

GLM : Generalized Linear Models

GLM - Une théorie unificatrice

intégrer
- régression
linéaire
- régression
logistique
- ...

1. Modèle linéaire généralisé - principe et ingrédients

On note Y la variable dépendante, avec n exemples Y_i , et p variables indépendantes

(explicatives) X_j

features

↑
à prédire

$$\hat{Y} = f(X)$$

modèle? prédiction
Erreur / coût $C(Y, \hat{Y})$

mesure de qualité
permet d'ajuster les paramètres
de f

1.1. Régression linéaire

Dans la régression linéaire, on cherche

$$\hat{Y}_i = \sum_{j=1}^p \beta_j X_{ij}$$

au sens des moindres carrés, et on se rappelle que ceci correspond à une hypothèse de bruit gaussien. Ainsi, Y est supposé gaussien, centré,

$$\eta_i = \sum_j \beta_j X_{ij}$$

Notation

$$Y_i = \sum_{j=1}^p \beta_j X_{ij} + e = \eta_i + e$$

$$= \hat{Y}_i + e = \eta_i + e$$

et sa *moyenne* $\mu_i = E[Y_i | X_{ij}]$ est un modèle linéaire. On a noté $\eta_j = \sum_{j=1}^p \beta_j X_{ij}$ la combinaison

linéaire des variables dépendantes. Bien évidemment on a ici $\eta_j = \mu_j$.

Le modèle linéaire porte sur la moyenne μ_i

$$E[Y_i] = \mu_i = \sum \beta_j E[X_{ij}]$$

$$\begin{aligned} E[Y_i] &= E\left[\sum_j \beta_j X_{ij} + e\right] \\ &= \sum_j E[\beta_j X_{ij}] + E[e] \\ E[Y_i] &= \sum_j \beta_j E[X_{ij}] \quad \overline{0} \end{aligned}$$

Rappel : lien régression linéaire et gaussiennes.

$$Y_i = \sum_{j=1}^p \beta_j X_{ij} + e_i$$

Supposons que e_i soit gaussien, alors

$$P(Y_i | X_{i1}, X_{i2}, \dots, X_{ip}) = P(e_i) = \frac{1}{\sqrt{2\pi}\sigma_e} e^{-\frac{e_i^2}{2\sigma_e^2}}$$

Proba de Y si les features sont connus

$$P(Y_1, \dots, Y_n | X) = \prod_{i=1}^n P(e_i) = \frac{1}{\sqrt{2\pi}\sigma_e} e^{-\frac{\sum e_i^2}{2\sigma_e^2}}$$

Par conséquent maximiser $P(Y|X)$ soit la vraisemblance
revient à minimiser $\sum_{i=1}^n e_i^2$

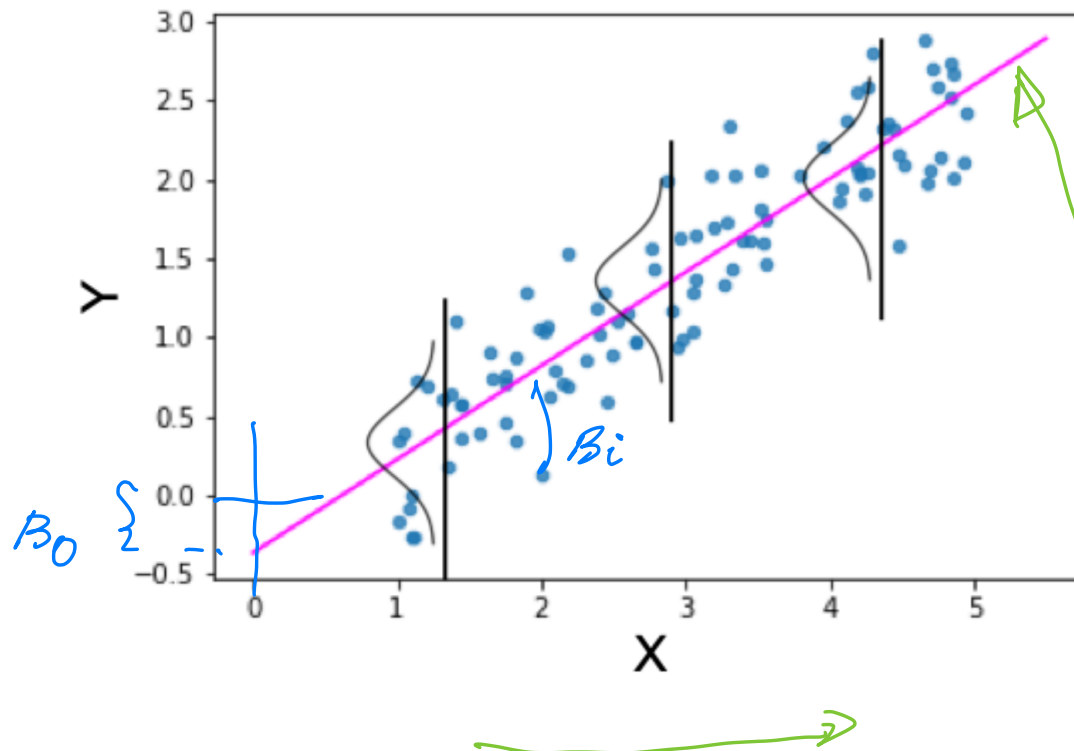
or $e_i = Y_i - \sum \beta_j X_{ij}$ donc chercher les β_j

