ETH zürich

Applied Generalized Linear Models (FS 20)

Introduction

Viviana Amati

What is this course about?

- A family of statical models
- Emphasis on basic ideas and explanation of methods from the point of view of an applied researcher
- Methods are illustrated using data sets from different disciplines mainly from the social, economic and behavioural sciences

What is a statistical model?

A statistical model is a simplified, analogous and necessary representation of the reality for a purpose

- simplified: expressing a complex reality in a parsimonious way
  (Occam's razor or lex parsimoniae)
- analogous: similar to the reality
- necessary: for understanding the reality
- representation: it stands for something in the real world
- purpose:
  - explanation
  - prediction

What is a statistical model?

A formula

A statistical model is a mathematical formula

$$Y = f(X_1, X_2, \ldots, X_p; \beta_0, \beta_1, \ldots, \beta_p) + \varepsilon \quad ,$$

where

- Y: a **dependent** (response, outcome) variable
- X_j: **explanatory/independent** variable (covariate, predictor)
- β_j: **parameter** to be estimated from the data
- f: function expressing the relation between Y and $X_1, X_2, \ldots, X_p$
- ε: random term (error term)

What is a statistical model?

A formula

A statistical model is a mathematical formula

$$Y = f(X_1, X_2, \ldots, X_p; \beta_0, \beta_1, \ldots, \beta_p) + \varepsilon \quad ,$$

The model structure is defined by

- the nature of Y and $X_1, X_2, \ldots, X_p$
- the data structure
- the function f

What is a statistical model?

A probability distribution

- A statistical model for Y is a family of probability distributions

$$\{P(y;\beta,x),\ y \in \mathcal{Y},\ \beta \in \mathcal{B}, x \in \mathcal{X}\} \tag{1}$$

  indexed by the parameter β

- $\mathcal{Y}$: support of Y (i.e. set of values taken by Y)
- The probability distribution in (1) assigns a probability to all the values $y \in \mathcal{Y}$

What is a statistical model?

A probability distribution

- A statistical model for Y is a family of probability distributions

$$\{P(y;\beta,x),\ y\in\mathcal{Y},\ \beta\in\mathcal{B},x\in\mathcal{X}\} \tag{1}$$

  indexed by the parameter β

- $\mathcal{Y}$: support of Y (i.e. set of values taken by Y)

- The probability distribution in (1) assigns a probability to all the values $y\in\mathcal{Y}$

Example: linear regression model

What is a statistical model?

A probability distribution

- A statistical model for Y is a family of probability distributions

$$\{P(y;\beta,x),\ y \in \mathcal{Y},\ \beta \in \mathcal{B}, x \in \mathcal{X}\} \tag{1}$$

  indexed by the parameter β

- $\mathcal{Y}$: support of Y (i.e. set of values taken by Y)
- The probability distribution in (1) assigns a probability to all the values $y \in \mathcal{Y}$

Example: linear regression model
The assumptions of the linear regression model implies that

$$Y|X \sim N\Big(\beta_0 + \sum_j \beta_j X_j, \sigma^2\Big)$$

Generalized linear models

GLMs (Nelder and Wedderburn, 1972) are a class of statistical models for the analysis of quantitative and qualitative data

A GLM consists of **three components:**

1. A ***random component*** ε determining the conditional distribution

$$Y|X_1,\ \ldots,\ X_p$$

2. A ***linear predictor*** η: a linear function of the explanatory variables

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

3. A linearizing ***link function***

$$g(E[Y|X_1,\ \ldots,\ X_p]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Generalized linear models

GLMs (Nelder and Wedderburn, 1972) are a class of statistical models for the analysis of quantitative and qualitative data

A GLM consists of **three components:**

1. A ***random component*** ε determining the conditional distribution

$$Y|X_1,\ \ldots,\ X_p$$

2. A ***linear predictor*** η: a linear function of the explanatory variables

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

3. A linearizing ***link function***

$$g(E[Y|X_1,\ \ldots,\ X_p]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Example: what are the three components in linear regression models?

