Modèles linéaires généralisés

Jean Bérard
jberard@unistra.fr

DUAS 2ème année

La régression

Objectif : modéliser la valeur d'une variable y en fonction des valeurs de m variables $x^{(1)}, \ldots, x^{(m)}$.

- y est la variable « réponse » (« expliquée »)
- $x^{(1)}, \ldots, x^{(m)}$ sont les « co-variables » (« explicatives »)

Exemple classique

$y =$ nombre/coût/survenance d'un sinistre futur
$x^{(1)}, \ldots, x^{(m)} =$ caractéristiques de la police (âge de l'assuré, zone géographique,...)

Nuances entre modéliser pour :

- prédire
- expliquer (interpréter, agir)
- décrire/explorer

La régression

On modélise en général les valeurs de y et $x = \left(x^{(1)}, \ldots, x^{(m)}\right)$ à l'aide d'un couple de variables aléatoires (Y, X), avec $X = \left(X^{(1)}, \ldots, X^{(m)}\right)$.

Dans ce cadre, la régression vise notamment à déterminer :

(1) l'espérance conditionnelle $\mathbb{E}(Y|X = x)$

(2) la variance conditionnelle $\mathbb{V}(Y|X = x)$

(3) la loi conditionnelle $\text{Loi}(Y|X = x)$

Bien entendu, l'objectif (3) contient en particulier les objectifs (1) et (2). On verra dans la suite que les efforts de modélisation portent souvent d'abord sur (1), les objectifs (2) et (3) étant (éventuellement) examinés dans un second temps.

La régression

La modélisation est généralement basée sur un jeu de données constitué d'observations conjointes des valeurs de y et de $x = \left(x^{(1)}, \ldots, x^{(m)}\right)$:

$$\left[y_i, x_i = \left(x_i^{(1)}, \ldots, x_i^{(m)}\right)\right]_{i=1,\ldots,N}$$

Il faut alors expliciter le lien entre le modèle que l'on veut construire pour (Y, X) et le jeu de données.

Hypothèse habituelle

On modélise les réponses $(y_i)_{i=1,\ldots,N}$ comme étant issues d'une suite de v.a. indépendantes $Y_1, \ldots, Y_N$ telles que $Y_i \sim \text{Loi}(Y|X = x_i)$.

L'hypothèse ci-dessus :

- ne comporte pas de modélisation explicite de la loi de X
- n'incorpore pas la dépendance que l'on s'attend *a priori* à trouver dans le cas de données structurées (ex. : suivi de la sinistralité d'une police sur une période de plusieurs années) si cette structure n'est pas reflétée dans les variables explicatives
- suppose que la liaison statistique modélisée entre réponse et covariables est identique dans le jeu de données et dans le couple (Y, X)

Méthodes de régression

Il existe de nombreuses méthodes de régression, correspondant à des hypothèses variées sur la loi de Y sachant $X = x$:

- modèles linéaires
- modèles linéaires généralisés
- modèles additifs généralisés
- lissage (noyaux, splines, régression locale)
- arbres de régression
- réseaux de neurones
- etc.

Tâches-clés

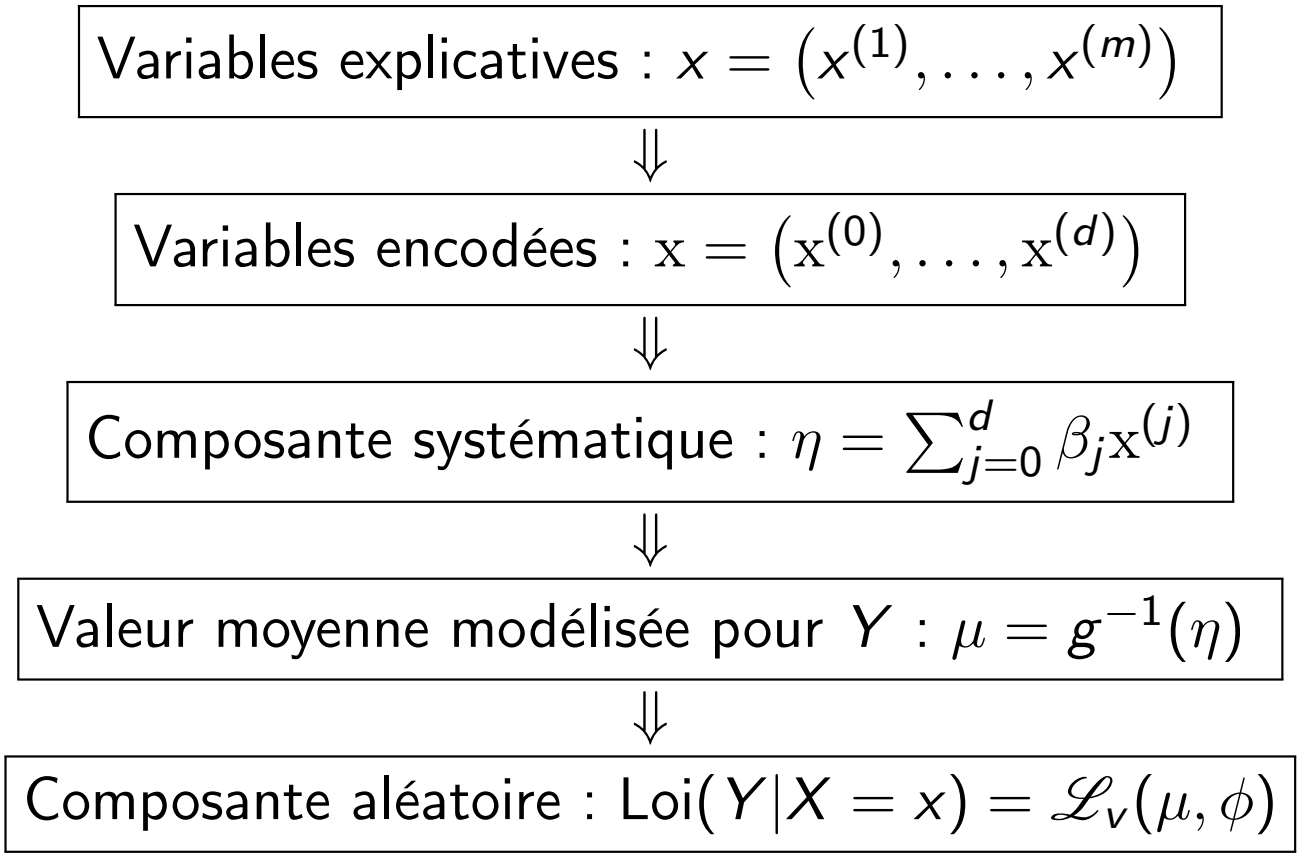
- Ajuster un modèle de régression aux données
- Effectuer des calculs à l'aide d'un modèle
- Evaluer la qualité d'un modèle
- Comparer des modèles entre eux

Structure d'un GLM

Un modèle linéaire généralisé est constitué de trois composantes essentielles :

- une **composante systématique** : fonction affine (d'un encodage) des variables explicatives $x = \left(x^{(1)}, \ldots, x^{(m)}\right)$;
- une **composante aléatoire** : spécification du type de Loi$(Y|X = x)$, au sein de la famille exponentielle ;
- une **fonction de lien** : spécification de la relation entre $\mathbb{E}(Y|X = x)$ et la composante systématique du modèle.

Structure d'un GLM



Encodage

Coefficients β_j

Fonction de lien g

Famille expo.

Famille exponentielle

La présentation donnée ci-après est relativement informelle. Pour un traitement encyclopédique de la question, voir [Jör97] ou le résumé donné au chapitre 2 de [Son07].