revient à
 $\arg \min_{\beta_j}$

$$\sum_{i=1}^n \left(Y_i - \sum_{j=1}^p \beta_j X_{ij} \right)^2$$

Critère des
moindres carrés

La figure suivante, tirée de [6], rend compte de la partie aléatoire du modèle



1 seule variable X

$$Y_i = B_0 + B_1 X_i + e$$

$$B_0 + B_1 X$$

$$\hat{Y}_i = B_0 + B_1 X_i$$

$$= \mu_i$$

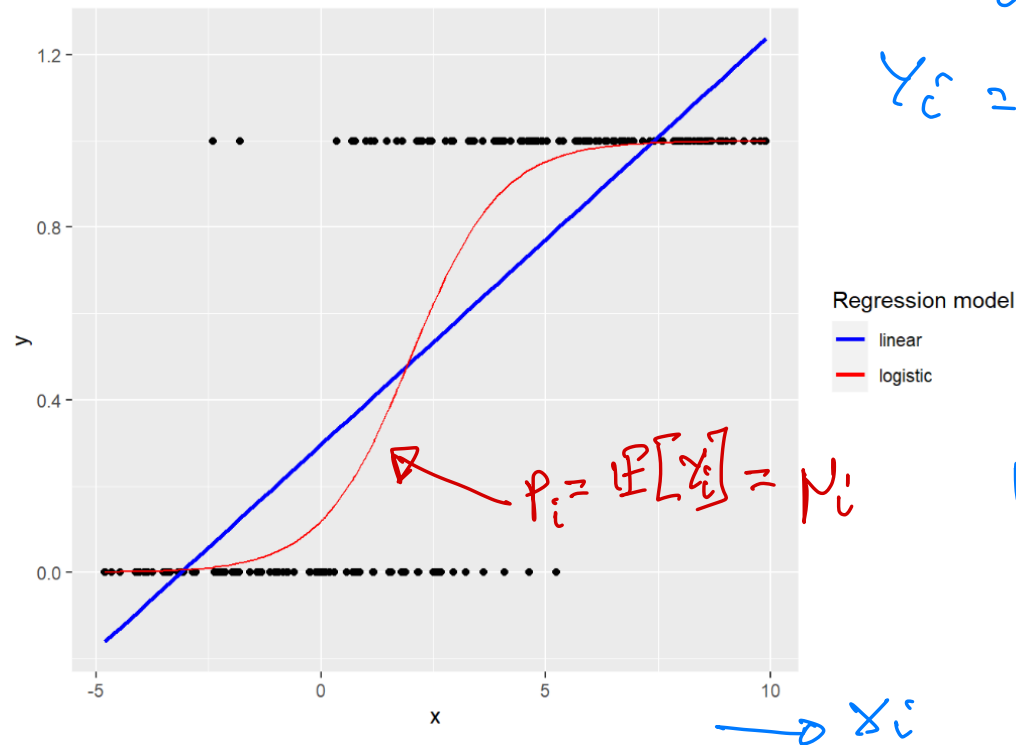
$$= \text{moyenne}$$

$$\text{d'une gaussienne}$$

$$(\text{dépendant de } i)$$

1.2. Régression logistique

La figure suivante, extraite de [1], compare une régression linéaire et une régression logistique



On modélise un Y_i qualitatif
 $Y_i = \begin{cases} 1 & \text{si classe 1 avec une proba } p \\ 0 & \text{si classe 0 avec } (1-p) \end{cases}$

modèle linéaire

$$\hat{Y}_i = B_0 + B_1 X_i$$

Modèle de Bernoulli

$$\begin{aligned} E[Y] &= p = \sum p_j Y_j \\ &= \sum 0 \times (1-p) + 1 \times p \end{aligned}$$

Y_i est modélisé comme une variable de Bernoulli, et on recherche le paramètre de cette loi de Bernoulli selon

$$E[Y_i | X_{ij}, j = 1..p] = p_i = \mu_i$$

Idee: modéliser
non pas Y_i , mais la
moyenne de Y_i

On modélise la moyenne

$$\mu_i = \mathbb{E}[Y_i] = g^{-1}(\sum \beta_j X_{ij}) = g^{-1}(\eta_i)$$

$$\mu_i = g^{-1}(\text{prédicteur linéaire})$$

On introduit la fonction de lien comme

$$\sum \beta_j X_{ij} = \eta_i = g(\mu_i) = \log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^p \beta_j X_{ij}$$

Dès lors, la régression logistique consiste à rechercher la moyenne d'une variable de Bernoulli, p_i ,

comme transformation $g^{-1}(\cdot)$ d'une combinaison linéaire des variables explicatives.

Dans le cas binaire, d'autres fonctions de lien sont possibles et entrent dans le champ des GLM.

Ici g est la fonction logistique et $g^{-1}(\cdot)$ est la sigmoïde.

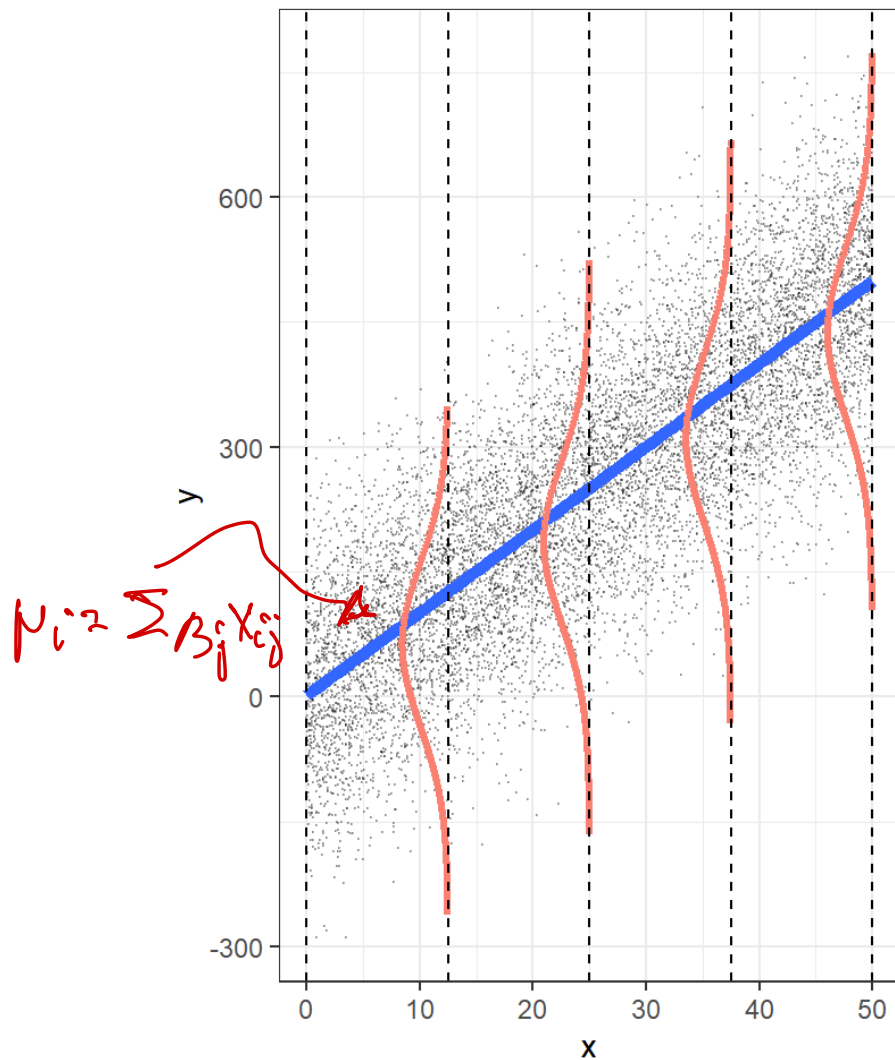
$\Rightarrow \mu_i$ est la sigmoïde du régresseur linéaire $\eta_i = \sum \beta_j X_{ij}$

