Generalized Linear Models (GLM)

(Extending the linear model)

N. Torelli, G. Di Credico, V. Gioia

2023

University of Trieste

Introduction

Generalized Linear Models: Basic ideas

Inference

Solution of the likelihood equations

Model Evaluation

Quasi-likelihood

GLM: Extensions and recent development

Introduction

Generalized Linear Models (GLMs) are a broad class of models that extend linear regression (LMs) to handle data that does not fit the assumptions of ordinary least squares (OLS) regression, such as non-normal response variables or relationships that are not linear.

While linear models work well for continuous data with constant variance (e.g., predicting weight from height), many real-world datasets involve binary outcomes, count data, or data with non-constant variance. GLMs provide a unified framework to model such scenarios.

Introduction

Generalized linear models (GLMs) encompass the statistical models reviewed so far

- The response variable in a GLM can be a quantitive variables (also considering the case when the response takes on only positive values), a dichotomous variable, a count variable.
- GLMs recognizes that for many models the idea that covariates effects can be summarized by a linear combination is useful and flexible enough.
- And it is recognized that the aim is to study how this linear combination of the covariates affects the expected value of the response variable.
- Usually the main interest is in estimating the unknown coefficients of the linear combination (the β parameters).
- Linear regression model, logistic and probit regression, Poisson regression (and in fact many other models) have this common structure

From LM to GLM

- Recall that Normal LMs, in matrix notation, are defined by

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where

1. $Y_i \sim N(\mu_i, \sigma^2)$, independent, where
   $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and
   $\boldsymbol{x}_i^T$ is $i-$th row of X, $i = 1, 2, \ldots, n$;

2. The density of $Y_i \sim N(\eta_i, \sigma^2)$ and covariates $\boldsymbol{x_i}$ appear through the **linear predictor**:

$$\eta_i = \sum_{j=1}^{p} x_{ij}\beta_j = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

;

- β e σ^2 are unknown parameters.

LIMITATIONS:

- IF THE RESPONSE VARIABLE IS BINARY
- -//- A COUNT
- THE VARIANCE OF THE RESPONSE IS NOT CONSTANT

Introducing GLMs

GLMs generalize LMs by:

- Y_i are assumed to be (independent) measurements from a distribution with density (probability) function from the **exponential dispersion family** (E.G., BINOMIAL, POISSON)
- Existence of the mean $E(Y) = \mu$ is assumed and μ is determined by η that is related to it by a suitable function

$$g(E(Y_i)) = g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \beta$$

  $g(\cdot)$ is called the **link function**. $g(\cdot)$ ENABLES MODELING OF NON-LINEAR RELATIONSHIPS
- in principle f could be any suitable density (or probability) function, but a family of distribution plays a key role:

The exponential (dispersion) family

- A random variable Y belongs to exponential (dispersion) family if its density (probability) function can be written as

$$f(Y;\theta,\phi) = \exp\left\{\frac{Y\theta - b(\theta)}{\phi} + c(Y,\phi)\right\}, \tag{1}$$

θ e ϕ are unknown scalar parameters, $b(\cdot)$ and $c(\cdot) > 0$ are known functions and domain of Y do not depend on θ or ϕ.

We will denote this by $Y \sim EF(b(\theta), \phi)$.

$b(\theta)$: DETERMINES THE MEAN $\mu = b'(\theta)$

$C(Y,\phi)$: NORMALIZING TERM

- θ is called the *natural or canonical parameter* of the exponential family.
- ϕ is called the *dispersion parameter*. It can be known in some cases. When it is unknown, the family is more properly called the exponential dispersion family. (E.G., VARIANCE SCALING)
- Many of the most common continuous and discrete distributions belong to this family (i.e. Normal, Gamma, Poisson, Binomial, etc)

Example: Poisson

- As we already noted it is the basic choice when modelling count data
- if $Y \sim \text{Poisson}(\lambda)$, its probability function is

$$\begin{aligned} f(Y;\lambda) &= \frac{e^{-\lambda}\lambda^{Y}}{Y!} \\ &= \exp\{Y \log \lambda - \lambda - \log Y!\}\ , \end{aligned}$$

  for $Y = 0, 1, \ldots,$.
- This shows that it is a member of (1) where $\theta = \log \lambda$ is the natural parameter, $\phi = 1$, $b(\theta) = \lambda = e^{\theta}$ and $c(Y, \phi) = -\log Y!$.
- We can write $Y \sim EF(e^{\theta}, 1)$.

Example: Binomial

- Standard distribution when modelling binary responses
- If $Y \sim \text{Bin}(n, \pi)$, its probability function is

$$\begin{aligned} f(Y;\pi) &= \binom{n}{Y}\pi^Y(1-\pi)^{n-Y} \\ &= \exp\{\log\binom{n}{Y} + Y\log\pi + (n-Y)\log(1-\pi)\} \\ &= \exp\left\{Y\log\frac{\pi}{1-\pi} + n\log(1-\pi) + \log\binom{n}{Y}\right\}, \end{aligned}$$

  for $Y = 0, 1, \ldots, n$.

- It belongs to (1) where $\theta = \log\frac{\pi}{1-\pi}$ natural parameter, $\phi = 1$,

$$b(\theta) = -n\log(1-\pi)\big|_{\pi=\frac{e^\theta}{1+e^\theta}} = n\log(1+e^\theta)$$

  and $c(Y, \phi) = \log\binom{n}{Y}$.

- $Y \sim EF(n\log(1+e^\theta), 1)$.