Famille exponentielle

Famille exponentielle

Une loi de probabilité sur $\mathbb{R}$ (discrète ou continue) appartient à la famille exponentielle si elle possède une densité de la forme :

$$f(y) = c(y, \phi) \cdot \exp\left(\frac{y\theta - b(\theta)}{\phi}\right),$$

où $\phi > 0$, $\theta \in I$ et $b : I \subset \mathbb{R} \to \mathbb{R}$ (supposée régulière sur l'intervalle I).

Rappel : Y suit la loi de densité f signifie

- dans le cas discret : $\mathbb{P}(Y = y) = f(y)$
- dans le cas continu : $\mathbb{P}(Y \in [s, t]) = \int_s^t f(y)dy$

Terminologie :

- θ est le « paramètre naturel »
- ϕ est le « paramètre de dispersion »

Partant de $b(\cdot)$, θ et ϕ, on obtient une densité ayant exactement la même forme en considérant $\tilde{b}(\cdot)$, $\tilde{\theta}$ et $\tilde{\phi}$ définis par : $\tilde{\theta} = \gamma\theta + \rho$, $\tilde{b}(\tilde{\theta}) = \gamma b((\tilde{\theta} - \rho)/\gamma) + \delta$, $\tilde{\phi} = \gamma\phi$. On vérifie que, génériquement, le choix de $b(\cdot)$, θ et ϕ est unique à une transformation de ce type près.

Famille exponentielle

Exemples classiques :

- loi normale $\mathcal{N}(m, \sigma^2)$
- loi de Poisson $\mathcal{P}(\lambda)$
- loi Gamma $\mathcal{G}amma(k, \alpha)$
- loi binomiale $\mathcal{B}(n, p)$

Exemples moins classiques :

- loi binomiale négative
- loi gaussienne inverse
- loi de Tweedie

Famille exponentielle

Expression pour l'espérance et la variance

Si Y suit la loi associée à $b(\cdot)$, θ et ϕ, on a :

$$\begin{cases} \mathbb{E}(Y) = b'(\theta) \\ \mathbb{V}(Y) = b''(\theta)\phi \end{cases}$$

- l'espérance ne fait intervenir que $b(\cdot)$ et θ
- le paramètre ϕ a un effet multiplicatif sur la variance

Famille exponentielle

On peut reparamétrer le modèle en fonction de $\mu = \mathbb{E}(Y)$, en posant :

- $\theta = (b')^{-1}(\mu)$
- $\mathbb{V}(Y) = \phi v(\mu)$, où $v(\mu) \overset{\text{déf.}}{=} b''((b')^{-1}(\mu))$

Les trois éléments (μ, ϕ) et $v(\cdot)$ caractérisent entièrement la loi de Y, que l'on note alors $\mathscr{L}_v(\mu, \phi)$. La fonction $v(\cdot)$ est appelée « fonction de variance ». Dans ce cadre, on a donc :

Expression pour l'espérance et la variance

Si Y suit la loi associée à (μ, ϕ) et $v(\cdot)$, c.à.d. $Y \sim \mathscr{L}_v(\mu, \phi)$, on a :

$$\begin{cases} \mathbb{E}(Y) = \mu \\ \mathbb{V}(Y) = \phi v(\mu) \end{cases}$$

Seules certaines fonctions $v(\cdot)$ sont effectivement associées à une loi de la famille exponentielle en tant que fonction de variance. De plus, il existe alors certaines restrictions sur les valeurs possibles de μ et ϕ. On ne peut pas choisir $v(\cdot)$ (et μ et ϕ) de manière totalement arbitraire !

Famille exponentielle : loi normale $\mathcal{N}(m, \sigma^2)$

On suppose que $Y \sim \mathcal{N}(m, \sigma^2)$. Alors la loi de Y se rattache à la famille exponentielle, avec les caractéristiques suivantes :

Nom	Normal
Modèle	$Y \sim \mathcal{N}(m, \sigma^2)$
μ	m
ϕ	σ^2
$v(\mu)$	1
θ	μ
$b(\theta)$	$\theta^2/2$

Famille exponentielle : loi de Poisson $\mathcal{P}(\lambda)$

On suppose que $Y \sim \mathcal{P}(\lambda)$. Alors la loi de Y se rattache à la famille exponentielle, avec les caractéristiques suivantes :

Nom	Poisson
Modèle	$Y \sim \mathcal{P}(\lambda)$
μ	λ
ϕ	1
$v(\mu)$	μ
θ	$\log(\mu)$
$b(\theta)$	e^{θ}

Famille exponentielle : loi $\mathcal{G}amma(\text{forme} = k, \text{éch.} = \alpha)$

On suppose que $Y \sim \mathcal{G}amma(\text{forme} = k, \text{éch.} = \alpha)$. Alors la loi de Y se rattache à la famille exponentielle, avec les caractéristiques suivantes :

Nom	Gamma
Modèle	$Y \sim \mathcal{G}amma(\text{forme} = k, \text{éch.} = \alpha)$
μ	$k\alpha$
ϕ	$1/k$
$v(\mu)$	μ^2
θ	$-1/\mu$
$b(\theta)$	$-\log(-\theta)$

Famille exponentielle : loi binomiale $\mathcal{B}(n, p)$

On suppose que $Z \sim \mathcal{B}(n, p)$, et l'on pose $Y = Z/n$. Alors la loi de Y se rattache à la famille exponentielle, avec les caractéristiques suivantes :

Nom	Binomial
Modèle	$nY \sim \mathcal{B}(n, p)$
μ	p
ϕ	$1/n$
$v(\mu)$	$\mu(1-\mu)$
θ	$\log(\mu/(1-\mu))$
$b(\theta)$	$\log(1+e^{\theta})$

Remarque : On peut directement écrire la loi de Z sous la forme $\mathcal{L}_v(\mu, \phi)$ en modifiant convenablement les paramètres, mais la forme ci-dessus est préférée.

Famille exponentielle : quelques exemples moins classiques

En annexe, on donne la description de trois lois un peu moins classiques appartenant à la famille exponentielle :

- loi binomiale négative (extension de la loi de Poisson pouvant par exemple être utilisée pour modéliser une sur-dispersion)
- loi gaussienne inverse (alternative possible à la loi Gamma pour modéliser des réponses continues positives)
- loi de Tweedie (modélise directement une loi de Poisson composée avec une loi Gamma, pouvant être utilisée pour un modèle de charge de sinistre sans passer par deux modèles séparés pour la fréquence et pour le coût)

Structure d'un GLM

Dans un modèle linéaire généralisé, on suppose donc que Loi$(Y|X = x)$ est donnée par une loi de la famille exponentielle, caractérisée par les paramètres (μ, ϕ) et la fonction de variance $v(\cdot)$ avec :

$$g(\mathbb{E}(Y|X = x)) = g(\mu) = \eta$$

où

$$\eta = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}$$

- La fonction g est la **fonction de lien** du modèle
- Les nombres $(\beta_0, \ldots, \beta_d)$ sont les **coefficients** du modèle
- les variables $\mathrm{x}^{(0)}, \ldots, \mathrm{x}^{(d)}$ **encodent** les variables explicatives $x^{(1)}, \ldots, x^{(m)}$
- La fonction $v(\cdot)$ et le paramètre ϕ sont fixés (on verra plus loin que le paramètre ϕ peut être modulé en fonction d'un « poids »)

Structure d'un GLM

Dans le modèle, l'espérance conditionnelle $\mathbb{E}(Y|X=x)$ est donc donnée par :

$$\mathbb{E}(Y|X=x) = \mu = g^{-1}(\eta) = g^{-1}\left(\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}\right).$$

