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,
September 2009

Adrian Barnett
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 ≥ 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

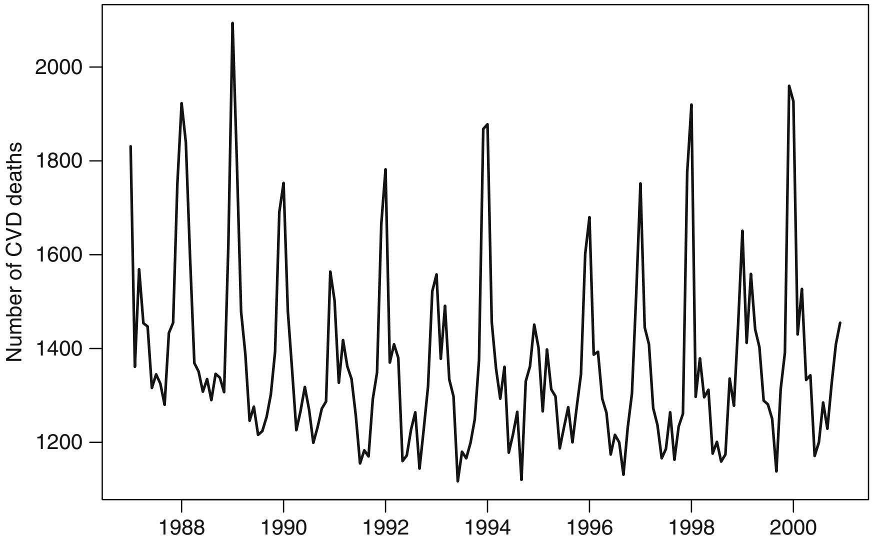

Fig. 1.1 Monthly counts of cardiovascular disease deaths in people aged ≥75 in Los Angeles for the years 1987–2000

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

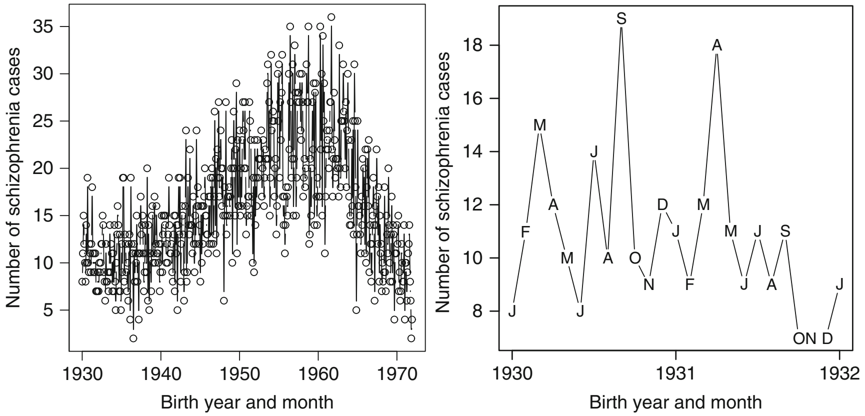

Fig. 1.2 Number of schizophrenia cases in Australia by date of birth from 1930 to 1971 (*left panel*), January 1930 to January 1932 (*right panel*)

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).

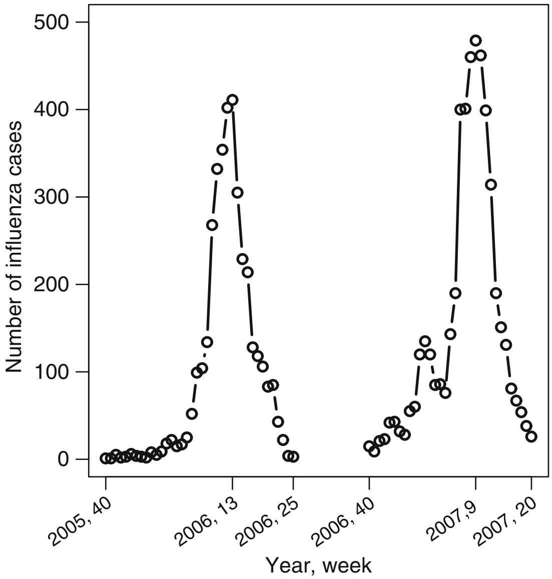

Fig. 1.3 Weekly number of positive influenza type B samples from the Centers for Disease Control and Prevention, United States, for the 2005/2006 and 2006/2007 flu seasons

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.

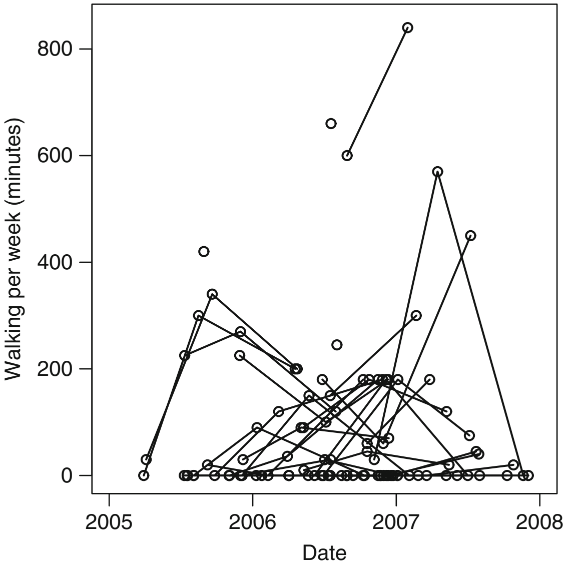

Fig. 1.4 Walking time in the last week (minutes) against date for 40 randomly selected subjects from the exercise study. Results from the same subjects are joined