Introduction to the time-series design

Aurelio Tobias

Spanish Research Council (CSIC)

Outline

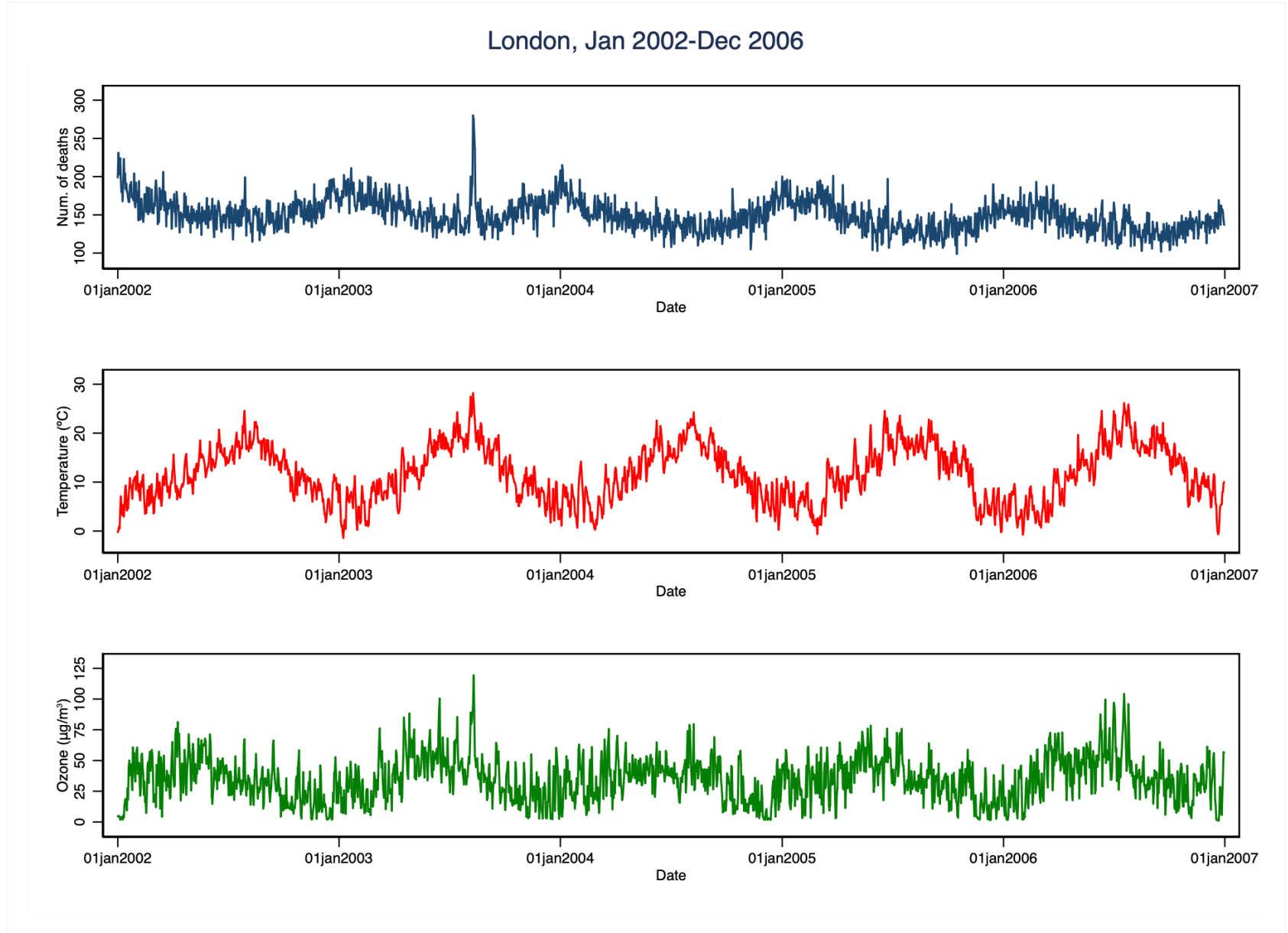
- Introduction
- Time-series design
- Principles of time-series regression
- Environmental risk exposures

Introduction

- **Health effects** are the changes in health status resulting from exposure to a given risk factor
  - **Short term effects –** Acute impact on health after an immediate exposure (time-series studies)
  - **Long term effects** – Chronic health effect after a cumulative exposure (cohort studies)
- **Health impact assessment** is the evaluation of potential health effects of proposed actions relative to a given exposure. The aim of HIA is to provide recommendations for decision-making process that will protect health

Introduction

- **Research question** – *"Is there an association between day-to-day variation in the environmental exposure and daily risk of health outcome"?*
  - Health outcomes and environmental exposures are characterized by **similar time-trends**
  - Measures of **individual predictors are usually not available**
  - We need a study design that relies on **between-day comparison within the same population** and able to control for time-trends
  - **Time-series data should be long enough** to identify day-to-day variation necessary to disentangle short-term effects from time-trends (e.g., at least 3 consecutive years for daily data)


London, Jan 2002-Dec 2006
Num. of deaths
100
150
200
250
300
01jan2002
01jan2003
01jan2004
01jan2005
01jan2006
01jan2007
Date
Temperature (°C)
0
10
20
30
01jan2002
01jan2003
01jan2004
01jan2005
01jan2006
01jan2007
Date
Ozone (µg/m³)
0
25
50
75
100
125
01jan2002
01jan2003
01jan2004
01jan2005
01jan2006
01jan2007
Date