Generalized Linear Models: Basic ideas

The structure of GLMs

- A general theory has been defined for GLMs also because this allowed to implement a single general procedure for estimating the parameters, checking their significance, evaluating the goodnes of fit of the model, selecting the "best" model, obtaining predictions and computing residuals.
- Most software packages have in fact been implemented to this aim.
- In GLMs the response variables are assumed to be distributed according to a more general family of random variables, the exponential dispersion family.
- This family includes, among the others, the Binomial, the Poisson, the Gamma and the Normal families.
- Consequently linear regression models (where the response are usually assumed to be Normal), logistic regression (where the response is binomial), Poisson regression (where the response is a count) can be written as GLMs.

The ingredients of GLMs

Components of GLMs are the following:

- NORMAL FOR CONTINUOS DATA
- BINOMIAL FOR BINARY //
- POISSON FOR COUNT //

1. **a response** Y_i distributed as a member of a quite comprehensive family of distributions the exponential dispersion family $Y_i \sim EF(b(\theta_i), \phi)$ where $E(Y_i) = \mu_i$ and whose variance is $V(Y_i) = \phi V(\mu_i)$. → RANDOM COMPONENT
2. **a linear predictor** $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$ $= X\beta$ → SYSTEMATIC COMPONENT
3. **a link function** $g(\cdot)$ assumed to be monotone and that relates the linear predictor η_i to μ_i so that $g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i)$, $i = 1, 2, \ldots, n$ and then

$$E(Y_i) = \mu_i = g^{-1}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})$$

and g^{-1} is called the response function

Example: Poisson

- As we already noted it is the basic choice when modelling count data
- if $Y \sim \text{Poisson}(\lambda)$, its probability function is

$$\begin{aligned} f(Y;\lambda) &= \frac{e^{-\lambda}\lambda^{Y}}{Y!} \\ &= \exp\{Y \log \lambda - \lambda - \log Y!\}\ , \end{aligned}$$

  for $Y = 0, 1, \ldots,$.
- This shows that it is a member of (1) where $\theta = \log \lambda$ is the natural parameter, $\phi = 1$, $b(\theta) = \lambda = e^{\theta}$ and $c(Y, \phi) = -\log Y!$.
- We can write $Y \sim EF(e^{\theta}, 1)$.

Example: Binomial

- Standard distribution when modelling binary responses
- If $Y \sim \text{Bin}(n, \pi)$, its probability function is

$$\begin{aligned} f(Y;\pi) &= \binom{n}{Y}\pi^Y(1-\pi)^{n-Y} \\ &= \exp\{\log\binom{n}{Y} + Y\log\pi + (n-Y)\log(1-\pi)\} \\ &= \exp\left\{Y\log\frac{\pi}{1-\pi} + n\log(1-\pi) + \log\binom{n}{Y}\right\}, \end{aligned}$$

  for $Y = 0, 1, \ldots, n$.

- It belongs to (1) where $\theta = \log\frac{\pi}{1-\pi}$ natural parameter, $\phi = 1$,

$$b(\theta) = -n\log(1-\pi)\big|_{\pi=\frac{e^\theta}{1+e^\theta}} = n\log(1+e^\theta)$$

  and $c(Y, \phi) = \log\binom{n}{Y}$.

- $Y \sim EF(n\log(1+e^\theta), 1)$.

Mean and variance for Exponential family

- The function $b(\cdot)$ is called the *cumulant function* and it is important in evaluating and interpreting first moments of the distribution.
- by using identities related to derivatives of log-likelihood function:

$$E\left(\ell_*(\theta)\right) = E\left(\frac{d}{d\theta}\ell(\theta; Y)\right) = 0$$

and

$$i(\theta) = var\left(\ell_*(\theta)\right) = E\left(-\ell_{**}(\theta)\right) = E\left(-\frac{d^2}{d\theta^2}\ell(\theta; Y)\right),$$

under usual regularity assumptions.
If Y is a r.v. member of the exponential family, log-likelihood for θ it follows that:

$$E\left(\frac{Y - b'(\theta)}{\phi}\right) = 0 \qquad \text{and} \qquad \boxed{E(Y) = \mu = b'(\theta)}$$

μ IS RELATED TO θ THROUGH $b(\theta)$

$$var\left(\frac{Y - b'(\theta)}{\phi}\right) = \frac{b''(\theta)}{\phi} \quad \Rightarrow \quad \boxed{var(Y) = \phi b''(\theta)}$$

Denote $V(\mu) = b''(\theta)$, we can write $\boxed{var(Y) = \phi V(\mu)}$

ϕ: CONSTANT (NOT ALWAYS TRUE)

$V(\mu)$: DEPENDS ON μ

- The function $V(\mu)$ is the so called *variance function* since it indicates how the variance depends on the mean of Y (GLM can be heteroscedastic). This becomes clear if we recall that μ is related to θ, i.e., $\mu = b'(\theta)$.

Some relevant member of the exponential family and their moments

Poisson

We have for a Poisson with mean λ

$$b(\theta) = e^{\theta} \quad \text{and} \quad \phi = 1 \qquad \text{and} \qquad E(Y) = b'(\theta) = e^{\theta} = \lambda \ .$$

$$var(Y) = b''(\theta) = e^{\theta} = \lambda \qquad \text{then} \qquad V(\mu) = \mu$$

Binomial

We have for a Binomial with parameters (n, π)

$$b(\theta) = n\log(1+e^{\theta}), \quad \phi = 1 \quad \text{then} \quad E(Y) = \mu = b'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}} = n\pi \ .$$

$$var(Y) = b''(\theta) = n\frac{e^{\theta}}{(1+e^{\theta})^2} = n\pi(1-\pi) \quad \text{and} \quad V(\mu) = \mu(1-\mu)/n \ .$$

The link function

- The second important step in specifying a GLM is the definition of the function relating μ_i and the linear predictor η_i.
- It is assumed that the link between μ_i, the mean of Y_i, and $\boldsymbol{x}_i^T$, the covariate vector, is

$$g(\mu_i) = \eta_i \qquad \text{and} \qquad \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} \ .$$

- $g(\cdot)$ is a known monotone and differentiable function. The function $g(\cdot)$ is the *link function* between μ_i and η_i.
- the inverse function $g(\cdot)^{-1} = r(\cdot)$ is also called the response function
- Covariates enter into the model by the linear predictor $\boldsymbol{x}_i^T \boldsymbol{\beta}$, but the μ_i and η_i are generally non linearly related.
- Appropriate choices of the link function are such that $\mu_i = g^{-1}(\eta_i)$ takes on values on the appropriate range.