Le paramètre naturel θ n'intervient pas explicitement ici, mais on peut l'exprimer :

$$\theta = (b')^{-1}(\mu) = (b')^{-1}\left(g^{-1}(\eta)\right).$$

Si l'on a $\theta \equiv \eta$, la fonction de lien g est dite **canonique**, ce qui simplifie certaines expressions. Cette condition se réécrit $\theta = \eta = g(\mu) = g(b'(\theta))$, soit

$$g^{-1} = b'.$$

Nom	Normal	Poisson	Gamma	Binomial
Lien canonique $g(\mu) =$	μ	$\log(\mu)$	$-1/\mu$	$\log(\mu/(1-\mu))$

Régression linéaire classique

La variable Y est à valeurs dans $\mathbb{R}$, et l'on suppose que :

$$\text{Loi}(Y|X = x) = \mathcal{N}(\mu, \sigma^2),\ \mu = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}$$

Ce modèle est un GLM avec :

- $g(\mu) = \mu$ (canonique pour la loi normale)
- $v(\mu) = 1$
- $\phi = \sigma^2$

Du fait de la fonction de lien identité, les variables explicatives encodées $\mathrm{x}^{(j)}$ ont un effet **additif** sur l'espérance de la loi normale :

$$\mathbb{E}(Y|X = x) = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}.$$

La variance $\mathbb{V}(Y|X = x)$ est la même pour toutes les valeurs de x.

Régression log-Poisson

La variable Y est à valeurs dans $\mathbb{N}$ (comptage), et l'on suppose que :

$$\text{Loi}(Y|X = x) = \mathcal{P}(\lambda = \mu),\ \mu = e^{\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}}$$

Ce modèle est un GLM avec :

- $g(\mu) = \log(\mu)$ (canonique pour la loi de Poisson)
- $v(\mu) = \mu$
- $\phi = 1$

Du fait de la fonction de lien logarithmique, les variables explicatives encodées $\mathrm{x}^{(j)}$ ont un effet **multiplicatif** sur l'espérance de la loi de Poisson :

$$\mathbb{E}(Y|X = x) = \prod_{j=0}^{d} e^{\beta_j \mathrm{x}^{(j)}}$$

On a $\mathbb{V}(Y|X = x) = \mathbb{E}(Y|X = x)$ pour tout x.

Régression log-Gamma

La variable Y est à valeurs dans $]0, +\infty[$ (quantité continue positive), et l'on suppose que :

$$\text{Loi}(Y|X = x) = \text{Gamma}(\text{forme} = k, \text{éch.} = \mu/k), \ \mu = e^{\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}}.$$

Ce modèle est un GLM avec :

- $g(\mu) = \log(\mu)$ (le lien canonique pour la loi Gamma est $g(\mu) = -1/\mu$)
- $v(\mu) = \mu^2$
- $\phi = 1/k$

Du fait de la fonction de lien logarithmique, les variables explicatives encodées $\mathrm{x}^{(j)}$ ont un effet **multiplicatif** sur l'espérance de la loi Gamma :

$$\mathbb{E}(Y|X = x) = \prod_{j=0}^{d} e^{\beta_j \mathrm{x}^{(j)}}.$$

Le coefficient de variation $\mathbf{CV}(Y|X = x)$ est le même pour toutes les valeurs de x.

Régression logistique

La variable Y est à valeurs dans $\{0, 1\}$ (dichotomique), et l'on suppose que :

$$\text{Loi}(Y|X = x) = \mathcal{B}(1, p = \mu),\ \mu = \frac{e^{\sum_{j=0}^{d} \beta_j x^{(j)}}}{1 + e^{\sum_{j=0}^{d} \beta_j x^{(j)}}}.$$

Ce modèle est un GLM avec :

- $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ (canonique pour la loi de Bernoulli)
- $v(\mu) = \mu(1 - \mu)$
- $\phi = 1$

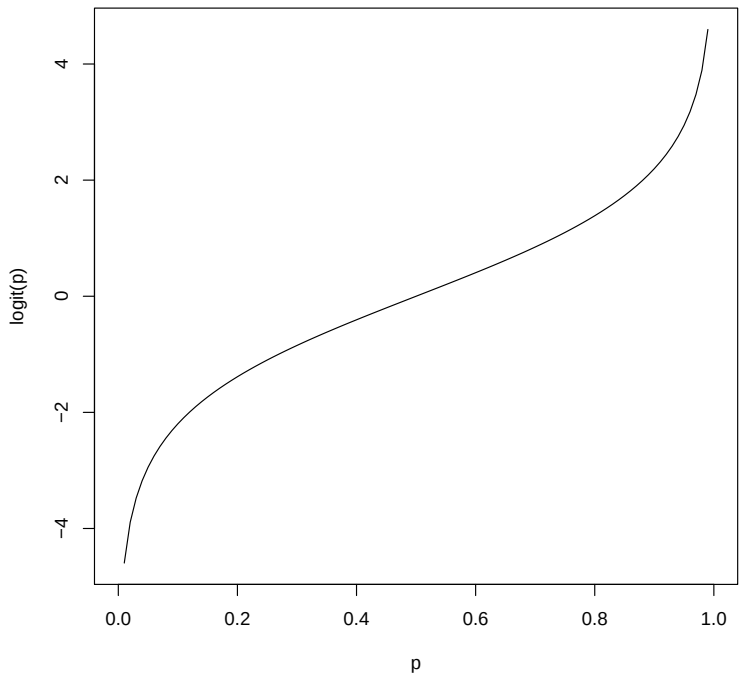
La fonction de lien utilisée s'appelle le **logit** :

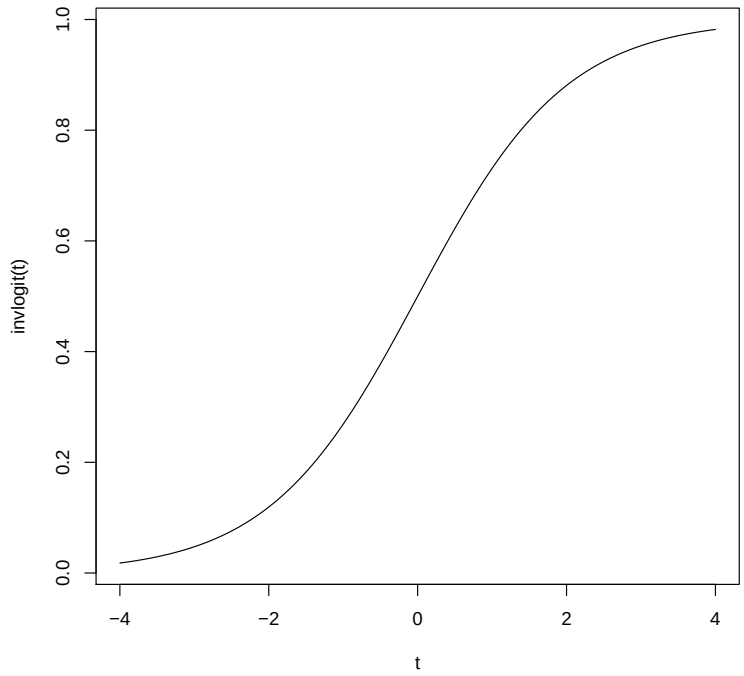
$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right),\ \mu \in]0, 1[$$

Pour $t \in \mathbb{R}$, l'inverse est donné par :

$$\text{invlogit}(t) \overset{\text{déf.}}{=} \frac{e^t}{1 + e^t} = \frac{1}{1 + e^{-t}}.$$

Régression logistique





Régression logistique

On peut reformuler le modèle en faisant intervenir la cote ou « odds » :

$$\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \frac{\mathbb{P}(Y=1|X=x)}{1-\mathbb{P}(Y=1|X=x)}.$$