1.3. Régression de Poisson

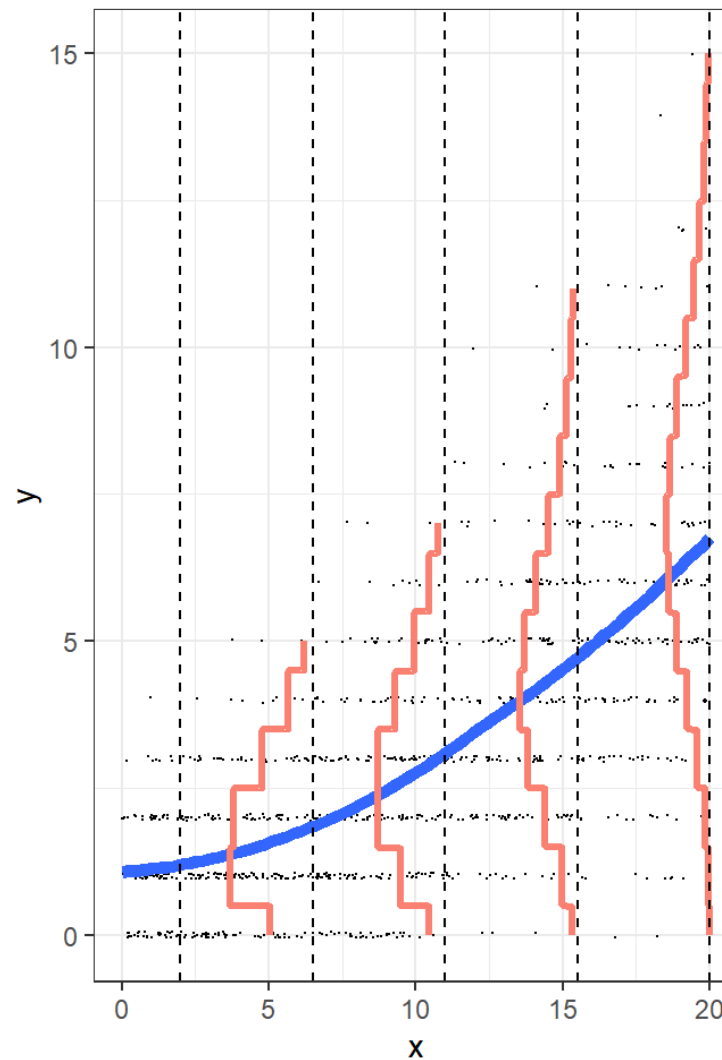
La figure ci-dessous, extraite de [1, chapitre 4], compare deux situations dans lesquelles un modèle gaussien et un modèle de Poisson sont pertinents.

Sur le panneau de gauche, on note (i) que la variable Y est discrète, et (ii) que les moyenne et variance dépendent de la variable dépendante X . Enfin, la relation entre Y et X ne semble pas linéaire mais ressemblerait plutôt à une exponentielle.

La question : comment choisir la fonction $g(\cdot)$?



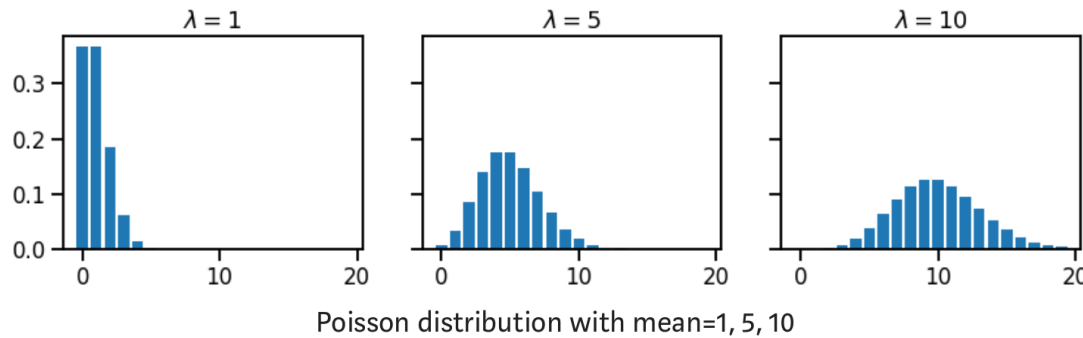
Exemple gaussien



Modèle de Poisson ?

- observations à valeurs discrètes (pour les entiers)
 - moyenne pas linéaire en fonction de x
 - variance augmente avec x .
- $\mu_i = e^{\beta_0 + \beta_1 x}$

Loi de Poisson dépend d'un paramètre λ



$$P(X=x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

et $E[X] = \lambda$

$$\text{var}[X] = \lambda$$

Dans ce cadre, on peut utiliser le modèle suivant. Y_i est une variable de Poisson, et on recherche la moyenne de cette loi de Poisson (qui est aussi son paramètre) selon

$$E[Y_i | X_{ij}, j = 1..p] = \lambda_i = \mu_i = \text{fonction de } x$$

On utilise ensuite la fonction de lien comme

$$\eta_i = g(\mu_i) = \log(\lambda_i) = \sum_{j=1}^p \beta_j X_{ij}$$

$$\lambda_i = \mu_i = e^{\sum_j \beta_j X_{ij}}$$

Dès lors, la régression de Poisson consiste à rechercher la moyenne d'une variable de Poisson, λ_i , comme transformation $g^{-1}(\cdot)$ d'une combinaison linéaire des variables explicatives. Ce type de régression entre dans la famille de régressions log-linéaires.

