $\exp(-Z)$ for a sample size greater than 50 [35].

Using the stillbirth data as an example gives $\overline{A} = 0.08$, $Z = 2.54$ and a p-value of 0.079 based on 352 stillbirths. This gives some indication that the number of stillbirths may have a sinusoidal seasonal pattern.

Adjusting for the At-Risk Population

The test of seasonality described in the previous section does not adjust for the population at risk. For the stillbirth example, we tested the seasonal distribution of the number of stillbirths without adjusting for the number of births. If birth numbers are seasonal then this may create a false pattern of seasonality in stillbirths, or mask a seasonal pattern (depending on the coincidence of the phases in the number of births and the risk of stillbirth).

Walter and Elwood created a test to adjust for the population at risk by grouping the data into months [86]. We define n_i as the number of trials (e.g., the number of births), and r_i as the number of "successes" (e.g., the number of stillbirths) in month i. The mean directions are

$$\overline{c} = \sum_{t=1}^{12} \sqrt{r_i}\cos(\omega_t) \Bigg/ \sum_{t=1}^{12} \sqrt{r_i}, \qquad i = 1, \ldots, 12,$$
$$\overline{s} = \sum_{t=1}^{12} \sqrt{r_i}\sin(\omega_t) \Bigg/ \sum_{t=1}^{12} \sqrt{r_i}, \qquad i = 1, \ldots, 12,$$

where $\omega_t = 2\pi f_t$ and f_t is based on (3.5). The mean directions for the population at-risk data are

$$\tilde{c} = \sum_{t=1}^{12} \sqrt{n_i}\cos(\omega_t) \Bigg/ \sum_{t=1}^{12} \sqrt{n_i},$$
$$\tilde{s} = \sum_{t=1}^{12} \sqrt{n_i}\sin(\omega_t) \Bigg/ \sum_{t=1}^{12} \sqrt{n_i}\ .$$

The test statistic is

$$X^2 = \left(\frac{\overline{c} - \tilde{c}}{\hat{\sigma}_c}\right)^2 + \left(\frac{\overline{s} - \tilde{s}}{\hat{\sigma}_s}\right)^2,$$

where the variances are

$$\hat{\sigma}_c^2 = \frac{1}{4}\sum_{t=1}^{12}\sqrt{r_i}\cos^2(\omega_t)\Bigg/\left[\sum_{t=1}^{12}\sqrt{Nm_i/M}\right]^2,$$
$$\hat{\sigma}_s^2 = \frac{1}{4}\sum_{t=1}^{12}\sqrt{r_i}\sin^2(\omega_t)\Bigg/\left[\sum_{t=1}^{12}\sqrt{Nm_i/M}\right]^2,$$

where $N = \sum r_i$ and $M = \sum n_i$.

Under the null hypothesis of no seasonal pattern, X^2 has a chi-squared distribution with two degrees of freedom. This is based on the assumption that $\overline{c}$ and $\overline{s}$ are Normally distributed.

For the stillbirth data, the test statistic is $X^2 = 6.90$ with p-value 0.032, indicating a significant seasonal pattern in stillbirth after adjusting for the seasonal pattern in the number of births.

3.2.1 Chi-Squared Test of Seasonality

The tests of seasonality described so far are designed to find a sinusoidal seasonal pattern, and are not designed to detect other seasonal patterns (such as sawtooth or spiked, Fig. 2.15). In this section we describe an alternative test that examines any departure from an evenly distributed pattern over 12 months.

For Binomial data we test the null hypothesis that the probability of success is the same in each month. The overall probability of success is estimated by $p = \sum_{i=1}^{12} r_i / \sum_{i=1}^{12} n_i$ where r_i is the observed number of successes in month i, and n_i the number of trials. The test statistic is then

$$X^2 = \sum_{i=1}^{12}\frac{(r_i - n_i p)^2}{n_i p\,(1-p)}. \tag{3.7}$$

Under the null hypothesis this has an approximate chi-squared distributed with 11 degrees of freedom. The statistic may not work well if the expected frequencies ($n_i p$) are less than 1 [26, Chap. 7].

For the stillbirth data, the observed statistic is $X^2 = 9.39$. A chi-squared distribution with 11 degrees of freedom has a critical value of 19.68 for a 5% significance level, and the p-value is 0.59. This result indicates that the risk in stillbirth does not change over the year.

3.2.1.1 Simulation Study Comparing Tests of Seasonality

To examine the differences in the tests of seasonality we created random Binomial data for 12 months using

$$r_i \sim \text{Bin}(n, p_i),$$
$$p_i = A\{(\cos[(i-1)2\pi/12] + 1)/2\}, \qquad i = 1, \ldots, 12.$$

This creates monthly data (for 1 year) with a sinusoidal seasonal pattern whose size is proportional to the amplitude (A). The study was based on monthly Binomial data using a fixed number of trials per month, n, with a total sample size of $N = 12n$.

To examine the ability of the tests to find a non-sinusoidal seasonal pattern we also generated data using the following monthly probabilities

$$p_i = \begin{cases} 0.5, & i = 1, \ldots, 11, \\ 0.5 + A, & i = 12, \end{cases}$$

which creates monthly data with a spiked seasonal pattern, the size of which is dependent on A. To examine the type I error probability of the tests we also generated random (non-seasonal) data using a fixed monthly probability of $p = 0.5$.

Table 3.1 shows the results of a simulation study comparing the power of Walter and Elwood's test, the cosinor test and the chi-squared test. The cosinor test had the best power to detect sinusoidal seasonal patterns, and had a power greater than or equal to the other two tests in every simulation.

The chi-squared test had the best power to detect a spiked seasonal pattern. For spiked patterns, the power of the chi-squared test was almost always greater than the cosinor test, and the power of the cosinor test was always greater than Walter and Elwood's test. The greater power of the chi-squared test for spiked seasonal data is because the alternative hypothesis of this test is any non-uniform pattern, whereas the other tests are specifically designed to detect a sinusoidal pattern.

Table 3.1 Power (%) of detecting seasonality for three tests using a simulation study for varying sample sizes and amplitudes. Results shown for sinusoidal and spiked seasonal patterns, and for random data. Based on 1,000 simulations for each sample size and amplitude combination

Trials per	Random	Sinusoidal amplitude (A)					Spike (A)				
month (n)	data	0.01	0.05	0.1	0.2	0.3	0.01	0.05	0.1	0.2	0.3
					Walter and Elwood's test						
50	0.6	0.5	1.4	7.0	58.5	97.7	0.0	0.0	0.8	2.3	3.8
100	0.3	0.3	3.6	23.5	95.4	100.0	0.1	0.8	1.0	4.3	14.2
200	0.4	0.5	6.1	55.4	99.9	100.0	0.3	0.6	1.5	12.6	42.5
					Cosinor test						
50	4.3	5.1	10.8	29.8	90.6	99.7	4.5	3.0	7.9	16.6	28.6
100	5.1	4.2	18.1	61.6	99.6	100.0	5.2	8.6	9.5	27.8	55.2
200	5.4	6.7	33.6	89.2	100.0	100.0	4.5	8.9	13.5	49.8	84.6
					Chi-squared test						
50	5.9	5.9	8.8	16.2	63.1	97.6	6.3	5.1	10.3	39.8	85.4
100	4.8	6.1	10.6	35.1	94.8	100.0	3.8	8.3	21.0	76.5	99.7
200	4.6	5.9	16.4	62.3	99.9	100.0	5.4	9.6	38.1	98.3	100.0

None of the tests had a good power to detect a very small seasonal pattern ($A = 0.01$). All three tests had close to perfect power for detecting a large sinusoidal seasonal pattern ($A = 0.3$).

All tests were based on a 5% rejection level. The cosinor and chi-squared tests gave type I errors closer to the correct 5% level, as shown by the results for random data ($A = 0$).

Based on these results we recommend the use of the cosinor test in place of Walter and Elwood's test. As the cosinor test can be applied as part of a GLM it is also a flexible test. The chi-squared test is preferable to the cosinor test when the shape of the seasonal pattern is not known, although the chi-squared test has less power to detect a sinusoidal seasonal pattern than the cosinor test.

3.2.2 Sample Size Using the Cosinor Test

We can calculate the sample size needed to detect a particular amplitude by exploiting the fact that the cosinor test corresponds to a linear regression model, and so use existing formulae for linear regression sample sizes [29].

The cosinor test involves two terms, the sine and cosine term. We can reduce this to one by assuming that the phase is zero, so that the sine term is zero. For sample size purposes only the size of the amplitude is important, and the time of the phase is immaterial. The regression equation is then

$$\begin{aligned} Y_t &= A\cos(\omega_t) + \varepsilon_t, \qquad t = 1, \ldots, n, \\ \varepsilon_t &\sim \mathrm{N}(0, \sigma_\varepsilon^2), \end{aligned}$$

where ε_t are uncorrelated noise with variance σ_ε^2.

To calculate the required sample size we need to specify the following five variables:

- The required power of the study
- The type I error probability, which we divide by 2 as the cosinor test involves two separate tests of seasonality
- The amplitude (A)
- The standard deviation of the errors σ_ε
- The standard deviation of the independent variable $\cos(\omega_t)$

As $n \to \infty$ the standard deviation $\mathrm{sd}[\cos(\omega_t)] \to \sqrt{0.5} \approx 0.707$. This assumes that the observations are reasonably spread over throughout the year, which should happen with time series data of a reasonable length, but may not always be true of survey data if responses are grouped at particular dates. The greater the variability in the independent variable (in this case time) the smaller the sample size needed for a linear regression model [29]. So having data equally spread over the dates of the year is optimal. If the responses are clustered around particular dates then $\mathrm{sd}[\cos(\omega_t)]$ will be less than 0.707, and hence the power will be reduced.

Table 3.2 Sample size needed to detect a sinusoidal seasonal pattern with amplitude A and error standard deviation σ_e using the cosinor test. For a power of 80% and two-sided significance level of 5%

	Amplitude (A)			
σ_ε	0.1	0.2	0.5	1
1	1,904	478	79	22
2	7,609	1,904	307	79
3	17,117	4,281	687	174

Table 3.2 shows the required samples sizes to detect a range of amplitudes and error standard deviations. These samples sizes can be applied to time series based on days or months. For example, to detect an amplitude of 0.2 with an error standard deviation of $\sigma_\varepsilon = 2$ would require $1,904/365.25 = 5.2$ years of daily data, or $1,904/12 = 158.7$ years of monthly data.

We can calculate the sample size needed for any combination of A and σ_ε using software such as PS [30].

3.3 Sawtooth Season

In contrast to the smooth sinusoidal pattern modelled by a cosinor, we can create a model with a sawtooth seasonal pattern (Fig. 2.1). Sawtooth seasonal patterns may be useful for modelling abrupt changes in exposure, or for modelling non-sinusoidal seasonal patterns.

For monthly data an example sawtooth model is

$$\begin{aligned} s_t &= \zeta_1(z_t - 5.5)/5.5, \qquad t = 1, \ldots, n, \\ z_t &= m_t - \zeta_2 + 12I(\zeta_2 > m_t), \end{aligned} \tag{3.8}$$

where m_t is the month at time t (Sect. 2.3.1), ζ_2 determines the month at which the seasonal pattern changes ($\zeta_2 \in \{1, 2, \ldots, 12\}$) and ζ_1 determines the height of the sawtooth pattern. $I(x)$ is an indicator function such that $I(x) = 1$ if $x > 0$, and $I(x) = 0$ if $x \leq 0$. z_t is the unstandardised sawtooth pattern, we subtract its mean and divide by its maximum (the mean and maximum both being 5.5), so that ζ_1 corresponds to the greatest absolute seasonal change. This also means that $\sum_{t=1}^{n} z_t = 0$.

Figure 3.5 shows two examples of sawtooth seasonal patterns created using model (3.8). The top panel shows a sawtooth pattern that increases sharply between May and June and then steadily declines. The bottom panel shows a pattern that drops sharply from October to November and then steadily rises. The maximum change in the seasonal pattern is determined by ζ_1 which is 1 in the top panel and 2 in the bottom panel. The sign of ζ_1 determines the direction of sawtooth pattern.

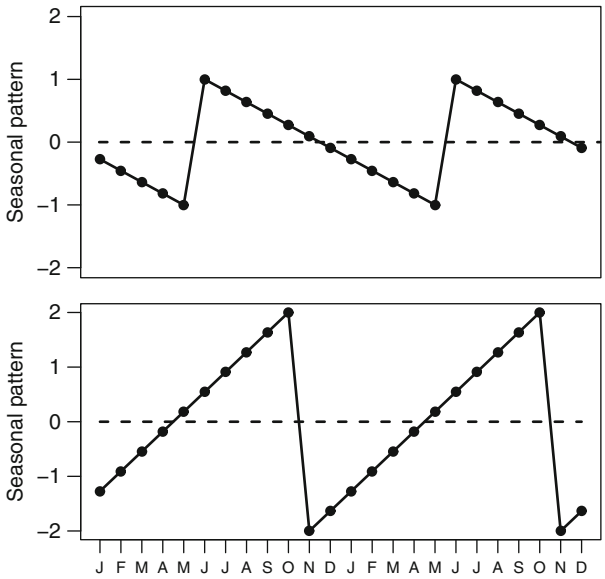


Fig. 3.5 Examples of sawtooth seasonal patterns created by model (3.8) plotted over two years: *top panel* $\zeta_1 = -1$, $\zeta_2 = 6$; *bottom panel* $\zeta_1 = 2$, $\zeta_2 = 11$; *dotted horizontal line* at zero

A negative sign for ζ_1 creates a seasonal pattern with a sharp increase and steady decline, whereas a positive sign creates a pattern with a steady increase and sharp decline. When ζ_1 is negative ζ_2 is the month with the highest level, whereas when ζ_1 is positive ζ_2 is the month with the lowest level.

3.3.1 Examples

3.3.1.1 Footballers

Football players' birthdays may have a sawtooth seasonal pattern (Fig. 1.6) as we suspect their distribution may be based on a social seasonal exposure (Sect. 2.1.1). A social seasonal exposure may well vary by country, and we can investigate this possible difference as we have data for Australia and the UK.

The total number of births in both Australia and the UK are seasonal (see Fig. 2.6 for the UK), and we need to adjust for this seasonal pattern when examining the seasonal pattern in players' births (as in Sect. 3.2). We first calculate the expected number of players born in each month

$$e_m = \frac{On_m}{\sum_{m=1}^{12} n_m}, \qquad m = 1, \ldots, 12,$$

where n_m is the total number of births in month m and O is the total number of players. We used data on the total number of births that covered the same time periods as the players' birthdays.

The model with this offset and a sawtooth seasonal pattern is then

$$\begin{aligned} o_m &\sim \text{Po}(\mu_m), \qquad m = 1, \ldots, 12, \\ \log(\mu_m) &= \log(e_m) + s_m, \\ s_m &= \zeta_1(z_m - 5.5)/5.5, \\ z_m &= m - \zeta_2 + 12I(\zeta_2 > m), \end{aligned}$$

where o_m is the number of footballers born in month m. The value $\exp(s_m)$ is then the relative risk of a footballer being born in month m assuming a sawtooth seasonal pattern.

We estimate model (3.8) using a Bayesian paradigm. We fit a separate model for all twelve values of $\zeta_2 = 1, 2, \ldots, 12$, and then select that model which has the lowest DIC (Sect. 1.6.2). We use a vague prior for $\zeta_1 \sim \text{N}(0, 10^6)$.

Figure 3.6 shows the DIC and estimated ζ_1 parameter over the 12 months for the AFL and EPL players. These estimates were based on a burn-in of 10,000 MCMC iterations, followed by a sample of 10,000. For the AFL players the lowest DIC was 75.6 in January, which was 12.6 lower than the next best DIC in August. In January the mean estimate of ζ_1 was negative, -0.29 with a 95% CI of -0.41 to -0.16. For the EPL players the lowest DIC was 73.7 in March, which was 8.1 lower than the next best DIC in September. In March the mean estimate of ζ_1 was positive, 0.44 with a 95% CI of 0.25 to 0.62.

We show the estimated sawtooth seasonal patterns in Fig. 3.7. For the AFL players we used a ζ_2 of January, and for the EPL players we used both March and

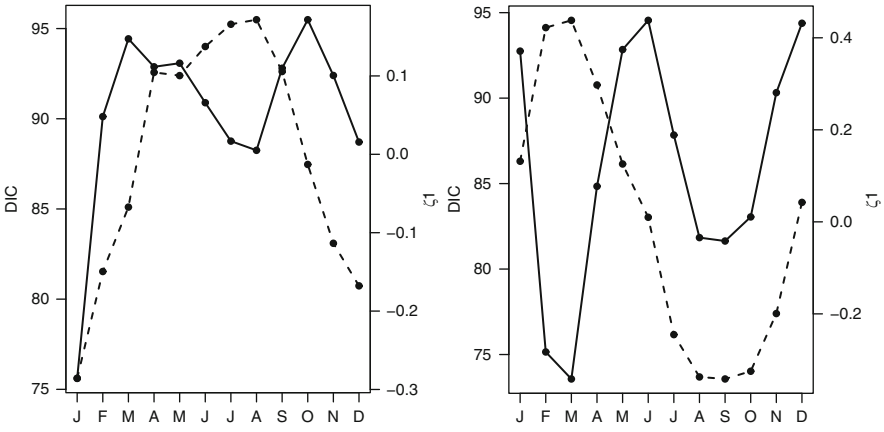


Fig. 3.6 Deviance information criterion for the 12 monthly values of ζ_2 (*solid line*) and mean estimate of ζ_1 (*dotted line*) for the AFL (*left*) and EPL (*right*) footballers data

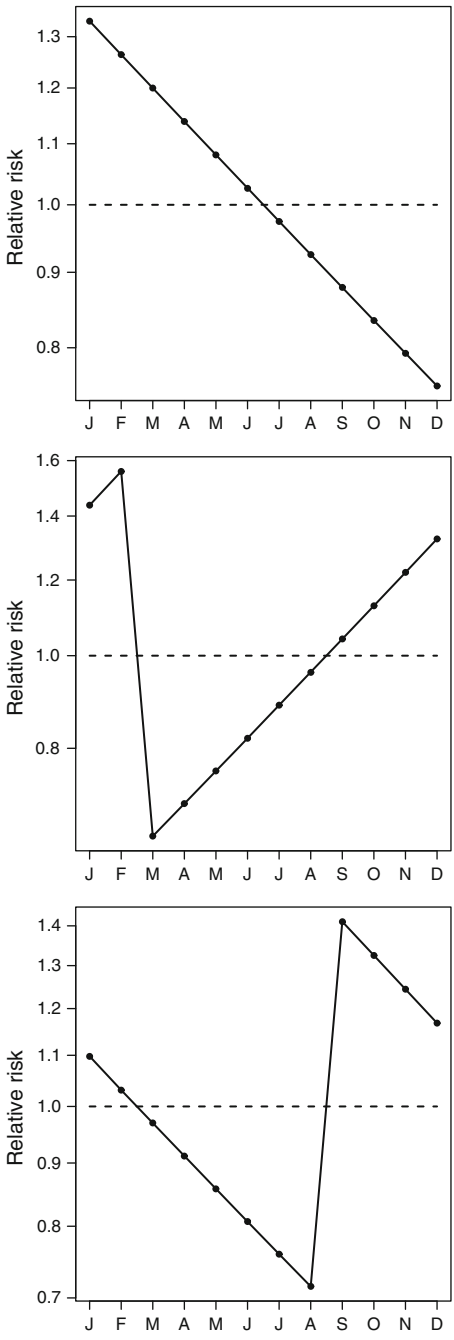


Fig. 3.7 Seasonal relative risks of birth for the footballers data using a sawtooth seasonal model. The *top panel* is for the AFL players using $\zeta_2 = 1$ (January), the *middle panel* is for the EPL players using $\zeta_2 = 3$ (March), and the *bottom panel* is for the EPL players using $\zeta_2 = 9$ (September). The y-axes are on a log scale

September. There was a peak in AFL players' birthdays in January with a steady decline to December. School enrolment in Australia is based on year of birth. So boys born at the start of the year will have almost of full year of extra growth compared with those born at the end of the year. This extra growth would give these boys an advantage in physical games. A similar pattern has been found in the birthdays of Canadian hockey players [43] and UK footballers [78].

In the UK school enrolment starts with children born in September. However, the optimal value for ζ_2 was March, which has no apparent association with the school year (Fig. 3.7). However, the second best value for ζ_2 according to the DIC was September (Fig. 3.6). The estimated sawtooth seasonal pattern based on September has a shape that could be explained by school year and the physical advantages of being the oldest in the class.

3.3.1.2 Cardiovascular Disease

It is possible to combine a sawtooth seasonal pattern with a sinusoidal pattern. As an example we fit both seasonal patterns to the monthly CVD death data. We modelled the number of deaths using a Poisson distribution and used an offset to account for the uneven number of days in the months (Sect. 2.2.1). We included a quadratic trend and modelled the over-dispersion. We used a burn-in of 10,000 MCMC iterations, followed by a sample of 10,000.

The WinBUGS code is shown below.

```
model{
 for (i in 1:N){
  cases[i]~dpois(mu[i]);
  log(mu[i])<-trend[i]+season[i]+log(offset[i])+disp[i];
  season[i]<-cosinor[i]+sawtooth[i];
  cosinor[i]<-(b.cos*cos(omega[i]))+(b.sin*sin(omega[i]));
  sawtooth[i]<-zeta_1*(z[i]-5.5)/5.5;
  z[i]<-(month[i]-zeta_2)+(12*step(zeta_2-month[i]-0.1))
  trend[i]<-beta[1]+(beta[2]*time[i])+(beta[3]*time2[i]);
  disp[i]~dnorm(0,tau.disp);
 }
 for (k in 1:3){
  beta[k]~dnorm(0,0.0001);
 }
 zeta_1~dnorm(0,0.0001);
 b.cos~dnorm(0,0.0001);
 b.sin~dnorm(0,0.0001);
 tau.disp~dgamma(0.0001,0.0001);
}
```

The parameters `b.cos` and `b.sin` are c and s respectively in the cosinor equation (3.1). The first two lines of the data in WinBUGS format are shown below.

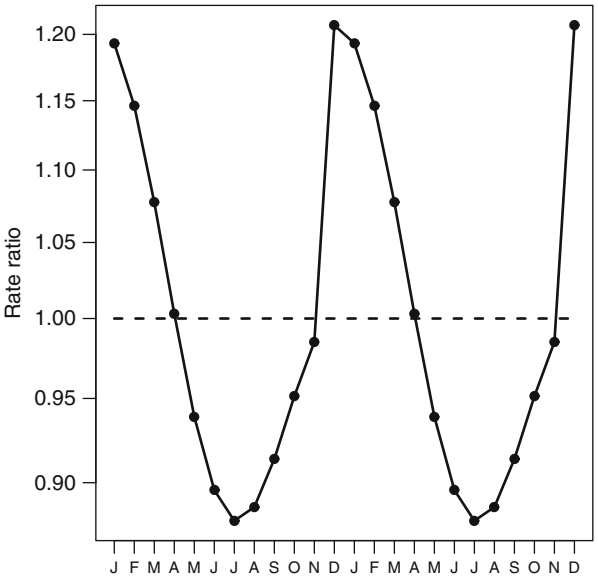


Fig. 3.8 Combined seasonal pattern of a sawtooth and cosinor using the CVD death data. Results are plotted for two years. *Dotted horizontal line* at a rate ratio of 1. The y-axis is on a log scale

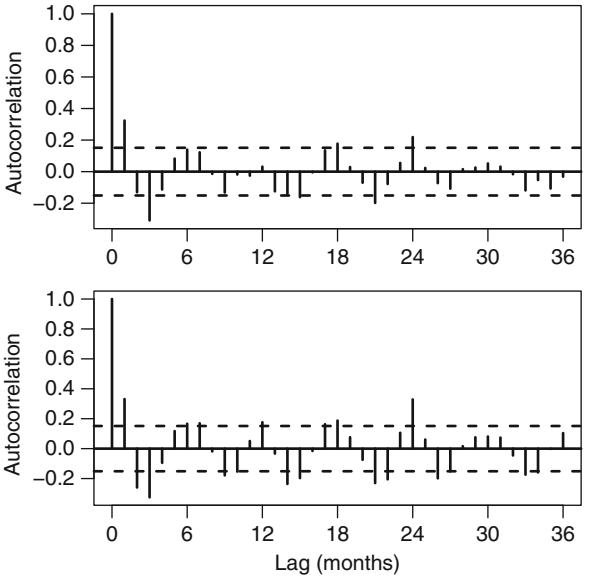


Fig. 3.9 Autocorrelation function of the residuals from the combined cosinor and sawtooth model (*top*) and the cosinor only model (*bottom*) for the CVD death data for the lags $k = 0, \ldots, 36$ months

```
cases[] offset[] month[] time[]     time2[]   omega[]
1831    1.01848  1       -0.9880952 0.9763322 0
1361    0.91991  2       -0.9761905 0.9529478 0.5235988
```

The variables `time` and `time2` are used to model the linear and quadratic trends, respectively. We used a centred version of time (Sect. 1.4.2) as centring often improves the convergence of the MCMC chain.

We used vague prior distributions for all the unknown parameters,

$$\zeta_1 \sim \mathrm{N}(0, 10^4), \quad c \sim \mathrm{N}(0, 10^4), \quad s \sim \mathrm{N}(0, 10^4).$$

The value for ζ_2 with the smallest DIC was December, which had an associated mean value for ζ_1 of -0.10 (95% CI: -0.27, -0.01). Figure 3.8 shows the estimated combined pattern for two years. The most noticeable difference between this seasonal pattern and a standard sinusoid is the sharp increase in risk between November and December. The sawtooth pattern combined with the sinusoidal pattern has made a non-symmetric seasonal pattern, as the decline in deaths from December to July takes 7 months, whereas the rise in deaths from July to December takes only 5 months.

Figure 3.9 shows the ACF (Sect. 1.2.1) of the residuals from this model and, for comparison, the residuals from a cosinor model (including a quadratic trend). The autocorrelation for the combined model is noticeably lower at 12 months, suggesting that it has better captured the seasonal pattern. However, some statistically significant autocorrelations remain, particularly at short lags.

Chapter 4
Decomposing Time Series

A useful way to view time series data is as a combination of trend, season and noise. For data that are equally spaced over time ($t = 1, \ldots, n$) an equation that splits the series into these three parts is

$$Y_t = \mu_t + s_t + \varepsilon_t, \qquad t = 1, \ldots, n, \tag{4.1}$$

where μ_t is the trend, s_t is the seasonal pattern and ε_t is the random noise (also known as the error or residuals). We assume the residuals are uncorrelated, with a zero mean and constant variance, so

$$\varepsilon_t \sim \mathrm{N}(0, \sigma_\varepsilon^2), \qquad \mathrm{cov}(\varepsilon_t, \varepsilon_k) = 0,\ t \neq k.$$

We can include multiple seasonal terms using the equation

$$Y_t = \mu_t + \sum_{j=1}^{k} s_{t,j} + \varepsilon_t, \qquad t = 1, \ldots, n. \tag{4.2}$$

For example, we could use this equation to model both an annual and biannual seasonal pattern.

As we show later (Sect. 4.4) we can also further split the season into an average seasonal pattern, and a year-to-year change from the average seasonal pattern using

$$s_t = \delta_{m(t)} + \nu_t, \qquad t = 1, \ldots, 12,$$

where $m(t)$ is the month at time t (Sect. 2.3.1), $\boldsymbol{\delta}$ represents the average seasonal pattern, and $\boldsymbol{\nu}$ the changes in this pattern. This is useful for examining changes over time in the seasonal pattern.

The technique of *decomposing* aims to split a time series into its constituent parts of trend, season and remaining noise. Figure 4.1 graphically shows the aim of decomposition. In this example of artificial data the trend is a linear increase and the season is sinusoidal.