Le modèle revient à poser que :

$$\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)}\right) = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}$$

Les variables explicatives encodées $\mathrm{x}^{(j)}$ ont un effet **multiplicatif sur les odds** :

$$\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=0|X=x)} = \prod_{j=0}^{d} e^{\beta_j \mathrm{x}^{(j)}}.$$

Expositions variables

Dans de nombreuses situations, une **mesure d'exposition** doit être prise en compte dans le modèle, en plus des variables explicatives.
On distingue alors la réponse brute et la réponse normalisée par l'exposition, selon le schéma suivant :

$$\left.\begin{array}{ll} \text{réponse brute} & Z \\ \text{exposition} & w \\ \text{var. explic.} & X = x \end{array}\right\} \Rightarrow \text{réponse normalisée } Y = \frac{Z}{w}$$

et l'approche classique consiste à supposer que :

$$\text{Loi}(Y|X = x) = \mathscr{L}_v(\mu, \phi/w),\ \mu = g^{-1}(\eta),\ \eta = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}$$

Cette approche revient donc à modéliser la réponse normalisée, en modulant le paramètre ϕ en fonction de l'exposition.
Dans ce cadre, on a donc $\mathbb{E}(Y|X = x) = \mu$ et $\mathbb{V}(Y|X = x) = \phi v(\mu)/w$.

Propriété d'agrégation

La modulation de ϕ en fonction de l'exposition décrite précédemment est naturelle dans le cadre GLM, du fait de la propriété d'agrégation suivante.

Propriété d'agrégation

Si $Y_1, \ldots, Y_n$ sont des variables aléatoires indépendantes telles que, pour tout i, $Y_i \sim \mathscr{L}_v(\mu, \phi/w_i)$, alors

$$\frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i} \sim \mathscr{L}_v\left(\mu, \frac{\phi}{\sum_{i=1}^n w_i}\right).$$

Régression log-Poisson avec exposition variable

Dans certains cas, la variable d'intérêt est un comptage correspondant à une exposition différente de l'unité. C'est le cas par exemple avec des polices d'assurance observées sur des durées variables, ou avec des données cumulées sur des groupes de polices correspondant à des expositions totales distinctes.

Pour une durée d'observation égale à t, on note Z le comptage brut, et $Y = Z/t$ le comptage normalisé par la durée. En prenant la durée t comme mesure d'exposition, la prise en compte d'expositions variables selon l'approche décrite précédemment revient, pour un modèle log-Poisson, à supposer que :

$$\text{Loi}(tY|X = x) = \mathcal{P}(\lambda = te^{\eta}),\ \eta = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}.$$

C'est un GLM avec Y comme variable réponse et :

$$\mu = e^{\sum_{j=0}^{d} \beta_j x^{(j)}},$$

- $g(\mu) = \log(\mu)$
- $v(\mu) = \mu$
- $\phi = 1/t$

Régression log-Poisson avec exposition variable : offset

Dans le cas log-Poisson, une approche alternative à la modulation de ϕ en fonction de l'exposition est l'utilisation d'un « offset » (voir l'annexe), qui consiste à introduire $\log(t)$ comme une variable explicative supplémentaire dont le coefficient est **fixé à** 1 dans le modèle log-Poisson de base, en utilisant comme variable réponse Y le comptage brut au lieu du comptage normalisé, en posant donc :

$$\text{Loi}(Y|X=x) = \mathcal{P}(\lambda = e^{\eta}),\ \eta = \sum_{j=0}^{d} \beta_j x^{(j)} + \log(t)\ .$$

C'est un GLM avec Y comme variable réponse et :

$$\mu = e^{\left(\sum_{j=0}^{d} \beta_j x^{(j)}\right) + \log(t)} = t \cdot e^{\sum_{j=0}^{d} \beta_j x^{(j)}},$$

- $g(\mu) = \log(\mu)$
- $v(\mu) = \mu$
- $\phi = 1$

Cette approche exploite à la fois les propriétés de la fonction log et de la loi de Poisson.

Régression log-Gamma : cas groupé

Dans certains cas, on travaille avec les valeurs cumulées de la variable d'intérêt sur un groupe au sein duquel les valeurs des variables explicatives sont les mêmes. C'est le cas par exemple avec des coûts cumulés sur un groupe de plusieurs sinistres.

Pour un groupe d'effectif total n, la valeur totale cumulée est notée Z, et la valeur normalisée au sein du groupe est $Y = Z/n$. En prenant l'effectif n du groupe comme mesure d'exposition, l'approche décrite précédemment revient, pour un modèle de régression log-Gamma, à supposer que :

$$\text{Loi}(Y|X = x) = \mathcal{G}amma(\text{esp.} = e^{\eta}, \text{forme} = k \cdot n), \ \eta = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}.$$

Ce modèle est un GLM avec Y comme variable réponse et :

$$\mu = e^{\sum_{j=0}^{d} \beta_j x^{(j)}},$$

- $g(\mu) = \log(\mu)$
- $v(\mu) = \mu^2$
- $\phi = 1/(k \cdot n)$

Régression logistique : cas groupé

Dans certains cas, la variable d'intérêt est une réponse dichotomique de type 0/1, mais on travaille avec les valeurs cumulées de celle-ci sur un groupe au sein duquel les valeurs des variables explicatives sont les mêmes. C'est le cas par exemple lorsque les données sont uniquement disponibles sous forme de valeurs cumulées sur de tels groupes.

Pour un groupe d'effectif total n, le nombre total de 1 obtenus est noté Z, et $Y = Z/n$ est la **proportion** de 1 au sein du groupe. En prenant l'effectif n du groupe comme mesure d'exposition, l'approche décrite précédemment revient, pour un modèle de régression logistique, à supposer que :

$$\text{Loi}(nY|X = x) = \mathcal{B}(n, p = \text{invlogit}(\eta)),\ \eta = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}$$

Ce modèle est un GLM avec Y comme variable réponse et :

$$\mu = \frac{e^{\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}}}{1 + e^{\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}}.}$$

- $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
- $v(\mu) = \mu(1-\mu)$
- $\phi = 1/n$

Régression logistique avec exposition variable

Ce qui précède permet de prendre en compte des groupes de taille (entière) variable en régression logistique. En revanche, cette solution n'est pas forcément bien adaptée au cas de réponses dichotomiques observées sur des durées (à valeurs dans $]0, +\infty[$) variables, par exemple en analyse de survie. On peut néanmoins l'appliquer en considérant des durées correspondant à des nombres entiers de « petits » pas de temps, mais il est alors plus cohérent d'utiliser une approche telle que le modèle à risques proportionnels de Cox.

Encodage des variables : de $x^{(1)}, \ldots, x^{(m)}$ à $\mathrm{x}^{(0)}, \ldots, \mathrm{x}^{(d)}$

Typologie des variables [1] :

- quantitative
  - continue (ex. âge)
  - discrète (ex. nombre d'enfants)
- catégorielle
  - ordinale avec échelle quantitative sous-jacente (ex. classes d'âge)
  - ordinale sans échelle quantitative sous-jacente (ex. satisfaction)
  - qualitative pure (ex. sexe)

Dans un GLM, l'effet des variables explicatives $x^{(1)}, \ldots, x^{(m)}$ sur la réponse est basé sur des combinaisons linéaires de la forme $\eta = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}$

- Les $\mathrm{x}^{(j)}$ **doivent** prendre leurs valeurs dans $\mathbb{R}$.
- Un encodage numérique approprié des variables catégorielles est donc **indispensable**.
- Un **ré-encodage** des variables quantitatives est très souvent utilisé.