The canonical link

IS MATHEMATICALLY CONVINIENT BECAUSE IT SIMPLIFIES CALCULATIONS AND ENSURES DESIRABLE STATISTICAL PROPERTIES

- A typical choice is to write directly the natural parameter θ as a linear function of the covariates Formally,

$$\eta = g(\mu) = g(b'(\theta)) = \theta \ ,$$

  $g(\cdot)$ is then the inverse function of $b'(\cdot)$. This choice of the link function is called *canonical link*.

- Some interesting properties derives from choosing a canonical link. Moreover the canonical link is the default link used in many softwares for estimation of GLMs (including R).

Inference

Inference in GLMs focuses on estimating the model parameters, assessing their uncertainty, and testing hypotheses. The primary tool for parameter estimation is maximum likelihood estimation (MLE), which maximizes the likelihood of observing the given data under the model. This involves solving the likelihood equations and leveraging properties like the Fisher information to quantify uncertainty.

Estimation of the parameters

- ML can be used since distributional assumptions on parameters are available (for the normal LM it coincides with LS).
- A property of the exponential families is that they satisfy enough regularity conditions to ensure that the MLE is given uniquely by the solution of the likelihood equations.
- Let us recall some important features of GLM:
  - $g(\mu_i) = \eta_i = \boldsymbol{x}_i^T \beta \Leftrightarrow \mu_i = g^{-1}(\boldsymbol{x}_i^T \beta)$;
  - $\mu_i = b'(\theta_i) \Leftrightarrow \theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(g^{-1}(\eta_i))$;
  - $var(Y_i) = \phi V(\mu_i)$, with $V(\mu_i) = b''(\theta_i)$.

θ DEPENDS ON β

- Assuming independence of $(y_1, \ldots, y_n)$, the log-likelihood $\ell(\beta, \phi)$ is simply given by

$$\ell(\beta, \phi) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi) = \sum_{i=1}^{n} \ell_i(\beta, \phi) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \right\}$$

where θ_i is a function of β through

SINCE β IS A VECTOR, I CAN TAKE THE DERIVATIVES

$$g(\mu_i) = g(b'(\theta_i)) = \eta_i = \boldsymbol{x}_i^T \beta \ .$$

Which Parameter Are We Using MLE On?
In the context of Generalized Linear Models (GLMs), maximum likelihood estimation (MLE) is used to estimate the parameters of the model, specifically:

Coefficients (β):
- These parameters determine the relationship between the predictors (covariates) and the response variable.
- The goal is to find β such that the likelihood of observing the data under the model is maximized.

Dispersion Parameter (ϕ) (if applicable):
- For some distributions in the exponential family (e.g., normal, gamma), there is an additional dispersion parameter ϕ that scales the variance of the response variable.

Why Use MLE?

MLE provides a systematic way to estimate β and ϕ based on the assumed distribution of the response variable:
+ GLMs assume that the response variable y follows a distribution from the exponential family.
+ MLE estimates are consistent, asymptotically normal, and efficient under the regularity conditions satisfied by the exponential family.

In practical terms:
+ MLE helps us find the best-fitting parameters β that describe how predictors influence the response y.
+ For example, in logistic regression (a GLM for binary outcomes), MLE estimates the β's that maximize the likelihood of the observed binary responses given the predictors.

Likelihood equations

- To obtain the MLE of β it is necessary to solve the *likelihood equations*:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = 0 \quad \text{for} \quad j = 1, 2, \ldots, p.$$

- Let us compute

$$\begin{aligned} \frac{\partial \ell_i}{\partial \beta_j} &= \frac{\partial \ell_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} \\ &= \frac{\partial \ell_i}{\partial \theta_i} \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} \frac{\partial \eta_i}{\partial \beta_j} , \end{aligned}$$

- where the terms can be written as

$$\begin{aligned} \frac{\partial \ell_i}{\partial \theta_i} &= \frac{y_i - b'(\theta_i)}{\phi} = \frac{y_i - \mu_i}{\phi} , \\ \frac{\partial \mu_i}{\partial \theta_i} &= b''(\theta_i) = \frac{var(Y_i)}{\phi} , \\ \frac{\partial \eta_i}{\partial \mu_i} &= g'(\mu_i) , \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} . \end{aligned}$$

Likelihood equations

- Thus, we have

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_j} &= \frac{y_i - \mu_i}{\phi} \frac{\phi}{var(Y_i)} \frac{1}{g'(\mu_i)} x_{ij} \\ &= \frac{(y_i - \mu_i) x_{ij}}{\phi V(\mu_i) g'(\mu_i)}.\end{aligned}$$

THIS EQ IS NONLINEAR AND NEEDS TO BE SOLVED ITERATIVELY

ϕ EXP

- The likelihood equations for β are then

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi V(\mu_i) g'(\mu_i)} x_{ij} = 0,$$

$j = 1, 2, \ldots, p$, where $\mu_i = g^{-1}(\boldsymbol{x}_i^T \beta)$.

- Note that the MLE of β for a fixed value of ϕ, does not depend on ϕ and coincides with the unconstrained MLE.