A note on causation and association

- A model describes the relationship between Y and X
- A causal relation usually has an asymmetry:

$$X \to Y$$

  X has an influence on Y but not viceversa

A note on causation and association

- A model describes the relationship between Y and X
- A causal relation usually has an asymmetry:

$$X \rightarrow Y$$

  X has an influence on Y but not viceversa
- A relationship to be causal must satisfy three criteria:
  - association between Y and X: as X changes, Y also changes
  - appropriate time order: the cause precede the effect
  - elimination of alternative explanations:
    association might be explained by other variables that may not have been measured
- With observational studies we cannot prove that one variable is a cause for another variable. This is however possible in randomized experiments

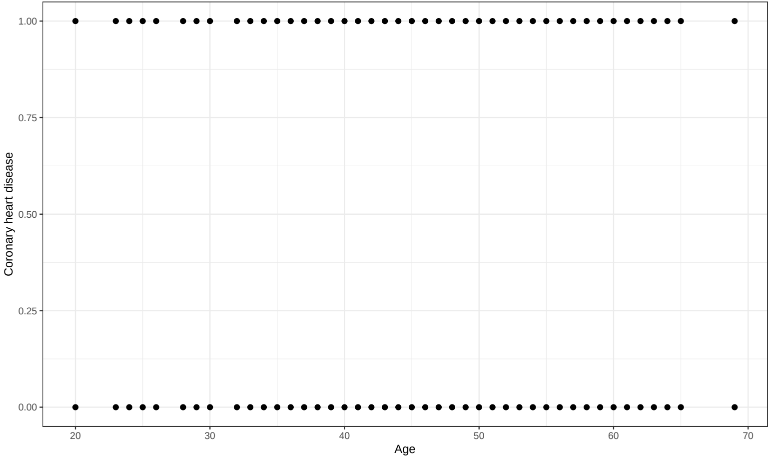
Generalized linear models (GLMs)

- Linear models for quantitative responses
  (multiple regression, analysis of variance and covariance)
- Models for binary response
  (binary logistic regression and probit models)
- Models for polythomous data
  (multinomial and ordered logistic regression and probit models)
- Log-linear models for count data
  (Poisson and Negative binomial regression models)
- Survival models for failure time data

This course covers the basic theory, methodology, and application of GLMs.
A few examples follow.

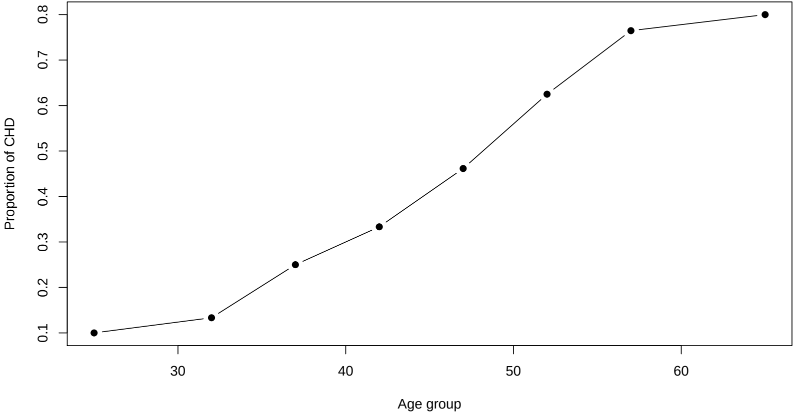
Example: binary logistic regression model (Hosmer et al., 2013)

Developing a coronary heart disease as a function of age



Example: binary logistic regression model (Hosmer et al., 2013)

Developing a coronary heart disease (CHD) as a function of age



Binary logistic regression model

- Y: binary response variable (y=1 success, y=0 failure)

- Bernoulli distribution

$$P(Y=y) = \begin{cases} \pi & \text{if } y = 1 \\ 1-\pi & \text{if } y = 0 \end{cases}$$