Time-series data

- A time-series is a sequence of **measurements equally spaced** through time *(Zeger et al. 2006)*
- The **unit of analysis use to be the day (t)**, not the individual person (i)
- However, it could be annual, monthly, weekly or hourly
- The **outcome is a count** (e.g., number of deaths)

- **Example** – First 10 rows of time-series data (London, Jan 2002 – Dec 2006)

obs	date	deaths	temp	ozone
1.	01jan2002	199	-0.2	4.6
2.	02jan2002	231	0.1	4.9
3.	03jan2002	210	0.9	4.7
4.	04jan2002	203	0.5	4.1
5.	05jan2002	224	4.2	2.0
6.	06jan2002	198	7.1	2.4
7.	07jan2002	180	5.2	4.1
8.	08jan2002	188	3.5	3.1
9.	09jan2002	168	3.2	2.1
10.	10jan2002	194	5.3	5.2

Time-series design

- **Strengths**
  - Use of administratively collected data
  - Same population is compared with itself, focus is day-to-day variation
  - Time-invariant or slowly varying individual risk factors controlled by design (e.g., gender, age, smoking)

- **Limitations**
  - Ecological design based on aggregated, not individual data
  - Not applicable to estimate log-term (chronic) effects
  - Sensitive to choices for modelling time-trends and exposure-response associations

Poisson regression

- Similar in principle to any regression analysis but with **some specific features**
- Poisson regression
  - $Y|x \sim Poisson(\mu)$
  - $E(Y|x) = \mu$
  - $log(Y) = \beta_0 + \beta_1 x$
    $with\ V(Y|x) = \mu$
- Why not to use ARIMA?
  - Lack of knowledge/training
  - Epidemiological interpretability
  - Non-normal distribution for most cause-specific health outcomes

Poisson regression

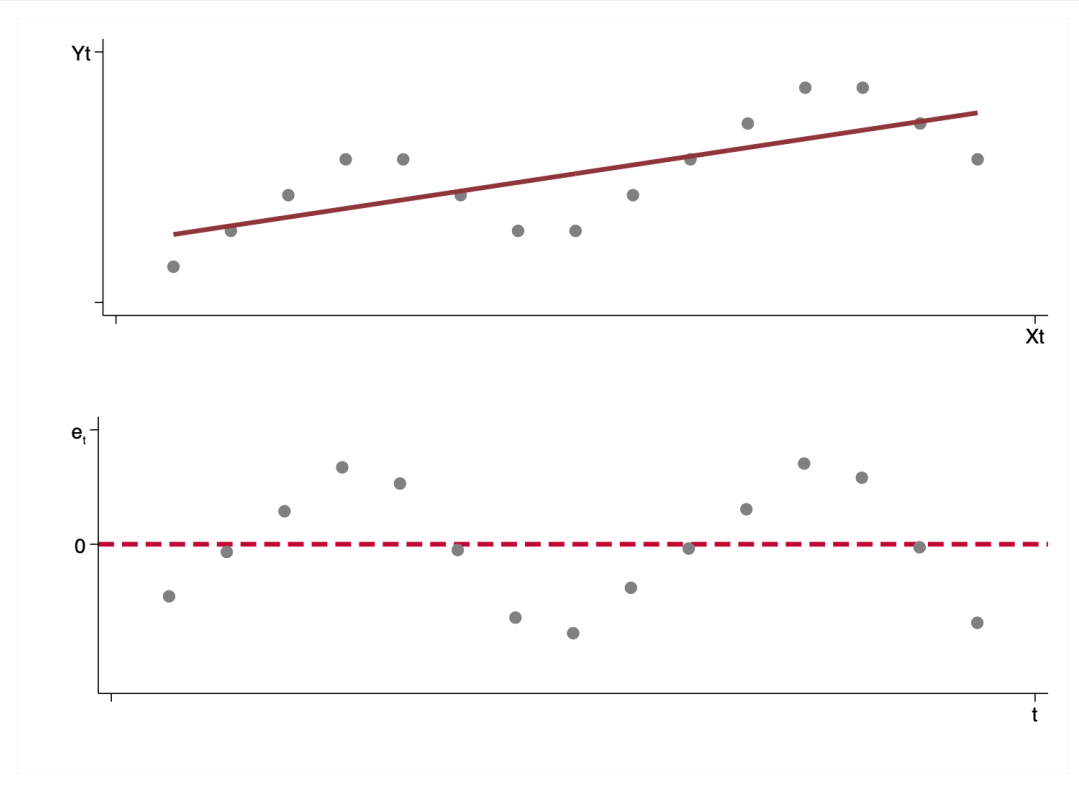
- Similar in principle to any regression analysis but with **some specific features**
- Poisson regression
  - $Y|x \sim Poisson(\mu)$
  - $E(Y|x) = \mu$
  - $log(Y) = \beta_0 + \beta_1 x$ $with\ V(Y|x) = \mu$
- Estimating effects
  - $exp(\beta_0) = \mu_0$ **is the baseline incidence rate**
  - $exp(\beta_1) = \mu/\mu_0$ **is the relative risk** (RR) for 1 unit increase of the exposure x
  - $(RR - 1)\times 100\%$ is the **percentage risk increase**

Model assumptions

- **Population size is not a big concern**
  - Assume that **today's population size in a city does not change much from yesterday's** population
  - **Denominator** (i..e, underlying population) **remains constant** over the study period
  - However, the **long-term change in the population size should be considered** in the model, if it exists
- **Overdispersion**
  - In Poisson regression we often find data with $V(Y|x)$ larger than $E(Y|x)$
  - **Underestimates the standard errors**
  - Usual **solution is to use quasi-Poisson**, where $V(Y|x) = \phi\mu$