1.4. Principe général des GLM

Comme on l'a vu au travers des 3 exemples précédents, les GLM étendent la régression linéaire classique de manière à prendre en compte des distributions de probabilité non gaussiennes pour la réponse Y , et des transformations non-linéaires de la moyenne.

On pose explicitement que ce qui est recherché c'est un modèle de la moyenne $E[Y]$ de la distribution, et que ce modèle est une certaine fonction $g(.)^{-1}$ d'un prédicteur linéaire.

Le GLM présente trois composantes.

1. **Une composante aléatoire**, qui spécifie la distribution de probabilité de la réponse y . Les différentes observations y_1, y_2, \dots, y_n sont supposées indépendantes. Pour les GLMs, la distribution est choisie comme un membre d'une *famille exponentielle*.

2. **Un prédicteur linéaire** $\eta_i = \sum_{j=1}^p \beta_j X_{ij}$, soit $\eta = X\beta$ sous forme compacte. Le caractère linéaire indique que la transformation est linéaire par rapport aux paramètres β_j . Les variables elles-mêmes peuvent être des transformations non-linéaires de variables sous-jacentes ou présenter des interactions.

3. **Une fonction de lien** qui associe la moyenne μ_i au prédicteur linéaire selon

$\eta_i = g(E[Y_i]) = g(\mu_i)$. Dans la famille exponentielle, lorsque qu'on choisit la fonction de lien tel que le paramètre naturel θ_i soit égal au prédicteur linéaire, alors on parle de *lien*

canonique $\eta_i = g(E[Y_i]) = g(\mu_i) = \theta_i = \sum_{j=1}^p \beta_j X_{ij}$.

1.5. Autres distributions et fonctions de liens possibles

D'autres modèles de distributions et de liens possibles (sans que ce soit exhaustif) sont présentés dans ces deux tableaux issus de [7]

Distribution	Range	Skewed
Gaussian	$(-\infty, +\infty)$	No
Binomial	$\{0,1\}$	NA
Gamma	$(0, +\infty)$	Yes
Inverse Gaussian	$(0, +\infty)$	Yes
Poisson	$\{0,1,2,3,4,5\}$	Yes

Name	Link Function	Mean	Range of Mean
Identity	$z = \mu$	$\mu = z$	$(-\infty, +\infty)$
Log	$z = \log(\mu)$	$\mu = \exp(z)$	$(0, +\infty)$
Inverse	$z = 1/\mu$	$\mu = \frac{1}{z}$	$(-\infty, +\infty)$
Inverse Squared	$z = 1/\mu^2$	$\mu = \frac{1}{\sqrt{z}}$	$(0, +\infty)$
Square root	$z = \sqrt{\mu}$	$\mu = z^2$	$(0, +\infty)$

[<https://bookdown.org/roback/bookdown-BeyondMLR/ch-glms.html#generalized-linear-modeling>,
<https://sdcastillo.github.io/PA-R-Study-Manual/generalized-linear-models-glms.html>]

2. Famille exponentielle

2.1. Définition

On appelle famille exponentielle l'ensemble des distributions de probabilité qui peuvent se mettre sous la forme suivante

$$\rightarrow f(y_i; \theta_i, \phi) = \exp\{[y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\}. \quad (4.1)$$

This is called the *exponential dispersion family*. The parameter θ_i is called the *natural parameter*, and ϕ is called the *dispersion parameter*. Often $a(\phi) = 1$ and $c(y_i, \phi) = c(y_i)$, giving the *natural exponential family* of the form $f(y_i; \theta_i) = h(y_i) \exp[y_i \theta_i - b(\theta_i)]$. Otherwise, usually $a(\phi)$ has the form $a(\phi) = \phi$ or $a(\phi) = \phi/\omega_i$ for $\phi > 0$ and a known weight ω_i . For instance, when y_i is a mean of n_i independent readings, $\omega_i = n_i$.

$$f(y_i; \theta_i, \phi) = e^{\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)}$$

paramètre naturel

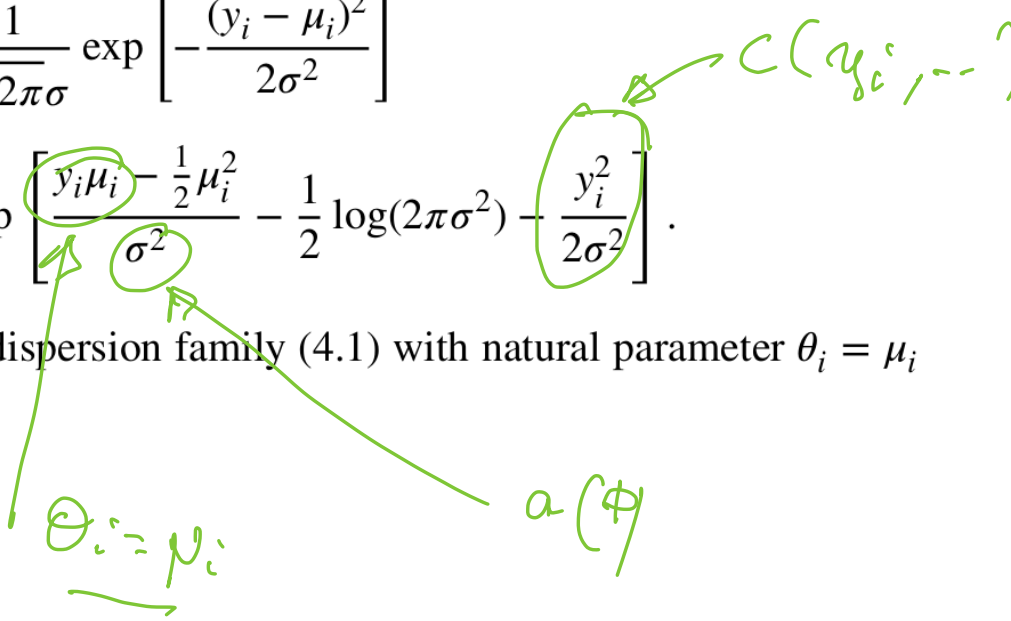
précision

produit variable par paramètre naturel

2.2. Exemples

2.2.1. Gaussienne

For the normal distribution, observation i has probability density function

$$\begin{aligned} f(y_i; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(y_i - \mu_i)^2}{2\sigma^2} \right] \\ &= \exp \left[\frac{y_i \mu_i - \frac{1}{2} \mu_i^2}{\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) - \frac{y_i^2}{2\sigma^2} \right]. \end{aligned}$$