- $E[Y] = \pi$

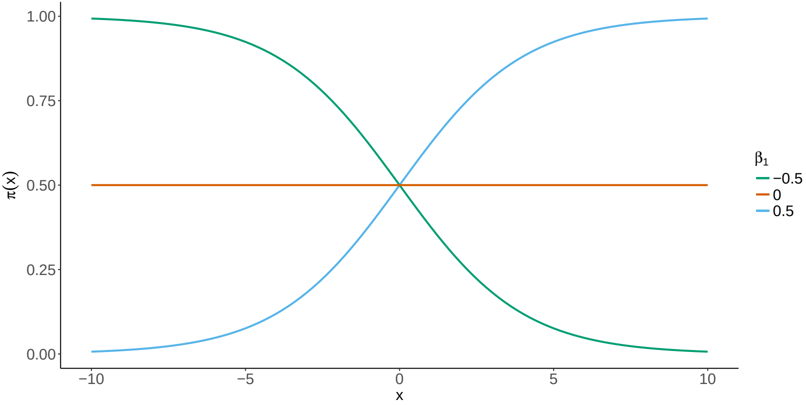
- We could mimic the linear regression model

$$E[Y|X] = \pi = \beta_0 + \beta_1 X_1$$

but this model has structural defects

Binary logistic regression model

$$\text{logit}[\pi] = \log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 X_1$$

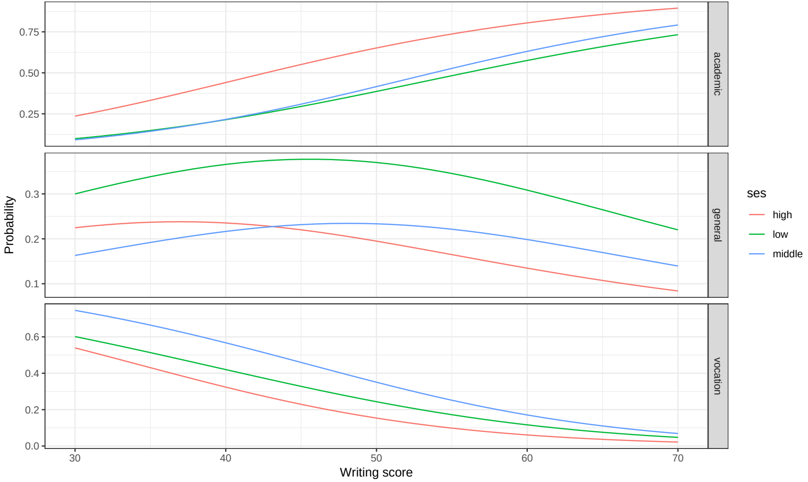


Binary logistic regression model

- probability that a subject is credit worthy given its credit score and payment history
- probability of a failure component based on environmental conditions (e.g., temperature)
- probability of inheriting an allele of one type based on phenotypic variables
- probability of a death penalty verdict based on race of the defendant and race of victims
- probability of migration based on socio-economic variables

Example: Multinomial logistic regression model

High school students' program choice, given writing score and social economic status



Multinomial logistic regression model

- Extension of binary logistic regression model
- Y is nominal and polytomous, i.e. has $M \geq 2$ nominal categories
- $\pi_m = P(Y = m)$
- Multinomial models pair each category with the baseline category M

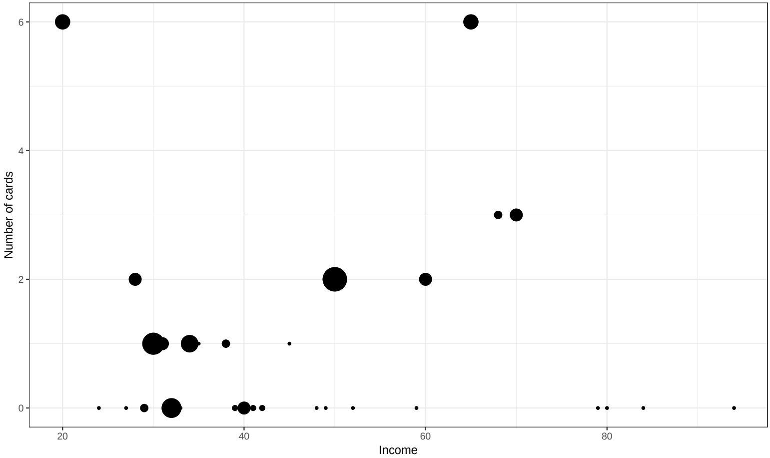
$$\text{logit}[\pi_m] = \log\left[\frac{\pi_m}{\pi_M}\right] = \beta_{0m} + \beta_{1m}X_1 + \ldots + \beta_{pm}X_p, \qquad m = 1, \ldots, M-1$$