Canonical link

- The use of the *canonical link* ($\eta_i = g(\mu_i) = g(b'(\theta_i)) = \theta_i$) produces some simplifications in the inference based on the log-likelihood $\ell(\beta, \phi)$.
- With the canonical link, we have $g'(\mu_i) = 1/V(\mu_i)$ and the first derivative reduces to

$$\sum_{i=1}^{n} \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi} \ .$$

- This result implies that the likelihood equations simplify and take the form

IT'S POSSIBLE TO FIND THE ZERO OF THIS →

$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \mu_i x_{ij} \ .$$

→ IN GM WE HAVE: $X^T y = X^T X \beta$

  Using matrix notation, $X^T \boldsymbol{y} = X^T \mu$.
- These equations agree with the general structure of the likelihood equations in exponential families: the observed value of the minimal sufficient statistic is equated to its expectation.
- As regards the existence and uniqueness of the MLE of β, if the link is the canonical one, the theory of exponential families applies.
- In general the likelihood equations for β are nonlinear and must be solved with iterative methods. To this end, the expected Fisher information for β is useful.

What Is Fisher Information?
Fisher information quantifies how much information the observed data provides about the unknown parameter β.
In mathematical terms:
+ Fisher information measures the curvature of the log-likelihood function at its maximum.
+ A steep curvature implies that the data provides precise information about β (i.e., smaller uncertainty), while a flat curvature implies less precise information.

Purpose of Fisher Information in GLMs:
1. Quantifying Uncertainty:
+The Fisher information matrix $I(\beta)$ is used to estimate the covariance matrix of the MLE β^.
+The diagonal elements of $I(\beta)$^−1 provide the variances of the estimated β's.

2. Hypothesis Testing and Confidence Intervals:
+The Fisher information allows us to construct confidence intervals and perform hypothesis tests for the β's using their standard errors.

3. Iterative Optimization:
+Fisher information is used in optimization algorithms (e.g., Newton-Raphson, Fisher Scoring) to solve the likelihood equations efficiently.

Fisher information

- Since β and ϕ are orthogonal, we can proceed as if ϕ were known and we can focus only on β.
- Let us consider the second derivatives of ℓ_i:

$$\begin{aligned} -E\left(\frac{\partial^2 \ell_i}{\partial\beta_j \partial\beta_k}\right) &= E\left(\frac{\partial \ell_i}{\partial\beta_j}\frac{\partial \ell_i}{\partial\beta_k}\right) \\ &= E\left(\left(\frac{(Y_i-\mu_i)x_{ij}}{\phi V(\mu_i) g'(\mu_i)}\right)\left(\frac{(Y_i-\mu_i)x_{ik}}{\phi V(\mu_i) g'(\mu_i)}\right)\right) \\ &= \frac{x_{ij}x_{ik}}{\phi^2 (V(\mu_i))^2 (g'(\mu_i))^2} E\left((Y_i-\mu_i)^2\right) \\ &= \frac{x_{ij}x_{ik}}{\phi V(\mu_i)(g'(\mu_i))^2}\ , \end{aligned}$$

which gives the (j,k)-element of the Fisher information matrix for β. Using matrix notation,

$$\boxed{i(\beta) = \frac{X^T W X}{\phi}\ ,}$$

with $W = \text{diag}(w_1, \ldots, w_n)$ and

$$w_i = \frac{1}{V(\mu_i)(g'(\mu_i))^2}\ ,$$

and X is the matrix of the explanatory variables.

Fisher information

- With the **canonical link**, the observed and the expected informations coincide and have (j,k)-element
$$\frac{x_{ij}x_{ik}V(\mu_i)}{\phi}\,.$$
In matrix form,
$$i(\beta) = j(\beta) = \frac{X^T V X}{\phi}\,,$$
with $V = \mathrm{diag}(V(\mu_i))$.
- Asymptotic normality of the MLE gives
$$\hat{\beta} \;\dot{\sim}\; N_p(\beta, \phi(X^T W X)^{-1})\,,$$
for large n.
- Therefore, a consistent estimate of the covariance matrix of β is $i(\hat{\beta}) = \phi(X^T \hat{W} X)^{-1}$, where $\hat{W}$ is the matrix W evaluated at $\hat{\beta}$.
- If ϕ is unknown, it should be replaced by a consistent estimator, such as the MLE or the estimator based on the method of moments.
- For normal distribution with identity link we have $g(\mu) = \mu$, so that $g'(\mu) = 1$. Moreover, $V(\mu) = 1$, $\phi = \sigma^2$ and $\mu_i = \boldsymbol{x}_i^T \beta$. The likelihood equations are $\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T\beta)x_{ij}}{\sigma^2} = 0$ that leads to usual LSE

Some models

Normal Linear Model

We have $g(\mu) = \mu$, so that $g'(\mu) = 1$. Moreover, $V(\mu) = 1$, $\phi = \sigma^2$ and $\mu_i = \boldsymbol{x}_i^T \beta$. The likelihood equations are

$$\sum_{i=1}^{n} \frac{(y_i - \boldsymbol{x}_i^T \beta) x_{ij}}{\sigma^2} = 0 \quad j = 1, 2, \ldots, p. \; .$$

Simplifying σ^2 and using matrix notation, the above equations reduce to the usual LS equations: $X^T(\boldsymbol{y} - X\beta) = 0$ or, equivalently,

$$X^T X\beta = X^T \boldsymbol{y} \quad \text{that leads to} \quad \hat{\beta} = (X^T X)^{-1} X^T \boldsymbol{y}$$

Poisson regression

We have $g(\mu) = \log \mu$, so that $g'(\mu) = 1/\mu$. Moreover, $V(\mu) = \mu$, $\phi = 1$ and $\mu_i = \boldsymbol{x}_i^T \beta$. The likelihood equations are

$$\sum_{i=1}^{n} (y_i - e^{\boldsymbol{x}_i^T \beta}) x_{ij} = 0 \; ,$$

which are generally nonlinear in β. In view of this, an explicit solution does not exist in general.