This satisfies the exponential dispersion family (4.1) with natural parameter $\theta_i = \mu_i$

2.2.2. Poisson

When y_i has a Poisson distribution, the probability mass function is

$$\begin{aligned} f(y_i; \mu_i) &= \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \exp[\underbrace{y_i \log \mu_i}_{\theta_i} - \underbrace{\mu_i}_{b(\theta_i)} - \underbrace{\log(y_i!)}_{c(\phi)}] \\ &= \exp[y_i \theta_i - \exp(\theta_i) - \log(y_i!)], \quad y_i = 0, 1, 2, \dots, \end{aligned} \quad (4.5)$$

where the natural parameter $\theta_i = \log \mu_i$. This has exponential dispersion form (4.1) with $b(\theta_i) = \exp(\theta_i)$, $a(\phi) = 1$, and $c(y_i, \phi) = -\log(y_i!)$. By (4.3) and (4.4),

$$\theta_i = \text{paramètre naturel} = \log(\mu_i) \\ (= \text{fonction de lien ?})$$

2.2.3. Binomiale

Next, suppose that $n_i y_i$ has a $\text{bin}(n_i, \pi_i)$ distribution; that is, here y_i is the sample *proportion* (rather than *number*) of successes, so $E(y_i) = \pi_i$ does not depend on n_i . Let $\theta_i = \log[\pi_i/(1 - \pi_i)]$. Then $\pi_i = \exp(\theta_i)/[1 + \exp(\theta_i)]$ and $\log(1 - \pi_i) = -\log[1 + \exp(\theta_i)]$. We can express

$$\begin{aligned} f(y_i; \pi_i, n_i) &= \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1 - \pi_i)^{n_i - n_i y_i}, \quad y_i = 0, \frac{1}{n_i}, \frac{2}{n_i}, \dots, 1, \\ &= \exp \left[\frac{y_i \theta_i - \log[1 + \exp(\theta_i)]}{1/n_i} + \log \left(\binom{n_i}{n_i y_i} \right) \right]. \end{aligned} \quad (4.6)$$

This has exponential dispersion form (4.1) with $b(\theta_i) = \log[1 + \exp(\theta_i)]$, $a(\phi) = 1/n_i$, and $c(y_i, \phi) = \log \left(\binom{n_i}{n_i y_i} \right)$. The natural parameter is $\theta_i = \log[\pi_i/(1 - \pi_i)]$, the *logit*.

$$\binom{n}{p} \pi^p (1-\pi)^{n-p}$$

$$ny = p$$

$$y = \frac{p}{n}$$

proportion
de succès.

$$\theta_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right) = \text{fonction logit}$$

$$\theta_i = \log \left(\frac{\mu_i}{1 - \mu_i} \right) = \text{fonction de lien ?}$$

fonction de lien : transformation paramètre naturel
et moyenne.

La question restant à traiter :
comment trouver les $\{\beta_j\}$ à partir
de $g(\mu_i) = \sum \beta_j x_{ij}$

2.3. Résultats

Plusieurs propriétés importantes caractérisent la famille exponentielle. La log-vraisemblance s'écrit

log-vraisemb : $L_i = \underbrace{[y_i \theta_i - b(\theta_i)] / a(\phi) + c(y_i, \phi)}$,

→ $\partial L_i / \partial \theta_i = [y_i - b'(\theta_i)] / a(\phi), \quad \partial^2 L_i / \partial \theta_i^2 = -b''(\theta_i) / a(\phi),$

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

= = ↑ =
 ?

On peut en déduire, ou par calcul direct, que

1. $\mu_i = E[y_i] = b'(\theta_i)$
2. $\text{var}[y_i] = b''(\theta_i) a(\phi)$

$$1) \int e^{\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)} dy_i = 1$$

$$e^{-\frac{b(\theta_i)}{a}} = \int e^{\frac{y_i \theta_i}{a} + c} dy_i$$

$$b(\theta_i) = -a \log \int e^{\frac{y_i \theta_i}{a} + c} dy_i$$

$$b'(\theta_i) = \frac{-a \int y_i e^{\frac{y_i \theta_i}{a} + c} dy_i}{\int e^{\frac{y_i \theta_i}{a} + c} dy_i} = E[y_i]$$

2.4. Canonical link

Comme vu précédemment, lorsque dans la famille exponentielle on choisit la fonction de lien telle que le paramètre naturel θ_i soit égal au prédicteur linéaire, alors on parle de lien canonique

$\eta_i = g(E[Y_i]) = g(\mu_i) = \theta_i = \sum_{j=1}^p \beta_j X_{ij}$. La table ci-dessous donne les liens canoniques pour les

distributions que nous avons rencontrées. Certaines simplifications ou “bonnes propriétés” apparaissent lorsqu’on utilise le lien canonique, mais il faut noter qu’il est cependant tout à fait possible d’utiliser d’autres types de liens !

