MA50260 Statistical Modelling

Lecture 1: Introduction

Ilaria Bussoli

February 6, 2024

General Information

Sessions:

- **Lectures:**
  - Tuesdays at 9:15-10:05 in CB 5.6, and
  - Fridays at 10:15-11:05 in CB 3.16
- **Problem Classes:**
  - Fridays at 14:15-15:05 in CB 3.16

Office Hours: Wednesday 15:30-17:30 (start in Week 2),
2 South 1.04A

Course materials: Lecture notes and problem sheets on Moodle

Assessment: 100% Closed-book exam in May/June

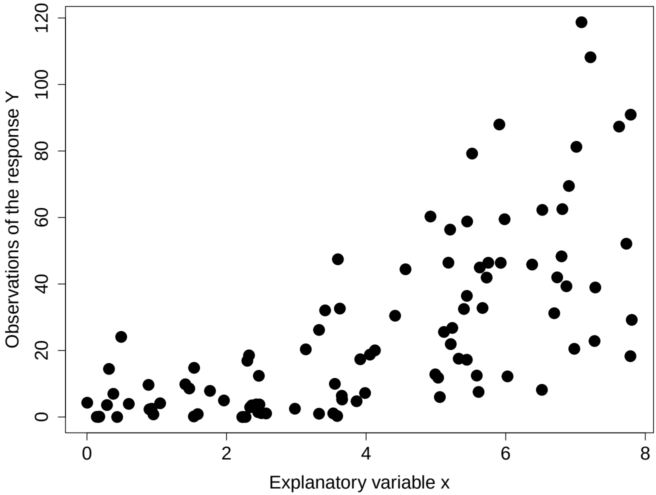
Statistical Models

- A **statistical model** incorporates random variation, which may
  - be intrinsic to the real-world process (biology, meteorology);
  - caused by imperfect measuring devices (physical sciences).

Statistical Models

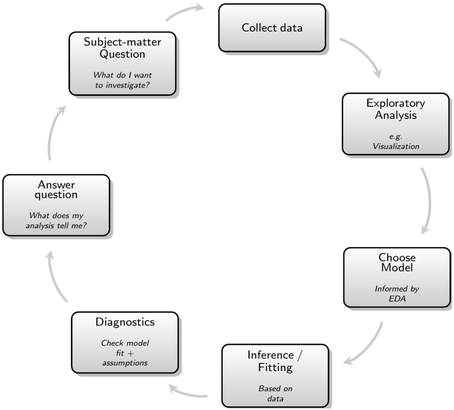
- A **statistical model** incorporates random variation, which may
  - be intrinsic to the real-world process (biology, meteorology);
  - caused by imperfect measuring devices (physical sciences).
- There is usually a variable of interest, the **response variable**.
- In this course, we seek to model the response variable conditional on a set of **explanatory variables**, which may help us to explain the variation in the response variable.
- We focus on **parametric statistical models**, which consist of a probability distribution F with unknown parameters $\underline{\theta}$.

Example



Observations of the response Y
0
20
40
60
80
100
120
0
2
4
6
8
Explanatory variable x

Philosophy of Statistical Modelling



This course introduces a range of models, their estimation and the subsequent diagnostics.

Remember

"All models are wrong, but some models are more useful than others."
— (George Box)

Types of Response Variable

Before defining any model, it is crucial to identify the variable types.

In this course, we use the following categorisation:

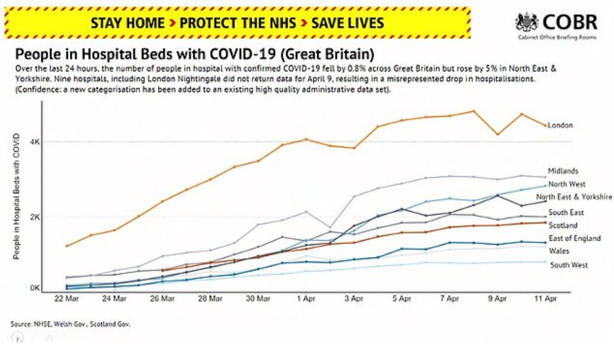
- **Continuous:** can take any decimal value.
- **Count:** represents a count (usually positive).
- **Categorical:** represents quality or preference.
- **Binary:** usually 0 and 1 and could represent a yes/no response.