Solution of the likelihood equations

An iterative algorithm

- Likelihood equations for GLMs do not usually have explicit solutions. They should be solved by iterative methods.
- For the GLM there exists the possibility to use a simple algorithm for the solution of the likelihood equations: the MLEs of the parameter $\boldsymbol{\beta}$ in the linear predictor can be obtained by iterative weighted least squares.
- Starting with appropriate initial value $\hat{\boldsymbol{\beta}}^{(0)}$ and obtaining a sequence $\hat{\boldsymbol{\beta}}^{(1)}, \hat{\boldsymbol{\beta}}^{(2)}, \ldots,$ using a rule to update $\hat{\boldsymbol{\beta}}^{(t+1)}$ with $\hat{\boldsymbol{\beta}}^{(t)}$, until that the value of

$$||\hat{\boldsymbol{\beta}}^{(t+1)} - \hat{\boldsymbol{\beta}}^{(t)}||$$

is sufficiently small ($< \epsilon$).

Newton-Raphson and Fisher scoring

- Let
$$\ell_* = \left(\frac{\partial \ell}{\partial \beta_1}, \ldots, \frac{\partial \ell}{\partial \beta_p} \right)^T$$
be the *score* vector. We want to solve the equation
$$\ell_* = \ell_*(\boldsymbol{\beta}) = 0 \ .$$

- The Newton-Raphson method is based on the updating rule at the $(t+1)-$th iteration
$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + (j(\hat{\boldsymbol{\beta}}^{(t)}))^{-1} \ell_*^{(t)} \ , \tag{2}$$
with $\ell_*^{(t)} = \ell_*(\hat{\boldsymbol{\beta}}^{(t)})$.

- The observed information can be replaced by the expected Fisher information $i(\boldsymbol{\beta})$. This algorithm takes the name of Fisher *scoring* method. This maintains the convergence of the algorithm and simplifies the expressions (if the canonical link function is used, the two expressions coincide).

Developing the algorithm

Expression (2) is equivalent to

$$i(\hat{\beta}^{(t)})\hat{\beta}^{(t+1)} = i(\hat{\beta}^{(t)})\hat{\beta}^{(t)} + \ell_*^{(t)} \ .$$

Remember that the (j,k)-th element of $i(\beta)$ is

$$\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \ ,$$

which gives $i(\beta) = \frac{X^T W X}{\phi}$, with $w_{ii} = \frac{1}{V(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$.

In view of this, the right hand term can be written as

$$(i^{(t)})\hat{\beta}^{(t)} + \ell_*^{(t)}$$

$$= \sum_{k=1}^{p}\sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \hat{\beta}_k^{(t)} + \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

$$= X^T W^{(t)} \boldsymbol{s}^{(t)} \ ,$$

Weighted Least Squares

- where $\boldsymbol{s}^{(t)}$ is a vector with elements

$$s_i^{(t)} = \sum_{k=1}^{p} x_{ik} \hat{\beta}_k^{(t)} + (y_i - \mu_i) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \ ,$$

  and all the involved quantities are evaluated at $\hat{\beta}$.

- Therefore, it is possible to arrive to the expression

$$X^T W^{(t)} X \hat{\beta}^{(t+1)} = X^T W^{(t)} \boldsymbol{s}^{(t)} \ . \tag{3}$$

- Clearly, the parameter ϕ simplifies.

- The above expression has the form of the normal equations for a LM obtained with weighted least squares, except that the equation above has to be solved iteratively because in general $\boldsymbol{s}$ and W depend on β.

Iterative Weigthed Least Squares (IWLS)

- Indeed, the Newton-Raphson iteration is

$$\hat{\beta}^{(t+1)} = (X^T W^{(t)} X)^{-1} X^T W^{(t)} \boldsymbol{s}^{(t)} \ . \tag{4}$$

- Each iteration of the algorithm is equivalent to a weighted least squares estimate, in which the adjusted dependent variable and the weights depend on the fitted values, for which only current estimates are available.
- The algorithm has two main steps:
  1. Given $\hat{\beta}^{(t)}$, compute $\boldsymbol{s}^{(t)}$ and $W^{(t)}$;
  2. Obtain $\hat{\beta}^{(t+1)}$ through (4).

To start the algorithm a simple and convenient choice of the starting values is $\boldsymbol{s}^{(0)} = g(Y_i)$ and $W^{(0)}$ equals to the identity matrix.

Estimating the dispersion parameter ϕ

- For the LM, the estimation of β is independent from the value of the variance σ^2. A similar situation holds for the dispersion parameter ϕ in GLMs.
- Obviously, the MLE of ϕ, with β replaced by $\hat{\beta}$, could be used.
- Also estimators based on the method of moments are often used for ϕ.
- Since $var(Y_i) = \phi V(\mu_i)$ or, equivalently, since $\frac{E((Y_i-\mu_i)^2)}{V(\mu_i)} = \phi$ if β is known, an unbiased estimator of ϕ is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{(y_i-\mu_i)^2}{V(\mu_i)} \ .$$

If the expected values μ_i are replaced with their estimates based on $\hat{\beta}$, then the following adjusted consistent estimator is obtained

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\widehat{\mu_i})^2}{V(\widehat{\mu_i})},$$

where

$$\widehat{\mu_i} = g^{-1}(\boldsymbol{x}_i^T\widehat{\beta}) \ .$$

Exploiting asymptotic normality of $\hat{\beta}$

- For n large, the asymptotic distribution of the MLE is

$$\hat{\beta} \dot{\sim} N_p(\beta, [i(\hat{\beta})]^{-1}) \quad \text{where} \quad i(\hat{\beta}) = \frac{X^T \hat{W} X}{\phi}$$

  with $\hat{W}$ computed at $\hat{\beta}$. The estimated asymptotic variances are the diagonal elements of the matrix $(X^T \hat{W} X)^{-1}\phi$.