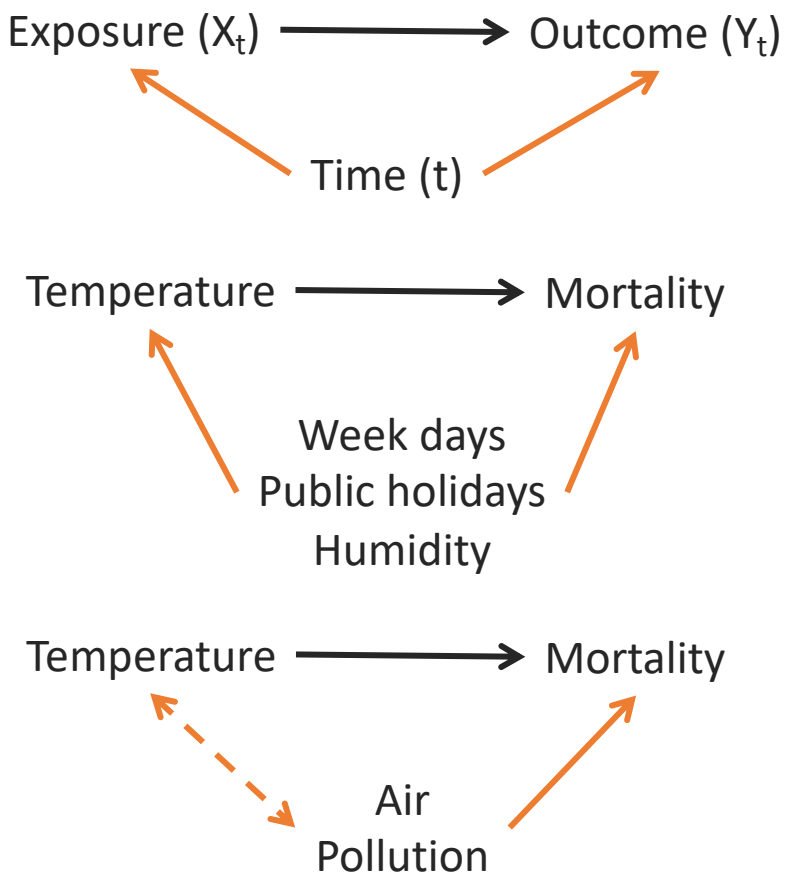
Model assumptions

- **Residual autocorrelation**
  - The **key assumption** for regression models is that the **outcome measures are independent** $(e_t \sim iid)$
  - However, in time-series, closer events tend to be more similar than those further apart in time
  - It also **affects standard errors** and should be accounted to make valid inferences

