Chapter 1

Scope and approach

This course requires a good understanding of the Part IB course *Statistics*. This course complements the Part II courses *Principles of Statistics* and *Mathematics of Machine Learning* by providing a more applied and computational perspective.

This year we will take a slightly different approach to statistical modelling. On the course website you will find the lecture notes from 2019, which take a more classical approach. Additionally, you might find the following books useful:

- A. Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley 2015. (Especially Chapters 2, 3, 4, 7.)
- P. MuCullagh, J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989. (A classic but lacks details and examples.)
- G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning (with Applications in R)*. Springer 2013. (Provides perspectives from machine learning.)
- D. Freedman. *Statistical Models: Theory and Practice*. Cambridge University Press, 2009. (Provides perspecties from causal inference and scientific applications.)

For mathematics students, it might not be obvious that *statistics is not a branch of mathematics.*[1] There is no consensus on the definition of statistics (especially with the rise of machine learning and data science), but the following definition in Wikipedia cannot be too wrong:

- *Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.*

Compare this with the following definition of mathematical statistics:

- *Mathematical statistics is the study of statistics from a mathematical standpoint, using probability theory as well as other branches of mathematics such as linear algebra and analysis.*

Another way to think about the difference is mathematics is mostly about deductive reasoning from a set of axioms and assumptions, while statistics is mostly concerned with inductive reasoning from empirical data.[2] Through exploring different statistical models and learning R, a great programming language for statistical computing, you will be exposed to both the mathematical and non-mathematical elements of statistics.

To understand how statistics are used in practice, the following quote by G. Box[3] may be illuminating:

> Scientific research is usually an iterative process. The cycle: conjecture–design–experiment–analysis leads to a new cycle of conjecture–design–experiment–analysis and so on.... The experimental environment ... and techniques appropriate for design and analysis tend to change as the investigation proceeds.

At one point, the dominant view was that statistical modelling is a critical step of "analysis" and the model is built after data are collected. However, modern statisticians (and in fact, many pioneers like Box and Fisher) view statistical model as an essential component of the scientific process that guides all steps of the cycle and is being continuously updated.

Another important realization in modern statistics is that statistical models may come at different levels:

(i) Models for conditional moments. For example, a *linear model for conditional expectation* assumes $\mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}^T\boldsymbol{\beta}$.

(ii) Models for joint or conditional distributions. For example, the *classical normal linear model* assumes $Y = \boldsymbol{X}^T\boldsymbol{\beta} + \epsilon$ where the noise variable $\epsilon \perp\!\!\!\perp \boldsymbol{X}$ and $\epsilon \sim \mathrm{N}(0, \sigma^2)$.

(iii) Structural or causal models that not only describe (associational) relationship for the data at hand but also (causal) relationship under counterfactual interventions. For example, the *linear structural equation model* assumes $Y^{(\boldsymbol{x})} = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon$, where $Y^{(\boldsymbol{x})}$ is the counterfactual value of Y under the intervention that sets $\boldsymbol{X}$ to $\boldsymbol{x}$ and ϵ is an independent noise variable.

This course will not consider the third type of statistical model; see Freedman's book for some good introduction to it.

This course will discuss the first two types of statistical models, which are often called *regression models*.[4] In particular, our focus will be on a class of models called *generalized linear models* (GLM), which extends the classical linear model by using a beautiful theory for exponential family distributions. In essence, a GLM assumes that the conditional distribution of Y given $\boldsymbol{X}$ is (almost) determined by the conditional expectation $\mathbb{E}[Y \mid \boldsymbol{X}]$.

Why do we care about regression problems? One obvious reason is their nearly ubiquitous presence in applications. Another reason is divide-and-conquer: the joint distribution of some random variables can always be factorized as a product of conditional distributions.

Why do we still care about (generalized) linear models, given the rise of machine learning algorithms that almost always have better prediction accuracy? Because GLMs are simple, elegant, and interpretable. Moreover, more complex models are often constituted by GLMs. For example, a neural network is essentially the composition of numerous GLMs (with the distributional assumptions stripped away).

Notation

Upper-case letters indicate matrices or random variables. Lower-case letters indicate fixed quantities. We use $\boldsymbol{I}_p$ to denote the $p \times p$ identity matrix, $\mathbf{1}_p$ to denote the p-vector of ones, and $\mathbf{0}_p$ the p-vector of zeros. Bolded symbols are vectors or matrices. Independent random variables (or vectors) X and Y are denoted as $X \perp\!\!\!\perp Y$. As a convention, we usually use subscript $i\ in\{1, \dots, n\}$ to index observations and $j \in \{1, \dots, p\}$ to index variables. "Independent and identically distributed" is abbreviated as "i.i.d.". The Euclidean norm of a vector $\boldsymbol{Y}$ is denoted as $\|\boldsymbol{Y}\|$. Convergence in distribution (weak convergence) is denoted as $\xrightarrow{d}$.

Notes

[1] Perhaps this is true for any non-statistician. When I told my neighbours that I am a statistician, most of their first reaction is that I do mathematics.

[2] Mathematics also involves induction, see G. Pólya's book *Mathematics and Plausible Reasoning*, but mathematical induction is a deductive method. Statistics also involves deductive reasoning (which is basically mathematical statistics).

[3] Abstracts. (1957). *Biometrics*, *13*(2), 238–246.

[4] The terminology "regression" was derived from a statistical phenomenon called "regression toward the mean" discovered by F. Galton. The original meaning of regression is no longer relevant today, but the terminology was kept for historical reasons.