Probability and Statistics 2 (GCED)
Generalized Linear Model

Marta Pérez-Casany and Jordi Valero Bayà

Department of Statistics and Operations Research
Technicat University of Catalonia

Facultat d'Informàtica de Barcelona, First Semester 2018

Generalized Linear Models: Motivation

EXAMPLE

Objective: To model the *length*, Y, of a plant as a function of the *days*, X, it has been planted.

- First attempt: Linear model $\mathbf{Length} = \beta_0 + \beta_1 \cdot \mathbf{days} + e$

  it is not appropiate if it doesn't exists a linear relation.

- Second attempt: Curvilinear model $\log(\mathbf{Length}) = \beta_0 + \beta_1 \cdot \mathbf{days} + e$

  it assumes that Lenght follows a Log-Normal distribution, and this may not be the case.

Generalized Linear Models: Motivation

Thus, a modelization technique that allows simultaneously to assume:

1) Normality for Y,
2) that a transformation of $\mu = E(Y)$ be linear in the covariates

is required.

Generalized Linear models do that and MUCH MORE!!!! since they allow that:

1) Y follows a more general probability distribution, not necessarily Normal,
2) the linearity is between a transformation of μ and the covariates.

Generalized Linear Models: Motivation

For the particulcar case of the Length of a plant example,

we need to assume:

$$\log(\mu) = \beta_0 + \beta_1 \cdot \mathrm{days},$$

where $\mu = E(Y)$ and Y is Normal distributed, and this can be done with Generalized Linear Models (GLM).

Generalized Linear Models: Definition

A GLM has three components:

1. **Random Component**: random vector $Y_{n\times 1} = (Y_1, Y_2, \cdots, Y_n)^t$,

$$Y_i \sim \exp\left(\frac{y\theta_i - b(\theta_i)}{a(\phi)} + c(y, \phi)\right). \tag{1}$$

   where θ_i changes by changing the covariates and ϕ is known as **dispersion parameter** also denoted by σ^2. $\mu_i = E(Y_i)$ depends on θ_i and ϕ.

2. **Deterministic component**: $X_{n\times p}\beta_{p\times 1}$, where $p < n$.

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

3. **Link Function**: Any monotone differentiable function η may be considered.

   The model is equal to:

   $$\eta = g(\mu) = X\beta, \quad \text{where } \mu = E(Y).$$

   and each component of Y verifies (1).

Generalized Linear Models: Random Component

The following probability distributions may be written as (1):

- **Normal**, $f(y;\mu,\sigma^2) = \exp\left(\frac{y\mu}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\Pi\sigma^2})\right)$;

  $$\theta = \mu,\ a(\phi) = \phi = \sigma^2,\ b(\theta) = \frac{\theta^2}{2},\ c(y;\phi) = \frac{-y^2}{2\sigma^2} - \log(\sqrt{2\Pi\sigma^2}).$$

- **Poisson**, $f(y;\lambda) = \exp\left(y\log(\lambda) - \lambda - \log y!\right)$;

  $$\theta = \log(\lambda),\ a(\phi) = \phi = 1,\ b(\theta) = e^\theta,\ c(y;\phi) = -\log y!.$$

Generalized Linear Models: Random Component

- **Binomial, with m known**, $f(y;p) = \exp\left(y\log(\frac{p}{1-p}) + m\log(1-p) + \log\binom{m}{y}\right)$;

$$\theta = \log\left(\frac{p}{1-p}\right),\ a(\phi) = \phi = 1,\ b(\theta) = m\log(1+e^{\theta})$$

and

$$c(y;\phi) = \log\binom{m}{k}.$$

- **Gamma**, $f(y;\rho,\phi^*) =$ $\exp\left(\left(\frac{-\rho}{\phi^*}y + \log\left(\frac{\rho}{\phi^*}\right)\right)\phi^* + (\phi^*-1)\log y - \log\Gamma(\phi^*) + \phi^*\log(\phi^*)\right)$;

$$\theta = \frac{-\rho}{\phi^*},\ \ b(\theta) = -\log(-\theta),\ \ a(\phi) = \phi = (\phi^*)^{-1},$$

$$c(y,\phi) = (\frac{1}{\phi}-1)\log y - \log\Gamma(1/\phi) - \frac{\log\phi}{\phi}.$$

Generalized Linear Models: Random Component

- **Inverse Gaussian**,

$$f(y;\mu,\lambda) = \exp\left(\frac{\frac{-1}{2\mu^2}y + \frac{1}{\mu}}{1/\lambda} - \frac{\lambda}{2y} + \frac{1}{2}\log(\lambda) - \frac{1}{2}\log(2\Pi y^3)\right);$$

$$\theta = \frac{-1}{2\mu^2},\; b(\theta) = \sqrt{-2\theta},\; a(\phi) = \phi = \frac{1}{\lambda},$$

$$c(y;\phi) = \frac{-\lambda}{2y} + \frac{1}{2}\log(\lambda) - \frac{1}{2}\log(2\Pi y^3).$$