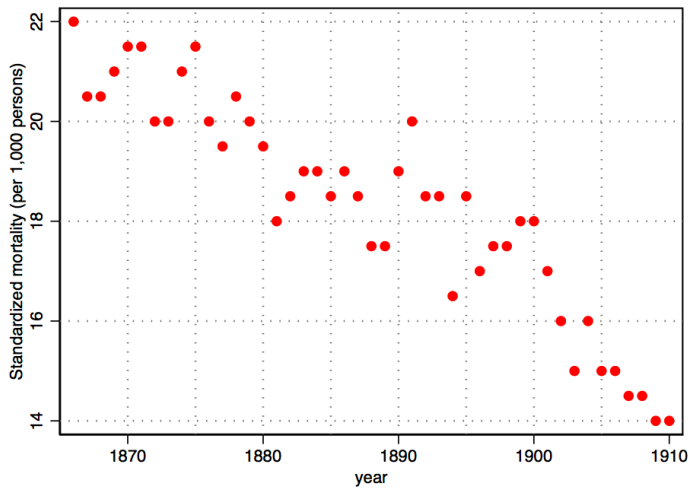


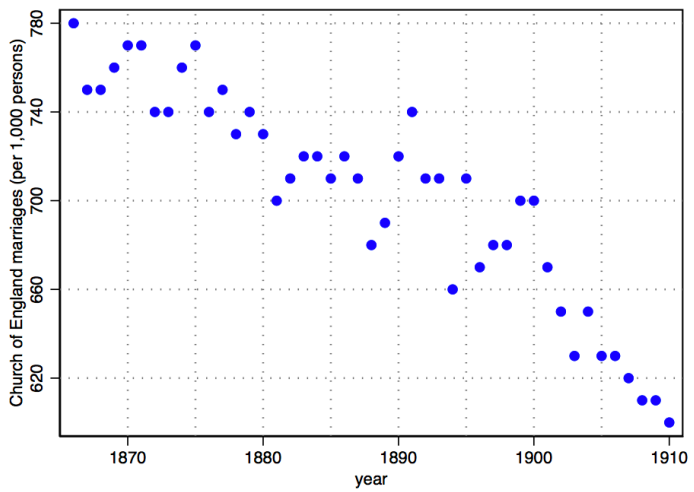
Confounding

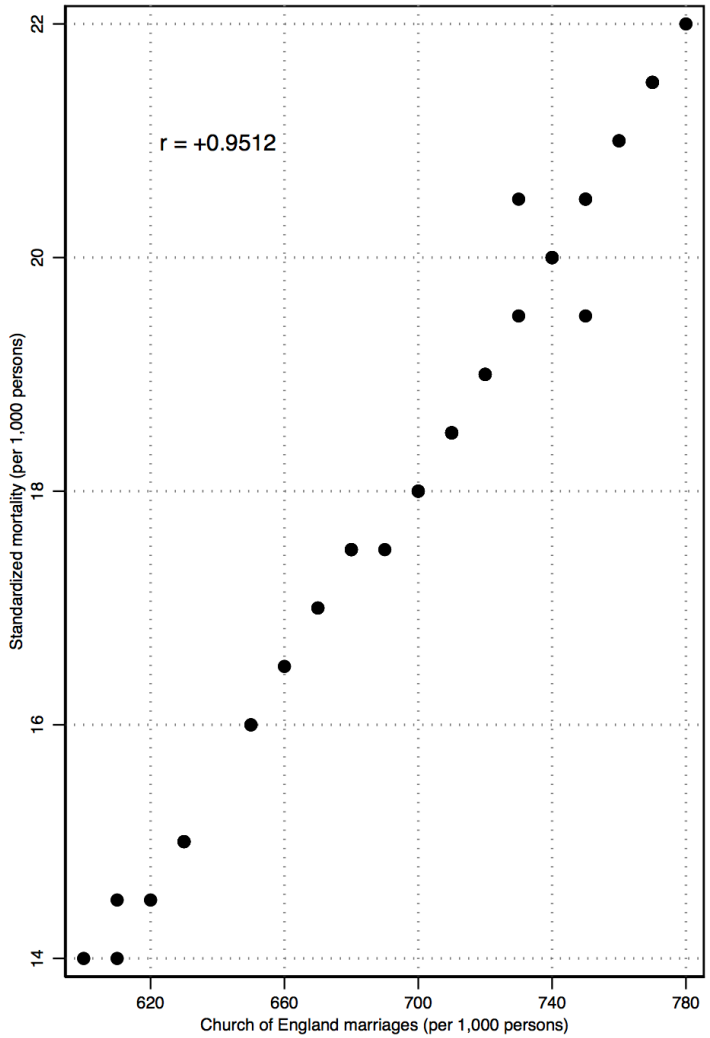
- It must be **associated with the exposure** (X) being investigated
- It must be independently **associated with the outcome** (Y) being investigated
- It must **not be on the causal pathway** between exposure (X) and outcome (Y)



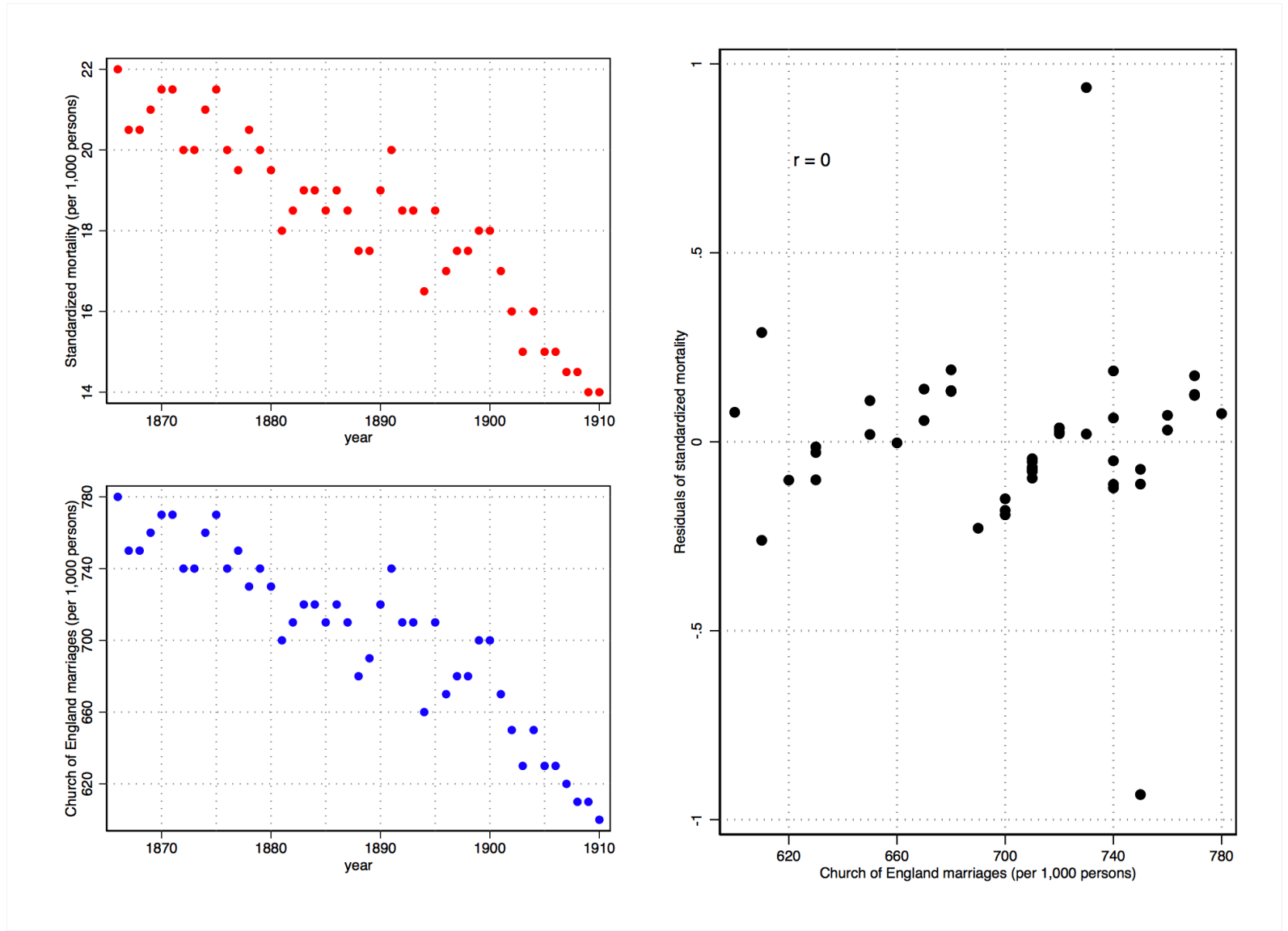
Yule GU. **Why do we sometimes get nonsense correlations between time series?** J Royal Stat Soc Sci. 1926;89:1-64.







