Lecture 3. GLM II.

Course Designer: Jesper Rydén, SLU, Uppsala, Sweden

Course Lecturer: Raazesh Sainudiin, UU, Uppsala, Sweden

Generalized Linear Models 1MS369 • Autumn 2018

Definition: The generalized linear model

DEFINITION. Assume that $Y_1, Y_2, \ldots, Y_n$ are mutually independent, and the densituy can be described by an ED family with the same variance function $V(\mu)$.

A **generalized linear model** for $Y_1, \ldots, Y_n$ describes an affine hypothesis for $\eta_1, \ldots, \eta_n$ where

$$\eta_i = g(\mu_i)$$

is a transformation of the mean values $\mu_1, \ldots, \mu_n$.

The hypothesis is of the form

$$H_0 : \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L,$$

where L is a linear subspace of $\mathbb{R}^n$ of dimension k and $\boldsymbol{\eta}_0$ denotes a vector of known off-set values.

Linear models vs. GLM

Classical linear models	**Generalized linear models**
Normal distribution	Exponential dispersion family
Mean value plus error	Distribution specified by mean value
Constant variance	Variance function of the mean
Easy to apply (in R: `lm`)	Almost as easy to apply (in R: `glm`)
Exact results	Approximate results

Dimension and design matrix of the GLM

DEFINITION. The dimension k of the subspace L for the GLM is the *dimension* of the model.

DEFINITION. Consider the linear subspace $L = \text{span}\{x_1, \ldots, x_k\}$, i.e. the subspace is spanned by k vectors ($k < n$) such that the hypothesis can be written

$$\boldsymbol{\eta} - \boldsymbol{\eta}_0 = \mathbf{X}\boldsymbol{\beta}$$

with $\boldsymbol{\beta} \in \mathbf{R}^k$, where $\mathbf{X}$ has full rank. The $n \times k$ matrix $\mathbf{X}$ is called the **design matrix**.

The ith row of the design matrix is given by the model vector

$$\mathbf{x}_i = (x_{i1}\ x_{i2}\ \cdots x_{ik})^{\mathsf{T}}$$

for the ith observation.

Specification of a GLM

Distribution/variance function. Specification of the distribution or the variance function $V(\mu)$.

Link function. Specification of the link function $g(.)$, which describes a function of the mean value which can be described linearly by the explanatory variables.

Linear predictor. Specification of the linear dependency for each case,

$$g(\mu) = \eta = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$$

Precision. If needed, the precision is formulated as known individual weights $\lambda_i = w_i$ or as a common dispersion parameter $\lambda = 1/\sigma^2$ or a combination $\lambda_i = w_i/\sigma^2$.

The link function

DEFINITION. The link function $g(.)$ describes the relation between the linear predictor η_i and the mean-value parameter $\mu_i = \mathsf{E}[Y_i]$. The relation is

$$\eta_i = g(\mu_i).$$

The inverse mapping $g^{-1}(.)$ thus expresses the mean value μ as a function of the linear predictor η:

$$\mu = g^{-1}(\eta),$$

that is,

$$\mu_i = g^{-1}(\sum_j x_{ij}\beta_j).$$

Canonical link

The **canonical link** g_c for a distribution is the link function such that

$$g_c(\mu_i) = \theta_i$$

where θ_i is the canonical parameter of the distribution.

The canonical link is thus the function which transforms the mean to the canonical parameter of the ED family.

Good property: μ stays within the range of the response variable.

The link function, canonical link and cumulant generator

Consider the function

$$\tau(\theta) = \kappa'(\theta)$$

which defines a one-to-one mapping $\mu = \tau(\theta)$ of the parameter space Ω for the canonical parameter θ on to a subset of the real line, called the mean-value space.

DEFINITION. The inverse mapping

$$\theta = \tau^{-1}(\mu)$$

is called the canonical link function.

Example: Canonical link for Poisson distribution

For $Y \sim \text{Po}(\mu)$, we have

$$f_Y(y;\mu) = \exp\big(y \ln \mu - \mu - \ln y!\big)$$

and identify the canonical parameter: $\theta = \ln \mu$.