1. Cette typologie, simplifiée, n'accorde pas de place spécifique aux variables structurées comme, par exemple, des variables spatiales : couple de coordonnées, ou catégorie représentant une zone géographique par exemple.

Le terme constant β_0

En général, on incorpore à la composante systématique du modèle un terme constant (dont la valeur n'est pas modulée en fonction des variables explicatives du modèle). Ce terme s'incorpore dans le cadre précédent en posant $\mathrm{x}^{(0)} \equiv 1$, si bien que le coefficient β_0 associé à $\mathrm{x}^{(0)}$ est en fait une constante qui vient s'ajouter au reste de la composante systématique du modèle :

$$\eta = \sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)} = \underbrace{\beta_0}_{\text{terme constant}} + \underbrace{\sum_{j=1}^{d} \beta_j \mathrm{x}^{(j)}}_{\text{terme dépendant des var. explic.}}$$

Dans la plupart des cas, un terme constant est inclus par défaut dans le modèle. Cependant, la formulation de certains modèles suppose de ne pas inclure ce type de terme. Dans tous les cas, l'inclusion/non-inclusion d'un terme constant **modifie** la forme du modèle,

Encodage des variables catégorielles

De manière générale, l'encodage d'une variable catégorielle dans un GLM s'effectue en passant par des variables indicatrices.

- Une variable $x^{(\ell)}$ avec K modalités $a_0, \ldots, a_{K-1}$ peut être encodée par les $K-1$ variables[2] indicatrices à valeurs dans $\{0,1\}$:

$$\mathrm{x}^{(\ell,1)} \overset{\text{déf.}}{=} \mathbf{1}(x^{(\ell)} = a_1), \ldots, \mathrm{x}^{(\ell,K-1)} \overset{\text{déf.}}{=} \mathbf{1}(x^{(\ell)} = a_{K-1}).$$

- Ceci suppose le choix d'une modalité de référence (ici : a_0)
- En l'absence d'interaction[3], la variable catégorielle $x^{(\ell)}$ intervient dans la composante systématique du GLM via les $K-1$ coefficients $\beta_{(\ell,1)}, \ldots, \beta_{(\ell,K-1)}$:

$$\eta = \underbrace{\sum_{k=1}^{K-1} \beta_{(\ell,k)} \mathbf{1}(x^{(\ell)} = a_k)}_{\text{effet de } x^{(\ell)}} + \underbrace{\cdots\cdots\cdots}_{\text{effet des autres var. explic.}}$$

2. On utilise ici une indexation double $\mathrm{x}^{(\ell,k)}$ qui peut être reconvertie en indexation simple pour retrouver la notation générique $\mathrm{x}^{(j)}$ employée ailleurs.
3. La notion d'interaction est discutée plus loin.

Encodage des variables catégorielles

La nécessité d'enlever une modalité dans les indicatrices provient de la relation de dépendance affine entre les indicatrices :

$$\sum_{k=0}^{K-1} \mathbf{1}(x^{(\ell)} = a_k) = 1,$$

qui rend le modèle non-identifiable si l'on inclut toutes les indicatrices $\mathbf{1}(x^{(\ell)} = a_0), \ldots, \mathbf{1}(x^{(\ell)} = a_{K-1})$ et un terme constant.

Interprétation des coefficients

En l'absence d'interaction, le coefficient $\beta_{(\ell,k)}$, pour $1 \leq k \leq K-1$, traduit le changement dans la valeur de η produit par le passage de la modalité de référence $x^{(\ell)} = a_0$ à la modalité $x^{(\ell)} = a_k$, *toutes choses égales par ailleurs*.

- Dans un modèle avec lien logarithmique, le passage de la modalité $x^{(\ell)} = a_0$ à la modalité $x^{(\ell)} = a_k$ **multiplie** l'espérance par $e^{\beta_{(\ell,k)}}$
- En régression logistique, le passage de la modalité $x^{(\ell)} = a_0$ à la modalité $x^{(\ell)} = a_k$ **multiplie** les odds par $e^{\beta_{(\ell,k)}}$.

Encodage des variables catégorielles

- Les modèles obtenus avec différents choix de la modalité de référence sont mathématiquement équivalents, mais l'interprétation des coefficients dépend de ce choix.
- L'incertitude sur l'estimation des coefficients varie avec le choix de la modalité de référence, et il est recommandé de choisir pour celle-ci une modalité qui soit suffisamment représentée dans les données.

On peut envisager d'autres types d'encodage, par exemple, pour une variable catégorielle ordinale avec K valeurs $\{0, \ldots, K-1\}$, un codage emboîté :

$$\mathbf{1}(x^{(\ell)} \geq 1), \ldots, \mathbf{1}(x^{(\ell)} \geq K-1).$$

De manière générale, l'utilisation d'une **matrice d'encodage**[4] permet de choisir des encodages divers et variés pour les variables catégorielles.


L'interprétation des coefficients **dépend** de l'encodage utilisé.

4. Voir par exemple le paquet `codingMatrices` de R.

Encodage des variables catégorielles

- L'encodage précédent permet de traiter séparément chacune des modalités de la variable catégorielle considérée, d'où la présence de $K-1$ coefficients.
- On est fréquemment amené à **regrouper** ou **fusionner** certaines modalités d'une variable catégorielle, en ne faisant plus de distinction entre certaines modalités.
- Etant donnée une variable $x^{(j)}$ avec K modalités $\{a_0, \ldots, a_{K-1}\}$, le fait de fusionner certaines modalités de $x^{(j)}$ revient à considérer au lieu de $x^{(j)}$ une nouvelle variable $h(x^{(j)})$, où $h : \{a_0, \ldots, a_{K-1}\} \to \{b_0, \ldots, b_{L-1}\}$ et $L < K$. Pour $k_1 \neq k_2$, les modalités a_{k_1} et a_{k_2} sont fusionnées lorsque $h(a_{k_1}) = h(a_{k_2})$.

Encodage des variables quantitatives : utilisation directe

- Il est **possible** d'utiliser directement une variable quantitative dans la composante systématique d'un GLM, avec donc $\mathrm{x}^{(\ell)} \overset{\text{déf.}}{=} x^{(\ell)}$.
- **Dans ce cas**, en l'absence d'interaction, la variable quantitative $x^{(\ell)}$ intervient dans la composante systématique du GLM via un seul coefficient β_ℓ :

$$\eta = \underbrace{\beta_\ell x^{(\ell)}}_{\text{effet de } x^{(\ell)}} + \underbrace{\cdots\cdots\cdots}_{\text{effet des autres var. explic.}}$$

Interprétation des coefficients

En l'absence d'interaction, le coefficient β_ℓ traduit le changement dans la valeur de η produit par l'augmentation d'une unité de la variable $x^{(\ell)}$, *toutes choses égales par ailleurs*.

- Dans un modèle avec lien logarithmique, l'augmentation d'une unité de la variable $x^{(\ell)}$ **multiplie** l'espérance par e^{β_ℓ}
- En régression logistique, l'augmentation d'une unité de la variable $x^{(\ell)}$ **multiplie** les odds par e^{β_ℓ}.

Encodage des variables quantitatives : utilisation directe

Utiliser directement une variable quantitative comme décrit précédemment est **possible, mais en aucun cas automatique** : rien ne garantit *a priori* que l'effet de cette variable sur la réponse soit ainsi modélisé de manière satisfaisante.

En revanche, utiliser directement une variable qualitative pure codée sous forme de nombres entiers est en général **absurde** sauf dans le cas dichotomique 0/1.

Utiliser directement une variable catégorielle ordinale codée sous forme de nombres entiers est **théoriquement possible, mais pas recommandé** *a priori*.