A.G. Barnett and A.J. Dobson, *Analysing Seasonal Health Data*, Statistics for Biology and Health, DOI 10.1007/978-3-642-10748-1_4,
© Springer-Verlag Berlin Heidelberg 2010

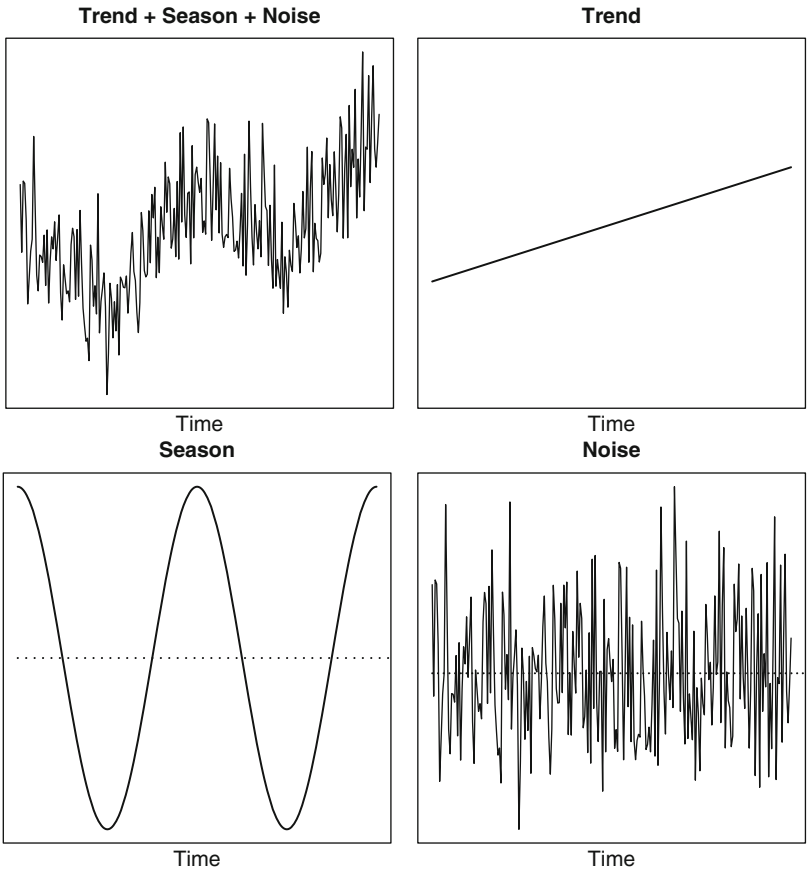


Fig. 4.1 Decomposing a time series (*top-left*) into the trend (*top-right*), season (*bottom-left*) and residuals (*bottom-right*). The scales on the y-axes are different. *Dotted horizontal lines* at zero in the season and residuals

Definition of a Trend

We assume that the trend in (4.1) represents the long-term change in disease. It is also known as the *secular* trend. This trend represents the gradual improvement or worsening in disease frequency, which may be linear or non-linear. These changes occur for many reasons. Examples include gradual improvements in health care, diet and working conditions.

The reasons behind the trend will depend on the time span of the study. Studies that cover decades are likely to include trends due to improvements in socio-economic conditions, although this depends on the disease being studied and the location. Trends may also occur because of a delay in case registration. For example, the plot of schizophrenia rates (Fig. 1.28) shows a drop in cases from 1960 to 1970. This drop is not because of a reduction in schizophrenia, but because at the

time the data were collected some people born in the 1960s have not yet lived long enough to be diagnosed with schizophrenia.

We assume that there are no sudden changes in the long-term trend. Such sudden changes may occur because of natural disasters, interventions or changes in the way that a disease is classified or recorded. These changes are also not part of the seasonal pattern, and to model them we would need to add another independent variable to (4.1) using

$$Y_t = \mu_t + s_t + \eta x_t + \varepsilon_t, \qquad t = 1, \ldots, n,$$

where η describes the sudden change in the mean at time d using the indicator variable

$$x_t = \begin{cases} 0, & t < d, \\ 1, & t \geq d. \end{cases}$$

This might occur if a disease is reclassified at some point in time, d, which we assume is known. The indicator variable can also be designed to model a temporary change in the mean

$$x_t = \begin{cases} 0, & t < d_1, \, t > d_2, \\ 1, & d_1 \leq t \leq d_2, \end{cases}$$

which defines a temporary change in the mean that starts at d_1 and ends at d_2. However, if the time period from d_1 to d_2 is short, it could be partially confounded with any seasonal change.

In some cases the trend may be the focus of the analysis, in which case (4.1) can be used to obtain estimates of the trend that are *seasonally adjusted*. For example, we may be interested in recent trends in mortality potentially due to global warming, but want to adjust for the seasonal pattern in mortality. In other cases the focus of the analysis is the seasonal pattern, and we want to remove the trend so that it can not influence (or become part of) the seasonal estimate. This is most likely to happen when the trend is highly changeable. For example, a disease increases over a year, and then decreases in the following year. Such relatively short-term changes can appear to be like seasonal changes, and become difficult to separate statistically. A common theme of the methods presented in this chapter is strategies that aim to prevent the trend and season from competing to "explain" the same variance.

An important constraint that helps to ensure the orthogonality of the trend and seasonal parts of model (4.1) is

$$\sum_{t=1}^{n} s_t = 0.$$

This means that, on average, the seasonal pattern is centred on the mean. We assume that n/l is an integer, where l is the length of the seasonal cycle (for an annual season pattern $l = 12$). This assumption ensures that the sum includes n/l complete cycles of the seasonal pattern.

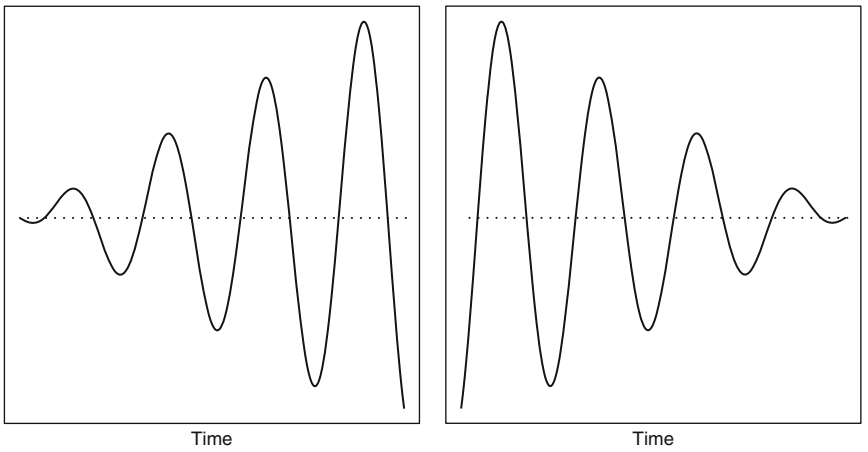


Fig. 4.2 Non-stationary seasonal patterns. *Dotted horizontal lines* at zero

Non-stationary Seasonal Patterns

An important distinction is whether the seasonal pattern is *stationary* or *non-stationary*. A stationary seasonal pattern is constant from season to season, whereas a non-stationary pattern changes over time. This is also known as *time-dependent* seasonality. We can also say that there is a trend in the seasonality, being careful to point out that this trend is distinct from the long-term trend in disease frequency.

The left panel of Fig. 4.2 shows a seasonal pattern with an increasing amplitude. The right panel of Fig. 4.2 shows a seasonal pattern with an decreasing amplitude. For both series the length of the seasonal cycle and phase has remained constant. These seasonal patterns are both non-stationary, whereas the seasonal pattern in Fig. 4.1 is stationary.

A non-stationary seasonal pattern may occur because of year-to-year changes in exposure. For example, an unusually mild winter may lead to a smaller number of cold-related CVD deaths compared to neighbouring years. Alternatively a gradual change over time in an exposure may lead to a gradual change in the seasonal pattern of a condition. For example, gradual changes in sunlight exposure (due to public health education programmes) may be responsible for changes in the seasonal pattern of birth weight [61]. Also, a gradual change in marriage trends may have reduced the seasonal pattern in UK birth numbers (Sect. 2.2.2).

4.1 Stationary Cosinor

In this section we present a very simple model of decomposition, based on a parametric trend and stationary cosinor to estimate the seasonal effect. This model is simpler than the models in later sections, but is easy to apply and can be a useful

comparison to more complex models. It may also be a sufficiently complex model for studies covering short periods (say 2–3 years), where there is less potential for complex trends and non-stationary seasonal patterns.

Using (4.1) we assume that the trend is a parametric function of time $\mu_t = f(t)$, for example a quadratic model, $\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$. The season is assumed to be a stationary cosinor, defined by

$$s_t = A\cos(\omega_t - P), \qquad t = 1, \ldots, n.$$

This is the cosinor model (Sect. 1.3.2). As with the previous cosinor model this model can be fitted as a GLM. So we can apply the model to data assuming a Normal, Poisson or Binomial distribution.

4.1.1 Examples

4.1.1.1 Cardiovascular Disease Deaths

We fitted a stationary cosinor model with trend to the cardiovascular disease death data. We first standardised the counts to adjust for the unequal number of days in the month (Sect. 2.2.1). We used linear and quadratic terms to model the trend.

The decomposition of the series into the trend, season, fitted values and residuals is shown in Fig. 4.3. Instead of showing a time series of the residuals we show the autocorrelation function of the residuals, as it is easier to judge the important assumption of independence at the seasonally important lag of 12 months.

The trend shows a strong quadratic U-shape over time. The seasonal pattern shows a symmetric oscillation with a phase of 1.3 months (early January) and an amplitude of 214 deaths per month. The fitted values in Fig. 4.3 can be compared with those in Fig. 3.2, which shows a stationary cosinor model without trend. The ACF values of the residuals are greatly reduced compared with the ACF of the monthly counts (Fig. 1.9), but still show a strong seasonal pattern.

The R code to fit a stationary cosinor model with a quadratic trend is

```
> CVD$yrmonc<-CVD$yrmon-1994
> CVD$yrmon2c<-CVD$yrmonc^2
> result<-cosinor(adj~yrmonc+yrmon2c,
type='monthly',date=month,data=CVD)
```

The first line centres the fraction of time variable `yrmon`; the second line creates a quadratic version for fitting the trend.

4.1.1.2 Schizophrenia

For the schizophrenia data we used a GLM assuming that the monthly number of cases followed a Poisson distribution. We used the log of the number of births divided by 1,000 as an offset. A cubic trend gave much better residuals than a quadratic trend.

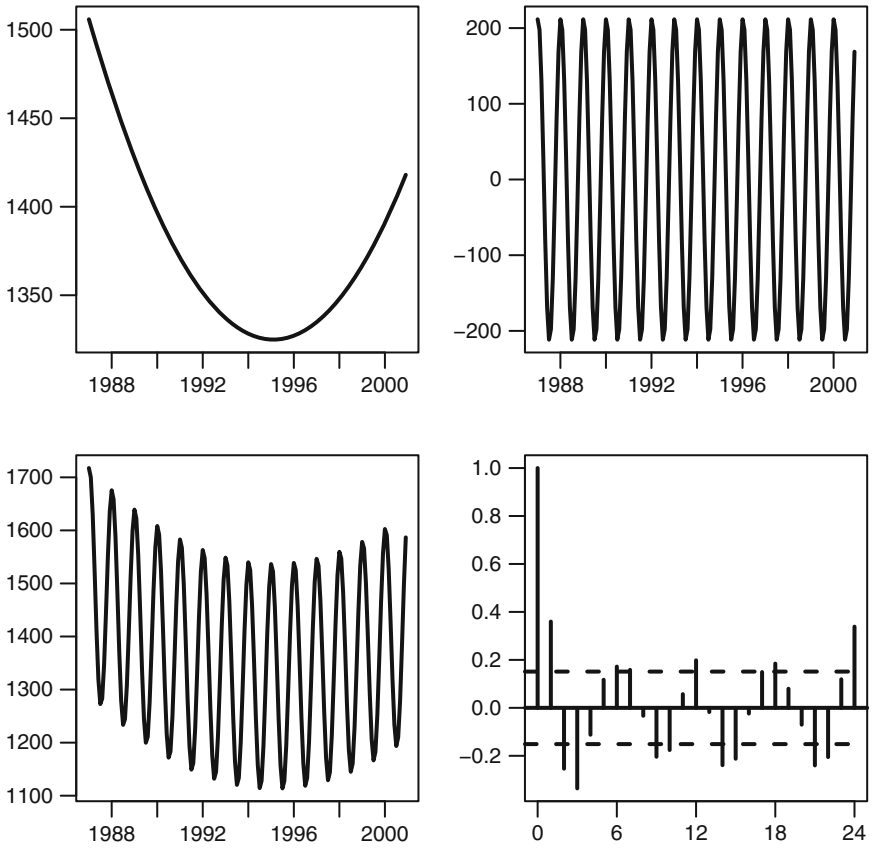


Fig. 4.3 Results from decomposing the monthly CVD time series using a stationary seasonal model. Quadratic trend (*top-right*), stationary seasonal component (*top-right*), fitted values (trend + season; *bottom-left*), and ACF of the residuals (*bottom-right*). The scales on the y-axes are different

The stationary seasonal estimate in terms of the rate ratio is shown in Fig. 4.4, together with the autocorrelation function of the residuals. The sinusoidal seasonal estimate had a phase of 6.9 months (end of June), with a peak in the rate ratio of 1.045. The autocorrelation in the residuals is small, suggesting that a stationary seasonal pattern (with a cubic trend) may be an adequate model.

4.2 Season, Trend, Loess

In this section we describe a decomposition method based on *loess* smoothing [17, 85]. It is given the acronym STL (season, trend, loess). The purpose of using loess smoothing is to model the trend and seasonal effects using purely numerical

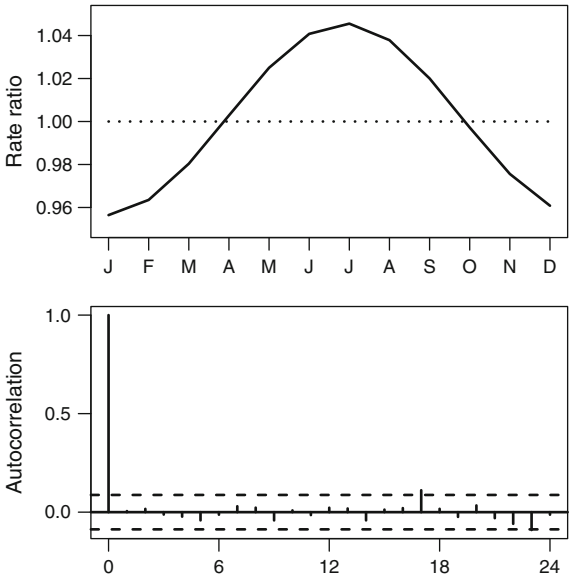


Fig. 4.4 Results from decomposing the schizophrenia time series using a stationary seasonal model. Stationary season (*top*) and ACF of the residuals (*bottom*)

methods based on the data rather than assuming mathematical models (such as the cosinor) or any particular probability distributions (e.g., the Poisson distribution for count data).

For a dependent variable **y** and independent variable **x** the stages of loess smoothing are:

1. For a value x_t define the window as $x_t - h$ to $x_t + h$ for some constant h. The pairs of dependent and independent variables within this window are $\mathbf{y}'$ and $\mathbf{x}'$, respectively.
2. Calculate weights for each point in the window $w_j = f(x'_j - x_t)$. These weights describe the distance between each point and the central point (x_t), so that values further from x_t are given less weight. A simple example is $f(u) = |u|$.
3. Fit a regression model $\mathbf{y}' = f(\mathbf{x}')$ using the weights **w**. This model is usually linear $f(u) = \beta_0 + \beta_1 u$, or quadratic $f(u) = \beta_0 + \beta_1 u + \beta_2 u^2$. It is also possible to fit a constant $f(u) = \beta_0$, which is equivalent to a *moving average*. The degree of the model is 0 for a moving average, 1 for a linear model, and 2 for a quadratic model.
4. Estimate the fitted value $s_t = \hat{y}'_t$ at x_t.

These four steps are then repeated for a range of values of **x** to build up the smooth seasonal curve (**s**).

The key choice for determining the smoothness of **s** is the size of h. Larger values of h increase the number of observations in the window and so give a smoother estimate. As h becomes very small the number of paired observations tends to 1, which gives a "smooth" that passes through each observation ($s_t = y_t$).

For monthly data with an annual seasonal pattern, the STL algorithm groups each of the months and applies a loess smooth to each group. So the results in every January are smoothed, then the results in every February, and so on. The window (h) is the number of neighbouring years included in the smooth. Selecting $h = 1$ means that only neighbouring years are used, whereas $h = 2$ uses two years either side. The overall size of the smoothing window is $w = 2h + 1$. This then estimates the long-term pattern in each month. The long-term seasonal pattern is then estimated by joining together these monthly estimates. This procedure treats the results from neighbouring months as independent, and so the STL does not constrain the seasonal pattern to take a particular form (e.g., sinusoidal).

The STL algorithm applies separate loess smooths to the trend and seasonal pattern, so we need to choose a window size for the trend (w_μ) and a separate window size for the season (w_s). We should choose w_μ so that it models the long-term variation in disease. For monthly data, the value for w_μ determines the number of months included in the window. A very small value for w_μ could therefore wrongly include seasonal variation. Guidelines for choosing w_s are w_μ are:

- w_s should be an odd integer, $\geq$7. The final choice of w_s should be based on knowledge about the expected variation in the seasonal pattern, and testing for remaining seasonal pattern in the estimated residuals.
- $w_\mu = \lfloor 1.5l/(1 - 1.5/w_s) \rfloor$ where $\lfloor x \rfloor$ means the smallest odd integer $\geq$x and l is the length of the seasonal cycle. For an annual seasonal pattern with monthly data $l = 12$.

The STL algorithm estimates the smooth trend and season by alternating between estimates of the season and trend, using updated estimates at each iteration [17]. It also applies a *low-pass filter* to the seasonal pattern in order to further smooth it by removing any erratic (high frequency) patterns.

A robustness step can be added to the algorithm, which multiplies the weight calculated in step 2 by a reliability weight. This weight is based on the estimated residuals

$$\hat{\varepsilon}_t = y_t - \hat{\mu}_t - \hat{s}_t, \qquad t = 1, \ldots, n.$$

The weight is designed to reduce the influence of observations with a relatively large $\hat{\varepsilon}_t$. The relative size of the residual is calculated using the absolute values of the residuals

$$\hat{e}_t = |\hat{\varepsilon}_t| / \left[6 \times \text{median}(|\boldsymbol{\varepsilon}|)\right].$$

The reliability weight is then

$$w_j^* = \begin{cases} (1 - \hat{e}_j^2)^2, & 0 \leq \hat{e}_j < 1, \\ 0, & \hat{e}_j \geq 1. \end{cases}$$

4.2.1 Examples

4.2.1.1 Cardiovascular Disease Deaths

Figure 4.5 shows the STL decomposition with window sizes, $w_s = 7$ and $w_\mu = 21$ using the cardiovascular disease data. We first adjusted the counts to account for the unequal number of days in the month (Sect. 2.2.1). We selected a reasonably small seasonal window ($w_s = 7$) because we believed the seasonal pattern in monthly CVD is non-stationary (based on the plot of the observed data, Fig 1.1). The degree of the loess model was 0 for season, and 1 for trend.

The trend in Fig. 4.5 shows the overall mean, which ranges from around 1,300–1,500 deaths per month. The trend is much smoother using the robustness step. Without using this step, the trend shows an oscillating pattern from high to low

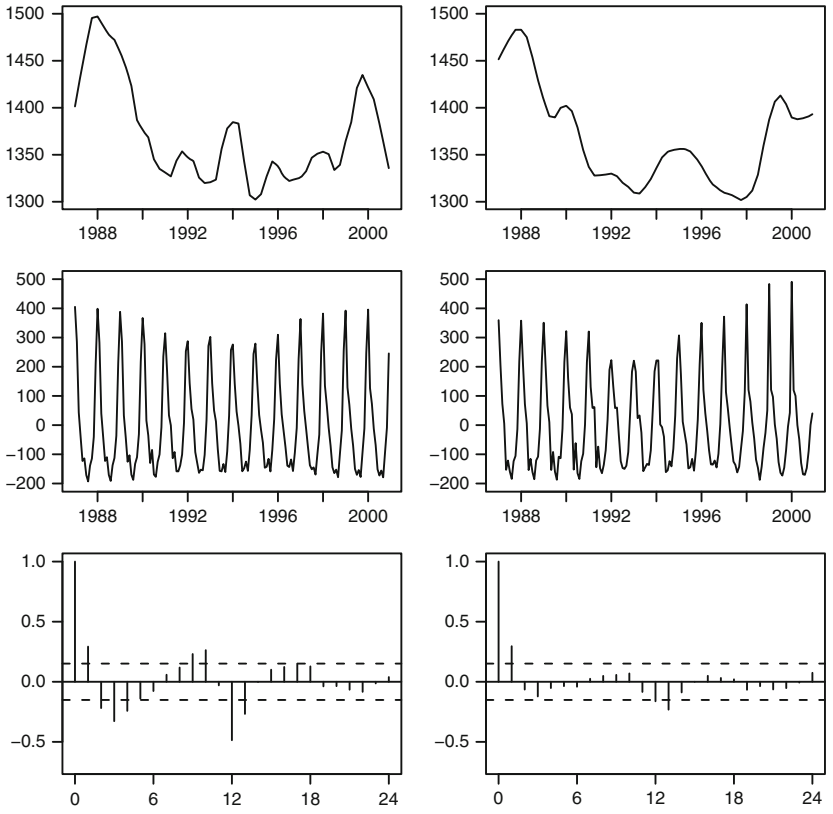


Fig. 4.5 Example of using STL for the monthly cardiovascular disease data. The rows show: the trend (*top*), season (*middle*) and autocorrelation of the residuals (*bottom*). The results are based on the windows $w_s = 7$, $w_\mu = 21$, the *right column* includes a robustness step to reduce the influence of large residuals

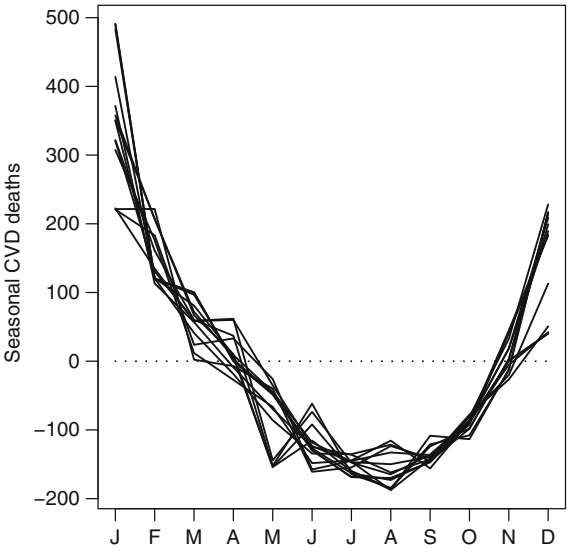


Fig. 4.6 Estimated annual seasonal pattern using STL for the monthly cardiovascular disease death data. Overlay of seasonal patterns from multiple years using windows of $w_s = 7$, $w_\mu = 21$ and a robustness step. *Dotted horizontal line* at zero

between 1990 and 1999. This is the seasonal pattern discussed in Sect. 2.2.2, and so the results with the robustness step are preferred.

The seasonal estimates have captured the skewed, non-sinusoidal, pattern in cardiovascular death, with its sharp January peak. This is an advantage of the STL method, as by smoothing in grouped months it places no restriction on the seasonal shape. A more detailed look at the seasonal pattern is shown in Fig. 4.6. The number of deaths peaks in January (200–500 more deaths than the average), and dips in July or August (100–200 fewer deaths).

The seasonal pattern also appears to be non-stationary as it changes over time. Both seasonal patterns in Fig. 4.5 remain relatively stable from 1987 to 1990, then reduce in amplitude from 1991 to 1994/1995, and then increase in amplitude to 2000. The non-stationary seasonal patterns in Fig. 4.5 are quite different to the stationary estimate shown in Fig. 4.3.

The STL allows a seasonal pattern that is non-sinusoidal and non-stationary. Despite this flexibility, the residuals from both models in Fig. 4.6 contain some remaining seasonality as indicated by the ACF, although this is much reduced using the robustness step. The positive autocorrelation at lag 1 may indicate an inadequate estimate for trend. Experimenting with different values of w_s and w_μ might eventually yield almost independent residuals. The autocorrelation in the residuals is greatly reduced compared with the stationary model in Fig. 4.3.

The R commands to fit a STL model are:

```
> ts<-ts(CVD$adj,frequency=12,start=c(1987,1))
> sm<-stl(ts,s.window=7,t.window=21,robust=TRUE)
> plot(sm)
```

where `s.window` is w_s, and `t.window` is w_μ. The first command creates a time-series object based on the adjusted monthly numbers of CVD deaths, starting in January, 1987, and with an 12-month frequency (annual cycle). A robust version of the STL is fitted by using the option `robust=TRUE`.

4.2.1.2 Schizophrenia

For the schizophrenia data (Sect. 1.1.2) we selected a seasonal window of $w_s = 11$ because we believed that any seasonal pattern would change relatively slowly over time. Figure 4.7 shows the seasonal pattern for the rate of schizophrenia per 1,000 births, for the four decades from the 1930s to the 1960s. To aid the visualisation of the annual season, dashed vertical lines have been added at the start of each year.

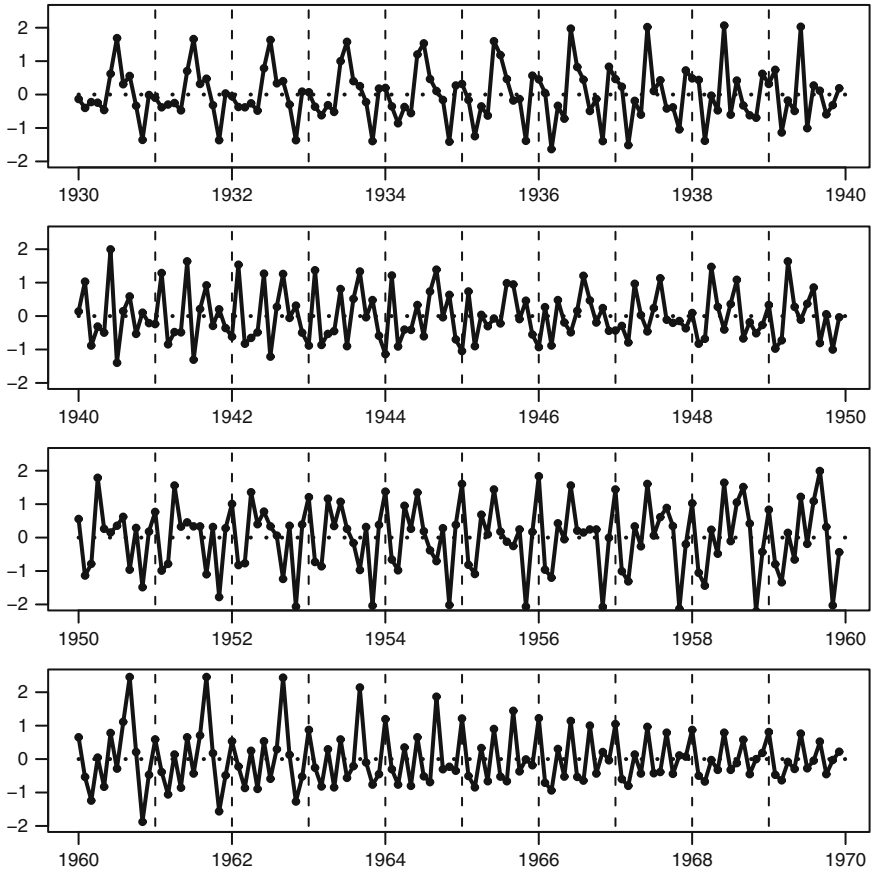


Fig. 4.7 Estimated annual seasonal pattern using STL for the monthly rates of schizophrenia per 1,000 births. The y-axes are on the scale of schizophrenia per 1,000 births centred at zero. Each panel shows a different decade. *Vertical dashed lines* are at the start of every year. The *dotted horizonal lines* are at zero

The estimated seasonal pattern in schizophrenia is strongly non-stationary, both in its phase and its amplitude. The pattern is relatively regular from 1930 to 1935, with a peak in risk in July births, and dip in November. The peak in risk then moved to June from 1935 to 1940. In the remainder of the 1940s it is difficult to see a clear seasonal pattern, in particular it is difficult to see a sinusoidal pattern. In the late 1950s and early 1960s the peak in risk was in September, with the dip in November. Over the 1960s there was a steady decline in this seasonal pattern. This decline corresponds to the decline in trend, which indicates that a multiplicative seasonal model may be more appropriate, in place of the additive model, (4.1).