Generalized Linear Models: Random Component

If $Y \sim \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right)$;

$$l(\theta; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)$$

Taking into account the properties of the score vector, one has that

$$0 = E\left(\frac{\partial l}{\partial \theta}\right) = E\left(\frac{y - b^{'}(\theta)}{a(\phi)}\right) \Longleftrightarrow E(Y) = b^{'}(\theta) = \mu;$$

$$\begin{aligned} E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -E\left(\left(\frac{\partial l}{\partial \theta}\right)^2\right) \quad &\Longleftrightarrow \quad \frac{b^{''}(\theta)}{a(\phi)} = E\left(\frac{Y - b^{'}(\theta)}{a(\phi)}\right)^2 \\ &\Longleftrightarrow \quad Var(Y) = a(\phi) b^{''}(\theta). \end{aligned}$$

The function $\mathbf{V}(\mu) = \mathbf{b}^{''}(\theta)$ is known as **Variance Function**.

Generalized Linear Models: Variance Function

Distribution	$E_\theta(Y) = b'(\theta)$	$Var_\theta(Y) = b^{''}(\theta)a(\phi)$	$V(\mu)$
Normal	$\theta = \mu$	σ^2	1
Poisson	$e^\theta = \mu$	$e^\theta = \mu$	μ
Binomial/m	$\frac{e^\theta}{1+e^\theta} = p = \mu$	$\frac{e^\theta}{(1+e^\theta)^2} = \mu(1-\mu)\,\frac{1}{m}$	$\mu(1-\mu)$
Gamma	$-\frac{1}{\theta} = \mu$	$\frac{1}{\theta^2}\phi = \mu^2\phi$	μ^2
Inversa Gausiana	$\frac{1}{\sqrt{-2\theta}} = \mu$	$\frac{1}{(\sqrt{-2\theta})^3}\phi$	μ^3

Generalized Linear Models: Canonical Link

Function η such that $\eta = \mathbf{g}(\mu) = \theta$ is denoted as **canonical link function** or **canonical link**

Distribution	$E_\theta(Y) = b'(\theta) = \mu$	Canonical Link
Normal	θ	μ
Poisson	e^θ	$\log(\mu)$
Binomial	$n\frac{e^\theta}{1+e^\theta}$	$\log\left(\frac{\mu}{n-\mu}\right)$
Gamma	$\frac{-1}{\theta}$	$\frac{1}{\mu}$
Inverse Gausian	$\frac{1}{\sqrt{-2\theta}}$	$\frac{1}{\mu^2}$

Generalized Linear Models: Canonical Link

To use the canonical link function has some advantatges:

1. $y^t X$ is a *sufficient* statistic.
2. The m.l.e is easier to be found (less computation)
3. The model is easier to be interpreted.

Generalized Linear Modes: m.l.e

To find the m.l.e of parameter vector β, the following iterative equation needs to be solve:

$$X^t W X b^{m+1} = X^t W Z; \tag{2}$$

where

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \text{ and } W_{ij} = 0 \text{ if } i \neq j$$

and

$$z_i = \sum_{j=1}^{p} x_{ij} b_j^m + (y_i - \mu_i) \frac{\partial \eta_i}{\partial \mu_i}.$$

(2) is equivalent to an iteratively **weighted least squares**.

Generalized Linear Modes: m.l.e

Each iteration may be described by means of the following figure:

$$\beta^m \longrightarrow \eta \longrightarrow \mu \longrightarrow \theta \longrightarrow Var(Y) \longrightarrow \left\{ \begin{array}{l} W \\ Z \end{array} \right. \longrightarrow \beta^{m+1}$$

To start the iteration,

- Initial value: μ_o equal to the observed values
- one starts the process by the second step.

Observation: If the canonical link is used, the two first steps are not required.

GLM: Predicted. Null and Full models

Once $\hat{\beta}$ is obtained, the predicted mean values are equal to:

$$\hat{y} = \hat{\mu} = g^{-1}(X\hat{\beta})$$

DEFINITIONS

The **null model** is defined to be the model with just one parameter (intercept).

The **full model** is defined to be the model with as many parameters as observations. The fit obtained with the full model is the **perfect fit**.

Observation: One wants to obtain a fit close to the perfect fit but with less parameters.