Multinomial regression model

- decision on shopping destination (A, B, C) based on retail, shopping opportunity, price of the trip (time and fuel)
- political ideology (very liberal, slightly liberal,moderate, slightly conservative, very conservative) by gender and political party affiliation
- belief in afterlife (yes, undecided, no) based on gender, religion and age
- alligator food choice (fish, reptile, bird, other) based on alligators' size and lake they live in

Example: Poisson regression model

Number of credit cards a person can have, given his/her income



Poisson regression model

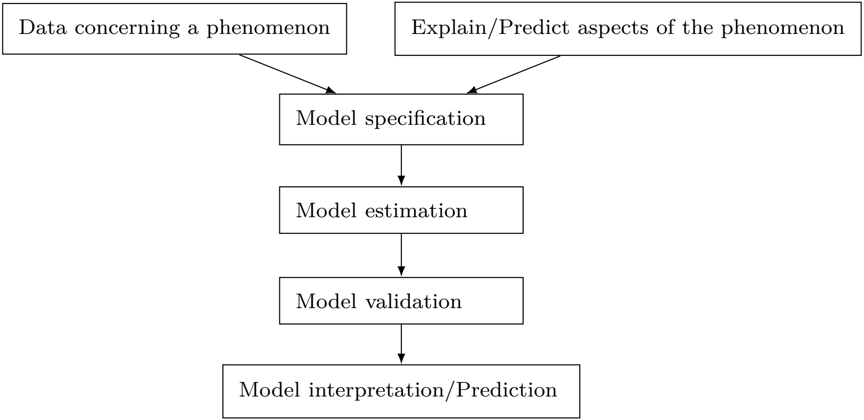
- Models for count data
- $Y = 0, 1, 2, \ldots$
- $Y \sim \text{Poisson}(\lambda)$
- Poisson regression model

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

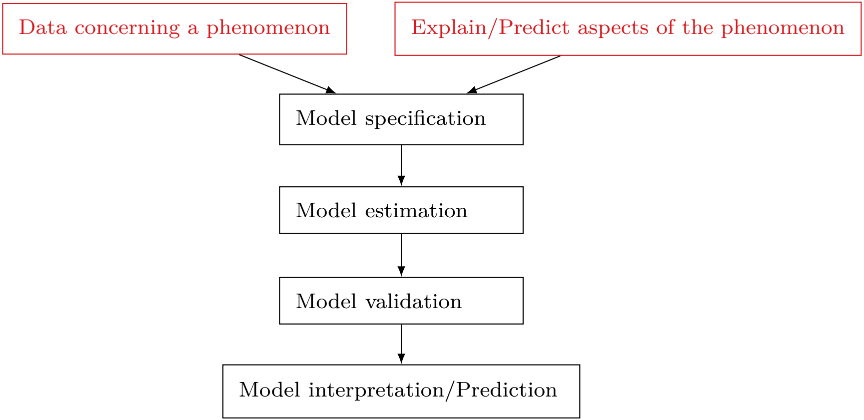
Poisson regression model

- Number of nests in a city based on the level of noise and pollution and the presence of parks
- Number of fatal injuries in car accidents based on the safety equipment in use (e.g. seat belt)
- Number of aggressive acts by children during a playground period based on age, gender, race
- Number of fissures that develop in turbine wheels