The multiplicative decomposition model is

$$Y_t = \mu_t \times s_t \times \varepsilon_t, \qquad t = 1, \ldots, n.$$

This can be fitted to the schizophrenia data by first log-transforming the rates of schizophrenia per 1,000 births. The results are shown in Fig. 4.8. The seasonal estimates are given as rate ratios, and range from around 0.7 to 1.4. The estimated seasonal pattern from the multiplicative model has a similar shape to that from the additive model shown in Fig. 4.7. The reduction in the seasonal variation over the 1960s using the multiplicative model is much less pronounced, although any seasonal pattern over this period is quite noisy, with erratic changes in risk from month to month.

4.3 Non-stationary Cosinor

In this section we show a method for decomposing a time series using a cubic spline for the trend, and a non-stationary seasonal pattern defined as

$$s_t = A_t \cos(\omega_t - P_t), \qquad t = 1, \ldots, n, \tag{4.3}$$

where $\omega_t = 2\pi f_t$ and f_t is the fraction of time as described in (3.4) for individual dates (e.g., birth dates) or (3.5) for monthly data. This model is parametric because we have constrained the pattern to be sinusoidal. It is non-stationary because the amplitude (A_t) and phase (P_t) are dependent on time. We refer to (4.3) as a non-stationary cosinor model.

The model, (4.1), with the seasonal pattern, (4.3), can be represented more generally, as in [5], as

$$\begin{aligned}
Y_t &= \mathbf{F}^T \mathbf{B}_t + \varepsilon_t, \qquad t = 1, \ldots, n, \\
\mathbf{B}_t &= \mathbf{G}\mathbf{B}_{t-1} + \mathbf{T} \cdot \mathbf{v}_t, \\
\mathbf{v}_t &\sim \text{MVN}(\mathbf{0}, \mathbf{V}), \\
\varepsilon_t &\sim N(0, \sigma_\varepsilon^2),
\end{aligned}$$

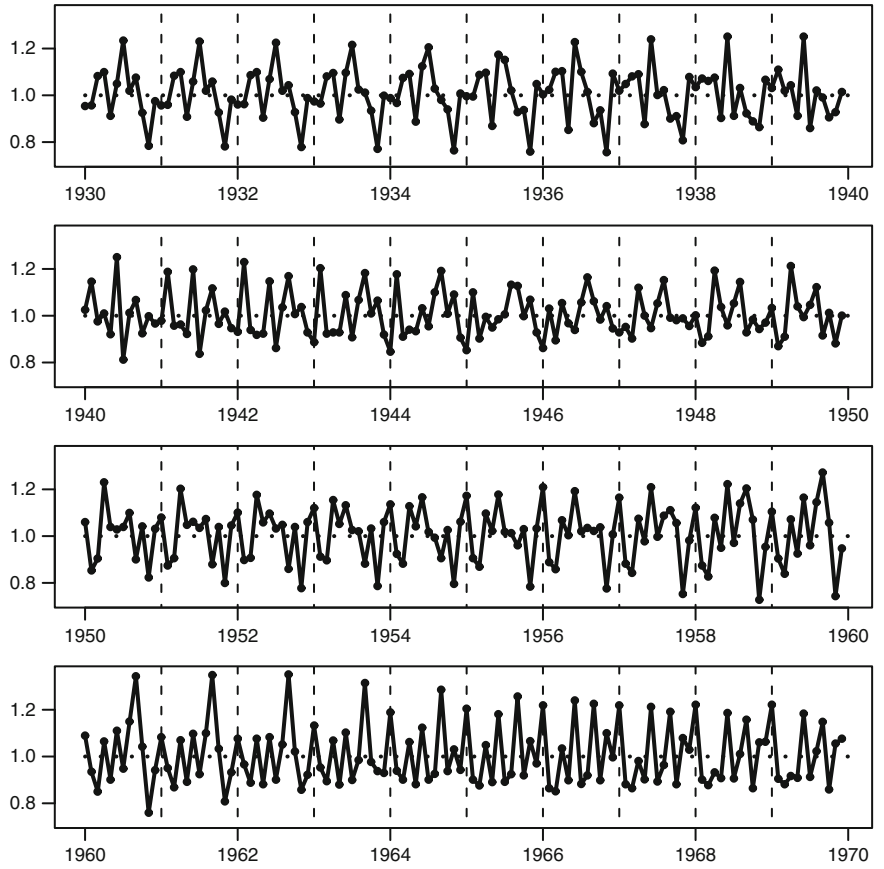


Fig. 4.8 Estimated annual seasonal pattern using STL for the log-transformed monthly rates of schizophrenia per 1,000 births. The y-axes show the rate ratios for schizophrenia per 1,000 births centred at zero. Each panel shows a different decade. *Vertical dashed lines* are at the start of every year. The *dotted horizonal lines* are at one

where $\cdot$ denotes element-by-element matrix multiplication and the matrices are

$$\mathbf{F} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{G} = \left[\begin{array}{cc|cc} 1 & \lambda & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & \theta & -1 \\ 0 & 0 & 1 & 0 \end{array}\right], \quad \mathbf{V} = \left[\begin{array}{cc|cc} \frac{\lambda^3}{3} & \frac{\lambda^2}{2} & 0 & 0 \\ \frac{\lambda^2}{2} & \lambda & 0 & 0 \\ \hline 0 & 0 & \sigma_s^2 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \quad \mathbf{T} = \begin{bmatrix} \tau_1 \\ \tau_1 \\ \tau_2 \\ \tau_2 \end{bmatrix},$$

and $\theta = 2\cos(2\pi/l)$ where l is the length of the seasonal cycle. For an annual seasonal pattern using monthly data $l = 12$. The term λ is the distance between observations (on a time scale), so for monthly data $\lambda = 1/12$, and for annual data $\lambda = 1/365$. We have added vertical and horizontal lines to the above matrices to show the partition between the trend (top-left) and seasonal (bottom-right) parts.

The matrix $\mathbf{B}_t$ defines a cubic spline for the trend and a non-stationary cosinor for the season. At each time point, t, the matrix is a combination of the matrix from the previous time, $\mathbf{B}_{t-1}$, and some change from the previous time, $\mathbf{v}_t$, which is scaled by τ_1. Larger values of τ_1 allow more change between observations and hence a greater potential flexibility in the spline. In this model, the value for τ_1 is fixed in advance, similarly to the degrees of freedom for a Generalized Additive Model (GAM) (Sect. 5.2) [45].

The flexibility in the non-stationary seasonal estimate is determined by the variance σ_s^2 and the smoothing parameter τ_2. This parameter τ_2 is fixed in advance, whilst the variance parameter σ_s^2 is estimated (see below).

One advantage of this model is that it is relatively easy to add multiple seasonal components, as in (4.2). For example, to fit two seasonal terms we expand the matrices thus,

$$\mathbf{F} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \mathbf{G} = \left[\begin{array}{cc|cc|cc} 1 & \lambda & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \theta_1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \theta_2 & -1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right], \mathbf{V} = \left[\begin{array}{cc|cc|cc} \frac{\lambda^3}{3} & \frac{\lambda^2}{2} & 0 & 0 & 0 & 0 \\ \frac{\lambda^2}{2} & \lambda & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \sigma_{s_1}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \sigma_{s_2}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \mathbf{T} = \begin{bmatrix} \tau_1 \\ \tau_1 \\ \tau_2 \\ \tau_2 \\ \tau_3 \\ \tau_3 \end{bmatrix},$$

where $\theta_1 = 2\cos(2\pi/c_1)$, $\theta_2 = 2\cos(2\pi/c_2)$ and c_1 and c_2 are the lengths of the seasonal cycles ($c_1 \neq c_2$).

4.3.1 Parameter Estimation

The estimates of $\mathbf{v}, \sigma_\varepsilon$ and $\boldsymbol{\sigma}_s$ in the non-stationary cosinor model are made using the *Kalman filter* [45]. We first use a forward and backward sweep of the Kalman filter to estimate $\hat{\mathbf{B}}_t$. The mean estimated trend and season are extracted using

$$\begin{aligned} \hat{\mu}_t &= [1\,0\,0\,0]\hat{\mathbf{B}}_t, \qquad t = 1, \ldots, n, \\ \hat{s}_t &= [0\,0\,1\,0]\hat{\mathbf{B}}_t, \qquad t = 1, \ldots, n, \end{aligned}$$

assuming a single seasonal component. The residuals are estimated as

$$\hat{\varepsilon}_t = y_t - \hat{\mu}_t - \hat{s}_t, \qquad t = 1, \ldots, n.$$

The variance σ_ε^2 is estimated using a randomly generated value from an inverse Gamma distribution with shape $(n/2) - 1$ and scale $\sum \hat{\varepsilon}_t^2/2$.

To estimate the variance in the seasonal estimate, σ_s^2, we first calculate the difference

$$\Delta\hat{s}_t = \hat{s}_t - \hat{s}_{t-1}, \qquad t = 2, \ldots, n.$$

The variance σ_s^2 is then estimated using a randomly generated value from an inverse Gamma distribution with shape $[(n-1)/2]-1$ and scale $\sum \Delta\hat{s}_t^2/2$. The value sampled from the inverse Gamma is divided by τ_2^2 to give σ_s^2.

The whole process is then repeated again using the updated estimates of σ_ε^2 and $\boldsymbol{\sigma}_s^2$. This is known as Markov chain Monte Carlo (MCMC) sampling. We store the estimates at the end of each iteration to give a vector of estimates, $\hat{\boldsymbol{\sigma}}_s = \hat{\sigma}_s^1, \hat{\sigma}_s^2, \ldots, \hat{\sigma}_s^M$, of which there are M in total. Mean estimates are then the sample means

$$\overline{\hat{\sigma}_s} = \frac{1}{M}\sum_{i=1}^{M}\hat{\sigma}_s^i, \qquad \overline{\hat{\sigma}_\varepsilon} = \frac{1}{M}\sum_{i=1}^{M}\hat{\sigma}_\varepsilon^i, \qquad \overline{\hat{\mu}}_t = \frac{1}{M}\sum_{i=1}^{M}\hat{\mu}_t^i, \qquad \overline{\hat{s}}_t = \frac{1}{M}\sum_{i=1}^{M}\hat{s}_t^i.$$

Similarly we can calculate confidence limits for the parameter estimates using the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles from the M MCMC samples. These confidence intervals have no distributional assumptions, so it not important if the MCMC samples are not Normally distributed. Being able to calculate confidence intervals for the trend and season is a distinct advantage of this method over the STL method.

The MCMC process is started by giving initial values for $\hat{\sigma}_\varepsilon$ and $\hat{\boldsymbol{\sigma}}_s$. The initial value for $\hat{\sigma}_\varepsilon$ is the standard deviation of the dependent variable. The initial values for $\hat{\boldsymbol{\sigma}}_s$ need to be specified. The convergence of the MCMC chains should be verified before creating any estimates (Sect. 1.6.1).

The values to be estimated are $\mathbf{v}$, σ_ε, $\boldsymbol{\sigma}_s$ and $\boldsymbol{\tau}$. The parameters $\boldsymbol{\tau}$ are fixed in advance of estimating the other parameters, but we can estimate their optimal values using a trial and error procedure, as we show in the examples below. We recommend starting with all values of τ equal to 1. A model can then be fitted using a small number of MCMC samples, say 2,000 samples with a burn-in of 500. The results of this initial fit should give an idea of whether the trend is too taut or too flexible. A second model can then be fitted by suitably adjusting τ_1. When choosing the next value it is worth bearing in mind that τ_1 is proportional to the variance of the dependent variable, $\text{var}(Y_t)$. We also recommend using initial values for $\boldsymbol{\sigma}_s$ based on the results of the first model. Further models should be fitted until the estimated trend appears to be modelling the long-term change in disease, and no seasonal components. Next we tune the value of τ_2 to control the flexibility in the seasonal estimate. We recommend trying both very small values that give an almost stationary fit, and large values that allow the seasonal pattern to be strongly non-stationary. A final value for τ_2 can be based on looking for any remaining seasonal pattern in the residuals.

The mean residuals are estimated as

$$\overline{\hat{\varepsilon}}_t = y_t - \overline{\hat{\mu}}_t - \overline{\hat{s}}_t, \qquad t = 1, \ldots, n.$$

These are the residuals that we test for any remaining trend or seasonal pattern.

4.3.1.1 Estimating the Amplitude and Phase

To estimate the amplitude and phase of the seasonal pattern we first calculate the periodogram of $\hat{s}_t$ (Sect. 1.3.3). We then select the estimates of $\hat{C}_j$ and $\hat{S}_j$ for the value of j closest to the cycle (l). We can then estimate the amplitude and phase using (Eqns. 3.2) and (3.3), respectively, with $c = C_j$ and $s = S_j$. We do this at every MCMC sample to give a chain of estimates for the amplitude, $\hat{\mathbf{A}} = \hat{A}^1, \hat{A}^2, \ldots, \hat{A}^M$, and phase, $\hat{\mathbf{P}} = \hat{P}^1, \hat{P}^2, \ldots, \hat{P}^M$.

The phase is a circular variable that is bounded between 0 and 2π. Hence estimating confidence intervals is complicated because the values 0 and $2\pi - a$ (for small a) are close on the circular scale, but far apart on the linear scale. The solution is to first rotate the phases ($\hat{\mathbf{P}}$) so that they are centred on π, estimate the confidence interval for the rotated phases, and then rotate these estimates back to their centre [35, Sect. 2.3.2]. As before the confidence interval is calculated using the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles.

Figure 4.9 demonstrates the issue for estimates of the phase centred close to zero. A confidence interval based on the unrotated estimates almost covers the entire range of possible phase values. After rotation, the mean is estimated as 0.32, and the 95% confidence interval as -0.15 to 0.84. The lower confidence interval is negative because it is relative to the mean. We can put the values on a more meaningful scale using the `invyrfracion` function in R. If the phase in Fig. 4.9 is for daily data, then the mean phase is 19 January, with a 95% confidence interval from 23 December to 18 February.

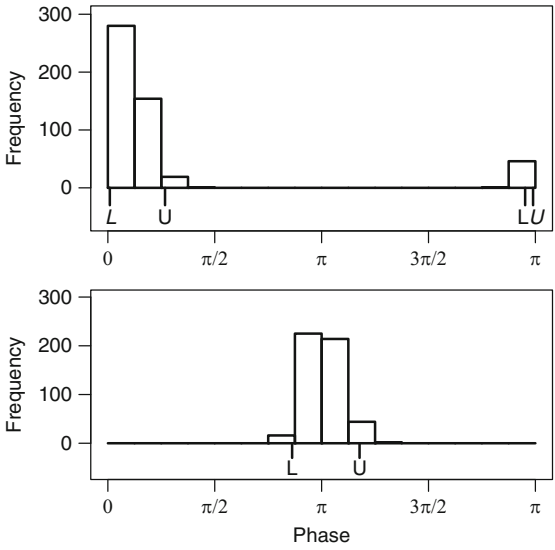


Fig. 4.9 Estimates of a 95% confidence interval for the phase based on 500 MCMC estimates. The lower (L) and upper (U) estimates are shown in italics for the unrotated estimates and in normal font for the rotated estimates. The *top panel* shows a histogram of the MCMC samples, the *bottom panel* shows the estimates after they have been rotated to have a centre on π

4.3.2 Examples

4.3.2.1 Cardiovascular Disease Deaths

Figure 4.10 shows the estimated trend, seasonal effect and residuals from the CVD death data. The estimates are based on a single seasonal cycle of 12 months. We show a histogram and box plot of the residuals against month. The results are based on $\tau_1 = 130$ and $\tau_2 = 10$. We used 5,000 MCMC samples, with a burn-in of 1,000.

The results in Fig. 4.10 include confidence intervals for the mean and seasonal estimate, whereas the STL only estimates the mean. This is an advantage of this method over the STL.

The seasonal pattern is reasonably consistent in its phase, the mean estimate is 2.3 months and 95% confidence interval is 2.1 to 2.5, corresponding to early February. This phase is surprising late, considering that the peak of deaths is usually in January (Fig. 2.3). The box plot of residuals (Fig. 4.10) shows that the model tends to underestimate mortality in January, with an inter-quartile range from around 100 to 300 deaths. The sinusoidal model has missed the peak in January deaths, and this is because it is based on the assumption of symmetry. For sinusoidal models the gap between the highest and lowest rates of disease will be 6 months apart. However, Fig. 2.3 shows a peak in deaths is in January and a that low that varies from July to September (6–9 months later). The estimated phase for the model shown in Fig. 4.10 has therefore been influenced more by the low in mortality than its peak. The box plot of the residuals confirms this, as the residuals for July to September, are comparatively good. Overall, there is still some seasonal pattern in the residuals, as shown by the ACF.

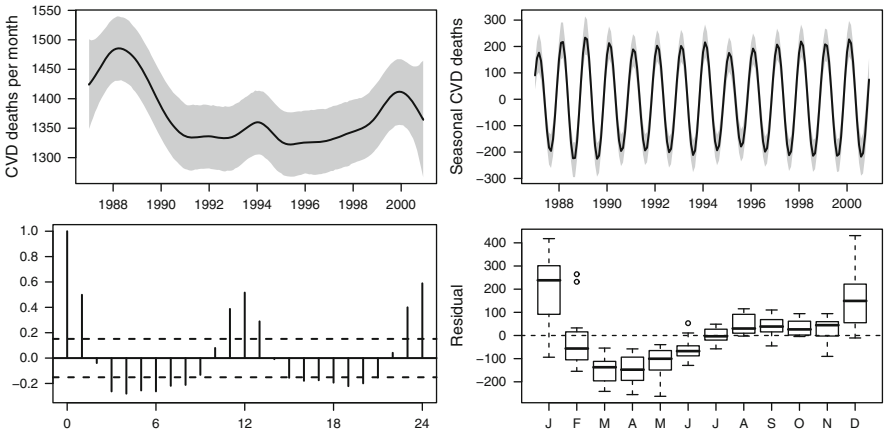


Fig. 4.10 Estimated trend (*top-left*) and season (*top-right*) over time, and residual autocorrelation (*bottom-left*) and box plot (*bottom-right*) from the CVD data using a non-stationary cosinor with smooth parameters $\tau_1 = 130$, $\tau_2 = 10$. In the *top row* the mean is a *solid line*, and the 95% confidence interval a *grey area*

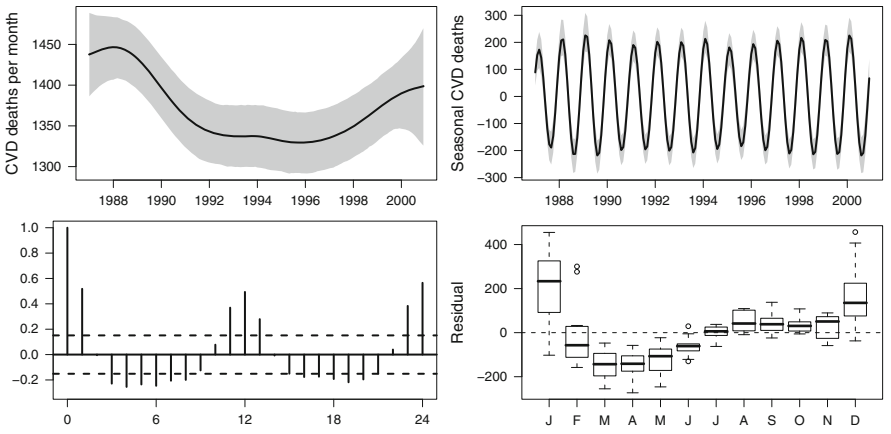


Fig. 4.11 Estimated trend (*top-left*) and season (*top-right*) over time, and residual autocorrelation (*bottom-left*) and box plot (*bottom-right*) from the CVD data using a non-stationary cosinor with smooth parameters $\tau_1 = 30$, $\tau_2 = 100$. In the *top row* the mean is a *solid line*, and the 95% confidence interval a *grey area*

The estimated amplitude does vary from year-to-year. The average amplitude is 207 deaths per month (95% CI: 183, 231). The mean estimate of σ_ε is 110.0, and of σ_s is 0.79.

The long-term trend shows clear peaks in 1989 and 2000. The small peak in trend in 1994 in January could be due to a seasonal increase. We want the trend to exclusively model the long-term changes in disease. We therefore re-ran the model with a smaller smoother parameter for the trend ($\tau_1 = 30$). At the same time we increased the smoothing parameter for the seasonal pattern in the hope that it would increase in flexibility and thus give a better fit to the seasonal pattern ($\tau_2 = 100$). The results are shown in Fig. 4.11.

The trend in Fig. 4.11 is much smoother than that shown in Fig. 4.10, so the reduction in the smoothing parameter has had the desired effect. In contrast, increasing the smoothing parameter has done little to change the seasonal estimate. The estimated mean phase and amplitude are almost identical. We increased the τ_2 smoother parameter further, but still found no change in the seasonal pattern. This suggest that the variance in the seasonal pattern is at its limit, and a better fit can only be achieved using a different model.

To see if we could improve the residuals, we added a second seasonal cycle at 6 months. This cycle was suggested by the periodogram in Fig. 1.18. The results are shown in Fig. 4.12.

The 6-month season amplitude is smaller than the 12-month amplitude, averaging 74 deaths per month (95% CI: 52, 93). Adding a 6-month cycle has not improved the overall fit, as the residuals still have a strong seasonal pattern.

The R commands to fit a model with two seasonal components with cycles of 6 and 12 months is:

```
> nscosinor(response=adj,cycles=c(6,12),tau=c(30,10,10),
  niters=5000,burnin=1000,data=CVD,inits=c(2,2))
```

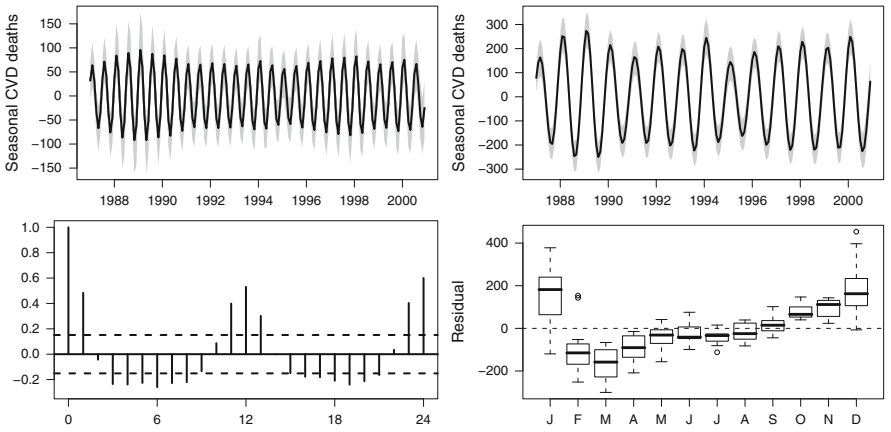


Fig. 4.12 Estimated 6-monthly season (*top-left*) and 12-monthly season (*top-right*) over time, and residual autocorrelation (*bottom-left*) and box plot (*bottom-right*) from the CVD data using a non-stationary cosinor with smooth parameters $\tau_1 = 30$, $\tau_2 = 100$, $\tau_3 = 100$. In the *top row* the mean is a *solid line*, and the 95% confidence interval a *grey area*

4.3.2.2 Schizophrenia

For the schizophrenia data we fitted a single annual seasonal pattern. Based on the experience with the STL (Sect. 4.2.1), we fitted the model to the log-transformed rates of schizophrenia per 1,000 births (i.e., multiplicative model). The results are based on 5,000 MCMC samples and a burn-in of 1,000.

The mean phase is at 6.4 months, with a wide 95% confidence interval of 4.6 to 8.5 months. The mean amplitude is a rate ratio of 1.048, with a 95% confidence interval of 1.016 to 1.088. This result is quite similar to the model using a stationary cosinor (Fig. 4.4), which had a phase of 6.9 months, with a rate ratio amplitude of 1.045. The estimated non-stationary season is plotted by decade in Fig. 4.13.

4.4 Modelling the Amplitude and Phase

The model in the previous section used a non-stationary cosinor for the seasonal pattern, Eqn. (4.3). The advantages of this were shown in the two examples, as we were able to see how the seasonal pattern had changed over time. However, the model did not give actual estimates of how the trend or phase had changed over time, except via the plot of the overall seasonal pattern. Also this model is restricted to a seasonal pattern with both a non-stationary phase and amplitude, but we may prefer to assume that only one of these parameters is non-stationary. In this section we describe another model that assumes a non-stationary seasonal pattern, but more explicitly models the non-stationarity.

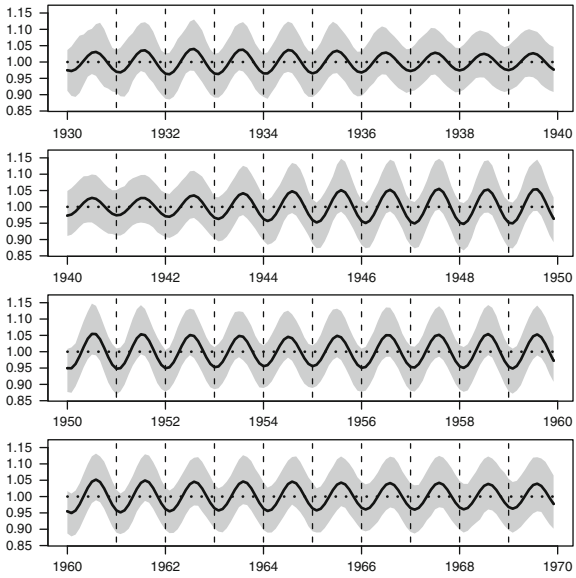


Fig. 4.13 Estimated annual seasonal pattern using a non-stationary cosinor with smooth parameters $\tau_1 = 0.01$, $\tau_2 = 2$, for the monthly rates of schizophrenia per 1,000 births. The y-axes are on the scale of schizophrenia rate ratios per 1,000 births. Each panel shows a different decade. *Vertical dashed lines* are at the start of every year. The *dotted horizonal lines* are at one. The mean is a *solid line*, and the 95% confidence interval a *grey area*

Eilers et al. used a similar model to examine non-stationary patterns in total mortality over the years 1960–2000 in 50 different age groups [32]. Their method smooths the cosine and sine parameters of a cosinor model over time and age using a two-dimensional spline. In their example they found non-stationarity in both the phase and amplitude that varied by age group.

We define two alternative versions of (4.3). A model with non-stationary amplitude and stationary phase is

$$s_t = A_t \cos(\omega_t - P), \qquad t = 1, \ldots, n. \tag{4.4}$$

A model with stationary amplitude and non-stationary phase is

$$s_t = A \cos(\omega_t - P_t), \qquad t = 1, \ldots, n. \tag{4.5}$$

The choice of whether the phase or amplitude (or both) is stationary is dependent on knowledge about the disease being studied.

A simple model for non-stationary amplitude is a linear model,

$$A_t = \alpha_0 + \alpha_1 t, \qquad t = 1, \ldots, n,$$

which would describe a steady increase or decrease in the amplitude, such as that shown in Fig. 4.1. We can test the hypothesis that the amplitude has remained constant by testing whether $\hat{\alpha}_1 = 0$. Similarly we could model a steady drift in the phase using the model

$$P_t = \phi_0 + \phi_1 t, \qquad t = 1, \ldots, n,$$

although as we show in the next section we need to restrict ϕ_0 and ϕ_1 because the phase is circular.