Yule GU. **Why do we sometimes get nonsense correlations between time series?** J Royal Stat Soc Sci. 1926;89:1-64.

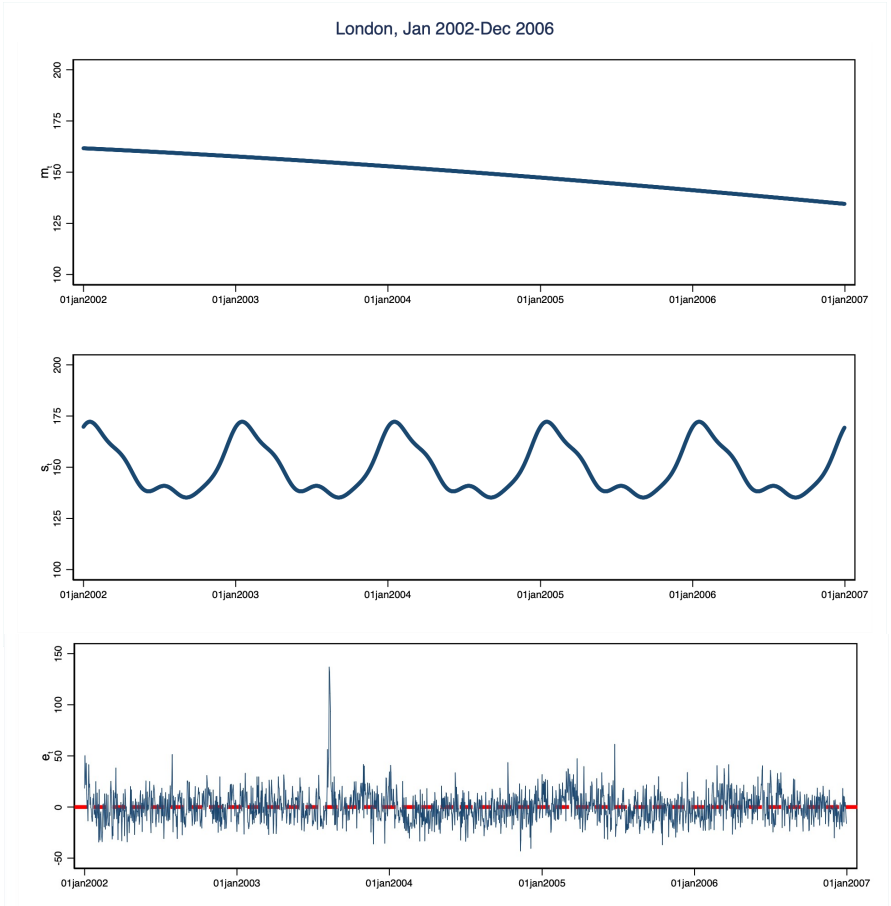


Modelling framework

- **Temporal decomposition**

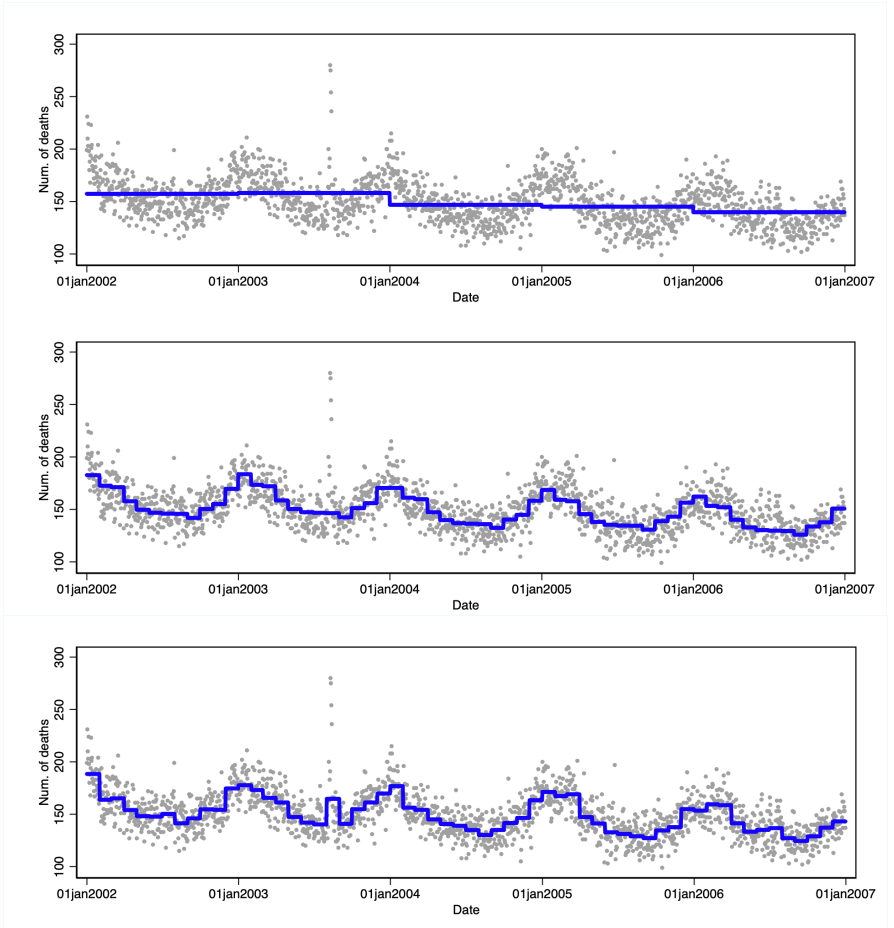
$$Y_t = m_t + s_t + e_t$$

  - *With m_t and s_t as time components (long trend and seasonality) and e_t as residual series*
- Underlying trends are **filtered-out** from the time series, allowing the inspection of associations at shorter time scale



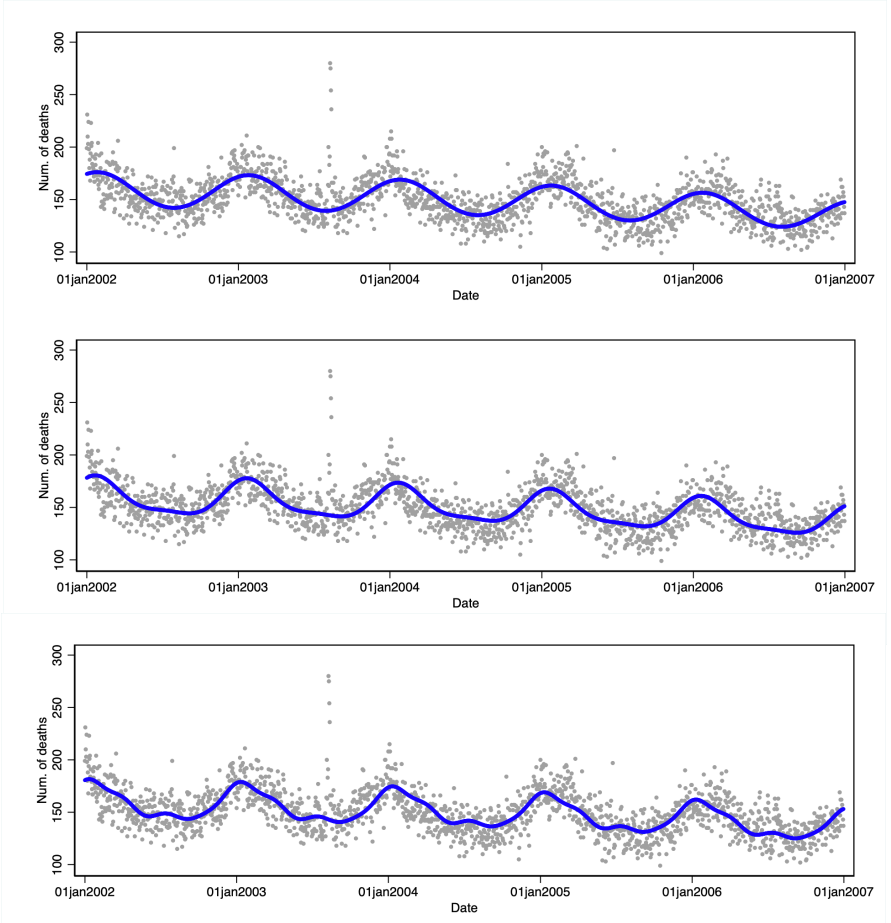
Time-stratified model

- Split the study period into time-intervals estimating a different baseline mortality risk
- Use of indicator variables for year and month
- $Y_t = \beta_0 + \sum \beta_i year_i + \sum \beta_j month_j$
  - *Easy to understand, and often captures main long-term patterns*
  - *Implicitly assumes biologically implausible jumps in risk between adjacent months*