- Using the asymptotic distribution of $\hat{\beta}$, a confidence interval for β_j with approximate level $1 - \alpha$ is

$$\hat{\beta}_j \pm z_{1-\alpha/2} \sqrt{\phi[(X^T \hat{W} X)^{-1}]_{j,j}} \; .$$

- and the statistic $\frac{\hat{\beta}_j}{\sqrt{\phi[(X^T \hat{W} X)^{-1}]_{j,j}}}$ can be used to test significance of a single β_j

Model Evaluation

Comparing nested models

- Let us start by considering two nested GLMs. Let denote the models by M_C and M_R, such that $M_R \subset M_C$. Specifically, the current model M_C contains p parameters and the reduced model M_R contains p_0 parameters, where $p > p_0$.
- Consider the following partition of $\boldsymbol{\beta} = (\boldsymbol{\beta}_{MR}, \boldsymbol{\beta}_{MC})$, where $\boldsymbol{\beta}_{MR} = (\beta_1, \ldots, \beta_{p_0})$ and $\boldsymbol{\beta}_{MC} = (\beta_{p_0+1}, \ldots, \beta_p)$. Suppose we want to test the following hypothesis

$$H_0 : \boldsymbol{\beta}_{MC} = 0 \quad \text{against} \quad H_1 : \boldsymbol{\beta}_{MC} \neq 0 \ .$$

- The criterion we will adopt to compare M_C and M_R is the likelihood ratio

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\} \ .$$

The deviance in LMs

- In normal LMs, with σ^2 known, the likelihood ratio is a function of the deviance (sum of square of residuals) $D = SSE = \sum_i (y_i - \hat{\mu}_i)^2$ of the two models. When comparing two nested models ($M_R \subset M_C$), the likelihood ratio criterion will lead to rejection of H_0 for large values of the following statistic

$$W = 2\{\ell(\hat{\beta}) - \ell(\hat{\beta}_{MR})\} = \frac{D_{MR} - D}{\sigma^2} \ ,$$

  where $D_{MR} = SSE_{H_0}$ and $D = SSE$ are sums of square of residuals in the reduced and current models respectively.

- When H_0 holds this statistic has a $\chi^2_{p-p_0}$ distribution.

LR test

- Like Normal LMs, we look for an interpretation of (log-)likelihood ratio in GLMs so that the relationship between the two classes of models is clear. It will help if we can define an analogous quantity as deviance in LMs.
- Log-likelihood for a GLM is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}) \ ,$$

  where

$$\ell_i(\boldsymbol{\beta}) = \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi) \ .$$

- With nested GLM, the statistic

$$W = 2\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}_{MR})\}$$

  is asymptotically distributed as a $\chi^2_{p-p_0}$ when H_0 holds.

The saturated model

- Analogy with Normal LM can be kept by introducing likelihood associated to a model where there are as many parameters as observations. This model will be denoted as **saturated** or **full**.
- At the other extreme there is a model as simple as possible, *i.e.*, a model where a single parameter represents a common μ for all the y_i

A "'good"' model usually stands between these two extremes since a saturated model is uniformative being unable to summarize data: it just repeats them in full, and a null model is usually too simple to be useful. We should seek a balance between conflicting goals of parsimony and goodness of fit.

- Saturated model is defined as:
  - $\mapsto$ a GLM having the same distribution and link function of the current model;
  - $\mapsto$ but a number of parameter equal to n (or to the number of different groups sharing the same $\boldsymbol{x}$ vector).
- We can evaluate likelihood function for the saturated model and the current model at the value of the MLE obtained in both cases ($\tilde{\theta}$ and $\hat{\theta}$ respectively). If the current model fits the data, $\ell(\tilde{\theta})$ should be very similar to $\ell(\hat{\theta})$. In case of a poor fit then $\ell(\hat{\theta})$ should be much smaller than $\ell(\tilde{\theta})$.

The deviance in GLMs

- Formally, the quantity

$$D(y;\hat{\theta}) \quad = \quad 2\phi\{\ell(\tilde{\theta}) - \ell(\hat{\theta})\} = \phi \sum_{i=1}^{n} D_i$$

with $D_i = 2\{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\}$, is called *deviance function* of the model and

$$\frac{D(y;\hat{\theta})}{\phi} = \sum_{i=1}^{n} D_i \tag{5}$$

is the *scaled deviance*: note that it is always non negative.

This quantity is small for good models and is large when the current model gives a poor fit. Behaviour of deviance is equivalent to that of SSE in LMs.

- $\ell(\tilde{\theta})$ is the log-likelihood obtained by letting $\mu_i = b'(\theta_i) = y_i (\Leftrightarrow (\partial \ell_i / \partial \theta_i) = 0)$, so the saturated model has $p = n$ parameters.
- The saturated model is useless but $\ell(\tilde{\theta})$ provides a benchmark to compare log-likelihood of the current model.

Example: Normal regression model

Since Normal LMs are GLMs with identity link functions we can show that calculating the above defined deviance we give in this case the same result obtained by standard theory for goodness of fit evaluation in Normal LMs.