GLM: Goodness of fit measures

Let $l(\hat{\mu}, \phi; y)$ and $l(y, \phi; y)$ be the values of the log-likelihood corresponding to our model and full model respectively,

The **scaled deviance** is defined as

$$D^*(y; \hat{\mu}) = 2(l(y; y) - l(\hat{\mu}; y)).$$

To compare

$$H_0 : \text{our model} \qquad vs \qquad H_1 : \text{full model},$$

one has that under H_0, assimptotically $D^*(y; \hat{\mu}) \sim \chi^2_{n-p}$, where p is the number of parameters of our model.

So, one rejects H_0 when $D^*(y; \hat{\mu}) \geq \chi^2_{\alpha, n-p}$.

GLM: Goodness of fit measures

In general, when $a(\phi) = \phi/w_i$, which is the case for instance of the Normal, Poisson and Binomial distributions ($w_i = 1$)

$$D^*(y;\mu) = \frac{\mathbf{D}(\mathbf{y};\mu)}{\phi}.$$

Function $D(y;\mu)$ is known as **deviance**.

For the Poisson and the Binomial distribuions, deviance and scaled deviance are equal.

For the Normal distribution the scaled deviance is the deviance divided by σ^2.

For the Gamma distribution the scaled deviance is equal to the deviance multiplied by one of the parameters of the distribution.

Distribución	Devianza
Normal	$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$
Binomial	$2\sum_{i=1}^{n}\{y_i \log(y_i/\hat{\mu}_i) + (n - y_i)\log[(n - y_i)/(n - \hat{\mu}_i)]\}$
Gamma	$2\sum_{i=1}^{n}\{-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i\}$
Inversa Gausiana	$\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)$

Observación: Si la distribución asumida es la de Poisson, D es el estadístico G^2 de Bishop et al (1975).

GLM: Goodness of fit measures

The scaled deviance allows to compare two **nested models**.
DEFINITION:

Given two models (mod1, mod2), it is said mod1 is **nested** in mod2 if, and only if, mod2 contains all the parameters in mod1 and some more.

Denoting by p_i the number ob parameters of mod$i$, and by D_i its corresponding scaled deviance,

to compare

$$H_0 : mod1 \quad vs \quad H_1 : mod2$$

one has that under H_0, assimptotically

$$D_1 - D_2 \sim \chi^2_{p_2 - p_1}$$

and we reject H_0 when $D_1 - D_2 \geq \chi^2_{\alpha, p_2 - p_1}$

GLM: Goodness of fit measures

The $\mathbf{X^2}$ **generalized Pearson Statistics** is defined as

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)};$$

where $V(\hat{\mu}_i)$ is the variance function at $\hat{\mu}_i$.

For the Normal distribution, X^2 is equal to the residual sum of squares.

If the model is appropiate and $a_i(\phi) = \phi$, the assimptotic distribution of $\frac{X^2}{\phi}$ is a χ^2 with $n - p$ degrees of freedom. Thus,

we **reject our model when** $X^2 \geq \chi^2_{\alpha, n-p}$

Observation: X^2 is a more intuitive measure but it doesn't allow to compare nested models.

GLM: Residuals

- **Pearson residual** $r_p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$.
- **Deviance residual** $r_D = \text{sing}(y_i - \hat{\mu}_i) d_i$ it is verified:

$$D(y;\mu) = \sum_{i=1}^{n} d_i^2$$

GLM: Dispersion parameter estimation

Given that if our model is appropiate, $\frac{X^2}{\phi} \sim \chi^2_{n-p}$, applying the moment method one has that:

$$\tilde{\phi} = \frac{X^2}{n-p},$$

which is a consistent estimator for ϕ.

If we assume Poisson or binomial response, $\tilde{\phi} \simeq 1$.

If this is not the case, the real dispersion of the data is larger than what it should be, which is known as **Overdispersion**.

Summarizing:

- GLM allow to model $\mu = E(Y)$ as a function of the covariates, relaxing the normality and homocedasticity hypothesis of the Linear model. Nevertheless, they require that:
  a) the distribution of Y be written in a given exponential form,.
  b) a link function be specified.

- In the GLM theory, the independence of the observations keeps being necessary.

- It may be necessary to deal with a dispersion parameter (ϕ) that plays the same role as σ^2 in LM.

- The punctual estimation of β is obtained by iteratively weighted least squares.
- Asumptotically $\hat{\beta}$ follows a Normal distribution.
- The deciance and the χ^2 generalized pearson statistic allow to study the goodnes-of-fit of the model. Asymptotically they are equivalent.
- The deviance also allows to compare nested models.
- The dispersion parameter estimation is done by means of $\chi^{\mathbf{2}}$ divided by its degrees of freedom.
- We have two types of residuals: the deviance residuals and the χ^2 residuals, that both asumptotically are nomal distributed.