Encodage des variables quantitatives

Plutôt qu'une utilisation directe telle que décrite précédemment, on est souvent amené à considérer d'autres encodages d'une variable quantitative, tels que :

- Changement d'échelle simple : $x^{(\ell)}$ est remplacé par $\log(x^{(\ell)})$, $(x^{(\ell)})^{\alpha}$,... puis utilisation directe de la variable ainsi transformée
- Expression dans une base de fonctions (polynômes, splines) de dimension D : $x^{(\ell)}$ intervient alors dans la composante systématique du GLM via les D variables $\mathrm{x}^{(\ell,k)} \overset{\text{déf.}}{=} f_k(x^{(\ell)})$, pour $k = 1, \ldots, D$, associées à D coefficients :

$$\eta = \underbrace{\sum_{k=1}^{D} \beta_{\ell,k} f_k(x^{(\ell)})}_{\text{effet de } x^{(\ell)}} + \underbrace{\cdots}_{\text{effet des autres var. explic.}}$$

- Catégorisation de $x^{(\ell)}$: le domaine de valeurs de $x^{(\ell)}$ est partitionné en sous-ensembles (typiquement des intervalles) $I_0, \ldots, I_{K-1}$, et $x^{(\ell)}$ est remplacée par la variable catégorielle donnant le numéro du sous-ensemble I_k contenant la valeur de $x^{(\ell)}$

Retour sur le terme constant β_0

- Dans un modèle ne comportant que des variables catégorielles encodées classiquement (indicatrices en enlevant une modalité de référence), le terme constant β_0 est simplement l'effet global des variables explicatives sur la composante systématique du modèle lorsque chaque variable est égale à sa modalité de référence.
- En présence de variables quantitatives (utilisées directement ou via des bases de fonction), l'interprétation de la valeur de β_0 peut être moins évidente (elle correspond au cas où la contribution globale de chacune des variables explicatives à la valeur de η est nulle).

Modèle linéaire et GLM

Classiquement, les hypothèses du **modèle linéaire** sont :

(i) normalité : Loi$(Y|X = x)$ est normale

(ii) homoscédasticité : $\mathbb{V}(Y|X = x)$ est la même pour toute valeur de x

(iii) linéarité : $\mathbb{E}(Y|X = x)$ est une fonction linéaire (affine) de x

Dans de nombreux cas, ces hypothèses ne sont manifestement pas satisfaites, et l'on peut tenter de s'en rapprocher en appliquant des transformations (ex. : $z \mapsto \log z, z^\alpha$) à la variable réponse et aux variables explicatives. Des transformations permettant d'avoir simultanément (i)-(ii)-(iii) n'existent pas forcément...

Dans l'approche GLM, on pose directement une loi appropriée pour la variable réponse (p. ex. : loi binomiale ou de Poisson pour un comptage). On conserve une transformation affine des variables explicatives, à laquelle on applique une transformation généralement non-linéaire (l'inverse de la fonction de lien) pour modéliser l'espérance de la variable réponse.

Pour bien comprendre la différence de structure, comparer par exemple :

- Modèle 1 : Modèle linéaire classique pour $\log(y)$

$$Y = \exp\left(\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)} + \sigma \varepsilon\right), \ \varepsilon \sim \mathcal{N}(0,1)$$

  Ce modèle suppose que Loi$(Y|X = x)$ est **log-normale** ;

- Modèle 2 : GLM basé sur une loi normale et fonction de lien logarithmique

$$Y = \exp\left(\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}\right) + \sigma \varepsilon, \ \varepsilon \sim \mathcal{N}(0,1)$$

  Ce modèle suppose que Loi$(Y|X = x)$ est **normale**.

- Vraisemblance
- Prédiction avec un GLM
- Exemple(s) sur des données d'assurance
- Déviance
- Estimation du paramètre de dispersion
- Exemple(s) sur des données d'assurance

Modèle et données

On considère un jeu de données[5] constitué de N observations conjointes des valeurs de y et des $x^{(1)}, \ldots, x^{(m)}$: $\left[y_i, x_i = \left(x_i^{(1)}, \ldots, x_i^{(m)} \right) \right]_{i=1,\ldots,N}$.
On modélise les réponses $(y_i)_{i=1,\ldots,N}$ comme étant issues d'une suite de v.a. indépendantes $Y_1, \ldots, Y_N$ telles que

$$\text{Loi}(Y_i) = \mathscr{L}_v(\mu_i, \phi / w_i),$$

où :

- $\mu_i = g^{-1}(\eta_i)$, et $\eta_i = \sum_{j=0}^{d} \beta_j \mathrm{x}_i^{(j)}$, avec l'encodage des variables explicatives : $(x_i^{(1)}, \ldots, x_i^{(m)}) \mapsto (\mathrm{x}_i^{(0)}, \ldots, \mathrm{x}_i^{(d)})$
- w_i est un poids[6] donné *a priori*

On note que $\mathbb{E}(Y_i) = \mu_i$ et $\mathbb{V}(Y_i) = \phi v(\mu_i)/w_i$.

5. Dans la suite, on verra celui-ci comme un tableau à N lignes et $m+1$ colonnes.
6. Pour la gestion de groupes de taille variable ou de différences d'exposition.

Modèle et données

Dans ce cadre, la densité de Y_i s'écrit donc :

$$f_{Y_i}(y) = c(y, \phi/w_i) \cdot \exp\left(\frac{y\theta_i - b(\theta_i)}{\phi/w_i}\right)$$

avec :

- $\theta_i = (b')^{-1}(\mu_i)$
- $\mu_i = g^{-1}(\eta_i)$
- $\eta_i = \sum_{j=0}^{d} \beta_j \mathrm{x}_i^{(j)}$

Calculs de vraisemblance

Dans le cadre du modèle, la log-vraisemblance associée

- au jeu d'observations $(x_i, y_i)_{1 \leq i \leq N}$
- au jeu de poids $(w_i)_{1 \leq i \leq N}$
- au jeu de paramètres $(\beta_0, \ldots, \beta_d, \phi)$

s'écrit donc :

$$\log L = \sum_{i=1}^{N} \log L_i, \text{ avec } \log L_i = \frac{y_i \theta_i - b(\theta_i)}{\phi / w_i} + \log c(y_i, \phi / w_i)$$

La vraisemblance exprimée ci-dessus n'inclut pas de loi de probabilité qui décrirait les valeurs observées des variables explicatives, celles-ci étant considérées dans notre modèle comme fixées *a priori* (« fixed design »). Si l'on introduit une loi de probabilité pour les variables explicatives, vues alors comme les réalisations de variables aléatoires $X_1, \ldots, X_N$, la vraisemblance ci-dessus devient une vraisemblance *conditionnelle* aux valeurs observées des variables explicatives, reposant sur l'hypothèse que, conditionnellement à $X_1 = x_1, \ldots, X_N = x_N$, la loi employée précédemment pour décrire les $y_1, \ldots, y_N$ demeure valable.