- $Y_i \sim N(\mu_i, \sigma^2)$, $b'(\theta) = \frac{\theta^2}{2}$, $\theta = \mu = b'(\theta)$ and $\phi = \sigma^2$.
- $\ell(\theta) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$
- For the saturated model $\tilde{\mu}_i = y_i$, and

$$\ell(\tilde{\theta}) = -\frac{n}{2}\log\sigma^2 \ .$$

- For the current model $\hat{\mu}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$, and

$$\ell(\hat{\theta}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

- Scaled deviance is

$$D(y;\hat{\theta}) = \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\sigma^2}$$

Poisson

- $Y_i \sim Poisson(\mu_i), b(\theta_i) = e^{\mu_i} = b'(\theta_i), \phi = 1, \log \mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$
- $\ell(\theta) = \sum_{i=1}^n y_i \log \mu_i - \sum_{i=1}^n \mu_i$
- For the saturated model $\tilde{\mu}_i = y_i$, and

$$\ell(\tilde{\theta}) = \sum_{i=1}^n y_i \log y_i - \sum_{i=1}^n y_i \ .$$

- For the current model $\log \hat{\mu}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$, and

$$\ell(\hat{\theta}) = \sum_{i=1}^n y_i \log \hat{\mu}_i - \sum_{i=1}^n \hat{\mu}_i \ .$$

- So deviance is $D(y; \hat{\theta}) = 2 \left(\sum_{i=1}^n y_i \log \frac{y_i}{\hat{\mu}_i} - \sum_{i=1}^n y_i + \sum_{i=1}^n \hat{\mu}_i \right)$

Binomial

- $Y_i \sim Bin(1, \pi_i)$, con $\pi_i = Pr(Y_i = 1) = E(Y_i) = \mu_i$
- $\ell(\theta) = \sum_{i=1}^{n} (y_i \log \pi_i + (1 - y_i) \log(1 - \pi_i))$
- For the saturated model $\tilde{\mu}_i = y_i$ and

$$\ell(\tilde{\theta}) = \sum_{i=1}^{n} (y_i \log y_i + (1 - y_i) \log(1 - y_i)) \ .$$

- For the current model $logit(\hat{\mu}_i) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$ and

$$\ell(\hat{\theta}) = \sum_{i=1}^{n} (y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)) \ .$$

- The deviance is

$$D(y; \hat{\theta}) \quad = \quad 2 \sum_{i=1}^{n} \left(y_i \log \frac{y_i}{\hat{\pi}_i} + (1 - y_i) \log \frac{1 - y_i}{1 - \hat{\pi}_i} \right) \ .$$

Comparing nested models

- Considering two nested models M_C and M_R, likelihood ratio test is

$$\begin{aligned} W &= 2\left\{\ell(\hat{\beta}) - \ell(\hat{\beta}_{MR})\right\} \\ &= \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi}, \end{aligned}$$

  as $n \to \infty$ it is distributed $\chi^2_{p-p_0}$ when H_0 holds.

- So to test if reduced model can be accepted we can compare

$$W = \frac{D(Y, \hat{\theta}_{MR}) - D(Y, \hat{\theta})}{\phi}$$

  with the quantiles of the distribution $\chi^2_{p-p_0}$. We reject H_0 for large values of the statistic (or for a small *p-value*).

Residual Deviance

- It is important to note that since also deviance is defined as a function of the difference arising from a log-likelihood ratio of two nested model one is tempted to use the same criteria for evaluating if deviance of the current model is significantly small. One can look if value of deviance is not large enough when compared to a χ^2_{n-p}.
- In this last case standard asymptotic theory could not work when the number of parameter in the saturated model is not fixed as n goes to infinity.

Nonetheless the criterion could work when the number of parameters is fixed: this is, for instance, the case of a binomial model for grouped data or a Poisson model with factors as the only covariates (as it happens in log linear model from contingency tables).

- In some cases (the most notable being binomial and Poisson) the dispersion parameter is fixed to 1.
- When dispersion parameter ϕ is not known another consistent estimate of it must be considered

$$\widehat{\phi} = \frac{D(Y, \hat{\theta})}{(n-p)}$$

and under mild conditions the result stated above still works.

Model selection

- Model selection strategies can exploit the tools defined above to explore which combination of explanatory variables leads to a satisfactory model.
- So one can consider a stepwise backward search by starting with a model that includes all the covariates and then consider a set of reduced sub models obtained by removing certain variables (backward selection). In order to choose among models, one can consider the sub-model obtained by deleting variables with a large p-value.
- A forward search starts from the null model (usually the one including only the intercept) and (groups of) variables are included if the p-values associated are small.
- A combination of the two strategies can also be considered.
- To compare models also the well known criteria AIC and BIC can be used. For instance, in this case $AIC = -2\ell(\hat{\theta}) + 2p$ where p is the number of parameters of the model (when dispersion parameter is known) and one chooses the model where AIC is smaller.

Residuals in GLM

- Let us recall the basic ideas in using residual analysis in LMs:
  - residuals are easily defined as the difference between the observed datum and the estimated systematic part of the model: this step is less natural in GLM.
  - residuals tell us if there are symptoms of systematic differences between observed and fitted values (i.e. plot of residuals against fitted values, or against covariates)
  - residuals help us recognizing discrepancies between few data and the rest (ouliers detection, evaluation of leverage: hat matrix, case deletion measures -Cook's distance-, jacknife residuals, etc.)
- Some of these ideas can be generalized in GLMs.
- A straight extension of the concept of standardized residual is given by

$$r_{Pi} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} V(\hat{\mu}_i)}} \,, \tag{6}$$

  called *Pearson residuals*. The definition (6) resembles that for residuals in LMs based on the estimation of the error term ϵ_i.

Deviance residuals

- Recall that in GLMs ϵ_i does not exist in general, so we can measure the contribution of each observation to deviance. This is analogous to LMs where SSE is defined as

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i^T \hat{\beta})^2 \ ,$$

  while in GLMs a similar quantity is the deviance. Recall that deviance is defined as

$$D(y, \hat{\theta}) = \sum_{i=1}^{n} D_i \ .$$

  Large individual contributions to total deviance D_i reflect data that are not properly reproduced by the model. Let us define

$$r_{Di} = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{D_i} \ ,$$

  that is called *deviance residual* of the model.

