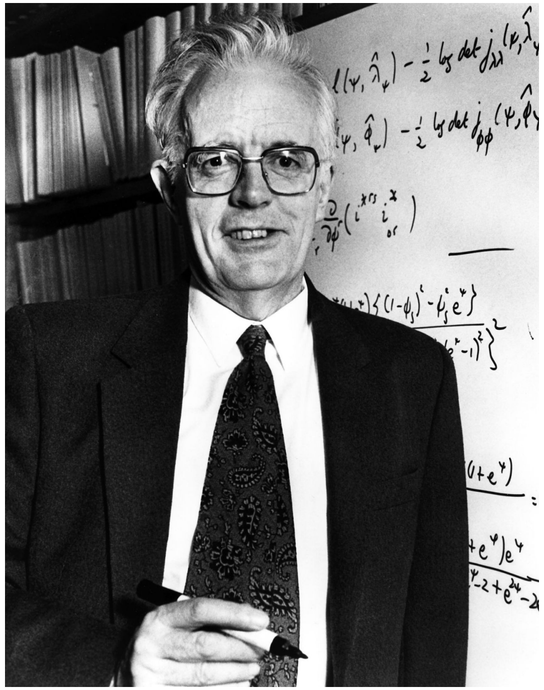
Introduction

Statistics III - CdL SSE

Tommaso Rigon
Università degli Studi di Milano-Bicocca

Homepage



"I would like to think of myself as a scientist, who happens largely to specialise in the use of statistics."

Sir David Cox (1924-2022)

- *Statistica III* is a monographic course on **Generalized Linear Models** (GLMs), a broadly applicable **regression** technique.
- This is a **B.Sc.-level** course, but there are some prerequisites: it is assumed that you have already been exposed to:
  - **Simple linear regression**, from *Statistica I*;
  - **Inferential statistics**, from *Statistica II*;
  - **Linear models**, from *Analisi Statistica Multivariata* and *Econometria*;
  - **R software**, from *Analisi Statistica Multivariata*.
- In *Statistica III* we extend linear models within a unified and elegant framework.
- Regression is such an important topic that the **tour will continue** at the **M.Sc. at CLAMSES**. In Data Mining I will cover penalized methods and nonparametric regression.
- Indeed, GLMs can be arguably regarded one of the most influential statistical ideas of the XX century.

Home page

Statistics of the 20th century

Biometrika (2001), **88,** 1, *pp.* 13–52

Printed in Great Britain

Biometrika Centenary: Theory and general methodology

BY A. C. DAVISON

Department of Mathematics, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland

anthony.davison@epfl.ch

SUMMARY

Contributions to statistical theory and general methodology published in *Biometrika*, 1901–2000, are telegraphically reviewed.

Some key words: Bayesian inference; Estimating function; Foundations of statistics; Generalised regression model; Graphical method; Graphical model: Laplace approximation; Likelihood; Missing data; Model selection; Multivariate statistics; Non-regular model; Quasilikelihood; Saddlepoint: Simulation; Spatial statistics.

- **Biometrika** is among the most prestigious journals in Statistics. Past editors include Karl Pearson, Sir David Cox, and Anthony Davison.

Early ideas

- Classical linear models and least squares began with the work of **Gauss** and **Legendre** who applied the method to astronomical data.
- Their idea, in modern terms, was to **predict the mean** of a normal, or Gaussian, distribution as a function of a **covariate**:

$$\mathbb{E}(Y_i) = \beta_1 + \beta_2 x_i, \qquad i = 1, \dots, n.$$

- As early as Fisher (1922), a more advanced non-linear model was introduced, designed to handle **proportion data** of the form S_i/m.
- Through some modeling and calculus, Fisher derived a **binomial** model for S_i, with

$$\mathbb{E}(S_i/m) = \pi_i = 1 - \exp\{-\exp(\beta_1 + \beta_2 x_i)\}, \qquad i = 1, \dots, n.$$

  where $\pi_i \in (0, 1)$ is the **probability of success** of a binomial distribution.
- The corresponding inverse relationship is known as the **complementary log-log** link function:

$$\beta_1 + \beta_2 x_i = \log\{-\log(1 - \pi_i)\}.$$

Early ideas II

- In the **probit model**, developed by Bliss (1935), a **Binomial** model for S_i is specified with

$$\mathbb{E}(S_i/m) = \pi_i = \Phi(\beta_1 + \beta_2 x_i), \qquad i = 1, \dots, n.$$

  where $\Phi(x)$ is the cumulative distribution function of a Gaussian distribution.

- Dyke and Patterson (1952) also considered the case of modelling proportions, but specified

$$\mathbb{E}(S_i/m) = \pi_i = \frac{\exp(\beta_1 + \beta_2 x_i)}{1 + \exp(\beta_1 + \beta_2 x_i)}, \qquad i = 1, \dots, n.$$

- The corresponding inverse relationship is known as the **logit** link function:

$$\beta_1 + \beta_2 x_i = \mathrm{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right).$$

  In fact, this approach is currently known as **logistic regression**.

- See Chapter 1 of McCullagh and Nelder (1989) for a more exhaustive **historical** account, including early ideas about the **Poisson** distribution, the multinomial, and more.