Calculs de vraisemblance

L'estimation du jeu de coefficients $\beta = (\beta_0, \ldots, \beta_d)$ par maximum de vraisemblance conduit, par différenciation, à l'équation

$$\nabla_\beta \log L = 0,$$

où

$$\nabla_\beta \log L = \begin{pmatrix} \frac{\partial \log L}{\partial \beta_0} \\ \vdots \\ \frac{\partial \log L}{\partial \beta_d} \end{pmatrix}.$$

En dérivant l'expression de $\log L$, on obtient la formule :

$$\frac{\partial \log L}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{N} \frac{w_i (y_i - \mu_i) \mathrm{x}_i^{(j)}}{v(\mu_i) g'(\mu_i)}.$$

On a utilisé pour ce calcul les identités :

- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''((b')^{-1})(\mu_i)} = 1/v(\mu_i)$
- $\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(g^{-1}(\eta_i))} = \frac{1}{g'(\mu_i)}$
- $\frac{\partial \eta_i}{\partial \beta_j} = \mathrm{x}_i^{(j)}$

Equations de vraisemblance

L'équation $\frac{\partial \log L}{\partial \beta_j} = 0$ se met donc sous la forme :

$$\sum_{i=1}^{N} \frac{w_i(y_i - \mu_i)x_i^{(j)}}{v(\mu_i)g'(\mu_i)} = 0.$$

On constate que :

- Le paramètre ϕ n'intervient pas dans l'équation
  $\rightarrow$ on peut donc estimer les coefficients β_j sans estimer ϕ
- La loi conditionnelle Loi$(Y|X = x_i)$ n'intervient dans l'équation que via son espérance μ_i et sa variance $\phi v(\mu_i)/w_i$
- Si la fonction de lien est canonique, l'équation se réécrit :

$$\sum_{i=1}^{N} w_i(y_i - \mu_i)x_i^{(j)} = 0.$$

Equations de vraisemblance

Lorsque la fonction de lien est canonique, l'équation précédente montre que :

- pour le terme constant $\beta_{\mathbf{0}}$, l'annulation du gradient de la log-vraisemblance par rapport à $\beta_{\mathbf{0}}$ se réécrit :

$$\underbrace{\sum_{i=\mathbf{1}}^{N} w_i \mu_i}_{\text{réponse moyenne totale modélisée}} = \underbrace{\sum_{i=\mathbf{1}}^{N} w_i y_i}_{\text{réponse totale observée}}$$

- pour une variable catégorielle $x^{(\ell)}$ à K modalités $\{a_{\mathbf{0}}, \ldots, a_{K-\mathbf{1}}\}$, encodée via les indicatrices des $K-1$ modalités $a_{\mathbf{1}}, \ldots, a_{K-\mathbf{1}}$, l'annulation du gradient de la log-vraisemblance par rapport au coefficient $\beta_{(\ell,k)}$ associé à la modalité a_k se réécrit :

$$\underbrace{\sum_{i=\mathbf{1}}^{N} w_i \mu_i \mathbf{1}(x_i^{(\ell)} = a_k)}_{\substack{\text{réponse moyenne totale modélisée} \\ \text{sur les cas où } x^{(\ell)} = a_k}} = \underbrace{\sum_{i=\mathbf{1}}^{N} w_i y_i \mathbf{1}(x_i^{(\ell)} = a_k)}_{\substack{\text{réponse totale observée} \\ \text{sur les cas où } x^{(\ell)} = a_k}}$$

- par différence avec le total (équation pour $\beta_{\mathbf{0}}$), on en déduit la même identité sur les cas où $x^{(\ell)} = a_{\mathbf{0}}$.

Les équations de vraisemblance se ramènent donc dans ce cas à des équations entre les totaux marginaux observés et moyens modélisés. Lorsque la fonction de lien n'est plus canonique, on obtient des équations différentes (voir l'annexe pour les cas log-Gamma et log-NB).

Vraisemblance et groupement de données

Considérons un groupe de données $(y_{i_1}, x_{i_1}), \ldots, (y_{i_n}, x_{i_n})$ au sein duquel les variables explicatives prennent exactement la même combinaison de valeurs : $x_{i_1} = \cdots = x_{i_n} = x_{\text{groupe}}$. On a donc $\mu_{i_1} = \cdots = \mu_{i_n} = \mu_{\text{groupe}}$, et la contribution de ce groupe au gradient (en β) de la log-vraisemblance $\frac{\partial \log L}{\partial \beta_j}$ est donnée par :

$$\frac{1}{\phi} \sum_{p=1}^{n} \frac{w_{i_p}(y_{i_p} - \mu_{i_p}) \mathrm{x}_{i_p}^{(j)}}{v(\mu_{i_p}) g'(\mu_{i_p})} = \frac{1}{\phi} \frac{w_{\text{groupe}}(y_{\text{groupe}} - \mu_{\text{groupe}}) \mathrm{x}_{\text{groupe}}^{(j)}}{v(\mu_{\text{groupe}}) g'(\mu_{\text{groupe}})},$$

avec $w_{\text{groupe}} = \sum_{p=1}^{n} w_{i_p}$ et $y_{\text{groupe}} = \frac{\sum_{p=1}^{n} w_{i_p} y_{i_p}}{w_{\text{groupe}}}$.
Par conséquent, du point de vue de l'estimation des coefficients β, le groupe de données est équivalent à une unique observation groupée $(y_{\text{groupe}}, x_{\text{groupe}})$, de poids total w_{groupe}.


Il n'y a plus équivalence si l'on considère l'estimation du paramètre ϕ, ou des questions de validation de modèle, pour lesquelles le remplacement du groupe par une unique observation entraîne une perte d'information.

Calculs de vraisemblance

On introduit les notations :

- $\mathbf{Y} = (Y_i)_{1 \leq i \leq N}$ et $\mathbf{y} = (y_i)_{1 \leq i \leq N}$
- $\mathbf{x} = (x_i)_{1 \leq i \leq N}$
- $\beta = (\beta_j)_{0 \leq j \leq d}$

La vraisemblance $L = L(\beta, \mathbf{y}, \mathbf{x})$ se présente comme une fonction[7] de β, $\mathbf{y}$ et $\mathbf{x}$.

La matrice d'information de Fisher $I = (I_{jk})_{0 \leq j,k \leq d}$ est la matrice $(d+1) \times (d+1)$ définie par :

$$\begin{aligned} I_{jk} &= -\mathbb{E}\left(\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k}(\beta, \mathbf{Y}, \mathbf{x}) \right) \\ &= \mathbb{E}\left(\left[\frac{\partial \log L}{\partial \beta_j}(\beta, \mathbf{Y}, \mathbf{x}) \right] \left[\frac{\partial \log L}{\partial \beta_k}(\beta, \mathbf{Y}, \mathbf{x}) \right] \right) \end{aligned}$$

Dans l'expression ci-dessus, l'espérance $\mathbb{E}$ est prise par rapport se réfère à la loi des variables aléatoires $Y_1, \ldots, Y_N$, supposées indépendantes et vérifiant $\text{Loi}(Y_i) = \mathscr{L}_v(\mu_i, \phi/w_i)$, où $\mu_i = g^{-1}\left(\sum_{j=0}^{d} \beta_j \mathrm{x}_i^{(j)}\right)$.

7. Entre autres : il reste encore bien entendu ϕ et $(w_i)_{i=1,\ldots,N}$.

Calculs de vraisemblance

Les coefficients de la matrice d'information de Fisher s'écrivent (voir annexe) :

$$I_{j,k} = \sum_{i=1}^{N} w_i \frac{\mathrm{x}_i^{(j)} \mathrm{x}_i^{(k)}}{\phi v(\mu_i)} \left(\frac{1}{g'(\mu_i)} \right)^2$$

Soit sous forme matricielle :

$$I = {}^t\mathfrak{X} \times W \times \mathfrak{X}$$

avec

- $\mathfrak{X} = (\mathrm{x}_i^{(j)})_{\substack{1 \leq i \leq N \\ 0 \leq j \leq d}}$ vu comme une matrice $N \times (d+1)$
- $W = \mathrm{diag}\left[\frac{w_i}{\phi v(\mu_i)} \left(\frac{1}{g'(\mu_i)} \right)^2,\ 1 \leq i \leq N \right]$

Estimation par maximum de vraisemblance

L'estimation des coefficients β se fait en général par maximum de vraisemblance, et conduit donc à chercher une estimation $\beta^{\text{est.}}$ des (supposés) vrais coefficients β via l'équation :

$$\nabla_{\beta^{\text{est.}}}(\log L) = 0$$

- Pas de solution explicite (en général).
- Résolution approchée par des méthodes numériques itératives [8].

