MATH 60604A
Statistical modelling
§ 4a - Generalized linear models

Léo Belzile

HEC Montréal
Department of Decision Sciences

Introduction

- Linear models are only suitable for data that are (approximately) normally distributed.
- However, there are many settings where we may wish to analyse a response variable which is not necessarily continuous, including when
  - Y is binary,
  - Y is a count variable,
  - Y is continuous, but non-negative,
- We consider particular distributions for binary/proportion and counts data, in order to do likelihood-based inference.

Binary response variables

- If the response variable Y takes values in $\{0, 1\}$, we may assume that Y follows a Bernoulli distribution, meaning

$$\mathsf{P}\left(Y = y\right) = \pi^y(1-\pi)^{1-y}, \quad y = 0, 1.$$

- For Bernoulli random variables, $\mathsf{E}\left(Y\right) = \pi$ and $\mathsf{Var}\left(Y\right) = \pi(1-\pi)$.
- By convention, failures (no) are zeros and successes (yes) ones.
- Potential research questions with binary responses include
  - Did a potential client respond favourably to a promotional offer?
  - Is the client satisfied with service provided post-purchase?
  - Will a company declare bankruptcy in the next three years?
  - Did a study participant successfully complete a task?

Aggregated binary response variables

If the data are aggregated independent binary events with Bernoulli distribution, the distribution of the number of successes Y out of m trials is Binomial, denoted $\mathrm{Bin}(m, \pi)$ with mass function

$$\mathsf{P}\left(Y = y\right) = \binom{m}{y}\pi^y(1-\pi)^{m-y}, \quad y = 0, 1, \ldots, m.$$

The likelihood is the same (up to a normalizing constant that does not depend on π) as that of m independent Bernoulli random variables and $\mathsf{E}\left(Y\right) = m\pi$, $\mathsf{Var}\left(Y\right) = m\pi(1-\pi)$.

Count response variables

- If the probability of an event is **rare**, we often assume that the number of successes in a given time interval Y follows a Poisson distribution,

$$\mathsf{P}\left(Y = y\right) = \frac{\exp(-\mu)\mu^y}{\Gamma(y+1)}, \quad y = 0, 1, 2, \ldots$$

- The parameter μ of the Poisson distribution characterizes both its mean and variance, meaning $\mathsf{E}\left(Y\right) = \mathsf{Var}\left(Y\right) = \mu$.
- Examples of response variables include the number of
  - insurance claims made by a policyholder over a year,
  - purchases made by a client over a month on a website,
  - number tasks completed by a study participant in a given time frame.

Notation for generalized linear models

- The starting point is the same as for linear regression:
  - We have a random sample of independent observations
    $$(Y_i, \mathrm{X}_{i1}, \ldots, \mathrm{X}_{ip}), \quad i = 1, \ldots, n$$
    where Y is the response variable and $\mathrm{X}_1, \ldots, \mathrm{X}_p$ are p explanatory variables or covariates which are assumed fixed (non-random).
  - The goal is to model the response variable as a function of the explanatory variables.
- Let μ_i denote the (conditional) mean of Y_i given covariates,
  $$\mu_i = \mathrm{E}\left(Y_i \mid \mathrm{X}_{i1}, \ldots, \mathrm{X}_{ip}\right).$$
- Let η_i denote the linear combination of the covariates that will be used to model the response variable,
  $$\eta_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}.$$

Definition of generalized linear model

- There are three building blocks to the generalized linear model:
  - A probability distribution for the outcome Y that is a member of the exponential family (normal, binomial, Poisson, gamma, ...).
  - The linear predictors $\boldsymbol{\eta} = \mathbf{X}\beta$.
  - A function g, called link function, that links the mean of Y_i to the predictor variables, $g(\mu_i) = \eta_i$.

Link function

- The link function connects the mean to the explanatory variables,

$$\begin{aligned} g(\mu_i) &= \eta_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip} \\ \Leftrightarrow \quad \mu_i &= g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 \mathrm{X}_{i1} + \cdots + \beta_p \mathrm{X}_{ip}). \end{aligned}$$

- In the ordinary linear regression model, we do not impose constraints on the mean μ_i and $\widehat{\mu}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \mathrm{X}_{i1} + \cdots + \widehat{\beta}_p \mathrm{X}_{ip}$ can take on any value in $(-\infty, \infty)$.
- For some response variables, we would need to impose constraints on the mean.
  - For Bernoulli responses, the mean $\mu = \pi$ must lie in the interval $(0, 1)$.
  - For Poisson responses, the mean μ must be positive.
- An appropriate choice of link function sets μ_i equal to a transformation of the linear combination η_i so as to avoid any parameter constraints on β.

Choice of link function

Certain choices of the link function facilite interpretation or make the likelihood function convenient for optimization.

- For the Bernoulli and binomial distributions, an appropriate link function is the logit function,

$$\mathrm{logit}(\mu) := \ln\left(\frac{\mu}{1-\mu}\right) = \eta \quad \Leftrightarrow \quad \mu = \frac{\exp(\eta)}{1+\exp(\eta)}.$$

- For the Poisson distribution, an appropriate link function is the natural logarithm,

$$\ln(\mu) = \eta \quad \Leftrightarrow \quad \mu = \exp(\eta).$$

- For the normal distribution, an appropriate link function is the identity function, $\mu = \eta$.

Generalized linear model: linear regression

- Ordinary linear regression is a special case of generalized linear models, with

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip} + \varepsilon_i, \qquad (i = 1, \ldots, n)$$

  where $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, i.e., $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distribution normal random variables with mean 0 and variance σ^2.
- This is equivalent to stating

$$Y_i \mid \mathbf{X}_i \stackrel{\text{ind}}{\sim} \mathrm{No}(\beta_0 + \beta_1 \mathrm{X}_{i1} + \ldots + \beta_p \mathrm{X}_{ip}, \sigma^2)$$

- Linear regression is a generalized linear model with
  - a normal distribution for the response and
  - the identity function as link function.