We can use more complex models for the amplitude, such as a quadratic change

$$A_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2, \qquad t = 1, \ldots, n,$$

but to obtain a more flexible fit for amplitude we can allow it to change linearly from year to year using

$$A_t = a_0 + \alpha_1 - \alpha_{y(t)} + \left[(\alpha_{y(t)} - \alpha_{y(t+12)})\Delta m(t)\right], \qquad t = 1, \ldots, n, \quad (4.6)$$

where a_0 is the mean amplitude, α_i is the amplitude in year $i = 1, \ldots, Y + 1$, $y(t)$ is the year at time t, and $\Delta m(t)$ is the scaled difference between months,

$$\Delta m(t) = \frac{m(t) - 1}{12},$$

where $m(t)$ is the month at time t (Sect. 2.3.1). We can smooth annual amplitudes by specifying a multivariate Normal distribution

$$\boldsymbol{\alpha} \sim \text{MVN}(\mathbf{0}, \mathbf{V}_\alpha),$$

where $\mathbf{0}$ is vector of zeros of length $Y + 1$, and $\mathbf{V}_\alpha$ is an $(Y + 1) \times (Y + 1)$ variance–covariance matrix, with (j, k)th element defined by

$$V_\alpha(j,k) = \begin{cases} \sigma_\alpha^2, & j = k, \\ \rho\sigma_\alpha^2, & |j - k| = 1, \\ 0, & \text{otherwise}, \end{cases} \quad (4.7)$$

where $0 \leq \rho \leq 1$ describes the correlation between neighbouring estimates of the amplitude, and σ_α^2 describes the overall variance of the estimates. The advantage of this model is that it produces a gradually changing amplitude, without putting any parametric restriction on its shape (e.g., a linear increase). We use a CAR model because the matrix $\mathbf{V}_\alpha$ is sparse (Sect. 2.3.3).

An advantage of this model over the previous decomposition models described in this chapter, is that it can be applied to irregularly spaced data. When using irregularly spaced data we need to modify the seasonal component described in (4.3) to

$$s(t_i) = A(t_i)\cos\left[w_i - P(t_i)\right], \qquad i = 1, \ldots, n,$$

where n is the number of observations, and t_i is the time of the ith observation. $\omega_i = 2\pi f_i$, where f_i is the fraction of time described in (3.4). $A(t')$ and $P(t')$ are the amplitude and phase at time t', respectively.

4.4.1 Parameter Estimation

We estimated this model using the Bayesian paradigm because:

- The sinusoidal model, (4.4), is non-linear in time (Sect. 3), and so standard estimation techniques (e.g., least squares) cannot be used.
- The CAR model needed for estimating $\mathbf{V}_\alpha$ is available in the Bayesian WinBUGS software.
- We are able to obtain an estimate of model fit from the deviance information criterion (Sect. 1.6.2), and the estimated number of parameters that we can use to calculate the degrees of freedom.

Another advantage of using a Bayesian model is that we are able to fit the model using a GLM framework. This means we can use the model for dependent data assuming a Normal, Poisson or Binomial distribution. The previous non-stationary cosinor model (Sect. 4.3) should only be applied to Normal dependent data.

For Normal data, a model with a stationary phase and a linearly changing amplitude is defined by

$$\begin{aligned} Y_t &\sim \mathrm{N}(m_t, \sigma_\varepsilon^2), \qquad t = 1, \ldots, n, \\ m_t &= \mu_t + s_t, \\ \mu_t &= \beta_0 + \beta_1 \tilde{t} + \beta_2 \tilde{t}^2, \\ s_t &= A_t \cos(\omega_t - P), \\ A_t &= \alpha_0 + \alpha_1 \tilde{t}. \end{aligned}$$

The overall mean, m_t, is split into the trend, μ_t, and season, s_t. The noise is assumed to be Normally distributed with variance σ_ε^2. In this model the trend is constrained to be quadratic. In place of time, $t = 1, \ldots, n$, we use a scaled version of time $\tilde{t} = [t - (n/2)]/(n/2)$, which is bounded between -1 and 1. This is because the MCMC chains generated by WinBUGS perform better with centred independent variables. The difference between two adjacent times is $\Delta\tilde{t} = 2/n$.

We use vague priors for all parameters, defined by

$$\begin{aligned} \alpha_j &\sim \mathrm{N}(0, 10^6), \qquad j = 0, 1, \\ \beta_j &\sim \mathrm{N}(0, 10^6), \qquad j = 0, 1, 2, \\ \sigma_\varepsilon &\sim \mathrm{U}(0, 10^6). \end{aligned}$$

A vague prior for a stationary phase is defined by

$$P \sim \mathrm{U}(0, 2\pi),$$

as this covers every possible point on the circle. However, this creates a difficulty when using MCMC estimation because when the phase is close to zero the chain may legitimately jump between 2π and 0. Such a large jump may appear to suggest

a bimodal likelihood, and hence a lack of convergence. Therefore to assess convergence we may first need to rotate the chain so that the mean, $\overline{\hat{P}}$, is π. Although we only need to do this if the chain has large jumps.

Another potential problem using MCMC estimation is caused because a sinusoidal seasonal pattern has two identical solutions as

$$-A\cos(t + P) = A\cos(t + P + \pi).$$

This could lead to the amplitude and phase chains jumping between these two equivalent solutions, which again complicates the assessment of convergence. To avoid this problem we restrict the amplitude to be positive ($A_t \geq 0$ for all t). We can do this by specifying

$$\begin{aligned} s_t &= A_t^* \cos(\omega_t - P), \\ A_t^* &= \max(0, A_t). \end{aligned}$$

When fitting the flexible model for the amplitude, (4.6), we can help to ensure that $A_t \geq 0$ by specifying that the mean amplitude is positive using the prior

$$a_0 \sim \mathrm{U}(0, 10^6).$$

In the model specified by (4.6), the smoothness of the year-to-year change in amplitude is determined by σ_α. Larger values of σ_α will result in greater year-to-year changes (greater non-stationarity). We can either estimate σ_α as part of the overall model by using a vague prior [21],

$$\sigma_\alpha \sim \mathrm{U}(0, 10^6),$$

or we can fix the value of σ_α in order to restrict the flexibility of A_t. Choosing the ideal value of σ_α is akin to choosing the $\boldsymbol{\tau}$ values for the model in Sect. 4.3. The degrees of freedom for a fixed value of σ_α can be estimated using the estimated number of parameters from the DIC (Sect. 1.6.2).

For the phase, we can specify a linear change over time using

$$P_t = \phi_0 + \phi_1 \tilde{t}, \qquad t = 1, \ldots, n,$$

For this model we use the priors

$$\begin{aligned} \phi_0 &\sim \mathrm{U}(0, 2\pi), \\ \phi_1 &\sim \mathrm{N}(0, \sigma_\phi^2). \end{aligned}$$

The prior for ϕ_0 is vague, and allows the phase at $t = 1$ to be any point on the circle. We use an informative prior for ϕ_1, to restrict the linear change in the phase to a plausible range of values. This is because very large changes in the phase are

unlikely (e.g., a change in the phase from January in one year to July in the next year). We restrict the linear change using the variance σ_ϕ^2. A standard deviation of $n\pi/288$ means the probability that the absolute change in the phase is greater than 1 month per year is 0.95 (using the scaled value of time, $\tilde{t}$). A larger standard deviation of $n\pi/144$ would mean that we believed that the change could be as large as 2 months per year.

4.4.2 Examples

4.4.2.1 Cardiovascular Disease Deaths

For the CVD data we assumed that the number of cases had a Poisson distribution. We used a log link, and the log-transformed numbers of days in the month as an offset. We used a burn-in of 20,000 MCMC samples, followed by 15,000 samples. The number of parameters and DIC for five different models are shown in Table 4.1.

A model with a stationary amplitude had a phase of 1.26 months (95% CI: 1.19, 1.33), with an amplitude risk ratio of 1.17 (95% CI: 1.16, 1.18). A linear model gave a worse fit compared with the stationary model, so it seems unlikely that the amplitude had a linear change over the years 1987–2000. The parameter for the linear change, $\hat{\alpha}_1$, had a mean of -0.0009, and 95% of -0.0110 to 0.0087, which supports this lack of linear change. The quadratic model was, likewise, no better.

Model (4.6) with a vague prior for σ_α gives a substantially better fit to the data as measured by the DIC. The mean value of σ_α was 0.134 (95% CI: 0.088, 0.207). We estimate the degrees of freedom for model (4.6) by subtracting its estimated number of parameters (p_D) from the estimated number of parameters for a stationary model. Using a fixed value of $\sigma_\alpha = 0.01$ resulted in 4.6 degrees of freedom for the non-stationary amplitude.

We show the non-stationary amplitudes for three models in Fig. 4.14. The model with the most flexible amplitude shows a large change in seasonality from year-to-year. The largest amplitude is a risk ratio of 1.29 in 1994 (95% CI: 1.26, 1.32), which was followed by the smallest of 1.03 in 1995 (95% CI: 1.01, 1.06). The pattern of

Table 4.1 Table of estimated number of parameters (p_D) and DICs for the cardiovascular disease death data for a range of different models for the amplitude. The lower the DIC, the better the model. The degrees of freedom for the amplitude is defined by df

Amplitude model	p_D	df	DIC
Stationary model, $A_t = \alpha_0$	5.0	1.0	2,757.2
Linear model, $A_t = \alpha_0 + \alpha_1\tilde{t}$	6.0	2.0	2,759.1
Quadratic model, $A_t = \alpha_0 + \alpha_1\tilde{t} + \alpha_2\tilde{t}^2$	7.0	3.0	2,757.8
Model (4.6), vague prior for σ_α	18.8	13.8	2,492.5
Model (4.6), fixed $\sigma_\alpha = 0.01$	9.6	4.6	2,680.9

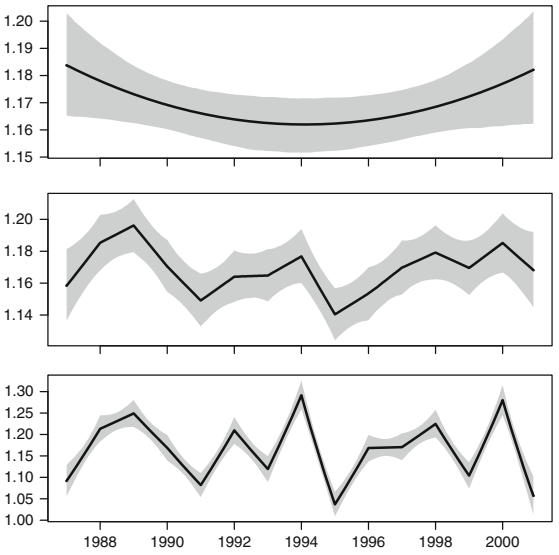


Fig. 4.14 Estimated non-stationary seasonal amplitude for the mortality risk ratio of the CVD data. Quadratic model (*top*), model (4.6) using a fixed $\sigma_\alpha = 0.01$ (*middle*), model (4.6) using a vague prior for σ_α (*bottom*). The the *solid line* is the mean estimate and the *grey area* the 95% credible interval. The y-axes show the risk ratios; the scales between panels are different

large amplitudes being followed by small amplitudes is also clear at other times. As discussed in Sect. 2.2.2 this may be because of the changes in the vulnerable pool.

We examined a non-stationary phase by first looking for a linear change, combined with model (4.6) using a vague prior for the change in amplitude (σ_α). The mean for ϕ_2 was -0.01, with a 95% credible interval of -0.06 to 0.04. The DIC for this model was 2,493.1 (based on 19.3 estimated parameters), which is slightly worse than the DIC of 2,492.5 for the model with a stationary phase. Hence there is little evidence of a linear change in the phase.

4.4.2.2 Exercise Data

An advantage of the model described in this section compared with the STL or non-stationary cosinor (Sect. 4.3) is that it can be applied to irregularly spaced data. We illustrate this by applying the model to the exercise data (Sect. 1.1.4).

A model with stationary amplitude and phase has an estimated phase of 1 July (95% CI: 28 June, 5 July), with an amplitude of 12 minutes (95% CI: 11, 13). This is a similar result to the cosinor model (Sect. 3.1). The DIC for this model is 203,481 based on 7.0 parameters.

We examined a model with a linear change over time in both the phase and amplitude. The linear change in the amplitude was estimated as 0.003, with a 95% CI of -0.011 to 0.016, hence there was little evidence of a change in the amplitude. The linear change in the phase was estimated as -1.018 (on a scale of radians), with a 95% CI of -1.082 to -0.956, hence there was strong evidence of a change in the phase. To understand this change we plot the phase and the seasonal pattern in Fig. 4.15. The phase in 2005 was after the middle of the year, the 2006 phase was

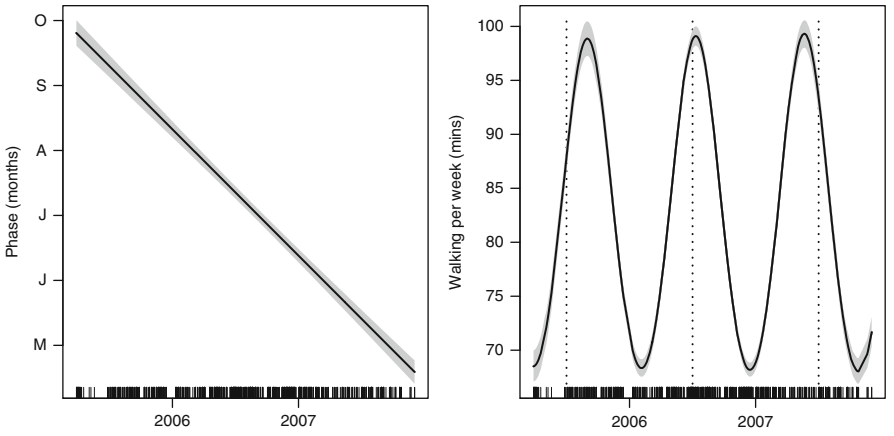


Fig. 4.15 Estimated linear change in the phase (*left*) and non-stationary seasonal pattern over time (*right*) for time spent walking. *Vertical dotted lines* are in the middle of each year (approximately 2 July)

close to the middle of the year, whereas the phase in 2007 was just before the middle of the year. The DIC for this model is 202,618 based on 9.1 parameters, which is a substantial improvement of 863 (with 2.1 extra parameters) compared with the model with a stationary phase.

4.5 Month as a Random Effect

In Sect. 2.3.3 we described a seasonal model that smoothed the results from neighbouring months using the CAR approach. In this section we expand on that model by including a trend for each month. This allows us to model a non-stationary seasonal pattern.

Again we use the model of trend, season and noise, (4.1), but this time the seasonal part of the model is defined by

$$\begin{aligned} s_t &= \delta_{m(t)} + \nu^*_{m(t)} y(t), \qquad t = 1, \ldots, n, \\ \nu^*_j &= \nu_j - \overline{\nu}, \qquad j = 1, \ldots, 12, \\ \nu_j &\sim \mathrm{N}(0, 10^6), \\ \boldsymbol{\delta} &\sim \mathrm{MVN}(\mathbf{0}, \mathbf{V}_\delta), \end{aligned} \tag{4.8}$$

where $m(t)$ and $y(t)$ are the month and year at time t, respectively. The variance–covariance matrix, $\mathbf{V}_\delta$, is defined by (2.2). The seasonal effect has been split into the monthly pattern, $\boldsymbol{\delta}$, and the change over time in this pattern, $\boldsymbol{\nu}^*$. We centre the estimates of $\boldsymbol{\nu}$ by subtracting the overall mean as otherwise these monthly trends

can become confounded with the overall trend ($\boldsymbol{\mu}$). We also used a centred version of year in (4.8),

$$\tilde{y}(t) = [y(t) - (Y/2)]/(Y/2),$$

where Y is the total number of years, in order to improve the convergence of the estimates of $\boldsymbol{\nu}^*$ when using MCMC estimation.

The model estimates the trend in each month independently of other months, which is similar to the STL (Sect. 4.2). Like the STL method, an important advantage of this model is that it does not assume a sinusoidal shape for the average seasonal pattern ($\boldsymbol{\delta}$). Also, it is possible to make future predictions with this model; a one season ahead prediction would be

$$\hat{s}_t = \hat{\delta}_{m(t)} + \hat{\nu}^*_{m(t)}\tilde{y}(t), \qquad t = n+1, \ldots, n+12.$$

Instead of restricting the change over time to be linear, we can smooth the estimates using a seasonal model defined by,

$$\begin{aligned} s_t &= \delta_{m(t)} + \nu_{\{m(t),y(t)\}}, \qquad t = 1, \ldots, n, \\ \boldsymbol{\delta} &\sim \text{MVN}(\mathbf{0}, \mathbf{V}_\delta), \\ \boldsymbol{\nu}_j &\sim \text{MVN}(\mathbf{0}, \mathbf{V}_\nu), \qquad j = 1, \ldots, 12, \end{aligned} \tag{4.9}$$

where $\mathbf{V}_\nu$ is a 12×12 variance–covariance matrix, with (j,k)th element defined by

$$V_\nu(j,k) = \begin{cases} \sigma^2_\nu, & j = k, \\ \rho\sigma^2_\nu, & |j-k| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

The mean $\overline{\boldsymbol{\nu}}_j$ is zero for every month. The values of $\overline{\boldsymbol{\nu}}$ describe the departures from the average seasonal pattern at each time. We use the same variance, σ^2_ν, for each month ($j = 1, \ldots, 12$), assuming that the overall departure from the mean seasonal estimate is constant across months.

4.5.1 Examples

4.5.1.1 Cardiovascular Disease Deaths

We applied a model with linear changes in each month to the CVD data. Figure 4.16 shows the average seasonal pattern, which is almost identical to the estimates from a stationary model but with narrower credible intervals (Fig. 2.21). The narrower credible intervals are due to some of the variability in season now being explained by the seasonal changes over time ($\boldsymbol{\nu}$).

The linear changes in each month on the scale of log relative risk are shown in Fig. 4.17. The relative risks estimates for February declined over time, whereas

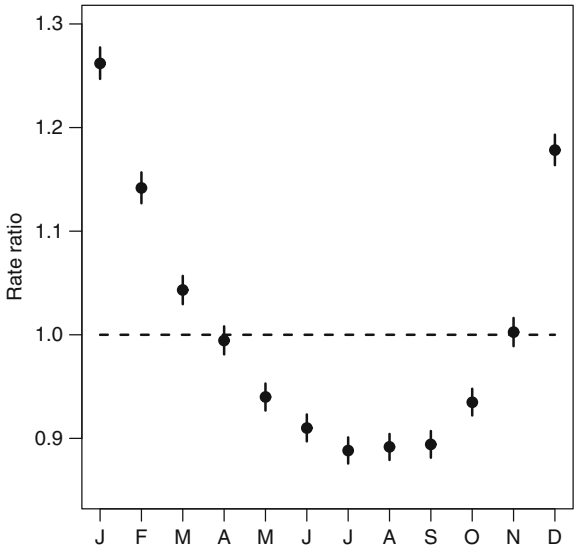


Fig. 4.16 Means and 95% credible intervals for the monthly relative risk estimates for the CVD death data. *Dashed horizontal line* at RR = 1

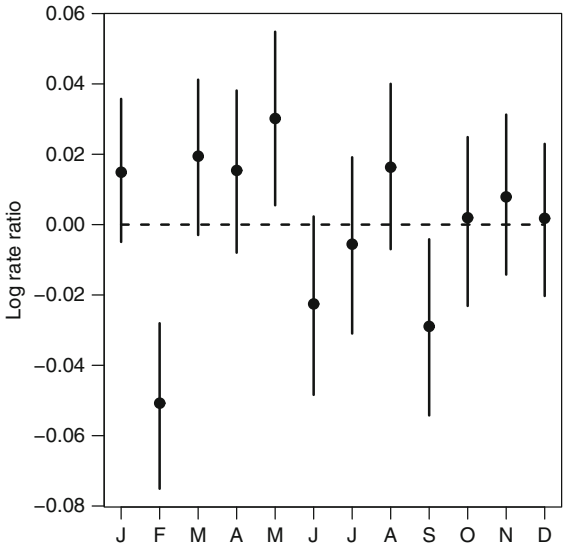


Fig. 4.17 Means and 95% credible intervals for the monthly estimates of linear change for the CVD death data ($\boldsymbol{\nu}^*$). *Dashed horizontal line* at zero indicates no change

those for May increased. To help interpret the changes over time in the seasonal pattern, we plot the mean seasonal pattern in Fig. 4.18.

The DIC for this model is 2,334.2 based on 24.9 parameters. This is better than any of the models based on a sinusoidal pattern shown in Table 4.1.

A linear change in the seasonal pattern over time may be too simple for these data. Accordingly we fitted model (4.9). Figure 4.19 shows the estimated changes in the seasonal pattern over time for each month ($\boldsymbol{\nu}$). The biggest changes were in January, February and December, and these changes were quite erratic from year

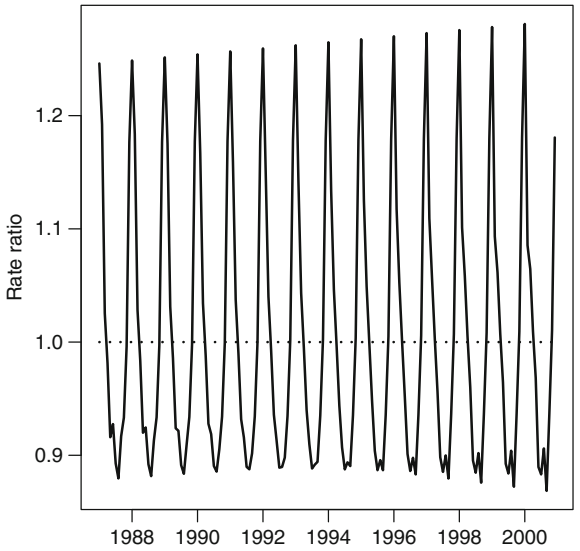


Fig. 4.18 Means seasonal pattern for the rate ratio of CVD death data using a non-stationary model 4.8

Table 4.2 Summary of the five different decomposition methods

Model (section)	Sinusoidal	MCMC estimation	CIs	GLM	Irregularly spaced data
Stationary cosinor (Sect. 4.1)	Yes	No	No	Yes	No
STL (Sect. 4.2)	No	No	No	No	No
Non-stationary cosinor (Sect. 4.3)	Yes	Yes	Yes	No	Yes
Modelling the amplitude and phase (Sect. 4.4)	Yes	Yes	Yes	Yes	Yes
Month as a random effect (Sect. 4.5)	No	Yes	Yes	Yes	No

CIs confidence/credible intervals, *GLM* generalized linear model

to year. This model had a DIC of 1,929.5 based on 182.9 parameters, indicating a big improvement in fit compared with the simpler model which assumed a linear change in months over time (but with a substantial increase in model complexity).

4.6 Comparing the Decomposition Methods

We summarise the five different methods described in this chapter in terms of five important characteristics in Table 4.2. Three of the methods assume a sinusoidal pattern in disease. This is a parametric assumption that may add strength to a weak signal, but it is also a fairly restrictive shape, that is symmetric about the x- and y-axes. As we have seen, this assumption does not fit the skewed seasonal pattern in CVD mortality.

Three of the models use MCMC estimation. One disadvantage of this method is the extra computation time. On the other hand, the extra computation means that confidence (or credible) intervals can be calculated for important parameters,

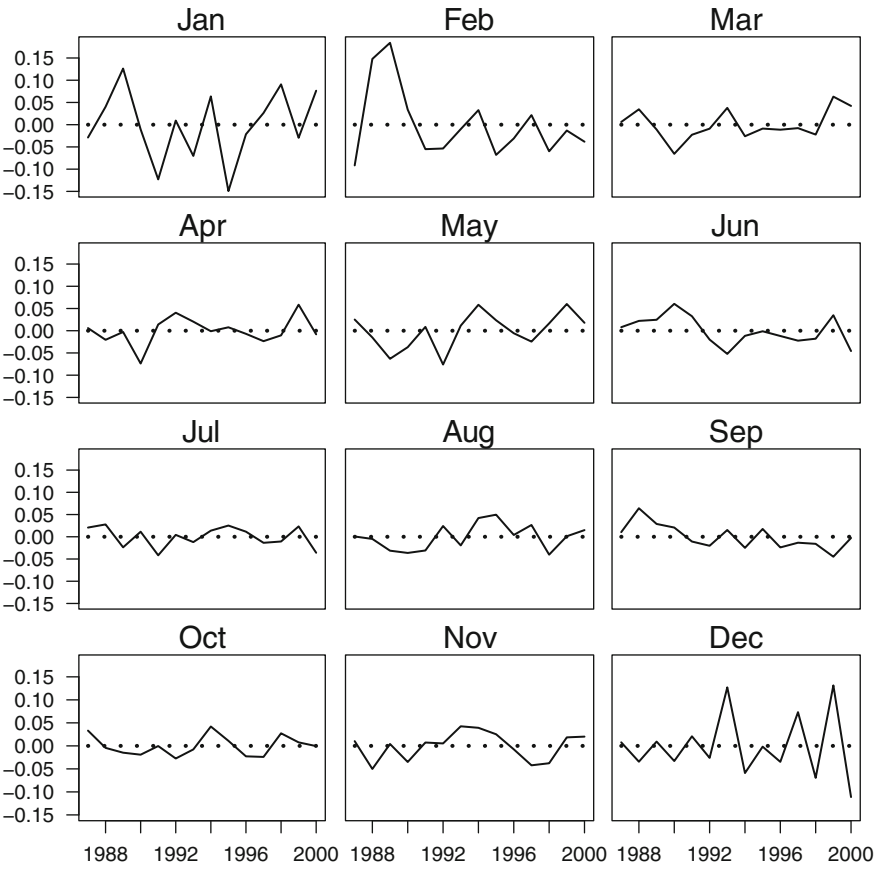


Fig. 4.19 Estimated changes over time in the monthly seasonal pattern for the CVD death data using model (4.9). y-Axes are on the same scale and show the log rate ratio. The *dotted horizontal lines* at a log rate ratio of zero (no change)

including the seasonal pattern. Three of the methods use a GLM framework, and hence are able to cope with non-Normal data (such as logistic or Poisson regression). Two of the methods are only able to be used on equally spaced time series data, whereas the others can be applied to unequally spaced survey data.

4.7 Exposures

So far, the methods in this chapter have been used to model the seasonal pattern in health outcomes without considering the seasonal pattern in the exposure (or exposures) behind the seasonal pattern in risk. In this section we describe methods for exploring the link between seasonal patterns in exposure and risk using

decomposition methods. We examine decomposing a putative exposure in the same way that we decomposed disease, and then comparing the trends and seasonal patterns. However, comparing seasonal patterns will not be straightforward if there is a non-linear relationship between risk and exposure.

4.7.1 Comparing Trends with Trends and Seasons with Seasons

For data on a daily timescale we can expand (4.1) to become

$$Y_t = \mu_t + s_t + d_t + \varepsilon_t, \qquad t = 1, \ldots, n,$$

where d_t is the short-term pattern in disease. The signal part of the time series is on three separate time scales: long-term trend (μ_t), season (s_t) and short-term (d_t). The short term variation encompasses any (non-noise) variation which is at a shorter than seasonal timescale. So it may include week-to-week or day-to-day variation.

The advantage of looking at the associations at different timescales is that associations at multiple levels increase the weight of evidence. Also, associations that may be confounded at one timescale, may not be at another. For example, the association between mortality and air pollution may be somewhat confounded by day of the week, as combustive pollutants increase on Mondays (when there are more cars on the road), and some causes of death also increase on Mondays. However, day of the week could not confound the long-term association between trends in mortality and air pollution, although it is possible that there are other confounders at this timescale, such as general improvements in health or medical treatment over time and reductions in the levels of some air pollutants.

In Chap. 5 we consider methods that control for season and trend, so that the short-term associations between exposure and disease can be estimated.

We present some examples comparing different timescales below. Other examples are given in [27] and [65, Sect. 5.3.1].