Blackboard

Canonical link functions for some widely used distributions

Distribution	Link $\eta = g(\mu)$	Name
Normal	μ	identity
Poisson	$\ln \mu$	logarithm
Binomial	$\ln(\mu/(1-\mu))$	logit
Gamma	$1/\mu$	reciprocal
Inverse Gauss	$1/\mu^2$	power (raised to 2)

Nelder on the canonical link

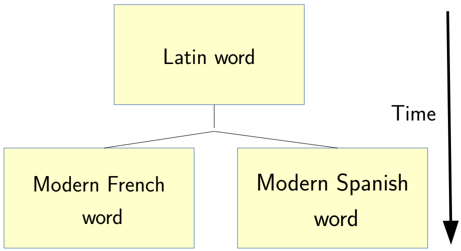
It should be stressed, however, that, although the canonical links lead to desirable statistical properties of the model, particularly in small samples, there is in general no *a priori* reason why the systematic effects in the model should be additive on the scale given by that link.

P. McCullagh and J.A. Nelder (1983). Generalized Linear Models (Section 2.2.4).

Example: Historical linguistics

Consider a language descendent of another language.

A simple model for the change in vocabulary: if languages are separated by time t, then the probability that they have cognate words for a particular meaning is $\exp(-\theta t)$, where θ is a parameter.



Example, ctd.

It is believed that θ is approximately the same for many commonly used meanings.

For test list of N different commonly used meanings suppose that a linguist judges, for each meaning, whether the corresponding words in two languages are cognate or not cognate.

Generalized linear model?

Example from Dobson and Barnett (2008)

Blackboard

Example: Mortality rates

Number of deaths Y in a population can be modelled by a Poisson distribution:

$$f(y;\mu) = \frac{\mu \mathrm{e}^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots$$

where $\mathsf{E}[Y] = \mu$ is the expected number of deaths in a specified time period (e.g. one year).

The parameter μ may depend on population size, period of observation and various characteristics of the population (e.g. age, sex, medical history).

Example of a model:

$$\mathsf{E}[Y] = \mu = n\lambda \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$$

where

n:	population size
$\lambda \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta}$:	rate per 100 000 people per year

Example, ctd.

Data: Number of deaths from coronary heart disease, population sizes by 5-year age groups for men in the Hunter region of New South Wales, Australia, in 1991.

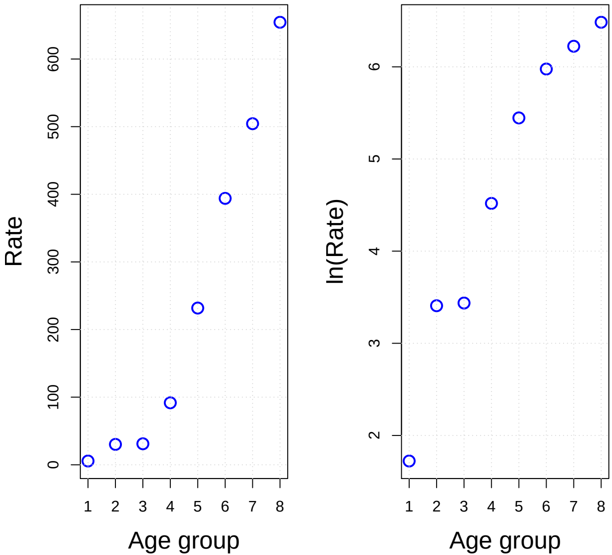
Age group	Nr of deaths y_i	Population size n_i	Rate
30-34	1	17,742	5.6
35-39	5	16,554	30.2
40-44	5	16,059	31.1
45-49	12	13,083	91.7
50-54	25	10,784	231.8
55-59	38	9,645	394.0
60-64	54	10,706	504.4
65-69	65	9,933	654.4

Rate: $y_i/n_i \times 100,000$, i.e. rate per 100,000 men per year.

Example from Dobson and Barnett (2008)

Blackboard

Example, ctd.