Density	Link: $\eta = g(\mu)$	Name
Normal	$\eta = \mu$	identity
Poisson	$\eta = \log(\mu)$	log
Binomial	$\eta = \log[\mu/(1 - \mu)]$	logit
Gamma	$\eta = 1/\mu$	inverse
Inverse Gauss	$\eta = 1/\mu^2$	1/mu^2

2.5. Table

Table 2.1 *Characteristics of some common univariate distributions in the exponential family[†]*

	<i>Normal</i>	<i>Poisson</i>	<i>Binomial</i>	<i>Gamma</i>	<i>Inverse Gaussian</i>
<i>Notation</i>	$N(\mu, \sigma^2)$	$P(\mu)$	$B(m, \pi)/m$	$G(\mu, \nu)$	$IG(\mu, \sigma^2)$
<i>Range of y</i>	$(-\infty, \infty)$	$0(1)\infty$	$\frac{0(1)m}{m}$	$(0, \infty)$	$(0, \infty)$
<i>Dispersion parameter: ϕ</i>	$\phi = \sigma^2$	1	$1/m$	$\phi = \nu^{-1}$	$\phi = \sigma^2$
<i>Cumulant function: $b(\theta)$</i>	$\theta^2/2$	$\exp(\theta)$	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
<i>$c(y; \phi)$</i>	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)$	$-\log y!$	$\log\left(\frac{m}{my}\right)$	$\nu \log(\nu y) - \log y - \log \Gamma(\nu)$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$
<i>$\mu(\theta) = E(Y; \theta)$</i>	θ	$\exp(\theta)$	$e^\theta/(1 + e^\theta)$	$-1/\theta$	$(-2\theta)^{-1/2}$
<i>Canonical link: $\theta(\mu)$</i>	identity	log	logit	reciprocal	$1/\mu^2$
<i>Variance function: $V(\mu)$</i>	1	μ	$\mu(1 - \mu)$	μ^2	μ^3

[†]The mean-value parameter is denoted by μ , or by π for the binomial distribution.

The parameterization of the gamma distribution is such that its variance is μ^2/ν .

The canonical parameter, denoted by θ , is defined by (2.4). The relationship between μ and θ is given in lines 6 and 7 of the Table.

2.7. Maximum de vraisemblance pour les GLM

2.7.1. Équations de vraisemblance

Pour la famille exponentielle, la vraisemblance s'écrit

$$L(\boldsymbol{\beta}) = \sum_{i=1}^n L_i = \sum_{i=1}^n \log f(y_i; \theta_i, \phi) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi).$$

et pour le *lien canonique*, on a $\theta_i = \sum_j \beta_j x_{ij} \approx g(\mu_i)$

Pour déterminer les paramètres au sens du maximum de vraisemblance, on doit calculer

$$\partial L(\boldsymbol{\beta}) / \partial \beta_j = \sum_{i=1}^n \partial L_i / \partial \beta_j = 0, \quad \text{for all } j.$$

Étant donné que

$$\partial L_i / \partial \theta_i = (y_i - \mu_i) / a(\phi), \quad \partial \mu_i / \partial \theta_i = b''(\theta_i) = \text{var}(y_i) / a(\phi).$$

- $\mu_i = b'(\theta_i)$

- $\text{var}(y_i) = b''(\theta_i)a(\phi)$

- $\eta_i = \sum_{j=1}^p \beta_j x_{ij}, \quad \partial \eta_i / \partial \beta_j = x_{ij}$

- $\eta_i = g(\mu_i)$

alors

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

$$= \frac{(y_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i}.$$

NB:

$$\frac{d\mu_i}{d\eta_i} = g'(\mu_i)$$

$$\frac{d\mu_i}{du_i} = \frac{1}{g'(p_i)} = \frac{1}{g'(g^{-1}(u_i))}$$

et on obtient au final, après sommation sur les n exemples

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\text{var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, 2, \dots, p,$$

avec

$$\eta_i = \sum_{j=1}^p \beta_j x_{ij} = g(\mu_i)$$

ce qui peut encore se mettre sous la forme compacte

$$X^T D V^{-1} (y - \mu) = 0.$$

Il s'agit d'un ensemble d'équations non linéaires qui doivent être résolues de manière numérique et itérative.

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

$$\frac{d\eta_i}{d\mu_i} = g'(\mu_i)$$

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

2.7.3. Incertitude et intervalle de confiance

Sous des conditions larges, le maximum de vraisemblance est asymptotiquement non biaisé et tend vers une loi normale. Ici,

Asymptotic distribution of $\hat{\beta}$ for GLM $\eta = X\beta$:

$\hat{\beta}$ has an approximate $N[\beta, (X^T W X)^{-1}]$ distribution,

where W is the diagonal matrix with elements $w_i = (\partial \mu_i / \partial \eta_i)^2 / \text{var}(y_i)$.

et on pourra évaluer la matrice de covariance en calculant W au point $\hat{\beta}$.

Par suite, on obtient la matrice de covariance des valeurs estimées

$$\text{var}(\hat{\eta}) = X \text{var}(\hat{\beta}) X^T \approx X (X^T W X)^{-1} X^T.$$

La méthode delta, qui consiste à linéariser la fonction g autour du point d'intérêt permet d'évaluer

$$\text{var}(\hat{\mu}) \approx D \text{var}(\hat{\eta}) D \approx D X (X^T W X)^{-1} X^T D.$$

où D est une matrice diagonale d'éléments $\frac{\partial \mu_i}{\partial \eta_i}$. Il peut-être plus simple de construire un intervalle de confiance pour η_i et d'appliquer $g^{-1}()$ aux bornes de l'intervalle.

2.7.4. Algorithmes de recherche du maximum de vraisemblance

2.7.4.1. Newton-Raphson

Mathematically, here is how the Newton–Raphson method determines the value $\hat{\beta}$ at which a function $L(\beta)$ is maximized. Let

$$\mathbf{u} = \left(\frac{\partial L(\beta)}{\partial \beta_1}, \frac{\partial L(\beta)}{\partial \beta_2}, \dots, \frac{\partial L(\beta)}{\partial \beta_p} \right)^T.$$