4.7.1.1 Cardiovascular Disease Deaths and Temperature

Day-to-day changes in mortality can be dependent on day-to-day changes in temperature [7]. We are interested in whether this association also appears at a seasonal and long-term level. Previously we looked at the association between monthly temperature and mortality using a scatter plot (Fig. 1.22). The plot showed a negative association, but this plot combines the long-term and seasonal variation.

Figure 4.20 shows the long-term estimate of temperature trend and the non-stationary seasonal pattern (based on monthly data). In general higher trends in temperatures (shown in Fig. 4.20) are associated with lower trends in cardiovascular deaths (shown in Fig. 4.11), and the Pearson's correlation is -0.84 (using monthly estimates of the trends). So there is evidence of higher temperatures being correlated

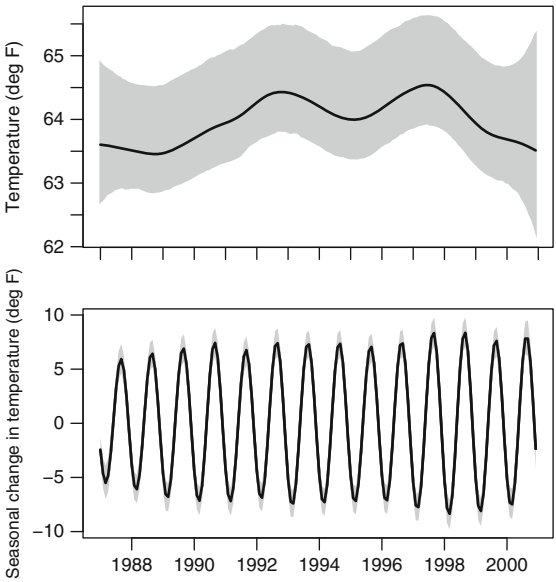


Fig. 4.20 Estimated trend (*top*) and annual season (*bottom*) over time in temperature (°F) from the CVD death data using a non-stationary cosinor with smooth parameters $\tau_1 = 1$, $\tau_2 = 30$ for temperature. The mean is a *solid line*, and the 95% confidence interval a *grey area*

with lower death rates in the long-term. This conclusion is in contrast to many fears about global warming and increased mortality. A plausible hypothesis is that higher temperatures cause lower blood pressures, and that lower blood pressure always prevents cardiovascular disease (no matter how low a person's level is already) [53].

To look for a correlation in seasons we first reduce the data by taking the highest mean seasonal temperature and lowest mean seasonal mortality in each year. This is because we are interested in the association between relatively warm years and relatively low levels of mortality. The Pearson correlation between these relative measures of season is 0.42 (based on 14 years). This is moderately suggestive that the warmest years were generally associated with lower numbers of deaths.

4.7.2 Exposure–Risk Relationships

In this section we examine exposure–risk relationships, and show that reasonably complex relationships between exposure and risk can lead to very complex seasonal patterns in risk [61].

We assume that the seasonal pattern in risk is due to a seasonal pattern in exposure described by the equation

$$s_t^r = f\left(s_t^e\right), \qquad t = 1, \ldots, n,$$

where s_t^r is the seasonal pattern in risk, s_t^e is the seasonal pattern in exposure, and $f(.)$ is the association between exposure and risk, the exposure–risk relationship (ERR). We assume that any trends in risk or exposure have been removed, and any noise. Both seasonal patterns are also subject to a zero sum constraint

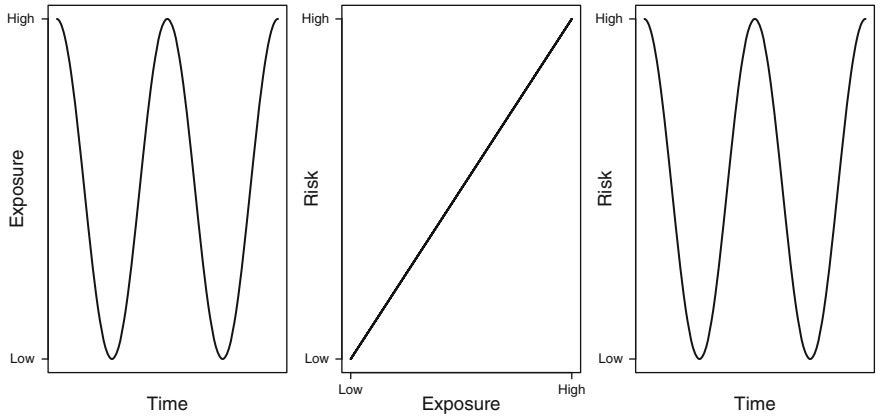


Fig. 4.21 Seasonal pattern in exposure over two years (*left panel*), negative linear exposure–risk relationship (*centre panel*), and seasonal pattern in risk (*right panel*)

$$\sum_{t=1}^{n} s_t^r = 0, \qquad \sum_{t=1}^{n} s_t^e = 0,$$

assuming that n/l is an integer, where l is the length of the seasonal cycle.

The seasonal pattern in risk will only match the seasonal pattern in exposure if $f(.)$ is the linear function, $f(x) = ax$, for some constant $a > 0$. Figure 4.21 shows this scenario with a sinusoidal seasonal exposure. High seasonal exposure is mirrored by high levels of risk, and low exposure by low levels of risk. The difference in the scales of exposure and risk is determined by a, if $a = 1$ then the scales are the same.

A linear exposure–risk relationship, $f(x) = ax$, but with a negative slope, $a < 0$, creates an inverse relationship between exposure and risk as shown in Fig. 4.22. The effect of the negative linear relationship is to delay the phase in risk by 6 months.

Non-linear exposure–risk relationships can lead to complex patterns in risk, as shown in Fig. 4.23. These risks were created using the same seasonal exposure shown in Fig. 4.22. A quadratic ERR creates a sinusoidal pattern in risk with double the number of cycles. A J-shaped ERR also creates double the number of cycles but with one amplitude much greater than the other. A threshold ERR creates a strongly non-linear seasonal pattern, with periods of no change in risk and seasonal "dips".

The results in Fig. 4.23 show that complex seasonal patterns in risk can be created by combining a relatively simple sinusoidal seasonal exposure with a non-linear ERR. Many researchers only examine the seasonal pattern in risk, and may not know or consider the seasonal pattern in exposure. A researcher finding a seasonal pattern in risk with two cycles per year may think that there are two separate seasonal exposures, one with a 12-month period, and another with a 6-month period. The results in Fig. 4.23 provide an alternative explanation, of one seasonal exposure with

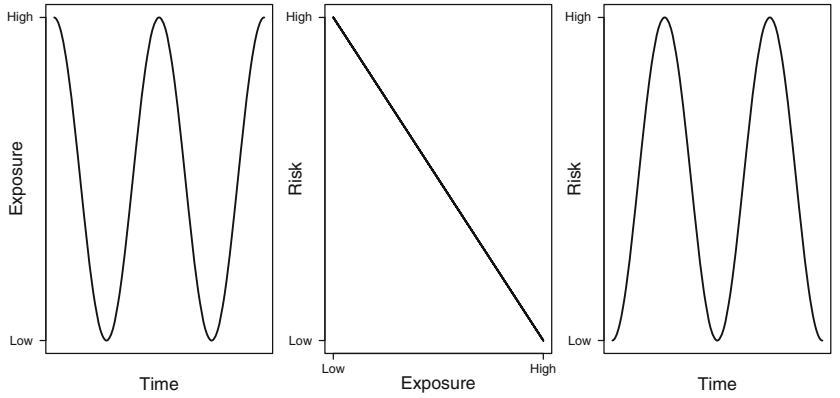


Fig. 4.22 Seasonal pattern in exposure over two years (*left panel*), positive linear exposure–risk relationship (*centre panel*), and seasonal pattern in risk (*right panel*)

a 12-month period and a non-linear ERR. This explanation is more parsimonious than assuming two separate exposures.

We made number of important assumptions to create these figures. We assumed that the seasonal exposure had a sinusoidal pattern, but different patterns would be observed if we used a sawtooth or spiked exposure (Sect. 2.1). We also assumed that there was no delay between exposure and risk. We could introduce a delay using the equation,

$$s_t^r = f\left(s_{t-k}^e\right), \qquad t = 1, \ldots, n,$$

where k is an integer, often referred to as the *lag*. This would have the effect of moving the phase in the risk to k time points after the phase in exposure.

An example of a sawtooth exposure and lagged pattern in risk is shown in Fig. 4.24. The seasonal pattern in risk has a sharp non-sinusoidal peak, that is out of phase with the peak in exposure.

4.7.2.1 Example

As an example of a non-linear ERR we consider the association between temperature and cardiovascular mortality. Temperature is a roughly sinusoidal exposure, but may have a non-linear ERR with mortality as shown in Fig. 1.22. The periodogram of CVD deaths shows a secondary peak at 6 months (Fig. 1.18), and the time series shows secondary peaks in mortality in the summer months (Fig. 1.1). These findings point towards a J-shaped association between temperature and risk [4]. Using the CVD death data a model with a J-shaped ERR is defined as

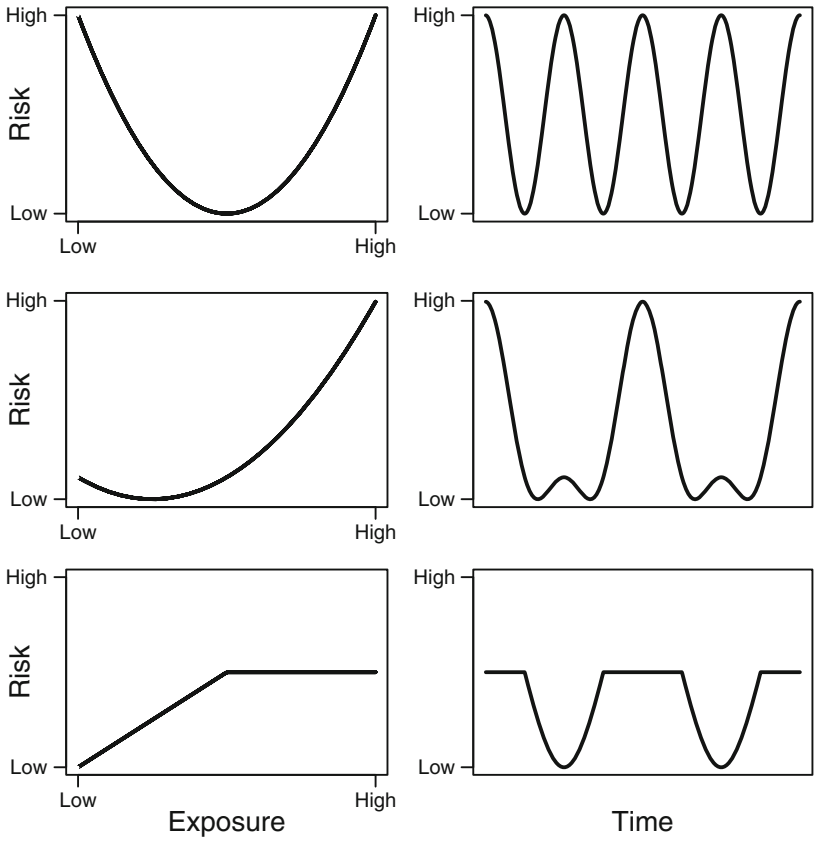


Fig. 4.23 Exposure–risk relationships (*left column*) and associated seasonal patterns in risk (*right column*) based on the seasonal exposure shown in Fig. 4.22

$$\begin{aligned} Y_t &\sim \mathrm{N}(m_t, \sigma_\varepsilon^2), \qquad t = 1, \ldots, n, \\ m_t &= \mu_t + \tilde{s}_t - \bar{\tilde{\mathbf{s}}}, \\ \tilde{s}_t &= s_t + \tau s_t^2, \\ s_t &= A_t \cos(\omega_t - P). \end{aligned}$$

In this model s_t is the seasonal exposure at time t and $\tilde{s}_t$ is the seasonal risk which is a quadratic function of the exposure. We subtract the mean seasonal risk ($\bar{\tilde{\mathbf{s}}}$) to maintain the zero sum constraint. We have created a quadratic ERR using only one extra parameter (τ). We gave this parameter a vague prior,

$$\tau \sim \mathrm{N}(0, 10^6).$$

We used a non-stationary amplitude (A_t) defined by (4.6) with a vague prior for σ_α, which is shown graphically in Fig. 4.14.

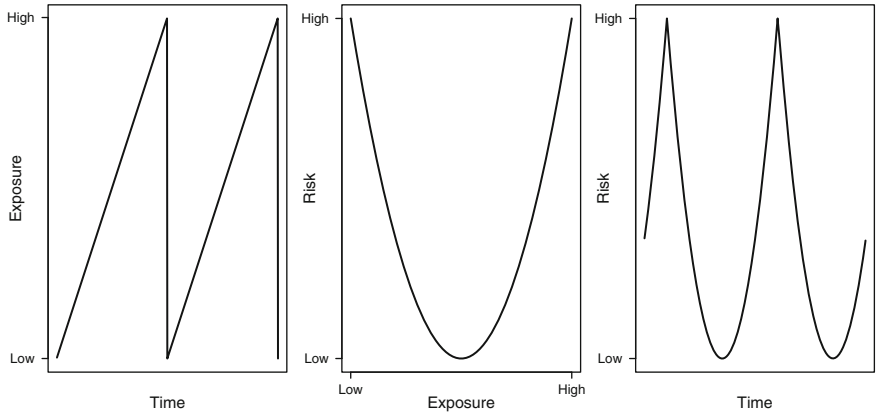


Fig. 4.24 Sawtooth seasonal exposure (*left panel*) quadratic exposure–risk relationship (*centre panel*) and associated lagged seasonal pattern in risk (*right panel*)

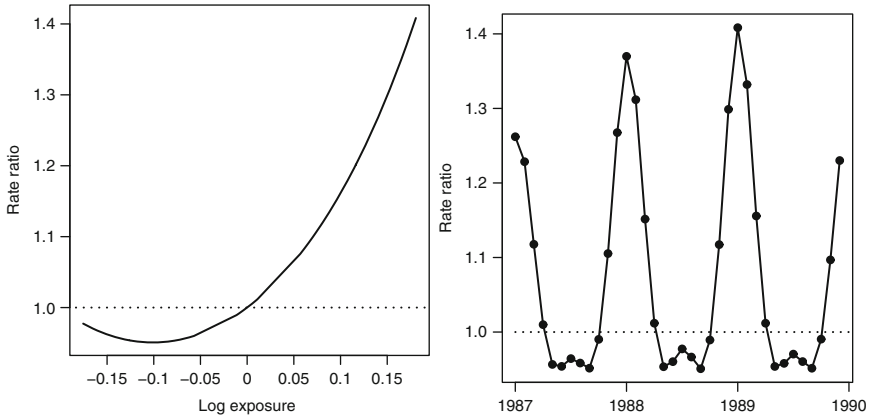


Fig. 4.25 Quadratic exposure–risk relationship between season and CVD death risk (*left*) and estimated seasonal pattern in CVD death risk for the years 1987–1989 (*right*)

The results of this ERR model are shown in Fig. 4.25. The ERR is J-shaped with a strong increase at high exposure (peak season) and lower increase at low exposure (low season). The effect on the estimated seasonal pattern is to produce a sharper peak in the increased risk in January, and a longer period of low risk from May to September which includes a slight increase in risk in July. The estimated seasonal pattern is now distinctly non-sinusoidal.

The mean estimate for τ was 5.0 (95% CI: 4.3, 5.6). The DIC for the ERR model is 2,276.5 based on 13.8 parameters. This is a substantial improvement of 404.4 from the DIC of 2,680.9 compared with a linear model (Table. 4.1). The large τ and big improvement in DIC provide strong evidence that a quadratic ERR is a better fit than a linear ERR.

Chapter 5
Controlling for Season

In some circumstances seasonality is not the focus of investigation, but is important because its effects need to be controlled for. This could be because both the outcome and the exposure have an annual seasonal pattern, but we are interested in associations at a different frequency (e.g., daily).

An example is the association between air pollution and death. Death has an annual seasonal pattern due to a complex array of causes (including temperature and influenza). Air pollution is a generic term for a mix of toxic substances found in the air. Ozone is just one of these substances, and it varies seasonally because it is formed by the combination of other air pollutants (often from traffic) and sunlight. As ozone has an association with sunlight then an important confounder is temperature. Ozone's association with traffic means that day of the week could also be a confounder.

The diagram in Fig. 5.1 shows these associations. In the following sections we demonstrate a number of different methods for modelling these associations at the same time as controlling for seasonal confounding. For illustrative purposes we only use one or two independent variables, whereas models in this area tend to include the delayed effects of pollution and temperature [65].

5.1 Case–Crossover

A useful method for controlling for seasonality in time series data is the *case–crossover* study design [58]. The idea is to compare "case" days when events occurred (e.g., deaths) with control days to look for differences in exposure that might explain differences in the number of cases. Control days are selected to be nearby to case days, which means that only recent changes in exposure are compared. This means that any long-term or seasonal variation in exposure can be eliminated. This elimination depends on the definition of nearby and on the seasonal and long-term patterns in exposure (we discuss these issues below).

In a case–crossover design each subject acts as their own control, and so any fixed characteristics (e.g., sex) are controlled for by design. This matching of subjects is akin to controlling for important confounders using a matched case–control design.

A.G. Barnett and A.J. Dobson, *Analysing Seasonal Health Data*, Statistics for Biology and Health, DOI 10.1007/978-3-642-10748-1_5,
© Springer-Verlag Berlin Heidelberg 2010

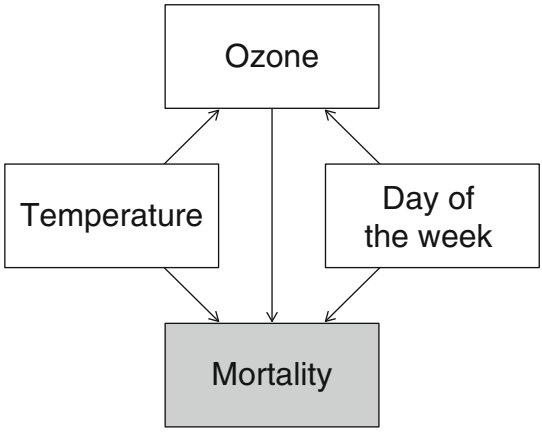


Fig. 5.1 Higher ozone is possibly associated with increased mortality and both have an annual seasonal pattern. Temperature and day of the week are important confounders

For both designs *conditional logistic regression* is used to calculate the odds ratio for cases compared with controls for a unit increase in exposure.

There are two main parts to a case–crossover analysis: (1) matching the controls days to the case days, (2) performing the conditional logistic regression. There are many different strategies for matching control days to case days. We use the *time-stratified* method because it has been shown to have good properties in terms of bias [47].

The first stage of the time-stratified method is to split the time series into equally sized non-overlapping strata covering the entire study time. Case days will only be compared with control days from the same strata. The stratum number for an observation at time t is given by

$$w_t = \lfloor t/l_s \rfloor + 1, \qquad t = 1, \ldots, n, \tag{5.1}$$

where l_s is the strata length, n is the length of the series and $\lfloor a \rfloor$ is the integer part of a. Unless n/l_s is an integer the last stratum will have fewer than l_s observations.

The length of the strata (l_s) needs to be short enough to remove the seasonal patterns from both the dependent and independent variables, but not so short that the exposure for the cases and controls is too similar. As an example of the effect of strata length, we look at the seasonal exposure of ozone and CVD mortality in the Los Angeles data. Ozone levels change from day-to-day as they depend on traffic and sunlight (amongst other things). The dependence on sunlight means that levels tend to peak in summer and dip in winter. However, for this example we are not interested in the seasonal exposure in ozone, but in whether day-to-day changes in ozone are associated with changes in mortality. Ozone is measured in parts per billion (ppb).

The left panel of Fig. 5.2 plots the daily 8-hour ozone levels and the mean 8-hour ozone levels in each 28-day stratum. The right panel shows the daily numbers of CVD deaths and the means in each 28-day stratum. The means have captured the strong seasonal pattern in ozone, but within each stratum there is still some variation

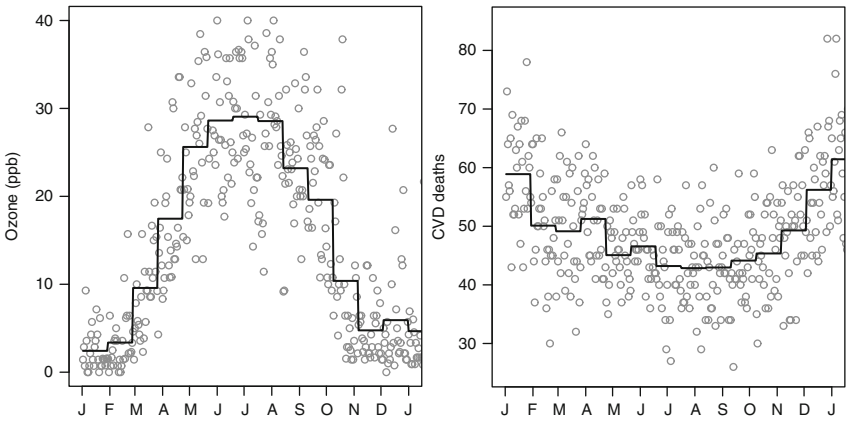


Fig. 5.2 Daily 8-hour ozone levels (*left panel*) and number of CVD deaths (*right panel*) in Los Angeles in 1987. Daily means are shown as *grey dots*, and the mean in each 28-day strata as a *solid line*

in ozone levels. It is this variation, not the larger seasonal variation, that the case–crossover design will use to relate to the de-seasoned variation in CVD mortality. The seasonal pattern in mortality is the reverse of the seasonal pattern in ozone, as it is lowest in summer.

For comparison we show the means using strata twice the length (56 days) in Fig. 5.3. The main seasonal pattern is still captured by the means, but the sharp rise in ozone levels from March to April has not been captured. Using this wider stratification would mean that some of the differences in exposure between cases and controls would be due to seasonal differences, which we are trying to eliminate.

Splitting the time series into strata follows the same principle as calculating the mean in each month, which we discussed in Sect. 2.2.1.

After selecting the stratum length, we need to decide which days to use as controls within the strata. The key consideration is the autocorrelation in exposure, as this determines the similarity in exposure for case days and control days. For the ozone example, we would expect the results to be positively autocorrelated for nearby days. If we select control days that are too close to the case day then we run the risk of over-matching. We can reduce the correlation by using an exclusion window around each case day.

As an example, Fig. 5.4 shows a case–crossover scheme with a stratum length of 28 days and an exclusion period of 4 days. In the top row the case day is the first day of the stratum, and this is compared to days 6–28 (23 controls). In the third row the case day in the fifth day of the stratum, and this is compared to day 1 and days 10–28 (20 controls).

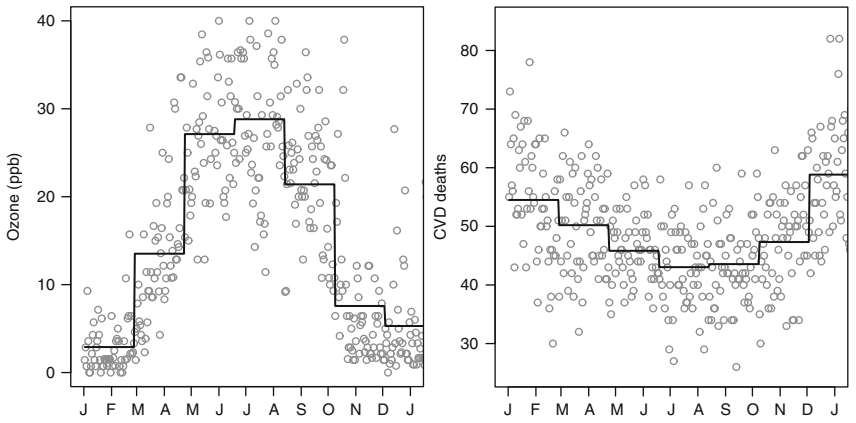


Fig. 5.3 Daily 8-hour ozone levels (*left panel*) and number of CVD deaths (*right panel*) in Los Angeles in 1987. Daily means are shown as *grey dots*, and the mean in each 56-day strata as a *solid line*

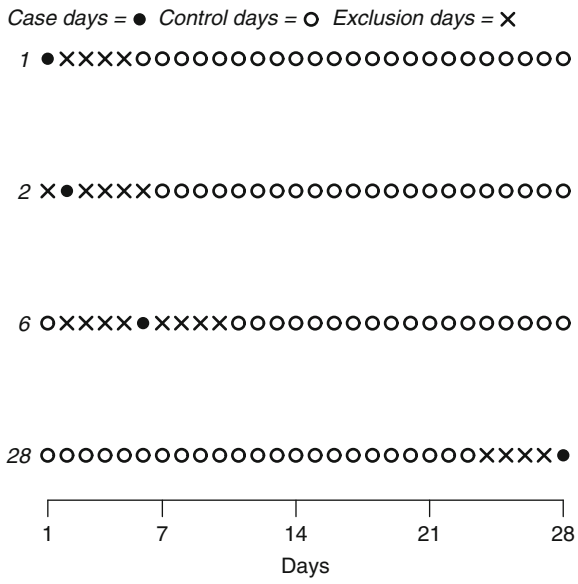


Fig. 5.4 Graphical representation of case and control days for a time-stratified case–crossover design using a stratum length of 28 days. Examples for four different case days. The number on the *left* indicates the case day

5.1.1 Matching Using Day of the Week

An alternative scheme is to match cases and controls by day of the week. Figure 5.5 shows some examples using a stratum length of 28 days. The four examples are based on cases in the first week of the stratum. Each case has three controls. When matching on day of the week, a consistent number of controls is achieved by using a stratum length (l_s) that is a multiple of seven.

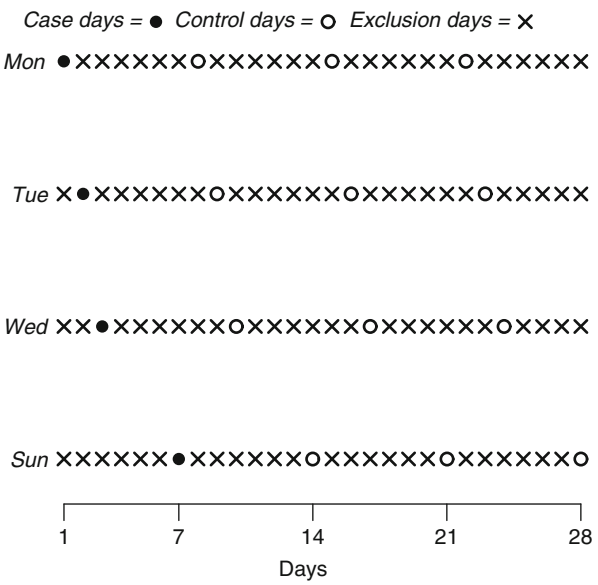


Fig. 5.5 Graphical representation of case and control days for a time-stratified case–crossover design using a stratum length of 28 days, and matched by the day of the week

An advantage of matching on day of the week is that the large gaps between observations will reduce the autocorrelation in exposure [47]. Also, this extra matching removes any confounding caused by day of the week. Such confounding is possible if the dependent and independent variable *are both* dependent on day of the week. Mortality has been shown to vary by day of the week, although for CVD mortality the variation is small and may be due to data quality issues [6]. Suicide deaths do vary strongly by day of the week, with big increases on Mondays possibly due to the pressures of work [49]. An example of an independent variable with a weekly exposure pattern is air pollution, as levels are generally higher on weekdays due to greater volumes of traffic.

5.1.2 Case–Crossover Examples

To illustrate the case–crossover design we use the daily CVD data (Sect. 1.1.1). We examine the association between ozone and mortality, both of which have a strong annual seasonal pattern, but with different phases resulting in a negative association overall which may mask any short-term positive association (Fig. 5.2).