Modeling framework

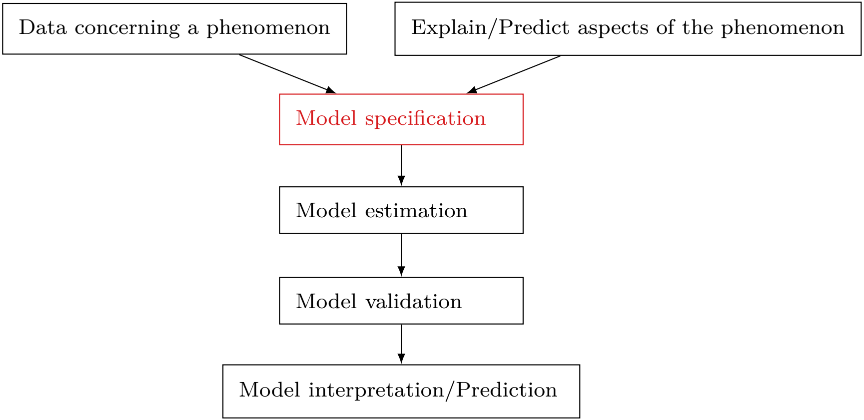


Modeling framework



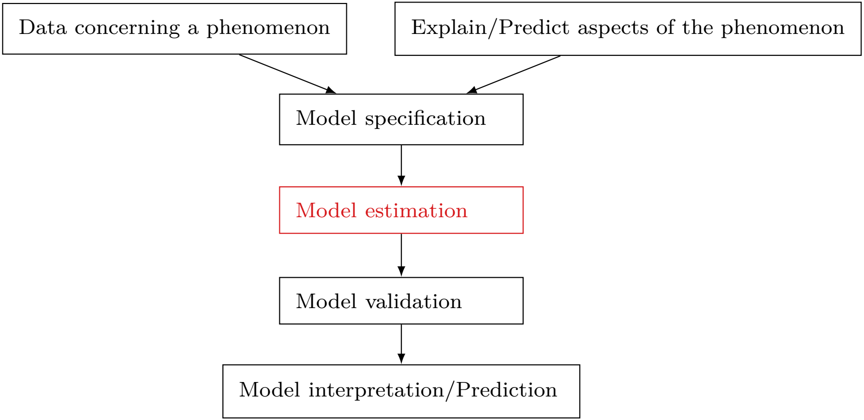
Input

Modeling framework



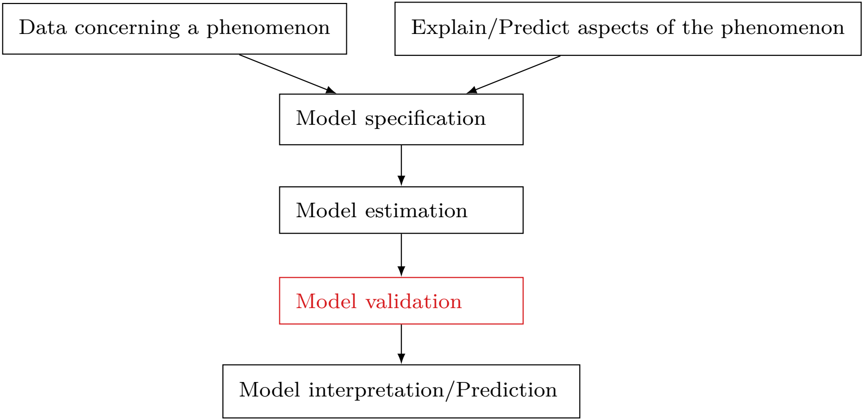
Select dependent and explanatory variables and identify ε and g

Modeling framework



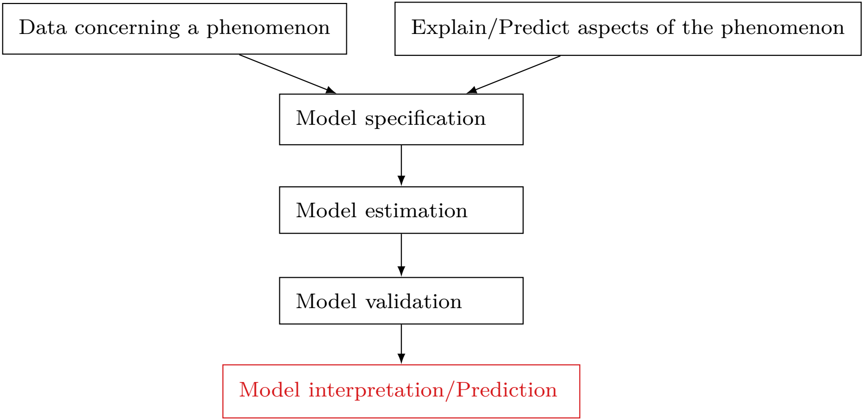
Estimate the parameter β using the available data