Exercise - Identify the type of the variable

1. **Does someone prefer cats or dogs?**

Exercise - Identify the type of the variable

2. Beds occupied due to Covid 19?



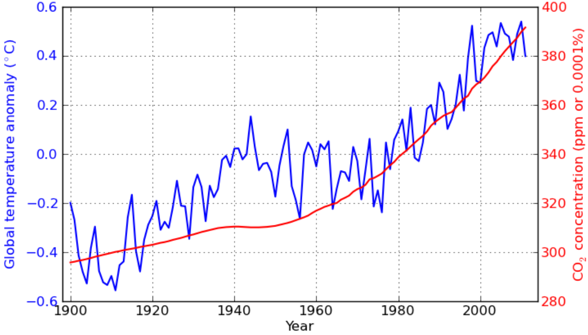
Exercise - Identify the type of the variable

3. Ratings in "Trustpilot"?

Lower	Upper	Stars	Star Label
1.0	1.2	1	Bad
1.3	1.7	1.5	Bad
1.8	2.2	2	Poor
2.3	2.7	2.5	Poor
2.8	3.2	3	Average
3.3	3.7	3.5	Average
3.8	4.2	4	Great
4.3	4.7	4.5	Excellent
4.8	5.0	5	Excellent

Exercise - Identify the type of the variable

4. CO_2 concentration in the atmosphere?



Statistical Background

Estimator and Estimate

Definition 1.1 (Estimator). An estimator $\hat{\theta}(Y_1, \ldots, Y_n)$ of θ is a function of the random variables $Y_1, \ldots, Y_n$.

Definition 1.2 (Estimate). An estimate $\hat{\theta}$ of θ is a function of the observed sample $y_1, \ldots, y_n$.

You may have already come across the **sample mean**

$$\hat{\mu}_n \;=\; \hat{\mu}(Y_1, \ldots, Y_n) \;=\; \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

and the **sample variance**

$$\hat{\sigma}_n^2 \;=\; \hat{\sigma}^2(Y_1, \ldots, Y_n) \;=\; \frac{1}{n-1} \sum_{i=1}^{n} \Big(Y_i - \bar{Y} \Big)^2 .$$

Properties of Estimators

Definition 1.3 (Unbiased). An estimator is unbiased if

$$\mathbb{E}\left[\hat{\theta}\left(Y_1, , \ldots, Y_n\right)\right] = \theta,$$

where θ is the unknown true value.

Definition 1.4 (Consistent). An estimator of a parameter θ is consistent if for all $\epsilon > 0$,

$$\mathbb{P}\left[\,\left|\hat{\theta}(Y_1, \ldots, Y_n) - \theta\right| > \epsilon\,\right] \to 0$$

as $n \to \infty$.

Parameter Estimation: Maximum Likelihood

Definition 1.5 (Likelihood function). For IID data $y_1, \ldots, y_n$ that arise from a population with pmf (or pdf) $f(\cdot)$, the **likelihood function** is defined as

$$L(\theta \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} f\left(y_i \mid \theta\right).$$

Parameter Estimation: Maximum Likelihood

Definition 1.6 (Maximum likelihood estimate). For $y_1, \ldots, y_n$, the **maximum likelihood estimate (MLE)** $\hat{\theta}$ is the value of θ that maximises $L(\theta \mid y_1, \ldots, y_n)$.

Definition 1.7 (Log-likelihood function). The **log-likelihood function** is

$$\ell(\theta \mid y_1, \ldots, y_n) = \log \left[L(\theta \mid y_1, \ldots, y_n) \right].$$

Asymptotic Distribution of the MLE

Let θ be a p-dimensional parameter vector. Then (subject to the likelihood function being smooth) as $n \to \infty$

$$\hat{\theta}(\mathbf{Y}) \sim \mathrm{MVN}_p\left(\theta, \mathcal{I}^{-1}(\theta)\right),$$

where

$$\mathcal{I}(\theta) \;=\; -\left[\mathbb{E}\left\{\frac{\partial^2 \ell(\theta \mid \mathbf{Y})}{\partial\theta_j\,\partial\theta_k}\right\}\right]_{j,k=1,\ldots,p}$$

is the **Fisher information matrix**.