The R code to run a time-stratified case–crossover analysis is

```
> CVDdaily$o3mean.10<-CVDdaily$o3mean/10
> CVDdaily$tmpd.5<-CVDdaily$tmpd/5
```

```
> model<-casecross(cvd~o3mean.10+tmpd.5
  +Mon+Tue+Wed+Thu+Fri+Sat,data=CVDdaily,
  stratalength=28,exclusion=4)
```

We have used a 28-day stratum length with an exclusion period of 4 days as shown in Fig. 5.4. We have attempted to control for the confounding effect of mean temperature (`tmpd`) by adding it as a linear independent variable. We have controlled for day of the week by adding separate binary variables for each day of the week except Sunday, which becomes the reference day. We first scaled the results so that the estimates for Ozone correspond to a 10 ppb increase, and for temperature to a 5°F increase (Sect. 1.4.2). This R output is shown below.

```
Number of case days 5114
Average number of control days per case day 19.7
                 coef exp(coef)    se(coef)          z       p
o3mean.10 -0.007128078 0.9928973 0.003660678 -1.947201 5.2e-02
tmpd.5     0.009956200 1.0100059 0.002986518  3.333715 8.6e-04
Mon        0.033562804 1.0341324 0.008007677  4.191329 2.8e-05
Tue        0.015620390 1.0157430 0.008091139  1.930555 5.4e-02
Wed       -0.015751216 0.9843722 0.008108262 -1.942613 5.2e-02
Thu       -0.011979920 0.9880916 0.008091265 -1.480599 1.4e-01
Fri        0.009771223 1.0098191 0.008067023  1.211255 2.3e-01
Sat        0.014914884 1.0150267 0.007879267  1.892928 5.8e-02
```

The number of case days is the length of the time series. The values in the `coef` column are the log odds ratio, and in `exp(coef)` the odds ratios. A 10 ppb increase in ozone is associated with a decreased risk of death (odds ratio = 0.993), and this decrease is of borderline statistical significance (p-value = 0.052). The odds ratio for Mondays, relative to Sundays, is 1.034, and this increase is strongly statistically significant (p-value < 0.001).

The results from matching on day of the week (using a stratum length of 28) are shown below.

```
Number of case days 5114
Average number of control days per case day 3
               coef exp(coef)    se(coef)         z      p
o3mean.10 -0.007686579 0.9923429 0.003886986 -1.977517 0.0480
tmpd.5     0.009497229 1.0095425 0.003106989  3.056731 0.0022
```

The average number of controls per case is now only 3, compared with 23.2 for the previous scheme. We do not have any estimates for the days of the week, as they are controlled for by design. The estimated effect of ozone and its standard error are very similar to the previous analysis.

This effect of ozone is unexpected, as every 10 ppb increase significantly reduces the risk of death, odds ratio = 0.992, 95% CI: 0.984, 0.999. This means that higher levels of ozone are associated with fewer deaths. This strange result may have been caused by some remaining confounding between ozone and temperature (Fig. 5.1). Higher ozone levels are associated with higher temperatures, and higher temperatures are generally associated with fewer CVD deaths. We discuss further ways to remove this possible confounding below.

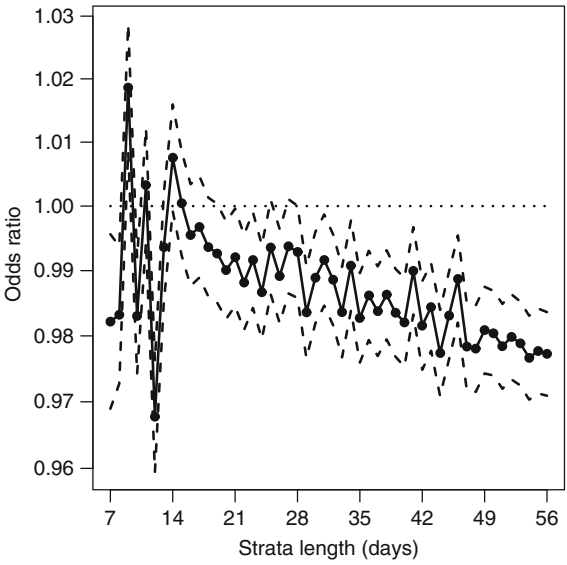


Fig. 5.6 Mean odds ratio (and 95% confidence intervals – *dashed lines*) from a case–crossover model for the effect of a 10 ppb increase in ozone on daily mortality for stratum lengths from 7 to 56 days. *Dotted horizontal line* at an odds ratio of 1 corresponds to the null hypothesis of no association with CVD mortality

5.1.3 Changing Stratum Length

One possible reason for the negative association between daily levels ozone and CVD mortality is that we have not sufficiently removed the seasonal variation in ozone. We can investigate the level of control for season by using different stratum lengths.

Figure 5.6 shows the odds ratio for the effect of a 10 ppb increase in ozone plotted against stratum length. For every stratum we used an exclusion period of 4 days. There are two main features in the plot: the erratic estimates for stratum lengths between 7 and 14 days, and the steady increase in the negative effect of ozone for stratum lengths from 14 to 56. The erratic estimates possibly occurred because these very narrow strata have very few controls, and a relatively small within-stratum variance in CVD mortality and ozone. The wider stratum lengths (from 14 days onwards) are associated with greater negative associations between ozone and mortality. As the strata become wider, the control for the seasonal pattern in ozone and mortality becomes weaker (compare Figs. 5.2 and 5.3). Hence there is a greater potential for the negative seasonal association to dominate any shorter-term daily association.

5.1.4 Matching Using a Continuous Confounder

In the previous example we controlled for the confounding effect of temperature on mortality by adding it as an independent variable. However, temperature and ozone are very strongly correlated, and hence there may be some *collinearity* between

these two independent variables. This collinearity should have been somewhat controlled for by the stratification, but even within 28-day periods there may be some correlation between temperature and ozone.

We can control more rigidly for the effects of temperature by only selecting control days within a similar temperature range as the case day. A narrower range corresponds to stricter matching, and this is more likely to remove the effect of the confounder, but it will also reduce the available number of control days.

The results of matching cases and controls by a difference of ± 1°F (using a stratum length of 28) are shown below. In this example we do not match on day of the week as that would further reduce the number of available controls.

```
Number of case days with available control days 4324
Average number of control days per case day 4.8
               coef exp(coef)    se(coef)        z     Pr(>|z|)
o3mean.10 0.02769832  1.028085 0.004503147 6.150880 7.705427e-10
Mon       0.06749914  1.069829 0.009781503 6.900692 5.174972e-12
Tue       0.04427209  1.045267 0.009978894 4.436573 9.140250e-06
Wed       0.02336136  1.023636 0.010047078 2.325189 2.006183e-02
Thu       0.03099919  1.031485 0.010002281 3.099212 1.940360e-03
Fri       0.04346522  1.044424 0.009728080 4.468016 7.894844e-06
Sat       0.03629306  1.036960 0.009544978 3.802319 1.433477e-04
```

The effect of a 10 ppb increase in ozone is now to increase the risk of death (OR = 1.028), and this increase is statistically significant (p-value < 0.001). The number of case days is 4324, which is smaller than the total number of days (5114), because there were 790 days (15.4%) for which a matching control day was not available. We do not include mean temperature as a dependent variable as it is controlled for by design.

An advantage of matching using a confounder is that the shape of the association between the confounder and the dependent variable is not important. This means the association can take any shape (e.g., non-linear).

Figure 5.7 shows the mean odds ratio for ozone using a model that matches on temperature for a range of stratum lengths. The smallest stratum length is 8 (whereas in Fig. 5.6 it is 7) as a model based on a stratum length of 7 did not have enough controls per case.

After matching on temperature to within a degree, ozone is almost always associated with an increased risk of mortality, regardless of the stratum length. The size of this increased risk becomes closer to the null as the stratum length increases. Again this possibly indicates a seasonal confounding that, in this case, is independent of temperature.

5.1.5 Non-linear Associations

The association between temperature and CVD mortality is known to be *non-linear*, with increases at hot and cold temperatures [4]. In most of the examples so far we have assumed that the effect of temperature is linear. The exception was in the previous example, when we matched on temperature.

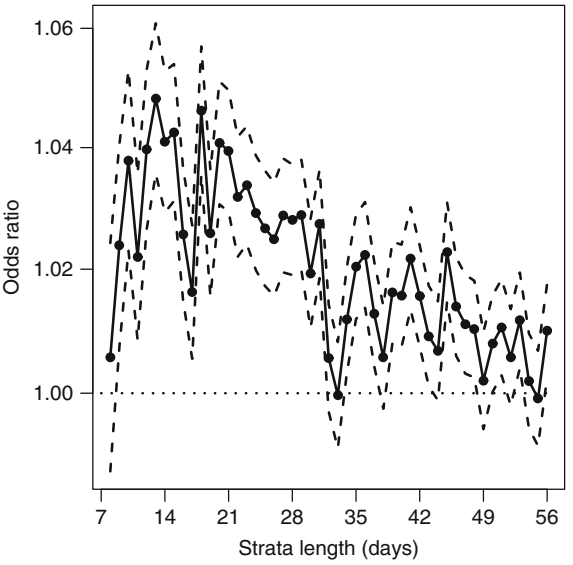


Fig. 5.7 Mean odds ratio (and 95% confidence intervals – *dashed lines*) from a case–crossover model for the effect of a 10 ppb increase in ozone on daily mortality for stratum lengths from 8 to 56 days. Control days were matched to case days by using a mean temperature difference $\pm 1^{\circ}$F. The *dotted horizontal line* at an odds ratio of 1 corresponds to the null hypothesis of no change in mortality

We can add a non-linear effect to a case–crossover analysis by including an interaction term. In the example below we have added an interaction between temperature and season (using the meteorological definition, Sect. 2.2.2). To make sure that the temperature estimates and strata overlap, instead of (5.1) we use a stratum number based on year and month

$$w_t = [y(t) - \min(\text{year})] \times 12 + m(t), \qquad t = 1, \ldots, n, \tag{5.2}$$

where $y(t)$ is the year of observation t and $m(t)$ is the numeric month (Sect. 2.3.1).

The results are shown below.

```
Number of case days 5114
Average number of control days per case day 22.1
                  coef exp(coef)   se(coef)         z       p
o3mean      -0.01397586 0.9861213 0.00363599 -3.843757 1.2e-04
tmpd.winter -0.01056041 0.9894952 0.00466591 -2.263313 2.4e-02
tmpd.spring  0.02210746 1.0223536 0.00590697  3.742607 1.8e-04
tmpd.summer  0.05040559 1.0516976 0.00905725  5.565219 2.6e-08
tmpd.autumn  0.01430658 1.0144094 0.00583980  2.449840 1.4e-02
Mon          0.03046140 1.0309301 0.00801143  3.802243 1.4e-04
Tue          0.02282887 1.0230914 0.00809855  2.818885 4.8e-03
Wed          0.00173084 1.0017323 0.00812288  0.213082 8.3e-01
Thu          0.00652781 1.0065492 0.00809858  0.806044 4.2e-01
Fri          0.01631883 1.0164527 0.00806924  2.022351 4.3e-02
Sat          0.01427699 1.0143794 0.00788437  1.810797 7.0e-02
```

So in winter an increase in temperature leads to a decrease in risk of CVD mortality (OR = 0.989), whereas in the other months, particularly summer, an increase in

temperature leads to an increase in risk. The differences in the temperature effects are relatively large, and this model probably controls better for the effect of temperature than the previous model assuming a linear risk. With this increased level of control for confounding, the effect of ozone has unexpectedly become more strongly protective.

5.2 Generalized Additive Model

In this section we show how to use a generalized additive model (GAM) to control for seasonal patterns. GAMs are semi-parametric models, which means they have parametric and non-parametric parts. It is this combination that makes them useful for flexibly modelling non-linearity. Here we aim to model the non-linear seasonal pattern in the dependent variable.

We do not cover the details of GAMs. For a more detailed description of GAMs see the books by Keele [52], Ruppert et al. [73] or Wood [88].

5.2.1 Definition of a GAM

We define a GAM by adding a smooth function to the equation for a GLM, (1.9),

$$g(Y_t) = \beta_0 + \beta_1 X_{1t} + \cdots + \beta_p X_{pt} + s(t, \lambda), \qquad t = 1, \ldots, n,$$

where $s(t, \lambda)$ is some smooth function of time (t), the smoothness of which is controlled by the degrees of freedom, $\lambda \geq 1$. Larger values of λ lead to more non-linear or flexible (bendy) functions; smaller values of λ lead to more linear functions, with $\lambda = 1$ corresponding to a linear function.

We will use the smooth function, $s(t, \lambda)$, to model the trend in Y_t and any annual seasonal pattern. We aim to design $s(t, \lambda)$ so that it removes all annual seasonal variation from the data, so that the other independent variables (e.g., X_{1t}) represent associations at a shorter time frequency. This will only be true if $s(t, \lambda)$ completely removes the annual seasonal variation. So choosing the correct value of λ is akin to choosing the correct stratum length in the case–crossover analysis. The difference is that the case–crossover design removes seasonality using a step-like function (Fig. 5.2), whereas the GAM uses a smooth function.

5.2.1.1 An Example of Applying a Seasonal GAM

In this section we show how to use a spline to model the roughly sinusoidal pattern in cardiovascular mortality (Fig. 1.1). As time is the independent variable, the degrees of freedom needed will depend on the length of time (longer time series needing more degrees of freedom). Also, because we know the seasonal pattern repeats every year it is helpful to define the degrees of freedom as the degrees of freedom per year.

This makes the degrees of freedom needed to model seasonality more comparable when using time series of different lengths.

To illustrate the effect of increasing the degrees of freedom, we show a fitted spline for the first three years of CVD data. We show the splines for one to six different degrees of freedom per year (Fig. 5.8). One degree of freedom per year gives an almost linear spline that only captures the long-term trend in mortality. Using two degrees of freedom per year gives a spline that has captured the basic seasonal pattern, but the phases of the spline do not coincide with the peaks in observed mortality in late January. Increasing the degrees of freedom gives a much closer fit to the seasonal pattern in the data. For five or six degrees of freedom per year, the spline has also captured the smaller summer increase in CVD deaths.

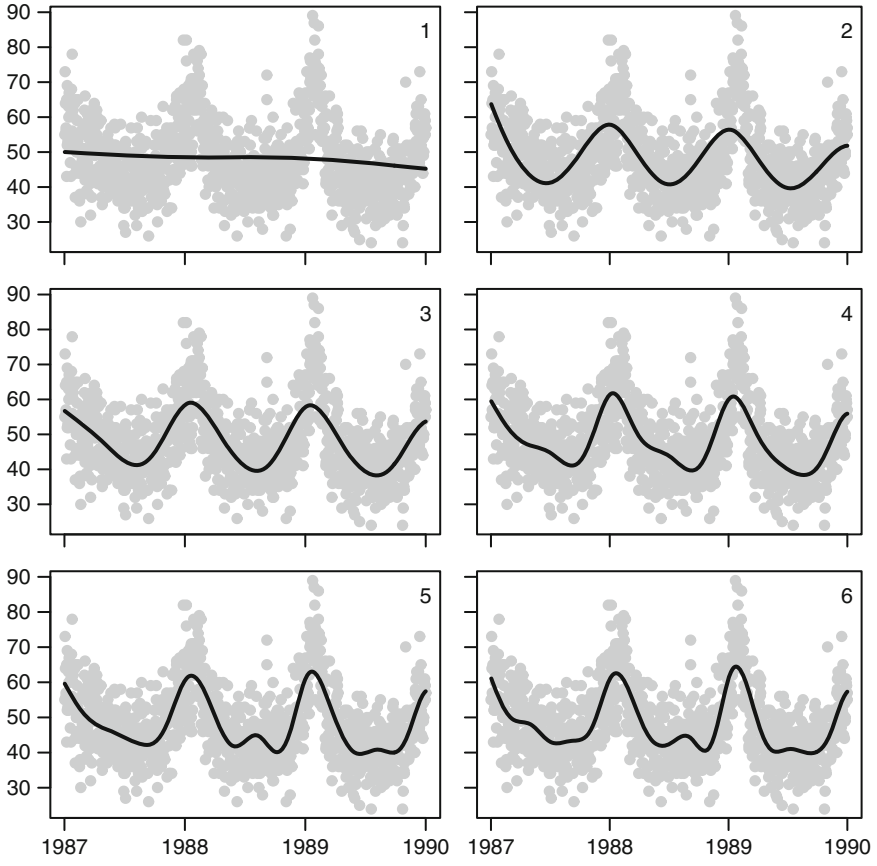


Fig. 5.8 Observed daily cardiovascular mortality (*grey dots*) and estimated spline (*solid line*) for increasing values of the number of degrees of freedom per year (number in *top-right corner*). First three years of data (1987–1989) for the CVD data. The y-axes show the daily number of deaths

To fit these splines we used the GAM functions available in the `mgcv` R library [88]. The R command is

```
> gam(cvd~s(as.numeric(date),k=df,bs='cr'),data=CVDdaily,
  family=poisson(),scale=-1)
```

The CVD data are the counts of daily deaths, which we assume follow a Poisson distribution and use the log-link. This means the estimated effects are given as rate ratios (Sect. 1.4.5). The `scale=-1` command means that we use an over-dispersed Poisson model (Sect. 1.4.5). The spline is a *penalized cubic regression* function; the cubic basis is specified using `bs='cr'`. This allows a cubic pattern between any two knots; in the previous section we used a quadratic basis (Fig. 1.29). The value `df` is the maximum degrees of freedom, calculated as the number of years multiplied by the degrees of freedom per year, plus one (so that the model is identifiable). The actual degrees of freedom (and hence the flexibility of the line) is chosen by penalizing the non-parametric part of the spline equation (1.10). The optimal size of the penalty is chosen by *generalized cross-validation* [73, 88]. Cross-validation is a robust estimation technique that aims to avoid giving estimates that are too closely tailored to the data. For a spline this would mean a very flexible estimate that had modelled idiosyncratic parts of the data. Cross-validation works by leaving out each observed value in turn and then predicting this value based on the remaining data. Minimising the squared residuals from these predictions gives estimated splines that are not overly flexible.

Next, we look at the effect of controlling for season on the estimated effect of daily ozone. In this model we included independent variables for mean daily ozone (`o3mean`), mean daily temperature (`tmpd`) and day of the week. Figure 5.9 shows the estimated rate ratios for ozone against the degrees of freedom per year for the spline using one to six degrees of freedom per year. As in Fig. 5.6, as we increase the seasonal control, the negative effect of ozone disappears. So greater seasonal control gives an estimated effect for ozone that is closer to the null value of no effect.

An advantage of GAMs over the case–crossover analyses is that it is possible to check the residuals of a GAM. This is not possible using the case–crossover because we cannot calculate the fitted values using conditional logistic regression (here the estimated daily number of deaths). We can check the residuals from the GAM for any remaining seasonal pattern using the ACF or cumulative periodogram. The residuals from a model with six degrees of freedom per year, and independent variable of daily ozone, temperature and day of the week, show a clear pattern using either the cumulative periodogram or ACF (Fig. 5.10). Something is still missing from the model, possibly an inadequate control for temperature. Adding temperature for the previous 1–3 days may help [65].

5.2.2 Non-linear Confounders

We can extend our previous GAM to incorporate non-linear estimates for independent variables other than time. The equation to include other smooth independent variables is

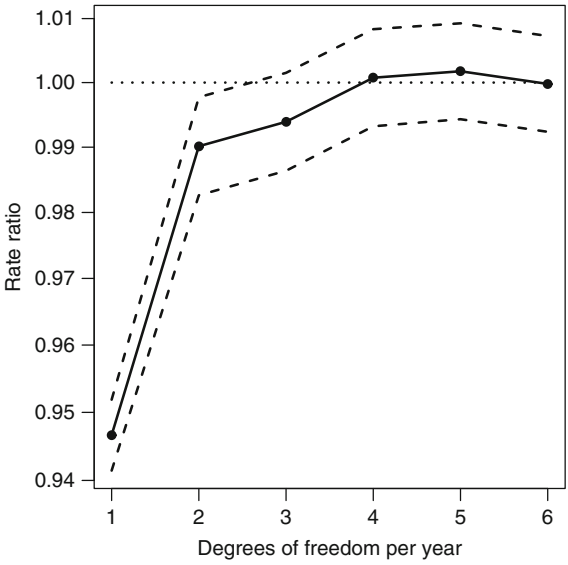


Fig. 5.9 Mean rate ratio (and 95% confidence intervals – *dashed lines*) from a GAM for the effect of a 10 ppb increase in daily ozone on daily CVD mortality against the degrees of freedom per year using a smooth effect of time. The *dotted horizontal line* at a rate ratio of 1 corresponds to the null hypothesis of no change in mortality

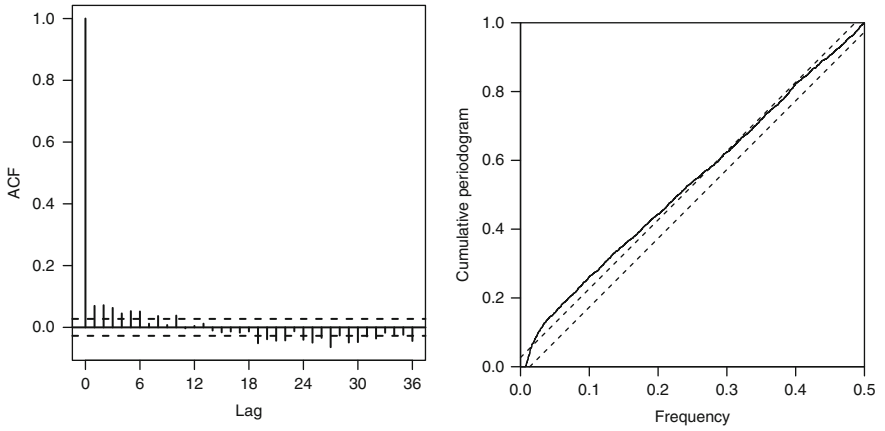


Fig. 5.10 ACF and cumulative periodogram of the residuals from a GAM with a dependent variable of daily CVD mortality, independent variables of daily ozone, temperature and day of the week, and using six degrees of freedom per year to control for trend and season

$$g(Y_t) = s(X_{1t}, \lambda_1) + \beta_2 X_{2t} + \cdots + \beta_p X_{pt} + s(t, \lambda_2), \qquad t = 1, \ldots, n.$$

In our example it could be particularly useful to fit the effect of temperature using a spline, as we know the effect is non-linear.

Figure 5.11 shows the estimated effect for temperature assuming three degrees of freedom. We choose 3 as we know that temperature has a roughly U-shaped association with mortality. We fitted a penalized spline using a cubic basis, and

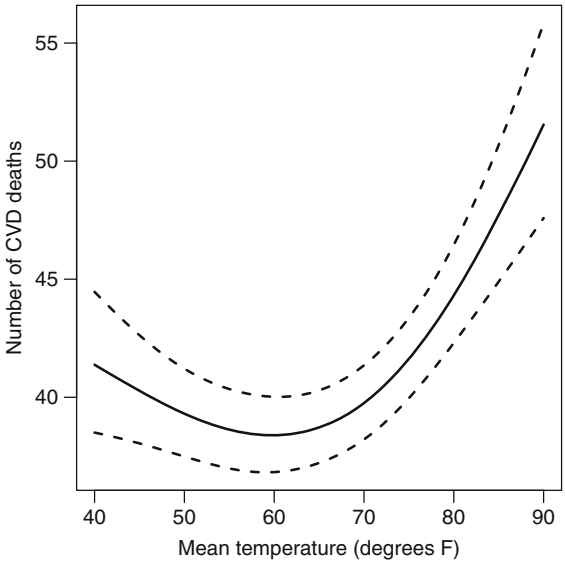


Fig. 5.11 Mean number of CVD deaths (and 95% confidence interval) plotted against mean temperature (°F): effect estimated using a penalized regression spline with three degrees of freedom

fitted the same type of spline for time, with four degrees of freedom per year (to control for season and trend).

The effect of temperature is strongly non-linear, with increases in mortality at cold and hot temperatures. The lowest average number of deaths occurs around 60°F. The AIC for a GAM using a linear effect for temperature is 35,002. The AIC using smoothed temperature is 34,923, indicating a big improvement in fit of 79.

5.3 A Spiked Seasonal Pattern

It is possible for a disease to have two or more seasonal patterns. In this section we show how to separate these patterns when one pattern is a short-term spike in disease and the other is a longer-term seasonal pattern. As an example, we examine the short-term increase in CVD at Christmas, after controlling for the longer-term seasonal increase in winter.

It is first useful to plot the daily number of CVD deaths over the Christmas period and surrounding days. We define the Christmas period as 10 days either side of Christmas day (15 December to 4 January), which includes New Year's day. Wider periods can be used, but they would put the Christmas increase in the wider context of the overall seasonal increase in deaths [68]. However, a shorter period means that we are comparing days with generally similar temperature and season. This definition of a comparison period is similar to choosing the length of the strata for the case–crossover analysis (Sect. 5.1.3).

Figure 5.12 is a box plot of the number of deaths over our defined Christmas period. To remove the effect of long-term trend, we first subtracted the mean number

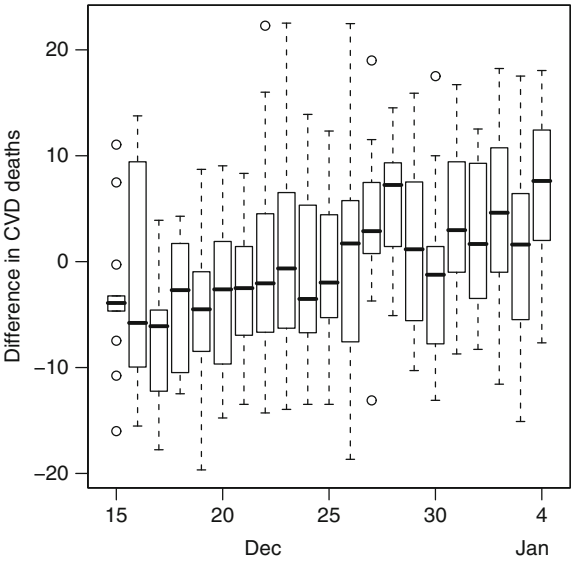


Fig. 5.12 Box plot of the difference in the daily number of CVD deaths 10 days either side of Christmas day (15 December to 4 January). The y-axis shows the difference in the daily number of deaths from the average number of deaths in the same Christmas season

of deaths in the same Christmas period. This is akin to the stratification used by the case–crossover analysis (Sect. 5.1).

The plot shows a generally increasing trend in deaths from mid-December to early January. There is no clear increase in deaths at Christmas, but there may be a small increase on the 27 and 28 December.

To further investigate the trend we consider the predicted values using five degrees of freedom from Fig. 5.8 and plot them over the Christmas period in Fig. 5.13. Over this 21-day period the longer term sinusoidal pattern in CVD deaths is approximately linear. The plot suggests that we could fit a *mixed model* with a random intercept and linear slope to control for the annual seasonal effect.

5.3.1 Modelling a Spiked Seasonal Pattern

As an initial model we investigate an increase on Christmas day only. We do this by adding another term to the GAM that we used to remove the effect of season in Sect 5.2. Christmas day is added as a binary independent variable. As before we fit independent variables for mean temperature (using a spline with three degrees of freedom) and day of the week.

The result is a mean RR for Christmas day of 1.036, 95% CI: 0.960, 1.119, p-value = 0.36. The confidence interval is relatively wide as there are only 14 Christmas days over the study period. Based on the results of the box plot (Fig. 5.12), we might instead define a binary variable to cover the Christmas period of 25 December to 28 December. The mean RR for this Christmas period is 1.100 per day, 95% CI: 1.059, 1.143, p-value < 0.001. So averaged between 25 and 28 December there is a 10% increase in mortality per day.