Modeling framework



Check the fit and the assumptions of the model

Modeling framework

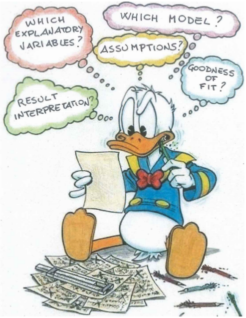


Interpret the results or use the model for prediction

Objectives

After the course, you should be able to:

- describe the various methods and the related theory
- identify adequate models for a given statistical problem
- use the R software to perform the analysis
- validate the model
- interpret the output
- make prediction
- apply the methods to your own data



Prerequisites

- Good knowledge of basic mathematical and statistical concepts
  A sound understanding of estimation methods, hypothesis testing and linear regression models (OLS) is required

- Strong mathematical soft skills
  e.g. ability to understand and work with mathematical definitions and equations, elements of linear algebra

- Strong statistical soft skills
  e.g. performing descriptive analysis (frequency distributions, mean and variance computation, data visualization) to better understand the data that will be analyzed using GLMs

Course structure

- Time: Tuesday, 17.15 to 19.00 (with a break)

- The course consists of lectures and practical parts

  – Lecture:
    A new instance of the GLM family is introduced
    Model definition, specification and parameter estimation

  – Practical:
    Applications of the introduced model are illustrated using the R software (please bring your laptop)
    Case studies drawn from social, economic, engineering, and behavioral sciences are used to illustrate the estimation, assessment and interpretation of GLMs

Course structure - schedule

Date	Topic	Assignment
18.02	Introduction to the course	Ass. 1 released on 25.02
25.02	Introduction to R and review of the linear regression model	
03.03	The general linear model: ANOVA and ANCOVA	
10.03	Practical: ANOVA and ANCOVA	
17.03	Binary outcomes: logistic regression and probit models	Ass. 2 released on 17.03
24.03	Practical: logistic regression and probit models	
31.03	Nominal outcomes: multinomial logistic regression	
07.04	Practical: multinomial logistic regression	
21.04	Ordinal outcomes: ordered logistic regression and probit models	Ass. 3 released on 21.04
28.04	Practical: ordered logistic regression and probit models	
05.05	Count outcomes: Poisson and negative binomial models	
12.05	Practical: Poisson and negative binomial models	
19.05	Survival models (lecture+practical)	
26.05	Exam	

- Assignments consist of 3 or 4 exercises involving output interpretation and analysis of datasets
- For solving assignments you are encouraged to work in groups of 3 or 4 people
- Assignments cover all the introduced models but the survival models

What do you need to do to pass the course?

- Evaluation:
  - Element A: 70% of the grade through three assignments
  - Element B: 30% of the grade through a (1-hour) written exam (on May 26 at 17.15)
  - You pass if you reach more than 50% of the points in each part

Material

- The course content is covered in lecture notes and R scripts that are made available in moodle after the lecture

  Please regularly check moodle for changes in the schedule of lectures and practical parts

- Assignments and corresponding data available in moodle

- Literature references for background information and a deeper study of theoretical foundations are listed in the lecture notes

Material

Useful books:

- Finlay, B., & Agresti, A. (1986). Statistical methods for the social sciences. Dellen.
- Fox, John. (2016). Applied regression analysis and generalized linear models (Third ed.). Los Angeles: SAGE.
- Fox, John, & Weisberg, Sanford. (2019). An R companion to applied regression (Third ed.). Los Angeles: SAGE.
- Hosmer, David W, Lemeshow, Stanley, & Sturdivant, Rodney X. (2013). Applied logistic regression. Hoboken: Wiley.
- Long, J. Scott. (1997). Regression models for categorical and limited dependent variables. Thousand Oaks, Calif: Sage Publications.

Next week (25.02.2019)

On Tuesday 25.02.2020 we will have a short introduction to R. Please

1. bring your laptop
2. install R
   `https://cran.r-project.org/`
3. install RStudio
   `https://www.rstudio.com/`

References

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.