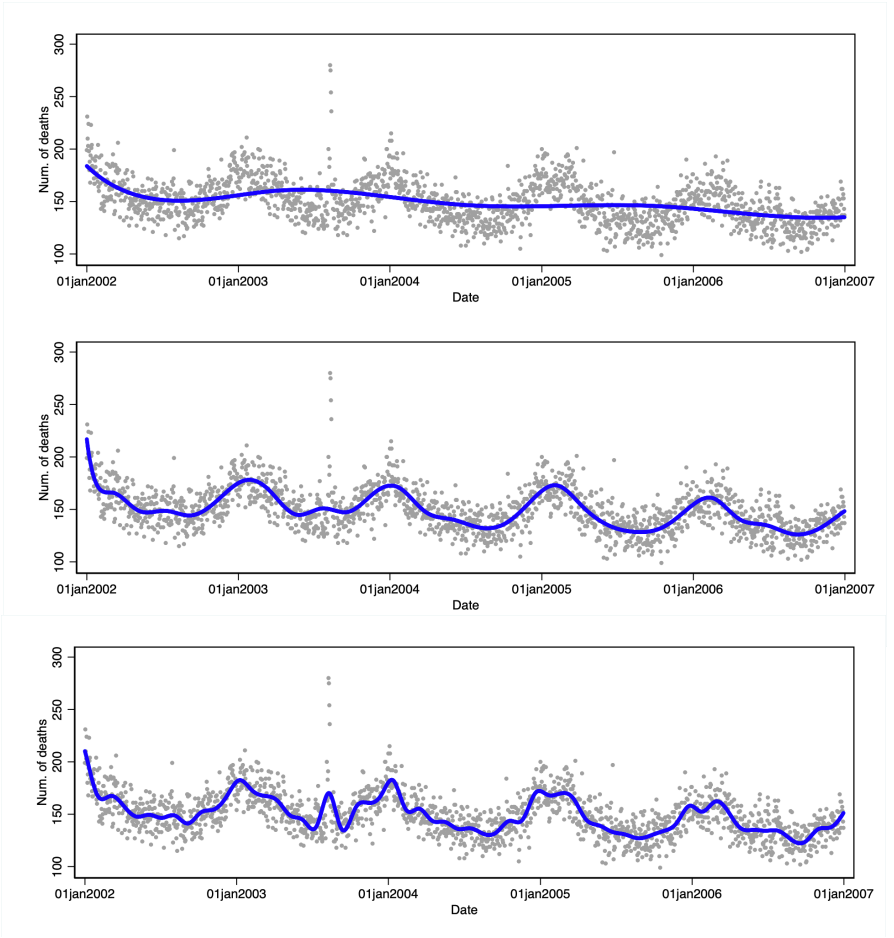
Periodic functions

- Fourier terms (pairs of sine and cosine functions of time) to model seasonal variation in the outcome as a regular wave each year
- $Y_t = \beta_0 + \beta_1 t + \sum \beta_i \sin\left(\frac{k\pi t}{T}\right) + \sum \beta_j \cos\left(\frac{k\pi t}{T}\right)$
  - *Suitable to capture very regular seasonal patterns*
  - *The modelled seasonal pattern is forced to be the same for each year, which may not reflect the data well*



Spline functions

- A set of polynomial curves (commonly cubic) that are joined smoothly end-to-end to cover the study period
- Basis variables are functions of the calendar time
- $Y_t = \beta_0 + f(t)$
  - *Capture seasonal patterns in a way allowed to vary from each year*
  - *It is necessary to decide how many knots there should be, which governs how flexible the curve will be*