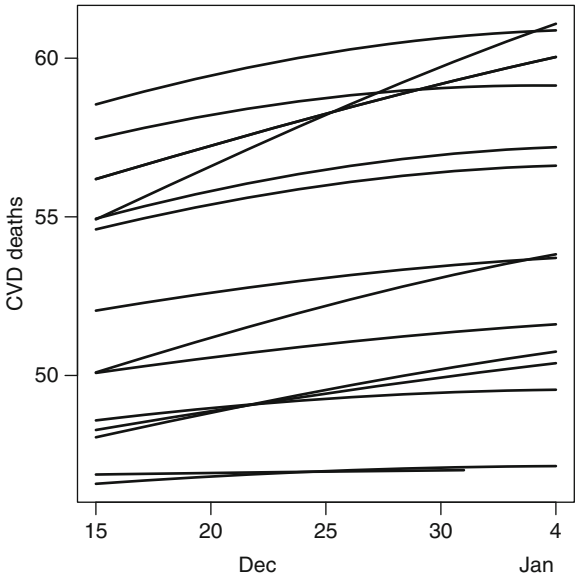


Fig. 5.13 Predicted number of CVD deaths from a GAM for the 10 days either side of Christmas day

A weakness of the above model is that it assumes equal risk for all days defined within the Christmas period. It might be more reasonable to assume that the spiked seasonal pattern has a rise-and-fall in risk. We can fit this pattern in a Bayesian framework by specifying a prior for the spiked pattern that has a multivariate Normal distribution

$$\boldsymbol{\beta} \sim \text{MVN}(\overline{\beta}, \mathbf{V}),$$

where $\boldsymbol{\beta} = \underline{\beta}_1, \ldots, \beta_m$ and m is the total number of days. The average increase on each day is $\overline{\beta}$, and $\mathbf{V}$ is an $m \times m$ symmetric variance–covariance matrix

$$\mathbf{V} = \sigma_\beta^2 \begin{bmatrix} 1 & \rho & 0 & \ldots & 0 & 0 & 0 \\ \rho & 1 & \rho & & 0 & 0 & 0 \\ 0 & \rho & 1 & & 0 & 0 & 0 \\ \vdots & & & \ddots & & & \vdots \\ 0 & 0 & 0 & & 1 & \rho & 0 \\ 0 & 0 & 0 & & \rho & 1 & \rho \\ 0 & 0 & 0 & \ldots & 0 & \rho & 1 \end{bmatrix}.$$

The correlation matrix has 1's along the diagonal and ρ's either side of the diagonal. This means that the estimates on the same day are perfectly correlated, and those from neighbouring days have a correlation of ρ, where $0 < \rho < 1$. The overall increase in risk over m days is then $\sum_m \boldsymbol{\beta}$. In practice, we fit this model using a conditional autoregressive framework (Sect. 2.3.3).

An advantage of this formulation is that it is *non-parametric*. So we have not constrained the spiked seasonal pattern ($\boldsymbol{\beta}$) to have any particular shape, but we

have acknowledged the correlation between neighbouring days, and so the daily estimates *borrow strength* from their neighbours. This is important here because the estimates are based on a relatively small number of days. We could have achieved a similar smoothing using a parametric model, for example, by specifying a quadratic shape over the m days. However, this restricts the spike to have a U-, or inverse U-, shape.

We fit this model in a Bayesian framework by only using the data from the Christmas period defined as the 10 days either side of Christmas day. This gives a sample size of 294 days, as we include the incomplete Christmas periods from January 1987 and December 2000 (the start and end of the time series). The advantage of this method is that we can control for the sinusoidal seasonal pattern using a linear model, as suggested by Fig. 5.13. The disadvantage is a loss of power, as the original sample size is 5,114 days (14 years of data).

We assume that the daily number of deaths follows a Poisson distribution and use a log-link. The CAR equation for the mean number of CVD deaths on day t in Christmas period i is then

$$\begin{aligned} Y_{it} &\sim \text{Poisson}(\mu_{it}), \qquad i = 1, \ldots, 15,\ t = 1, \ldots, 21, \\ \log(\mu_{it}) &= \gamma_{i1} + \gamma_{i2} t + \alpha_3 X_t + \boldsymbol{\zeta} \mathbf{X}_d + \boldsymbol{\beta} \mathbf{X}_s, \\ \gamma_{i1} &\sim \text{N}(\alpha_1, \sigma_1^2), \\ \gamma_{i2} &\sim \text{N}(\alpha_2, \sigma_2^2), \end{aligned}$$

where $t = 1$ corresponds to 15 December and $t = 21$ corresponds to 4 January. For the first Christmas period ($i = 1$) there are only data for $t = 18, \ldots, 21$ (1–4 January). For the last Christmas period ($i = 15$) there are only data for $t = 1, \ldots, 17$ (15–31 December).

This is a *mixed model*, as it includes both fixed and random effects. γ_{i1} is a random intercept, used to model the mean differences between Christmas periods, as shown by the variation in intercepts in Fig. 5.13. γ_{i2} is a random slope, used to model the mean differences in trends between Christmas periods, which is also visible in Fig. 5.13. The mean intercept is α_1, and the mean slope α_2. x_t is the temperature on day t, so $\exp(\alpha_3)$ is the relative risk for a 1°F increase in temperature. The matrix $\mathbf{X}_d$ has six columns and is designed so that $\boldsymbol{\zeta}$ models the effect of day of the week (with Sunday as a reference category). The matrix $\mathbf{X}_s$ has m columns and is designed so that $\boldsymbol{\beta}$ models the effect of a spiked seasonal pattern.

The estimated spiked seasonal pattern for 25–29 December is shown in Fig. 5.14. We also show the "flat" spike in risk estimated using $\boldsymbol{\beta} = \overline{\beta}$.

We can compare the fit of the models using the deviance information criteria (Sect. 1.6.2). The "flat" model uses 34.9 parameters for a DIC of 2,081.9. The CAR model uses 37.6 parameters (2.7 more) for a DIC of 2,082.2 (0.3 worse). So there is no clear difference between the two models, and their estimates of average daily estimated increase in risk is similar, as shown by the mean in Fig. 5.14.

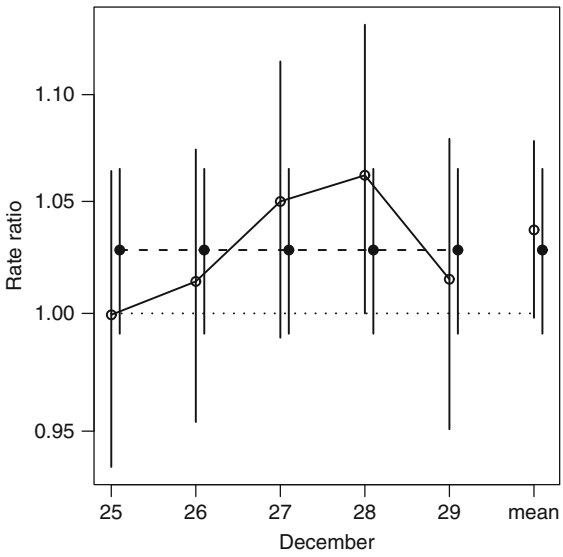


Fig. 5.14 Mean relative risk of the spike in CVD deaths over the Christmas period (25–29 December) using a CAR model (*open circles with solid line*) and flat model (*closed circles with dashed line*), plus the daily mean estimates. The *vertical solid lines* show the 95% credible intervals. The *dotted horizontal line* at RR = 1 corresponds to the null hypothesis of no increase in mortality

Arguably the CAR model is preferable as it gives more information on the spiked seasonal pattern. It shows a rise-and-fall that seems to agree with the pattern shown in Fig. 5.12.

5.4 Adjusting for Seasonal Independent Variables

In this section we show how controlling for a seasonal pattern in an independent variable can be important when analysing a dependent variable that is collected some time later. We consider the situation of predicting long-term survival based on an independent variable collected at baseline. For example, a study may consider survival time (in years) after discharge from hospital based on blood pressure on admission (baseline). If the study recruits subjects over a long enough period then baseline blood pressure will have a seasonal pattern. Thus some of the variation in blood pressure will be due to seasonality and so the time effect of blood pressure will be a mix of true risk and season. The seasonality in the independent variable can be thought of as *measurement error*, and the bias it causes is similar to the regression dilution bias [36].

We illustrate the problem, and the effect of seasonal adjustment, using artificial data in Fig. 5.15. The unadjusted responses are shown in the left-hand panel. There is a strong underlying sinusoidal pattern, which makes those results collected in the first half of the year generally higher than those in the second half. A subject's value depends on the time they were recruited into the study, and a subject recruited in the first half of the year would have had a different measured value if they had been recruited in the second half of the year.

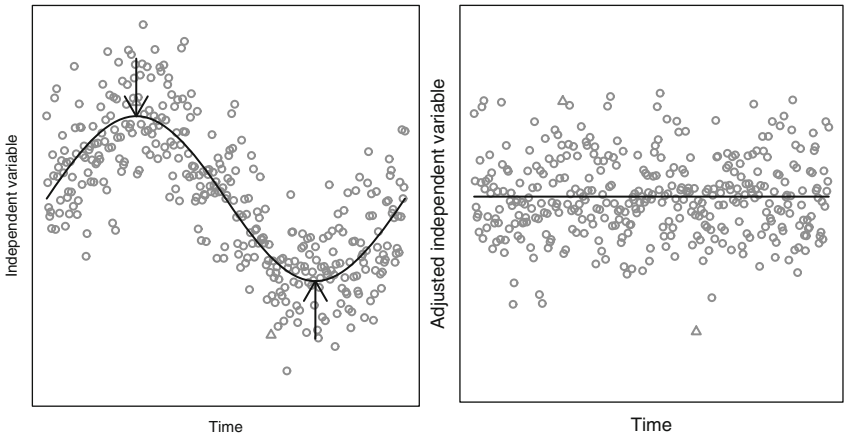


Fig. 5.15 Example of a seasonal pattern in an independent variable and the effect of adjusting (artificial data over one year). In the *left panel*, the *grey dots* show the observed independent variable values, the smallest and largest values are shown as *triangles*, the *black line* shows the underlying sinusoidal seasonal pattern, and the *arrows* show the direction of the seasonal adjustment. In the *right panel*, the *grey dots* show the seasonally adjusted independent variable. The seasonal pattern has been flattened

This seasonal variation is a nuisance when we are interested in long-term risk. The right-hand panel of Fig. 5.15 shows the effect of removing the seasonal variation using a cosinor (Chap. 3). The highest and lowest of the unadjusted values are plotted using triangles in both panels. The highest value is still relatively high, but has been reduced. The lowest value is no longer the smallest one after adjustment, although it is still relatively low. It is also worth noting that a low unadjusted value around the seasonal phase (under the downward pointing arrow in the left-hand panel), is roughly equal to the highest unadjusted values at the seasonal low (above the upward pointing arrow). After adjustment, the low and high values are now *relative* to the season they were collected in.

This type of seasonal variation is not a nuisance when we are interested in short-term risk, for example, examining the risk of death in the next month. In this situation the seasonal variation may be of great interest and importance. For example, the seasonal increase in blood pressure is likely to be one of the major contributors to the seasonal increase in mortality [9].

5.4.1 Effect on Estimates of Long-term Risk

We now consider the effect that seasonal adjustment will have on estimates of long-term risk. In the formula for a generalized linear model, (1.9), we assume there is a seasonally varying independent variable (X_t). The parameter estimated using the seasonally varying independent variable will approximately be

$$\begin{aligned}\beta^* &= \beta \frac{\sigma^2(X^*)}{\sigma^2(X)}, \\ &= \beta \left[\rho(X^*, X)\right]^2,\end{aligned} \tag{5.3}$$

where β is the "true" risk, $\sigma^2(X)$ is the variance of the unadjusted independent variable, and $\sigma^2(X^*)$ is the variance of the seasonally adjusted independent variable (X_t^*). The ratio of variances in (5.3) is called the *reliability ratio* [37, Chap. 1]. We have also shown the formula using the correlation between the unadjusted and adjusted variable, $\rho(X^*, X)$.

If there is no change in the independent variable after seasonal adjustment, then $\rho(X^*, X) = 1$ and $\beta^* = \beta$. The larger the change in the independent variable after seasonal adjustment, the smaller the correlation, and hence the greater the underestimation of risk from using an unadjusted variable. Figure 5.16 shows the bias in the estimated parameter for a range of correlations assuming the true risk is 1. Only at very high correlations will the bias caused by seasonality be negligible; at low correlations the bias will be substantial.

For the example in Fig. 5.15, the correlation between the adjusted and unadjusted responses is 0.35. We would therefore expect to see a greatly reduced estimate of risk when using an unadjusted independent variable, and almost eight times the estimated risk after seasonal adjustment.

The approximate bias, (5.3), applies to any model fitted using a generalized linear model framework, so it applies to Normal, logistic or Poisson regression. For logistic or Poisson regression it applies to the parameter estimates on the logit or log scale, respectively. The bias equation assumes that the independent variable, and the adjusted version, are roughly Normally distributed.

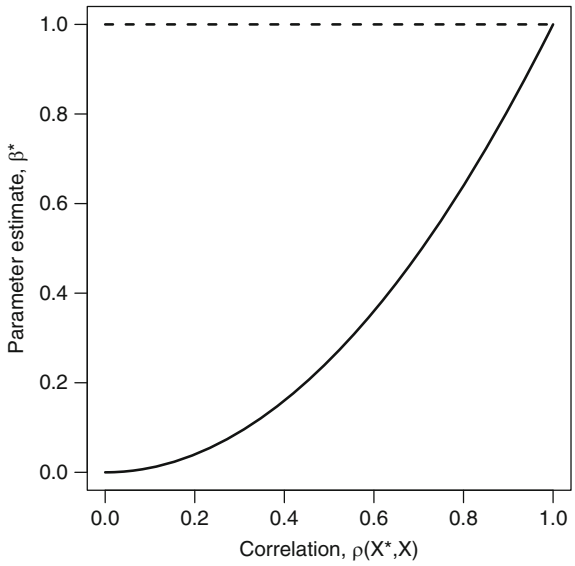


Fig. 5.16 Approximate bias in the estimated parameter depending on the correlation between the unadjusted and seasonally adjusted independent variable, when the true risk is $\beta = 1$ (*dashed horizontal line*)

A further assumption is that the appropriate seasonal adjustment can be made using a cosinor model, which is based on a sinusoidal seasonal pattern. Other more complex seasonal patterns, e.g., sawtooth, would require other types adjustment. For irregular seasonal patterns, we recommend examining the residuals after fitting a spline for time (Sect. 5.2).

5.4.1.1 Example Using Long-term Survival After Hospital Discharge

We give an example of seasonal adjustment based on real data. The data are from a study of long-term survival after discharge from the Prince Charles Hospital, Brisbane. On admission patients had their creatinine levels measured (μmol/L^{-1}), as it is a useful measure of renal function, and is thought to be important for long-term health. There were 5,559 admissions to hospital from 2001 to 2005 with creatinine levels recorded. The study has a total of 16,625 person years of follow-up, and 429 patients died (7.7%).

As the creatinine data are strongly positively skewed, we first applied a log-transformation. A cosinor model (Sect. 3) showed that log creatinine levels have a statistically significant annual seasonal variation, with an amplitude of 0.043 and a phase of 16 March.

A *Cox proportional hazards model* based on the unadjusted creatinine levels gives a hazard ratio of 3.47 (95% CI: 2.86, 4.20). So a one unit increase in log creatinine is associated with more than treble the risk of death. A model based on the seasonally adjusted creatinine levels gives a hazard ratio of 3.45 (95% CI: 2.85, 4.18). So in this example, despite the statistically significant seasonal variation in creatinine levels, the estimate of its long-term risk is not affected by seasonal adjustment. The Pearson correlation between the unadjusted and seasonally adjusted creatinine data is 0.991. Hence the very small change in the hazard ratio after adjustment.

5.5 Biases Caused by Ignoring Season

In this section we describe the potential biases caused by ignoring season in observational studies. We examine a study design that measures subjects at baseline, applies an intervention, and then measures them again at a follow-up visit [71, 74]. We examine the situation where the seasonal pattern in the main dependent variable is ignored, but show how to control for this bias by design.

The left-hand panel of Fig. 5.17 presents artificial data for an extreme case where a study is started at the phase of an annual seasonal pattern and the follow-up is six months later. The solid line shows the mean annual seasonal pattern, which peaks at the beginning of the study and dips six months later. The baseline observations were recorded in the first two weeks, and the follow-up observations six months later. The

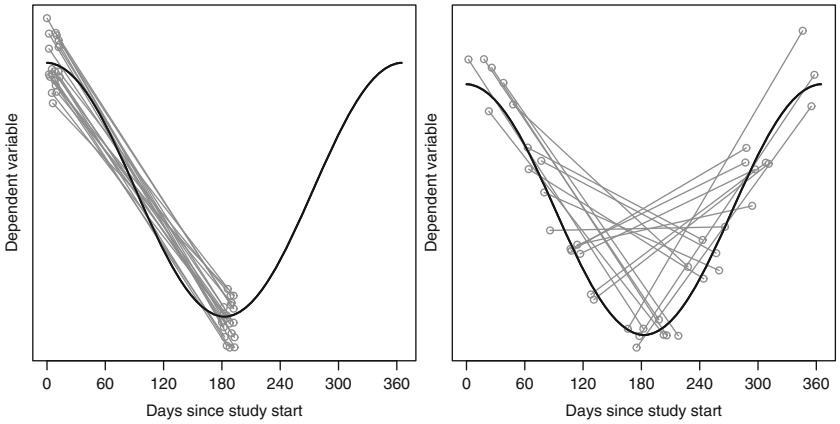


Fig. 5.17 Simulated data from an observational study with a baseline observation and 6-month follow-up. The *grey circles* show the observed responses (from 20 subjects), baseline and follow-up responses are joined by *grey lines*. The *solid black line* shows the underlying seasonal pattern. The subjects are recruited in 2 weeks (*left-panel*) and 6 months (*right-panel*)

individual observations were generated randomly using a Normal distribution with a mean centred on the mean seasonal pattern.

This example shows how the seasonal variation could be confused with an (almost linear) change in response. In this example there is no intervention effect, but the investigator is likely to conclude that the intervention has caused a strong decrease in the response. If the intervention truly caused a negative change in response, then this would look stronger by ignoring the seasonal variation. If the intervention truly caused a positive change in response, then this would look weaker by ignoring the seasonal variation (and possibly even cause a type II error). For this example the seasonal change is completely confounded with the intervention effect, and could not be controlled for statistically.

The problem could be avoided by design. One design is to include a control group (without the intervention), which has the advantage of controlling for other unmeasured confounding. If its not possible to have a control group, the bias can be "averaged out" by staggering the recruitment of subjects over a 6-month period. This design is shown in the right-hand panel of Fig. 5.17. In this case the seasonal pattern still causes strong changes between baseline and follow-up, but these differences are averaged out when looking at the mean difference between baseline and follow-up. Importantly, this design will work regardless of the phase.

Although the simulated example in Fig. 5.17 uses Normally distributed data, similar biases would apply to dependent data with a Poisson or Binomial distribution.

Chapter 6
Clustered Seasonal Data

In this section we show how to model clustered seasonal data. We look at two types of clustered data: longitudinal data and spatial data.

Longitudinal data are time series data from multiple subjects or clusters (e.g., cities). They may be equally spaced (e.g., every week) or irregularly spaced. If each subject has the same number of responses then the data are *balanced*, otherwise they are unbalanced.

The data on walking time are an example of unbalanced longitudinal data, as subjects have 1, 2 or 3 responses (Sect. 1.1.4). For this example, we might be interested in whether certain types of people show a greater seasonal variation in walking time (e.g., older vs. younger subjects). We might be interested an individual seasonal pattern, or the average seasonal pattern. As we show in the next section, these can be quite different.

The NMMAPS data is an example of *spatial data*, as they are available for 108 cities throughout the US. This is an example of *balanced* data, as the monthly (or daily) results are available during the same time span for each city (years 1987–2000). For this example, we might be interested in whether the seasonal variation in CVD mortality shows any geographic variation.

We only deal with seasonality in longitudinal or spatial data. For a more detailed discussion of all aspects of longitudinal analyses see Diggle et al. [24], and for spatial analyses see Elliot et al. [33].

6.1 Seasonal Heterogeneity

When considering longitudinal data it is important to consider whether there is any between-subject variation (or heterogeneity) in the seasonal pattern, and also whether the focus of the analysis is on the average seasonal pattern, or the individual seasonal pattern.

Figure 6.1 shows four hypothetical examples, using subject data that have a sinusoidal seasonal pattern. In this example there is no between-subject variation in amplitude (which is always 1), but there is some between-subject variation in phase. If there is no variation in the phase, then the average seasonal pattern and individual

A.G. Barnett and A.J. Dobson, *Analysing Seasonal Health Data*, Statistics for Biology and Health, DOI 10.1007/978-3-642-10748-1_6,
© Springer-Verlag Berlin Heidelberg 2010

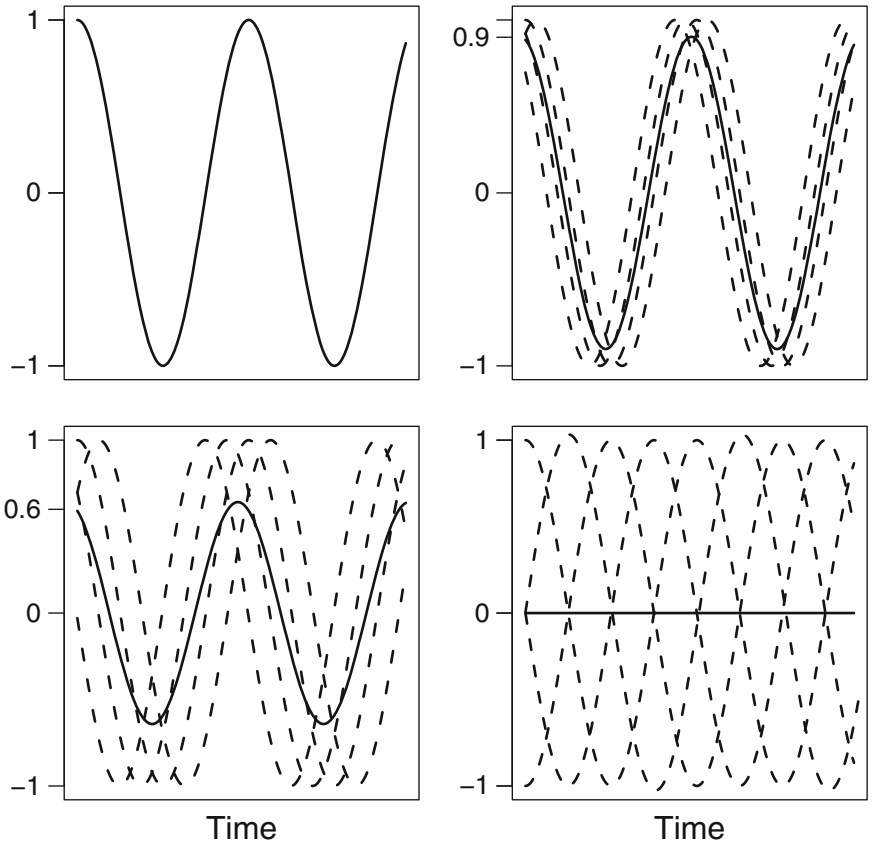


Fig. 6.1 Individual seasonality for four subjects (*dashed lines*) and mean seasonality (*solid line*) for four different levels of between-subject variability in phase. No variability, individual and mean seasonality are the same (*top-left*); moderate variability (*top-right*); large variability (*bottom-left*); maximum variability (*bottom-right*)

seasonal pattern are the same, and both have an amplitude of 1. When there is some variation in phase, the mean seasonal pattern has a smaller amplitude than the individual pattern. The larger the variation in subjects' phases, the smaller the average seasonal amplitude. In the extreme case, if the phases in individual seasonal patterns are spread evenly throughout the year, then the average seasonal pattern will be flat.

As an example, imagine we are interested in the seasonal pattern in time spent walking (Sect. 1.1.4). Public health practitioners would be interested in the mean seasonal pattern, with a view to launching campaigns to improve levels of walking in the whole community when mean levels are lowest, or when they begin to increase. An individual subject would be more interested in the average individual change, which would also be more useful for predicting the individual change in weight.

We can fit models to capture either the average seasonal pattern, or the individual pattern. The choice of model depends on whether the focus of the analysis is the average or the individual. This choice, and reduction in the mean estimate compared with the individual estimate (Fig. 6.1), is an analogous to the difference between a marginal and random effects logistic regression model [23, Sect. 7.4].

6.2 Longitudinal Models

The most important aspect of longitudinal data is that the repeated results from the same subject can no longer assumed to be independent. A statistical model that we can use to account for this dependence is a *mixed model*. A mixed model has both fixed and random effects (Sect. 2.3).

We start by combining a generalized linear model, (1.9), with a Fourier model, (1.4), and adding an index for each subject

$$g(\mu_{it}) = \beta_0 + \beta_1 X_{1it} + \cdots + \beta_p X_{pit} + A\cos(\omega w_{it} - P), \qquad (6.1)$$
$$i = 1, \ldots, m,\ t = 1, \ldots, n_i,$$

where m is the number of subjects, n_i is the number of observations for subject i, and w_{it} is the time of the response number t for subject i. β_0 is the intercept. For data that are both equally spaced and balanced, $w_{it} = t$. The constant, $\omega = 2\pi/l_s$, transforms the times into radians, where l_s is the length of the seasonal cycle. The seasonal pattern is controlled by the amplitude (A) and phase (P), assuming a sinusoidal seasonal pattern. The $\boldsymbol{\beta}$ parameters model the non-seasonal aspects of the data (e.g., trend).

Using a generalized linear model the response could follow, for example, a Normal distribution

$$y_{it} \sim \mathrm{N}(\mu_{it}, \sigma^2),$$

using a identity link function, or a Poisson distribution

$$y_{it} \sim \mathrm{Po}(\mu_{it}),$$

using a log link function.

We can add a random intercept to the model using

$$g(\mu_{it}) = \gamma_{i0} + \beta_1 X_{1it} + \cdots + \beta_p X_{pit} + A\cos(\omega w_{it} - P), \qquad (6.2)$$
$$\gamma_{i0} \sim \mathrm{N}(\beta_0, \sigma_0^2), \qquad i = 1, \ldots, m,$$

so each subject has their own intercept (γ_{i0}). These intercepts are constrained to follow a Normal distribution, centred on the mean intercept (β_0) and with variance σ_0^2. A random intercept would be useful for modelling between-subject differences that are constant over time. The estimated seasonal pattern is now expressed in terms

of the difference from the average, rather than being centred on the overall mean. Because the model combines a generalized linear model and a mixed model it is a *generalized linear mixed model* (GLMM).

To add subject-specific phases we use

$$\begin{aligned} g(\mu_{it}) &= \gamma_{i0} + \beta_1 X_{1it} + \cdots + \beta_p X_{pit} + A\cos(\omega w_{it} - P_i), \qquad (6.3)\\ P_i &= \overline{P} + (\tilde{P}_i - \pi), \qquad i = 1, \ldots, m,\\ \tilde{P}_i &\sim \mathrm{N}\left[\pi, \left(\frac{a\pi}{12}\right)^2\right],\\ \overline{P} &\sim \mathrm{U}(0, 2\pi), \end{aligned}$$

where P_i is the estimated phase for subject i. These terms are modelled by $\tilde{P}_i$ which is centred on π, the furthest point on the circle from the border at 0 and 2π. This avoids the problem of the MCMC chain jumping between two identical solutions. We used a similar restriction in Sect. 4.4.1. The value a controls the variance of the Normal distribution, and will approximately allow phases of up to $2a$ months either side of the mean phase.

In this model each subject has their own seasonal pattern (defined by A and P_i), which is centred on the mean individual pattern (defined by A and $\overline{P}$). In Fig. 6.1, the seasonal estimates from model (6.3) are the dashed lines, whereas those from (6.2) are the solid line. The seasonal pattern in model (6.3) is for an average individual, whereas the pattern in model (6.2) is the average of the individuals.