For large n it is possible to show that $r_{Pi} \approx r_{Di}$.
Other residuals, such as Anscombe residuals, are also defined for GLMs.

Residual analysis

- Actually if the model is valid, residuals of any type, possibly scaled by $\hat{\phi}$, will have a distribution that can be (loosely) approximated by a $N(0,1)$. This suggest to use standard graphical tools, like

  $\mapsto$ normal probability plot of the residuals;

  $\mapsto$ plot of residuals against the fitted values $\hat{Y}_i$;

  $\mapsto$ plot of residuals against explanatory variables

  to check assumptions.

- It is also possible to generalize the Hat matrix H to check influence and leverage of residuals. Recall that H in LMs is such that $\hat{\boldsymbol{y}} = \mathbf{H}\boldsymbol{y}$ and
  $H = X(X^T X)^{-1} X^T$.

- Generalized hat matrix is similarly obtained as
  $H = W^{\frac{1}{2}} X(X^T W X)^{-1} X^T W^{\frac{1}{2}}$ where W is substituted by $\hat{W}$.

- A generalization of the Cook's distances is also possible.

Quasi-likelihood

More on quasi-likelihood

- For LMs the method of LS allows to obtain estimates of the regression parameters without the specification of a probabilistic model.
- The method of LS requires only the specification of the relation between the expected value of the response variable and the linear predictor, and the specification of the variance of the error term, which is not related to the expected value:

$$E(Y_i) = \mu_i = \eta_i \quad var(Y_i) = \sigma^2$$

- Also for the GLMs it is possible to specify only these two relations (assuming that the variance function $V(\mu_i)$ is known).
- Indeed, the likelihood equation for β

$$\sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i) g'(\mu_i)} x_{ij} = 0 \ , \quad j = 1, \ldots, p \ ,$$

  is an unbiased estimating equation provided that $E(Y_i) = \mu_i = g^{-1}(\eta_i)$.
- In other words, this means that the parametric assumption $Y_i \sim EF(\cdot, \phi)$ could not even be satisfied. Only the assumption about expectations is essential: $\mu_i = E(Y_i) = g^{-1}(\eta_i)$
- The only distributional feature that must be known in order to calculate the estimating equation is the variance function $V(\mu)$.

Quasi-likelihood model

- Under suitable regularity conditions, the likelihood equations for a GLM give estimates for the coefficients β which maintain several properties, also if the parametric assumptions of Y_i are substituted with weaker **second order assumptions**:

  1. $g(\mu_i) = g(E(Y_i)) = \eta_i, \qquad i = 1, \ldots, n$
  2. $var(Y_i) = \phi V(\mu_i), \qquad i = 1, \ldots, n$
  3. $cov(Y_i, Y_j) = 0$, if $i \neq j$.

- The semi-parametric statistical model specified by assumptions 1–3 is called **quasi-likelihood model**.
- If $V(\mu) = 1$ and $g(\mu) = \mu$, the assumptions 1–3 match the usual second order assumptions of the classical LM.
- On the other hand, if $V(\mu) = \mu^2$ we obtain a multiplicative model, $Y_i = \mu_i \epsilon_i$, with $E(\epsilon_i) = 1$ and $var(\epsilon_i) = \phi$.

Quasi-likelihood equations

- Gauss-Markov (BLUE) optimality of LS extends to quasi-likelihood estimates and it has minimum asymptotic variance among estimating equations that are linear (in Y) and unbiased
- Indeed, the likelihood equation for β

$$q(y;\beta) = \sum_{i=1}^{n} q(y_i;\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)g'(\mu_i)} x_{ij} = 0 \ , \quad j = 1, \ldots, p \ ,$$

  behaves like a score vector. Specifically:

$$E(q(Y;\beta)) = 0, \quad \text{and} \quad var(q(Y;\beta)) = -E(\partial q(Y;\beta)/\partial\beta) \ .$$

- Quasi likelihood estimators shares many properties of a proper likelihood: the quasi-MLE β is asymptotically normal, the quasi-likelihood ratio statistic has a null chi-squared distribution.

Quasi-likelihood and overdispersion

- The assumptions 1–3 offer an increase in flexibility with respect to the usual parametric specifications based, respectively, on the Poisson, binomial or exponential distributions.
- In practice, there are situations in which the dispersion parameter does not agree with the assumed exponential family.
- For example, for the binomial or Poisson distributions we have $\phi = 1$, but data could show agreement with $\phi > 1$.
- In this case we have *overdispersion*, i.e. the variance of Y is greater than its theoretical value, and it is more plausible to assume $var(Y_i) = \phi V(\mu_i)$, with $\phi > 1$. For example, for proportions, it can be assumed that $var(Y) = \phi n\pi(1-\pi) > n\pi(1-\pi)$, with $\phi > 1$, where $n\pi(1-\pi)$ is the variance of a binomial distribution.
- In general, the quasi-likelihood approach allows to deal with *overdispersion problems*: it is possible to specify $var(Y_i)$ so that there is more variability with respect to the exponential family.

GLM: Extensions and recent development

GLM: Extensions and recent development

- Generalized linear models and the relevant theory have been introduced in the "80ies.
- They have been extended in many directions to take into account more complex data structures
- It has been introduced the use of quasi-likelihood estimators that allows to specify only the mean and the variance of the response Y. This leaves more flexibility and is a simple strategy to cope with overdispersion.
- It has been extended to multivariate responses.
- it has been estended to take into account nested structure of the data and the lack of independence that can arise in those cases (Generalized Linear Mixed Models GLMM).
- Procedures for regularization, such as the LASSO, can be adopted also in case of GLMs
- it has been estended to take into account non linear functions of the covariates and to estimate this functions non parametrically (Generalized additive models GAMs).
- Use of a different inferential approach, such as Bayesian inference, have been considered.