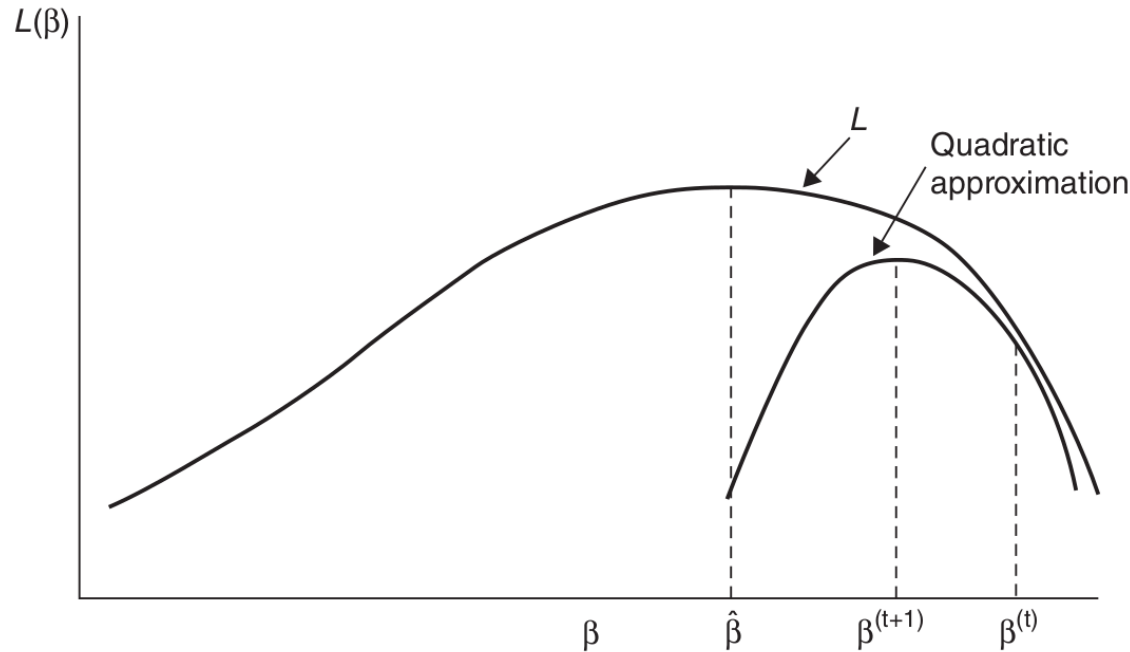
Let \mathbf{H} denote¹⁵ the matrix having entries $h_{ab} = \partial^2 L(\beta) / \partial \beta_a \partial \beta_b$, called the *Hessian matrix*. Let $\mathbf{u}^{(t)}$ and $\mathbf{H}^{(t)}$ be \mathbf{u} and \mathbf{H} evaluated at $\beta^{(t)}$, approximation t for $\hat{\beta}$. Step t in the iterative process ($t = 0, 1, 2, \dots$) approximates $L(\beta)$ near $\beta^{(t)}$ by the terms up to the second order in its Taylor series expansion,

$$L(\beta) \approx L(\beta^{(t)}) + \mathbf{u}^{(t)T}(\beta - \beta^{(t)}) + \left(\frac{1}{2} \right) (\beta - \beta^{(t)})^T \mathbf{H}^{(t)} (\beta - \beta^{(t)}).$$

Solving $\partial L(\beta) / \partial \beta \approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)}(\beta - \beta^{(t)}) = \mathbf{0}$ for β yields the next approximation,

$$\beta^{(t+1)} = \beta^{(t)} - (\mathbf{H}^{(t)})^{-1} \mathbf{u}^{(t)}, \quad (4.23)$$

assuming that $\mathbf{H}^{(t)}$ is nonsingular.



2.7.4.2. Fisher scoring

Let $\mathcal{J}^{(t)}$ denote approximation t for the ML estimate of the expected information matrix; that is, $\mathcal{J}^{(t)}$ has elements $-E(\partial^2 L(\beta) / \partial \beta_a \partial \beta_b)$, evaluated at $\beta^{(t)}$. The formula for Fisher scoring is

$$\beta^{(t+1)} = \beta^{(t)} + (\mathcal{J}^{(t)})^{-1} u^{(t)}, \quad \text{or} \quad \mathcal{J}^{(t)} \beta^{(t+1)} = \mathcal{J}^{(t)} \beta^{(t)} + u^{(t)}. \quad (4.24)$$

2.7.4.3. Moindres carrés “reweighted” (repondérés)

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Solution :

$$(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{z}.$$

On itère

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \mathbf{z}^{(t)}.$$

$$\mathbf{W} = \text{diag}\{(\partial \mu_i / \partial \eta_i)^2 / \text{var}(y_i)\}$$

The vector $\mathbf{z}^{(t)}$ in this formulation is an estimated linearized form of the link function g , evaluated at \mathbf{y} ,

$$g(y_i) \approx g(\mu_i^{(t)}) + (y_i - \mu_i^{(t)}) g'(\mu_i^{(t)}) = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}} = z_i^{(t)}. \quad (4.25)$$

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

2.7.5. Déviance

[https://en.wikipedia.org/wiki/Deviance_\(statistics\)](https://en.wikipedia.org/wiki/Deviance_(statistics))

La déviance est une mesure de l'écart entre le modèle courant et un modèle "saturé" (n paramètres pour n observations) qui correspond au modèle où les données sont exactement satisfaites

$$-2 \log \left[\frac{\text{maximum likelihood for model}}{\text{maximum likelihood for saturated model}} \right] = -2[L(\hat{\mu}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})]$$

A partir de l'expression générale de la log-vraisemblance pour une famille exponentielle

$$L(\beta) = \sum_{i=1}^n L_i = \sum_{i=1}^n \log f(y_i; \theta_i, \phi) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi).$$

on a

$$\begin{aligned} & -2[L(\hat{\mu}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})] \\ & = 2 \sum_i [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) - 2 \sum_i [y_i \hat{\theta}_i - b(\hat{\theta}_i)]/a(\phi). \end{aligned}$$

Dans la mesure où $L(\mathbf{y}; \mathbf{y})$ ne dépend pas des paramètres, maximiser la vraisemblance est équivalent à minimiser la déviance.