6.2.1 Example

We highlight the differences between marginal and individual models using the exercise data (Sect. 1.1.4). We assume that the walking time (number of minutes per week) follows a Poisson distribution. We therefore define a marginal model as

$$\begin{aligned} y_{it} &\sim \mathrm{Po}(\mu_{it}), \qquad i = 1, \ldots, 434,\ t = 1, \ldots, n_i,\\ \log(\mu_{it}) &= \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + c\cos(\omega w_{it}) + s\sin(\omega w_{it}), \end{aligned}$$

where $n_i \in \{1, 2, 3\}$, $\mathbf{X}_1$ is a binary covariate which is 1 for the second visit and 0 otherwise, and $\mathbf{X}_2$ is a binary covariate which is 1 for the third visit and 0 otherwise. We add these independent variables to model the changes in activity from baseline. We assume a stationary seasonal pattern, as c and s do not depend on t. We estimate the amplitude and phase using Eqns. (3.2) and (3.3), respectively.

The mean rate ratio amplitude is 1.146 (95% credible interval: 1.136, 1.156). The phase is 5 July (95% credible interval: 1 July, 9 July). The seasonal pattern is plotted in the left panel of Fig. 6.2. The right panel shows the mean seasonal pattern using a model with a random intercept. The mean rate ratio amplitude for this model is 1.202 (95% credible interval: 1.190, 1.213). The phase is 8 July (95%

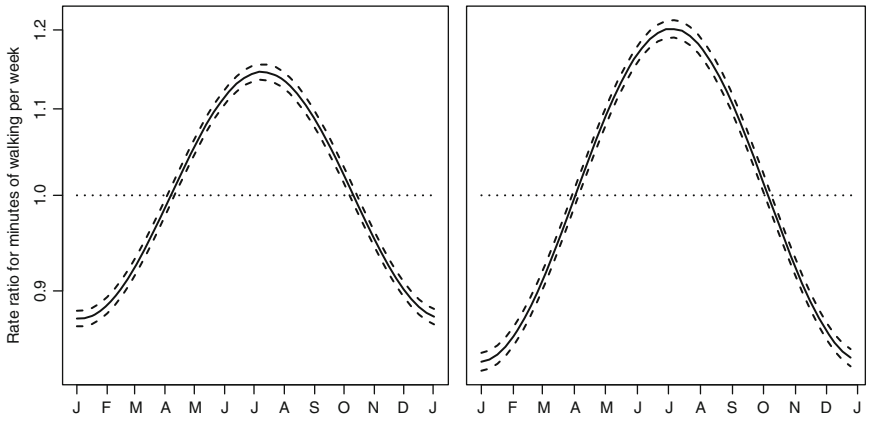


Fig. 6.2 Mean seasonal pattern in walking time (*solid line*) and 95% credible intervals (*dashed lines*) from a marginal model (6.1) (*left panel*), and model with a random intercept model (6.2) (*right panel*)

credible interval: 4 July, 11 July). The peak of activity in the Australian winter is not surprising considering the heat of a Brisbane summer and the mildness of winter, and we saw the same pattern when smoothing the data in Sect. 2.2.4.

For the model with random seasonal effects (6.3) we decided to allow subjects' phases to differ by no more than a month, and we did so by setting $a = 0.5$. The mean subject specific rate ratio amplitude is 1.378 (95% credible interval: 1.362, 1.394). The mean phase is 8 July (95% credible interval: 3 July, 11 July). Individual phases ranged from 19 June to 26 July (roughly within a month of each other).

The DIC for the marginal model is 204,475 based on 5.0 parameters. The DIC for the model with a random intercept is 62,475 based on 395.0 parameters (142,000 better than the marginal model). The DIC for the model with a random seasonal effects is 59,975 based on 398.4 parameters (2,500 better than the marginal model). So a random intercept greatly helped to explain the variation between subjects, and even allowing a relatively small variation in the phase gave a substantially better model.

6.3 Spatial Models

Seasonal diseases that have some association with temperature or sunlight are also likely to have some geographic variation. As the seasons are more extreme the further we move from the equator, we might expect greater seasonal amplitudes in disease with increasing north or south latitude.

We can use geographic variation in a seasonal pattern to give clues about the seasonal exposure. If the seasonal pattern in disease was due to weather, we would expect the risk to vary smoothly over space. This is because weather variables (such

as average sunlight hours and temperature) generally vary gradually, although there can be more sudden changes along coastlines. In contrast, if the seasonal pattern in disease was due to a social factor (Sect. 2.1.1), we might expect to see no geographic variation, or more sudden changes along political borders. For example, a seasonal pattern in risk due to the change in daylight saving time in Australia might be different in New South Wales (which uses daylight savings) compared with its neighbour Queensland (which does not).

6.3.1 Example

As an example we look at the geographic variation in CVD mortality in the US using the NMMAPS data. We used the monthly counts of CVD deaths for people aged $\geq$75 years old. We fitted a model with a stationary seasonal amplitude and stationary phase, and a quadratic trend (Sect. 4.1). We accounted for the unequal number of days in the month and population size using an offset. We fitted the model in each of the 108 cities.

Figure 6.3 maps the seasonal amplitude expressed as a percent change in CVD mortality (Sect. 1.4.5). The most noticeable spatial pattern is the group of large amplitudes in California and Arizona.

Figure 6.4 maps the phase in months. The latest phase was 2.3 months (early February) in Biddeford, Maine (top-right corner of Fig. 6.4). The phase was also 2.3 months in Honolulu. The earliest phase was 12.5 months (mid December) in Anchorage.

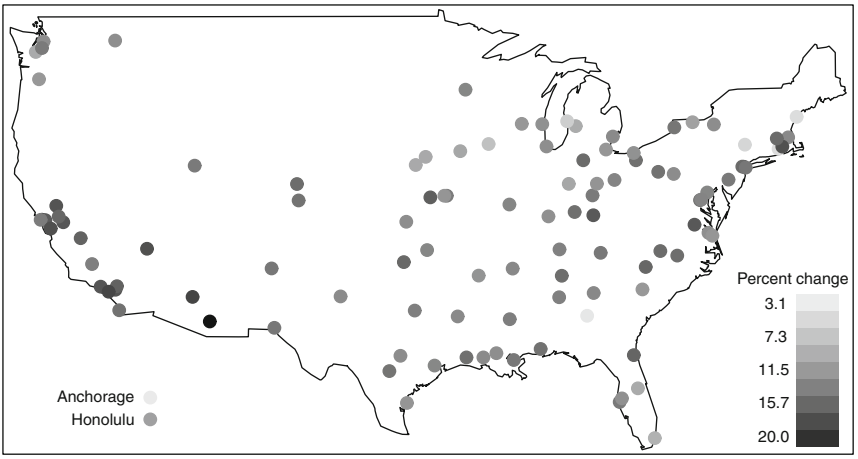


Fig. 6.3 Seasonal amplitude in CVD mortality expressed as a percent change by location (108 cities). Results for Anchorage and Honolulu shown in *bottom-left corner*

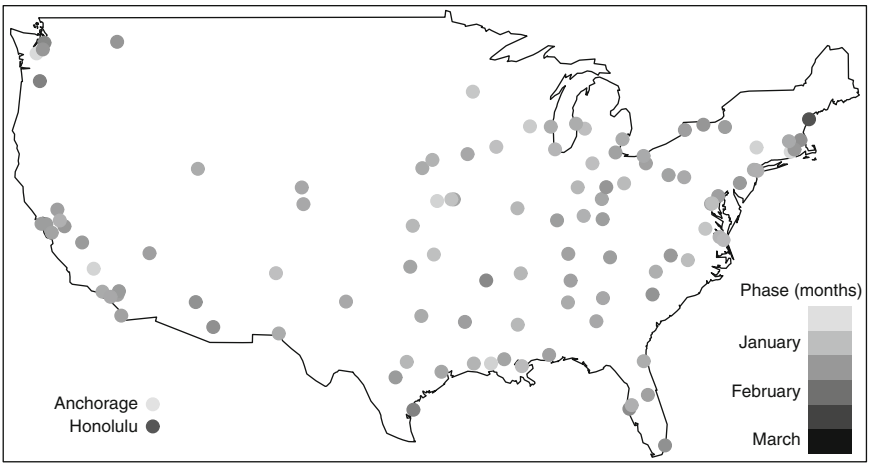


Fig. 6.4 Seasonal phase in CVD by location (108 cities). Results for Anchorage and Honolulu shown in *bottom-left corner*

Fig. 6.5 Circular plot of the phases in CVD mortality for the 108 NMMAPS cities

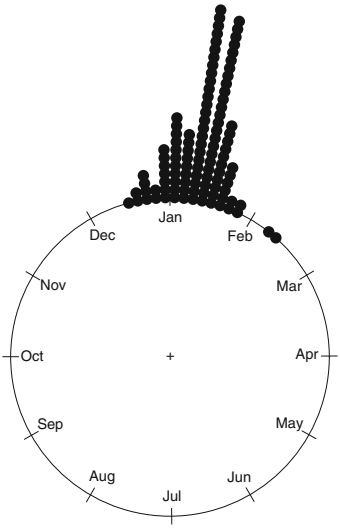


Figure 6.5 shows a circular plot of the phases, and demonstrates the predominance of phases in January.

To produce a smoothed picture of the geographic patterns in the amplitude and phase we used *kriging* interpolation [24]. We used the `surf.gls` function from the `spatial` R library [85]. We excluded the results from Anchorage and Honolulu, as these cities are too distant from mainland USA for smoothing purposes. The smooth amplitude in Fig. 6.6 shows the largest seasonal changes occurred in California. There is a steep decline in amplitude with increasing northerly latitude,

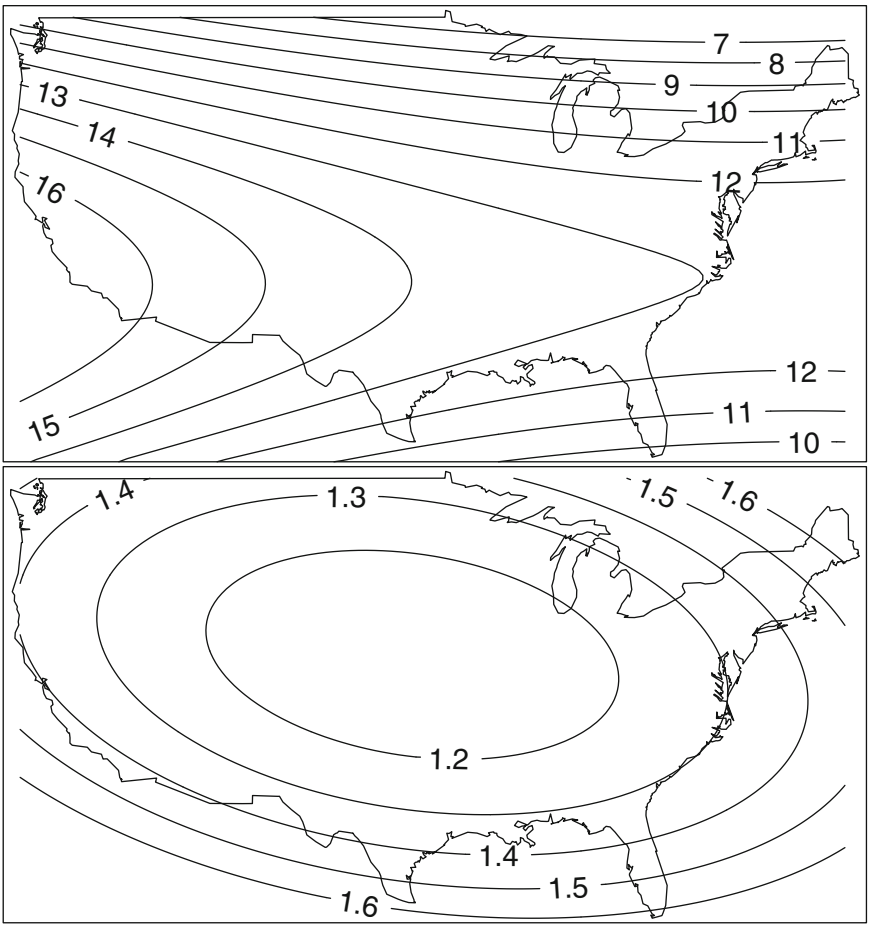


Fig. 6.6 Estimated smooth seasonal amplitude expressed as a percentage increase (*top*) and phase in months (*bottom*) in CVD mortality in mainland USA. Results from Anchorage and Honolulu were excluded

which matches the theory that people living in cold climates are better prepared to deal with relatively cold temperatures [82]. There is also a decline in amplitudes moving south through Florida. This is possibly because winter temperatures in Florida are not cold enough to cause large increases in CVD [8].

The smooth phase shows that the earliest phases occurred in central USA (1.2 months, corresponding to early January), with increasingly later phases as we move towards the US coast and border regions (1.6 months, corresponding to mid January).

References

1. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**(6), 716–723 (1974)
2. Altamura, C., VanGastel, A., Pioli, R., Mannu, P., Maes, M.: Seasonal and circadian rhythms in suicide in Cagliari, Italy. J. Affect. Disord. **53**(1), 77–85 (1999)
3. Babiker, H.A., Abdel-Muhsin, A.M., Ranford-Cartwright, L.C., Satti, G., Walliker, D.: Characteristics of *Plasmodium falciparum* parasites that survive the lengthy dry season in eastern Sudan where malaria transmission is markedly seasonal. Am. J. Trop. Med. Hyg. **59**(4), 582–590 (1998)
4. Ballester, F., Corella, D., Perez-Hoyos, S., Saez, M., Hervas, A.: Mortality as a function of temperature. A study in Valencia, Spain, 1991–1993. Int. J. Epidemiol. **26**(3), 551–561 (1997)
5. Barnett, A.G., Dobson, A.J.: Estimating trends and seasonality in coronary heart disease. Stat. Med. **23**(22), 3505–3523 (2004)
6. Barnett, A.G., Dobson, A.J.: Is the increase in coronary heart disease on Mondays an artifact? Epidemiology **15**(5), 583–588 (2004)
7. Barnett, A.G., Dobson, A.J., McElduff, P., Salomaa, V., Kuulasmaa, K., Sans, S.: The WHO MONICA Project: cold periods and coronary events: an analysis of populations worldwide. J. Epidemiol. Community Health **59**(7), 551–557 (2005)
8. Barnett, A.G., Fraser, J.F., de Looper, M.: The seasonality in heart failure deaths and total cardiovascular deaths in Australia. Aust. N. Z. J. Public Health **32**(5), 408–413 (2008)
9. Barnett, A.G., Sans, S., Salomaa, V., Kuulasmaa, K., Dobson, A.J.: The WHO MONICA Project: the effect of temperature on systolic blood pressure. Blood Press. Monit. **12**(3), 195–203 (2007)
10. Barnett, A.G., Wolff, R.C.: A time-domain test for some types of non-linearity. IEEE Trans. Signal Process. **53**(1), 26–33 (2005)
11. Bartlett, M.S.: An Introduction to Stochastic Processes with Special Reference to Methods and Applications, 2nd edn. Cambridge University Press, Cambridge (1966)
12. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer, New York (1985)
13. Burnham, K.P., Anderson, D.R.: Model Selection and Multi-model Inference, 2nd edn. Springer, New York (2002)
14. Burns, J.C., Cayan, D.R., Tong, G., Bainto, E.V., et al.: Seasonality and temporal clustering of Kawasaki syndrome. Epidemiology **16**(2), 220–225 (2005)
15. CDC: Flu activity and surveillance. Centers for Disease Control and Prevention, Atlanta, GA (2008). http://www.cdc.gov/flu/weekly/fluactivity.htm
16. Chatfield, C.: The Analysis of Time Series: Theory and Practice. Monographs on Applied Probability and Statistics. Chapman and Hall, London (1975)
17. Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I.: A seasonal-trend decomposition procedure based on loess. J. Off. Stat. **6**(1), 3–73 (1990)
18. Cooper, H.: On the relative prevalence of diseases in Hull, and the effects of season upon disease. J. Stat. Soc. London **16**(4), 352–355 (1853)

19. Coren, S.: Daylight savings time and traffic accidents. N. Engl. J. Med. **334**(14), 924–925 (1996)
20. Cowgill, U.M.: Season of birth in man. Contemporary situation with special reference to Europe and the southern hemisphere. Ecology **47**(4), 614–623 (1966)
21. Crainiceanu, C., Ruppert, D., Wand, M.P.: Bayesian analysis for penalized spline regression using WinBUGS. J. Stat. Softw. **14**(14) (2005)
22. Davies, G., Welham, J., Chant, D., Torrey, E.F., McGrath, J.: A systematic review and meta-analysis of Northern hemisphere season of birth studies in schizophrenia. Schizophr. Bull. **29**(3), 587–593 (2003)
23. Diggle, P., Heagerty, P., Liang, K.Y., et al.: Analysis of Longitudinal Data, 2nd edn. Oxford Statistical Science Series. Oxford University Press, New York (2002)
24. Diggle, P., Ribeiro, P.J.: Model-Based Geostatistics. Springer Series in Statistics. Springer, New York (2007)
25. Diggle, P.J.: Time Series: A Biostatistical Introduction. Oxford University Press, Oxford (1990)
26. Dobson, A.J., Barnett, A.G.: An Introduction to Generalized Linear Models, 3rd edn. Texts in Statistical Science. Chapman and Hall/CRC, Boca Raton, FL (2008)
27. Dominici, F., McDermott, A., Zeger, S.L., Samet, J.M.: Airborne particulate matter and mortality: Timescale effects in four US cities. Am. J. Epidemiol. **157**(12), 1055–1065 (2003)
28. Duncan, D.E.: The Calendar: The 5000-year Struggle to Align the Clock and the Heavens, and What Happened to the Missing Ten Days. Fourth Estate, London (1998)
29. Dupont, W.D., Jr., W.D.P.: Power and sample size calculations for studies involving linear regression. Control. Clin. Trials **19**(6), 589–601 (1998)
30. Dupont, W.D., Plummer, W.D.: PS power and sample size program available for free on the Internet. Control. Clin. Trials **18**(3), 274 (1997). http://biostat.mc.vanderbilt.edu/PowerSampleSize
31. Eakin, E., Reeves, M., Lawler, S., Graves, N., et al.: Telephone counselling for physical activity and diet in primary care patients. Am. J. Prev. Med. **36**(2), 142–149 (2009)
32. Eilers, P.H.C., Gampe, J., Marx, B.D., Rau, R.: Modulation models for seasonal time series and incidence tables. Stat. Med. **27**(17), 3430–3441 (2009)
33. Elliot, P., Wakefield, J., Best, N., Briggs, D.: Spatial Epidemiology: Methods and Applications. Oxford University Press, Oxford (2000)
34. Emch, M., Feldacker, C., Islam, M.S., Ali, M.: Seasonality of cholera from 1974 to 2005: a review of global patterns. Int. J. Health Geogr. **7**(1), 31 (2008)
35. Fisher, N.I.: Statistical Analysis of Circular Data. Cambridge University Press, Cambridge (1993)
36. Frost, C., Thompson, S.G.: Correcting for regression dilution bias: Comparison of methods for a single predictor variable. J. R. Stat. Soc. A Stat. Soc. **163**(2), 173–189 (2000)
37. Fuller, W.A.: Measurement Error Models. Wiley Series in Probability and Statistics. Wiley, New York (1987)
38. Fuller, W.A.: Introduction to Statistical Time Series, 2nd edn. Wiley Series in Probability and Statistics. Wiley, New York (1996)
39. Garry, V.F., Harkins, M.E., Erickson, L.L., Long-Simpson, L.K., et al.: Birth defects, season of conception, and sex of children born to pesticide applicators living in the Red River Valley of Minnesota, USA. Environ. Health Perspect. **110**(suppl 3), 441–449 (2002)
40. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: Bayesian Data Analysis, 2nd edn. Chapman and Hall/CRC, Boca Raton, FL (2004)
41. Ghysels, E., Osborn, D.R.: The Econometric Analysis of Seasonal Time Series. Cambridge University Press, Cambridge (2001)
42. Gilks, W.R., Richardson, S., Spiegelhalter, D.J.: Markov Chain Monte Carlo in Practice. Chapman and Hall, London (1996)
43. Gladwell, M.: Outliers. Hachette Book Group, New York (2009)
44. Green, A., Patterson, C.C., Group, E.T.S.: Trends in the incidence of childhood-onset diabetes in Europe 1989–1998. Diabetologia **44**(suppl 3), B3–B8 (2001)
45. Iannaccone, R., Coles, S.: Semiparametric models and inference for biomedical time series with extra-variation. Biostatistics **2**(3), 261–276 (2001)

46. James, W.H.: Social class and season of birth. J. Biosoc. Sci. **3**, 309–320 (1971)
47. Janes, H., Sheppard, L., Lumley, T.: Case–crossover analyses of air pollution exposure data: referent selection strategies and their implications for bias. Epidemiology **16**(6), 717–726 (2005)
48. Jessen, G., Jensen, B.F., Arensman, E., Bille-Brahe, U., et al.: Attempted suicide and major public holidays in Europe: findings from the WHO/EURO multicentre study on parasuicide. Acta. Psychiatr. Scand. **99**(6), 412–418 (1999)
49. Johnson, H., Brock, A., Griffiths, C., Rooney, C.: Mortality from suicide and drug-related poisoning by day of the week in England and Wales, 1993–2002. Health Stat. Q. **27**, 13–16 (2005)
50. Kao, C., Huang, J., Ou, L., See, L.: The prevalence, severity and seasonal variations of asthma, rhinitis and eczema in Taiwanese schoolchildren. Pediatr. Allergy. Immunol. **16**(5), 408–415 (2005)
51. Kasper, S., Wehr, T.A., Rosenthal, N.E.: Season-related forms of depression. I. Principles and clinical description of the syndrome. Nervenarzt **59**(4), 191–199 (1988)
52. Keele, L.: Semiparametric Regression for the Social Sciences. Wiley, Chichester (2008)
53. Law, M.R., Morris, J.K., Wald, N.J.: Use of blood pressure lowering drugs in the prevention of cardiovascular disease: meta-analysis of 147 randomised trials in the context of expectations from prospective epidemiological studies. BMJ **338**(191), b1665 (2009)
54. Lawlor, D.A., Leon, D.A., Davey Smith, G.: The association of ambient outdoor temperature throughout pregnancy and offspring birthweight: findings from the Aberdeen children of the 1950s cohort. BJOG **112**(5), 647–657 (2005)
55. Le, C.T., Liu, P., Lindgren, B.R., Daly, K.A., Giebink, G.S.: Some statistical methods for investigating the date of birth as a disease indicator. Stat. Med. **22**(13), 2127–2135 (2003)
56. Lincoln, D., Morgan, G., Sheppeard, V., Jalaludin, B., Corbett, S., Beard, J.: Childhood asthma and return to school in Sydney, Australia. Public Health **120**(9), 854–862 (2006)
57. MacIntosh, D.L., Spengler, J.D., Özkaynak, H., hui Tsai, L., Ryan, P.B.: Dietary exposures to selected metals and pesticides. Environ. Health Perspect. **104**(2), 202–209 (1996)
58. Maclure, M.: The case–crossover design: a method for studying transient effects on the risk of acute events. Am. J. Epidemiol. **133**(2), 144–153 (1991)
59. Mardia, K.V.: Statistics of Directional Data. Academic, London (1972)
60. Matthews, C.E., Freedson, P.S., Hebert, J.R., Stanek, E.J., et al.: Seasonal variation in household, occupational, and leisure time physical activity: longitudinal analyses from the seasonal variation of blood cholesterol study. Am. J. Epidemiol. **153**(2), 172–183 (2001)
61. McGrath, J.J., Barnett, A., Eyles, D., Burne, T., et al.: The impact of nonlinear exposure–risk relationships on seasonal time-series data: modelling Danish neonatal birth anthropometric data. BMC Med. Res. Methodol. **7**(45) (2007)
62. Meehl, G.A., Stocker, T.F., Collins, W.D., Friedlingston, P., et al.: Global climate projections. In: S. Solomon, et al. (eds.) Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge (2007)
63. Moan, J., Porojnicu, A., Lagunova, Z., Berg, J.P., Dahlback, A.: Colon cancer: prognosis for different latitudes, age groups and seasons in Norway. J. Photochem. Photobiol. B **89**(2–3), 148–155 (2007)
64. Ogle, W.: Suicides in England and Wales in relation to age, sex, season, and occupation. J. Stat. Soc. London **49**(1), 101–135 (1886)
65. Peng, R.D., Dominici, F.: Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health. Springer, New York (2008)
66. Peng, R.D., Welty, L.J.: The NMMAPSdata package. R News **4**(2), 10–14 (2004). http://CRAN.R-project.org/doc/Rnews/
67. Pfeiffer, D.U., Robinson, T.P., Stevenson, M., Stevens, K.B., et al.: Spatial Analysis in Epidemiology. Oxford University Press, Oxford (2008)
68. Phillips, D.P., Jarvinen, J.R., Abramson, I.S., Phillips, R.R.: Cardiac mortality is higher around Christmas and New Year's than at any other time. Circulation **110**(25), 3781–3788 (2004)

69. Ponsonby, A.L., Dwyer, T., Jones, M.E.: Sudden infant death syndrome: seasonality and a biphasic model of pathogenesis. J. Epidemiol. Community Health **46**(1), 33–37 (1992)
70. Rau, R.: Seasonality in Human Mortality A Demographic Approach. Springer, New York (2007)
71. Rochester, D., Jain, A., Polotsky, A.J., Polotsky, H., et al.: Partial recovery of luteal function after bariatric surgery in obese women. Fertil. Steril. **92**(4), 1410–1415 (2009)
72. Rocklöv, J.P., Forsberg, B., Meister, K.: Winter mortality modifies the heat-mortality association the following summer. Epidemiology **19**(6), S87–S88 (2008)
73. Ruppert, D.R., Wand, M.P., Carroll, R.J.: Semiparametric Regression. Cambridge University Press, New York (2003)
74. Saigh, O., Triola, M.M., Link, R.N.: Failure of an electronic medical record tool to improve pain assessment documentation. J. Gen. Intern. Med. **21**(2), 185–188 (2006)
75. Samet, J., Dominici, F., Zeger, S., Schwartz, J., Dockery, D.: The National Morbidity, Mortality, and Air Pollution Study. Part I: methods and methodologic issues. Res. Rep. Health Eff. Inst. **94**(pt 1), 5–14 (2000)
76. Scragg, R.: Seasonal variation of mortality in Queensland. Community Health Stud. **6**(2), 120–128 (1982)
77. Selevan, S.G., Borkovec, L., Slott, V.L., Zudová, Z., et al.: Semen quality and reproductive health of young Czech men exposed to seasonal air pollution. J. Stat. Soc. London **108**(9), 887–894 (2000)
78. Simmons, C., Paull, G.C.: Season-of-birth bias in association football. J. Sports Sci. **19**(9), 677–686 (2001)
79. Spiegelhalter, D.: The BUGS project – DIC. http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/dicpage.shtml, MRC Biostatistics Unit, Cambridge, UK (2008)
80. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A.: Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. B **64**(4), 583–640 (2002)
81. Sterne, J.A.C., Smith, G.D., Cox, D.R.: Sifting the evidence – what's wrong with significance tests? Another comment on the role of statistical methods. BMJ **322**(7280), 226–231 (2001)
82. The Eurowinter Group: Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. Lancet **349**(9062), 1341–1346 (1997)
83. Thomas, A., O'Hara, B., Ligges, U., Sturtz, S.: Making BUGS open. R News **6**(1), 12–17 (2006). http://cran.r-project.org/doc/Rnews/
84. Trewin, D.: Information paper: census of population and housing – socio-economic indexes for areas, Australia, 2001. Commonwealth of Australia (2003). ABS Catalogue No 2039.0
85. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002)
86. Walter, S.D., Elwood, J.M.: A test for seasonality of events with a variable population at risk. Br. J. Prev. Soc. Med. **29**(1), 18–21 (1975)
87. West, M., Harrison, J.: Bayesian Forecasting and Dynamic Models, 2nd edn. Springer Series in Statistics. Springer, New York (1997)
88. Wood, S.N.: Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC Press, Boca Raton, FL (2006)
89. Yiin, L.M., Rhoads, G.G., Lioy, P.J.: Seasonal influences on childhood lead exposure. Environ. Health Perspect. **108**(2), 177–182 (2002)

Index