Generalized linear models

10. GENERALISED REGRESSION

10·1. *Generalised linear models*

One of the most important developments of the 1970s and 1980s was the unification of regression provided by the notion of a generalised linear model (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989) and its associated software, though the concept had appeared earlier (Cox, 1968). In such models the response Y is taken to have an exponential family distribution, most often normal, gamma, Poisson or binomial, with its mean μ related to a vector of regressor variables through a linear predictor $\eta = x^{\mathrm{T}}\beta$ and a link function g, where $g(\mu) = \eta$. The variance of Y depends on μ through the variance function $V(\mu)$, giving $\mathrm{var}(Y) = \phi V(\mu)$, where ϕ is a dispersion parameter. Special cases are:

for fitting regression models. The estimating equations for a generalised linear model for independent responses $Y_1, \ldots, Y_n$ and corresponding covariate vectors $x_1, \ldots, x_n$ may be expressed as

$$\sum_{j=1}^{n} x_j \frac{\partial \mu_j}{\partial \eta_j} \frac{Y_j - \mu_j}{V(\mu_j)} = 0, \tag{10}$$

or in matrix form

$$D^{\mathrm{T}} V^{-1}(Y - \mu) = 0, \tag{11}$$

where D is the $n \times p$ matrix of derivatives $\partial \mu_j / \partial \beta_r$, and the $n \times n$ covariance matrix V is diagonal if the responses are independent but not in general. Taylor expansion of (11)

- The **pivotal paper** by Nelder and Wedderburn (1972) unified all these approaches.

Quasi likelihoods

10·2. *Quasilikelihood*

Data are often overdispersed relative to a textbook model. For example, although the variance of count data is often proportional to their mean, the constant of proportionality ϕ may exceed the value anticipated under a Poisson model, so $\text{var}(Y) = \phi\mu$ for $\phi > 1$. One way to deal with this is to model explicitly the source of overdispersion by the incorporation of random effects; see § 3·5. The resulting integrals can considerably complicate computation of the likelihood, however, and a simpler approach is through quasilikelihood (Wedderburn, 1974).

Quasilikelihood is perhaps best seen as an extension of generalised least squares. To see why, note that (11) is equivalent to $U(\beta) = 0$, where $U(\beta) = \phi^{-1} D V^{-1}(Y - \mu)$. Asymptotic properties of $\hat{\beta}$ stem from the relations $E(U) = 0$ and $\text{cov}(U) = -E(\partial U/\partial\beta)$, corresponding to standard results for a loglikelihood derivative. However, these properties do not depend on a particular probability model, requiring merely that $E(Y) = \mu$ and $\text{cov}(Y) = \phi V(\mu)$, subject also to some regularity conditions. Hence $\hat{\beta}$ has the key properties of a maximum likelihood estimator, namely consistency and asymptotic normality, despite not being based on a fully-specified probability model. Moreover, it may be computed simply by solving (11), that is, behaving as if the exponential family model with variance function $V(\mu)$ were correct. The scale parameter ϕ is estimated by $\hat{\phi} = (n-p)^{-1}(Y - \hat{\mu})^{\mathrm{T}} V(\hat{\mu})^{-1}(Y - \hat{\mu})$, and the asymptotic covariance matrix of $\hat{\beta}$ is $\hat{\phi}(D^{\mathrm{T}} V D)^{-1}$ evaluated at $\hat{\beta}$. A unified asymptotic treatment of such estimators from overdispersed

The content of this course

- **General theory**
  - Linear models and misspecification
  - Generalized Linear Models (GLMs)
- **Notable models**
  - Binary and binomial regression
  - Poisson regression
- **Advanced topics**
  - Quasi likelihoods

- Unfortunately, due to time constraints, we will **not** cover:
  - Contingency tables and log-linear models;
  - Multinomial response and ordinal response models;
  - Models with correlated responses (random effects);
  - Nonparametric regression.
- These topics will be covered e.g. in **Statistica Multivariata** and Data Mining at CLAMSES.

Textbooks

We will use several textbooks throughout this course — some more specialized than others. They are listed in order of importance:

1. The book by Salvan et al. (2020), **in Italian**, is the **main textbook**. Most of the material covered in these slides can be found there. I will also try to follow its notation as closely as possible.

2. The book by Azzalini (2008), **in Italian**, is more concise but very enjoyable to read. I highly recommend browsing through it.

3. The book by Agresti (2015), **in English**, is comprehensive and extremely well-written. It was the one I consulted most while preparing this course. Its only "drawback" is that it is in English.

4. The book by McCullagh and Nelder (1989), **in English**, is an **advanced and authoritative textbook** intended for experienced statisticians (at least at the M.Sc. level). Feel free to explore it out of curiosity, but it is not a main reference.

Exam

- The written exam has two parts, held on the same day:
  - **Theory and exercises**: questions to assess understanding of concepts and the ability to correctly set up a statistical model.
  - **Data set analysis**: applied analysis of a dataset using R.
- The **overall mark** is the average of the two parts.
  - You must pass both parts (each ≥ 18).
- The **oral exam** is optional:
  - Can be requested by the student or the teacher
  - Final mark $=$ average of written and oral marks
- The exam is **closed-book and closed-notes**, except for the **R scripts** provided at the beginning of the test.

References

Agresti, A. (2015), *Foundations of Linear and Generalized Linear Models*, Wiley.

Azzalini, A. (2008), *Inferenza statistica*, Springer Verlag.

McCullagh, P., and Nelder, J. A. (1989), *Generalized linear models*, Chapman & Hall/CRC.

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

Salvan, A., Sartori, N., and Pace, L. (2020), *Modelli lineari generalizzati*, Springer.