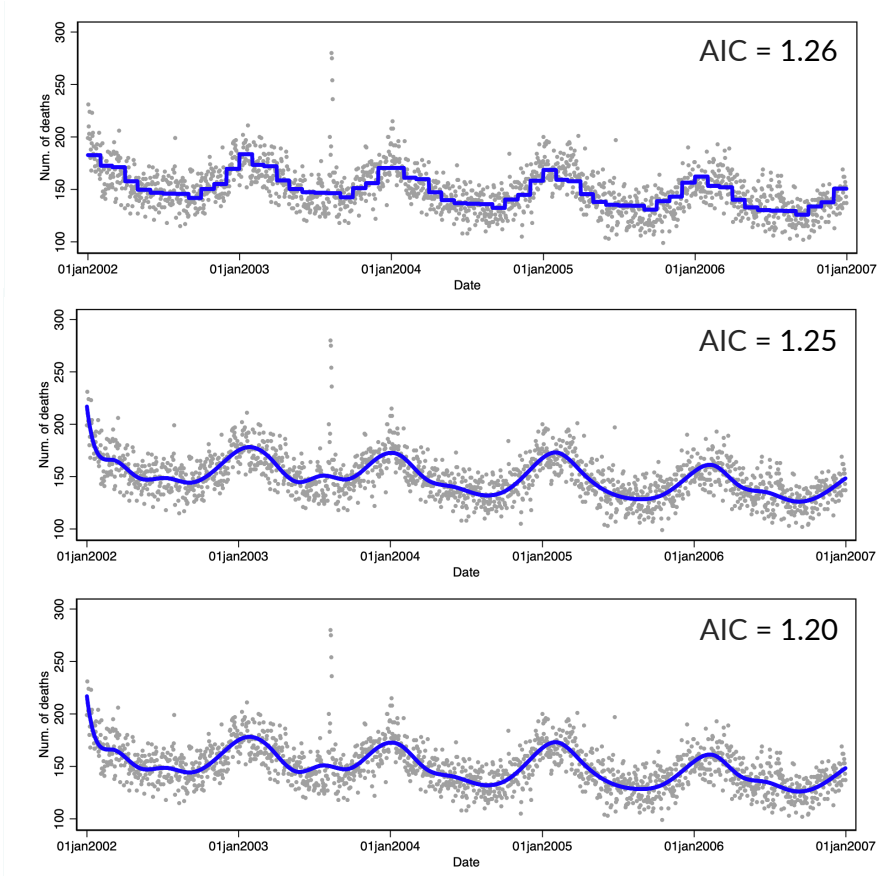


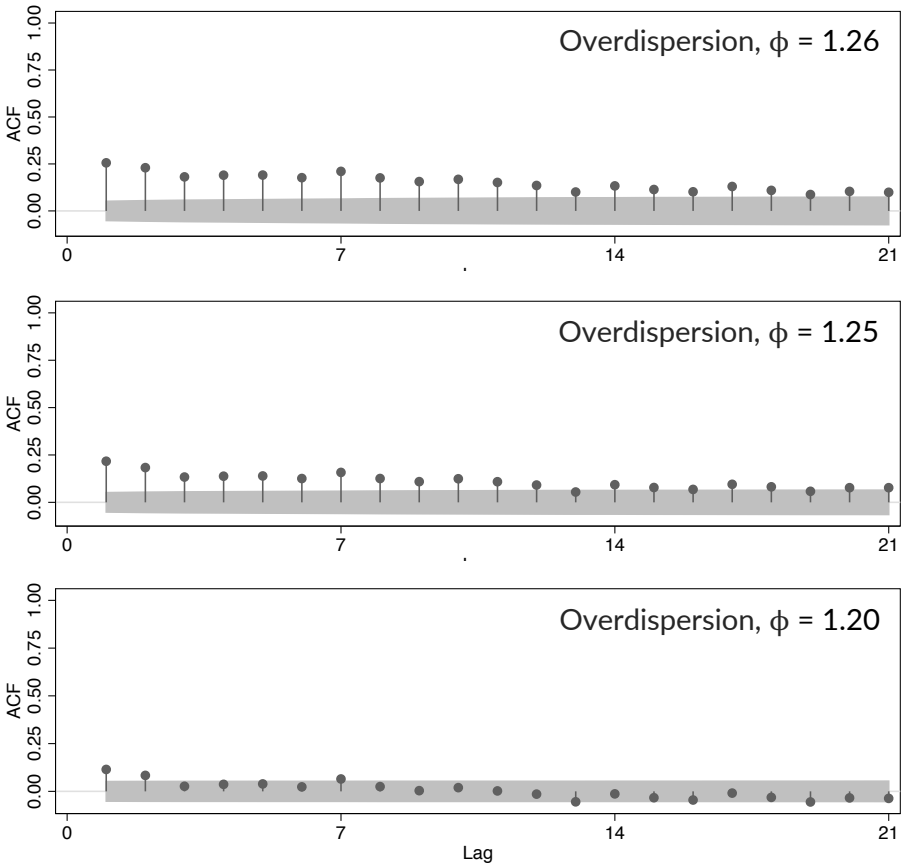
Criteria for model selection

- How to select the **“right” model** among the alternatives? (e.g., smoothing degree for time-trend)
- **Regression diagnostics** (e.g., residuals) might be of help in identifying deviations from model assumptions
- **Statistical criteria**, such as information indices (e.g., log likelihood, AIC), minimizing autocorrelation
- None has been proved as a general reliable method. **A priori assumptions based on epidemiological literature** and sensitivity analysis are recommended

Model selection

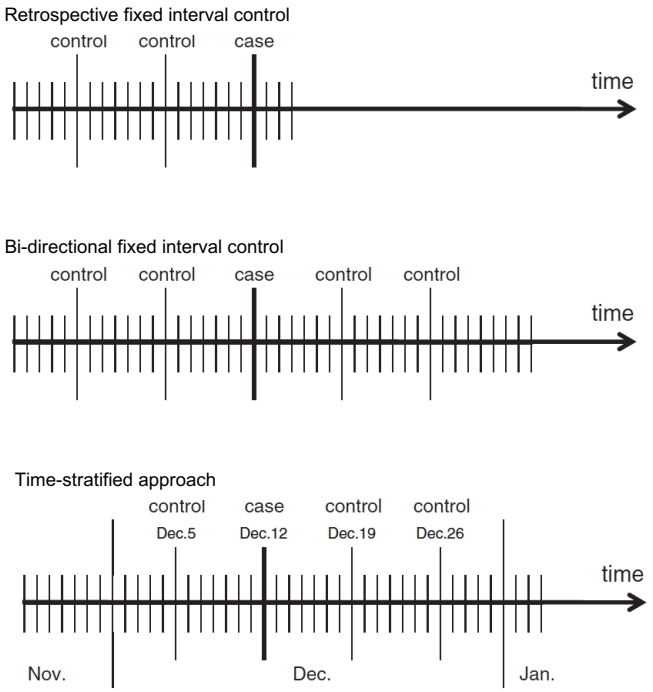
- Time-stratified model
- Periodic functions
- Spline functions





Case-crossover design

- Compare the levels of environmental exposure of the current day with that of the previous (fixed-interval) and later (bi-directional) days
- It allows to **control long trend and seasonality by design**, by matching the same days of the week, month and year (time-stratified)
- $Y_t = \beta_0 + \sum \beta_i year_i + \sum \beta_j month_j + \sum \beta_k dow_k + \sum \beta_{ij} year_i \times month_j + \sum \beta_{ik} year_i \times dow_k + + \sum \beta_{jk} month_j \times dow_k + \sum \beta_{ijk} year_i \times month_j \times dow_j$
- However, it is much more efficient the use **Conditional Poisson regression** models



Environmental risk exposures

- Associations between environmental exposures and health outcomes
  - Can show different type of shapes, usually a linear association with air pollution and non-linear with temperature
  - Are often characterized by lagged effects
- We need to model potentially complex temporal patterns of risk due to time-varying exposures
- It requires a knowledge previous knowledge about the shape of the exposure-response function and the lagged effects

Summary

- Time-series studies provide evidence on short-term associations between environmental exposures and health outcomes
- Time-series regression is similar in principle to any regression analysis but with some specific features
  - Residual autocorrelation
  - Control for time-trends and seasonality

References

- Erbas B, Hyndman R. Data visualisation for time series in environmental epidemiology. J Epidemiol Biostat 2001;6:433-43
- Zeger S, et al. On time series analysis of public health and biomedical data. Annu Rev Public Health 2006;27:57-79
- Bhaskaran K, et al. Time series regression studies in environmental epidemiology. Int J Epidemiol 2013;4:1187-95
- Madaniyazi L, et al. Should We Adjust for Season in Time-Series Studies of the Short-Term Association Between Temperature and Mortality? Epidemiol 2023;34:313-317
- Peng R, Dominici F. Statistical Methods for Environmental Epidemiology with R – A Case Study in Air Pollution and Health. Springer: New York, 2008.