Rate
0
100
200
300
400
500
600
1 2 3 4 5 6 7 8
Age group
ln(Rate)
2
3
4
5
6
1 2 3 4 5 6 7 8
Age group

ML estimation of βs

Recall notions:

$\eta_i = \mathbf{x}^\mathsf{T}\boldsymbol{\beta}$	The linear predictor
$\mu_i = \mathrm{E}[Y_i]$	The expected value of y_i.
$\eta_i = g(\mu_i)$	The link function, translates between linear predictor and mean of Y_i.
$g^{-1}(\eta_i) = \mu_i$	The inverse link function, the mean as a function of the linear predictor

Think of a sequence:

- η_i depends on $\mathbf{x}$ and $\boldsymbol{\beta}$
- μ_i depends on η_i (via the inverse of the link function)
- θ_i depends on μ_i via distribution specific linkage
- $f(y_i; \theta_i)$ depends on θ_i

ML estimation of βs

Consider a GLM with $Y \sim \text{ED}(\mu, V(\mu)/\lambda)$.
Recall the specification of the linear part, for each case:

$$g(\mu) = \eta = \mathbf{x}^\mathsf{T}\boldsymbol{\beta}.$$

Our aim is to estimate the parameters $\boldsymbol{\beta}$ (and later on, perform inference).

Recall GLM fact I:

$$\text{E}[Y_i] = \mu_i = \kappa'(\theta_i) = \tau(\theta_i).$$

We will first elaborate the score equation with respect to β_j (termwise), before presenting a compact matrix expression.

Blackboard

The local design matrix

The mappings

$$\mu_i = g^{-1}(\eta_i)$$

amd the linear mapping

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

defines $\boldsymbol{\mu}$ as a function of $\boldsymbol{\beta}$. The function is a vector function $\mathbb{R}^k \to \mathbb{R}^n$.

The matrix

$$\mathbf{X}(\boldsymbol{\beta}) = \frac{\partial \mu}{\partial \beta} = \left[\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right]^{\mathsf{T}} \frac{\partial \eta}{\partial \beta} = \mathrm{diag}\left\{\frac{1}{g'(\mu_i)}\right\}\mathbf{X}$$

is called the **local design matrix** corresponding to the parameter value $\boldsymbol{\beta}$.

The canonical link

For the canonical link we have

$$g'(\mu) = \frac{1}{V(\mu)}$$

and then

$$\left[\frac{\mathrm{d}\mu}{\mathrm{d}\eta}\right] = \mathrm{diag}\left\{\frac{1}{g'(\mu_i)}\right\} = \mathrm{diag}\left\{V(\mu_i)\right\}.$$

The local design matrix:

$$\mathbf{X}(\boldsymbol{\beta}) = \mathrm{diag}\left\{V(\mu_i)\right\}\mathbf{X}.$$

The score functions for ML estimation

THEOREM. The ML estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is found as the solution to

$$[\mathbf{X}(\boldsymbol{\beta})]^{\mathsf{T}}\mathbf{i}_{\mu}(\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu}) = \mathbf{0}$$

where $\mathbf{X}(\boldsymbol{\beta})$ is the local design matrix and $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\beta})$ is given by

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

Madsen & Thyregod, Theorem 4.1

Note. This system of equations is non-linear in $\boldsymbol{\beta}$, iterative procedures are needed.

$\boxed{\text{Blackboard: Proof}}$

The score equations for the canonical link

For the canonical link, the local design matrix is

$$\mathbf{X}(\beta) = \text{diag}\{V(\mu_i)\}\mathbf{X}.$$

Score equations simplify to the mean-value equation

$$\mathbf{X}^\mathsf{T}\text{diag}\{w_i\}\mathbf{y} = \mathbf{X}^\mathsf{T}\text{diag}\{w_i\}\boldsymbol{\mu}$$

which for an un-weighted model reduces to

$$\mathbf{X}^\mathsf{T}\mathbf{y} = \mathbf{X}^\mathsf{T}\boldsymbol{\mu}.$$

Blackboard

Example. Poisson distribution (Ex. 4.7)

In some simple situations, the score equations can be solved without iteration.

We will study a model for Poisson regression with with only one covariate – an intercept.

Consider the number of accidents with personal injury during daylight hours in January quarter, years 1987-1990. (Danish roads.) The *traffic index* is a measure for the amount of traffic on the roads (index for 1987 set to 100).

Year	1987	1988	1989	1990
Accidents y_i	57	67	54	59
Traffic index x_i 100	111	117	120	

Assume that $Y_i \sim \text{Po}(\mu_i)$ where

$$\mu_i = \gamma x_i,$$

thus γ is a rate.

Blackboard