2.7.5.1. Cas gaussien...

For normal GLMs, by Section 4.1.2, $\hat{\theta}_i = \hat{\mu}_i$ and $b(\hat{\theta}_i) = \hat{\theta}_i^2/2$. Similarly, $\tilde{\theta}_i = y_i$ and $b(\tilde{\theta}_i) = y_i^2/2$ for the saturated model. So the deviance equals

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_i \left[y_i(y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right] = \sum_i (y_i - \hat{\mu}_i)^2.$$

2.7.5.2. Cas Poisson...

For Poisson GLMs, from Section 4.1.2, $\hat{\theta}_i = \log \hat{\mu}_i$ and $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$. Similarly, $\tilde{\theta}_i = \log y_i$ and $b(\tilde{\theta}_i) = y_i$ for the saturated model. Also $a(\phi) = 1$, so the deviance and scaled deviance (4.15) equal

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_i [y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i].$$

2.7.5.3. Table des déviations pour quelques lois standard [3]

Normal	$\sum (y - \hat{\mu})^2,$
Poisson	$2 \sum \{y \log(y/\hat{\mu}) - (y - \hat{\mu})\},$
binomial	$2 \sum \{y \log(y/\hat{\mu}) + (m - y) \log[(m - y)/(m - \hat{\mu})]\},$
gamma	$2 \sum \{-\log(y/\hat{\mu}) + (y - \hat{\mu})/\hat{\mu}\},$
inverse Gaussian	$\sum (y - \hat{\mu})^2 / (\hat{\mu}^2 y).$

2.7.6. Tests

Deux questions peuvent se poser lorsqu'on calcule des modèles : (1) tous les coefficients sont-ils significatifs (2) comment comparer deux modèles.

(1) Lorsqu'on a estimé un jeu de coefficients, on peut former un test de Wald

$$Z = \frac{\hat{\beta} - \beta_0}{SE},$$

où β_0 est l'hypothèse nulle, et SE est l'écart type estimé à partir de la matrice d'information de

Fisher dans laquelle on a utilisé les estimées $\hat{\beta}$ courantes. On montre que z est approximativement une gaussienne standard. Dès lors, on peut calculer une p-value pour la valeur estimée.

(2) Quand on dispose de plusieurs modèles potentiels, qu'il s'agisse de jeux de variables explicatives différentes ou d'un modèle probabiliste différent, on peut s'appuyer sur la "*puissance de prédiction*", par exemple le R^2 ou le R^2 ajusté. On peut également utiliser une mesure informationnelle, le critère AIC d'Akaike, défini comme $AIC = -2[L(\hat{\beta}) - M]$, où $L(\hat{\beta})$ est la log-vraisemblance du modèle et M son nombre de paramètres. Il s'agit, parmi plusieurs modèles en compétition, de retenir celui qui présente l'AIC le plus faible. On note que le nombre de paramètres introduit une pénalisation des modèles avec des nombres de paramètres élevés.

3. Implémentations

3.1. En R

```
> mice.glm <- glm(formula = resp ~ conc,  
+                 family = binomial(link = logit),  
+                 weights = NULL,  
+                 data = mice  
+                 )
```

$$y \sim x_1 + x_2 + x_3 * x_4$$

- **formula**; as in general linear models
- **family**
 - `binomial`(link = `logit` | `probit` | `cauchit` | `log` | `cloglog`)
 - `gaussian`(link = `identity` | `log` | `inverse`)
 - `Gamma`(link = `inverse` | `identity` | `log`)
 - `inverse.gaussian`(link = `1/mu^2` | `inverse` | `identity` | `log`)
 - `poisson`(link = `log` | `identity` | `sqrt`)
 - `quasi`(link = `...` , variance = `...`))
 - `quasibinomial`(link = `logit` | `probit` | `cauchit` | `log` | `cloglog`)
 - `quasipoisson`(link = `log` | `identity` | `sqrt`)

```
##
## Call:
## glm(formula = target ~ AGE + GENDER + MARRIED + CAR_USE + BLUEBOOK +
##      CAR_TYPE + AREA, family = binomial(link = "logit"), data = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8431  -0.8077  -0.5331   0.9575   3.0441
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.523e-01  2.517e-01  -1.400  0.16160
## AGE            -2.289e-02  3.223e-03  -7.102 1.23e-12 ***
## GENDERM        -1.124e-02  9.304e-02  -0.121  0.90383
## MARRIEDYes      -6.028e-01  5.445e-02 -11.071 < 2e-16 ***
## CAR_USEPrivate -1.008e+00  6.569e-02 -15.350 < 2e-16 ***
## BLUEBOOK       -4.025e-05  4.699e-06  -8.564 < 2e-16 ***
## CAR_TYPEPickup -6.687e-02  1.390e-01  -0.481  0.63048
## CAR_TYPESedan  -3.689e-01  1.383e-01  -2.667  0.00765 **
## CAR_TYPESports Car 6.159e-01  1.891e-01   3.256  0.00113 **
## CAR_TYPESUV     2.982e-01  1.772e-01   1.683  0.09240 .
## CAR_TYPEVan     -8.983e-03  1.319e-01  -0.068  0.94569
## AREAUrban       2.128e+00  1.064e-01  19.993 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 9544.3  on 8236  degrees of freedom
## Residual deviance: 8309.6  on 8225  degrees of freedom
## AIC: 8333.6
##
## Number of Fisher Scoring iterations: 5
```

Documentation

glm R → <https://www.statmethods.net/advstats/glm.html>

les familles et liens → <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/family>

3.2. En Python

Python <https://www.statsmodels.org/stable/glm.html#technical-documentation>

(exemple) <https://www.statsmodels.org/stable/examples/notebooks/generated/glm.html>

4. Bibliographie

- [1] P. R. and J. Legler, *Beyond Multiple Linear Regression*.
<https://bookdown.org/roback/bookdown-BeyondMLR>, 2020.
- [2] A. Agresti, *Foundations of Linear and Generalized Linear Models*, 1st edition. Hoboken, New Jersey: John Wiley & Sons Inc, 2015.
- [3] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd edition. Boca Raton: Chapman and Hall/CRC, 1989.
- [4] J. A. Nelder and R. W. M. Wedderburn, 'Generalized Linear Models', *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972, doi: [10.2307/2344614](https://doi.org/10.2307/2344614).
- [5] Madsen, Henrik; Thyregod, Poul (2011). *Introduction to General and Generalized Linear Models*. Chapman & Hall/CRC. ISBN 978-1-4200-9155-7.
- [6] Y. Kida, 'Generalized linear models', *Medium*, Sep. 24, 2019.
<https://towardsdatascience.com/generalized-linear-models-9cbf848bb8ab> (accessed Nov. 21, 2020).
- [7] *10 Generalized linear Models (GLMs) | Exam PA Study Guide, Fall 2020*.
<https://sdcastillo.github.io/PA-R-Study-Manual/generalized-linear-models-glms.html>, (accessed Nov. 21, 2020).