Les deux approches classiques sont :

- Newton-Raphson
- Fisher scoring

8. Voir par exemple [Woo17] sec. 2.1.2 pour une discussion plus approfondie.

Newton-Raphson/Fisher scoring

Idée : pour $\beta^{\text{approx.}} \approx \beta^{\text{est.}}$, on a la relation

$$\underbrace{\nabla_{\beta^{\text{est.}}}(\log L)}_{=0} = \nabla_{\beta^{\text{approx.}}}(\log L) + H_{\beta^{\text{approx.}}}(\log L)(\beta^{\text{est.}} - \beta^{\text{approx.}}) + o\left(\beta^{\text{est.}} - \beta^{\text{approx.}}\right),$$

où H_β désigne la matrice Hessienne (formule explicite en annexe) :

$$H_\beta(\log L) = \left(\frac{\partial^2 \log L}{\partial\beta_j \partial\beta_k}\right)_{0\leq j,k\leq d}$$

si bien que

$$\beta^{\text{est.}} - \beta^{\text{approx.}} \approx \underbrace{-H_{\beta^{\text{approx.}}}(\log L)^{-1}\nabla_{\beta^{\text{approx.}}}(\log L)}_{=G(\beta^{\text{approx.}})}.$$

Méthode de Newton-Raphson

A partir d'une valeur initiale $\beta^{[0]}$, on calcule une suite d'estimations $\beta^{[r]}$ définie par :

$$\beta^{[r]} = \beta^{[r-1]} + G(\beta^{[r-1]})$$

jusqu'à obtenir $||\nabla_{\beta^{[r]}}(\log L)|| \leq \epsilon$, où $\epsilon << 1$ est un seuil fixé à l'avance.

Newton-Raphson/Fisher scoring

Fisher scoring

L'itération est similaire à celle de Newton-Raphson, mais en remplaçant la Hessienne par une approximation obtenue à partir de l'estimation courante de l'information de Fisher :

$$\beta^{[r]} = \beta^{[r-1]} + I(\beta^{[r-1]})^{-1} \nabla_{\beta^{[r-1]}} (\log L)$$

- Représentation de l'itération du Fisher scoring sous forme d'une itération de problèmes de moindre carrés pondérés (Iteratively Reweighted Least Squares, voir annexe).
- Lorsque la fonction de lien est canonique, les deux méthodes sont exactement identiques (et $\log L$ est une fonction convexe des coefficients).
- La fonction `glm` de R utilise la méthode de Fisher scoring.
- Ces méthodes se révèlent en général efficaces d'un point de vue numérique, ne nécessitant qu'un nombre limité d'itérations pour obtenir des valeurs satisfaisantes.

Comportement dans la limite d'un grand jeu de données

On note $\hat{\beta}$ la variable aléatoire correspondant à la valeur de $\beta^{\text{est.}}$ lorsque les réponses sont données par des v.a. $Y_1, \ldots, Y_N$.
On suppose que les v.a. $Y_1, \ldots, Y_N$ sont indépendantes et telles que, pour tout i, $Y_i \sim \mathscr{L}_v(\mu_i, \phi/w_i)$, où $\mu_i = g^{-1}(\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)})$.
On suppose en outre :

- un grand nombre de données ($N >> 1$) ;
- certaines hypothèses de régularité (voir annexe).

La théorie asymptotique de l'estimation par maximum de vraisemblance fournit alors les approximations suivantes (qui deviennent exactes dans la limite $N \to +\infty$) :

- $\hat{\beta} \approx \beta$
- Loi$(\hat{\beta} - \beta) \approx \mathcal{N}(0, \Sigma = I(\beta)^{-1}) \approx \mathcal{N}(0, \Sigma = I(\hat{\beta})^{-1})$, où I est la matrice d'information de Fisher.

On peut en particulier utiliser la normalité asymptotique pour construire des intervalles de confiance (approchés) pour les coefficients.

Comportement dans la limite d'un grand jeu de données

Lorsque le modèle est valide, c'est-à-dire que l'on peut effectivement considérer les réponses $y_1, \ldots, y_N$ comme des réalisations des variables aléatoires $Y_1, \ldots, Y_N$ possédant toutes les propriétés précédentes, l'estimation par maximum de vraisemblance produit donc, dans la limite d'un grand jeu de données (et sous certaines hypothèses de régularité) :

- une approximation des coefficients β, dont le modèle suppose l'existence, par les coefficients estimés ;
- une estimation (via des intervalles de confiance approchés) de l'incertitude affectant cette approximation.

Le paramètre ϕ intervient également dans l'expression de I. Cela ne pose pas de problème lorsque sa valeur est supposée connue *a priori* (par exemple en régression de Poisson ou binomiale, où $\phi = 1$). Dans les autres cas (par exemple en régression Gamma), le paramètre ϕ doit être estimé, et cette estimation utilisée pour calculer une valeur approchée de I. Les méthodes d'estimation du paramètre ϕ seront discutées plus tard !

Notations

Dans la suite, nous aurons régulièrement l'occasion d'utiliser plusieurs notations distinctes (mettant en valeur des points de vue différents) pour des objets semblables :

- la notation $A^{\text{est.}}$ pour désigner la valeur estimée de la quantité A après ajustement du modèle sur les données $(y_i, x_i)_{1 \leq i \leq N}$
- la notation $\hat{A}$ pour désigner la variable aléatoire correspondant à la valeur de $A^{\text{est.}}$ lorsque les y_i sont remplacées par des variables aléatoires $Y_1, \ldots, Y_N$.
- la notation $\hat{A}^{\text{obs.}}$ pour désigner la valeur effectivement observée de la variable aléatoire $\hat{A}$, les valeurs effectivement observées des v.a. Y_i étant les y_i.

On a donc l'identité :

$$A^{\text{est.}} = \hat{A}^{\text{obs.}}.$$

- Vraisemblance
- **Prédiction avec un GLM**
- Exemple(s) sur des données d'assurance
- Déviance
- Estimation du paramètre de dispersion
- Exemple(s) sur des données d'assurance

Valeur prédite

But : Etant donnée une combinaison de valeurs des variables explicatives $x = (x^{(1)}, \ldots, x^{(m)})$, fournir une prédiction sur la réponse (non-encore observée) y.
Les prédictions du modèle prennent par exemple les formes suivantes :

- **Valeur de l'espérance :** la valeur modélisée de $\mathbb{E}(Y)$
- **Valeur de la variance :** la valeur modélisée de $\mathbb{V}(Y)$
- **Loi de la réponse :** la loi modélisée de Y
- **Intervalle de prédiction :** étant donné un seuil de probabilité u (p. ex. 99%), les bornes y_- et y_+ telles que, d'après le modèle, $\mathbb{P}(Y \in [y_-, y_+]) \approx u$, ou (selon les besoins) $\mathbb{P}(Y \leq y_+) \approx u$ ou encore $\mathbb{P}(Y \geq y_-) \approx u$.

 La valeur de la réponse est vue dans le modèle comme une **variable aléatoire**, et, en général, il n'y a aucune raison pour que l'on ait $Y \approx \mathbb{E}(Y)$. Il ne faut pas confondre la valeur prédite pour l'espérance $\mathbb{E}(Y)$ avec une valeur « typique » prédite pour Y.

Valeur prédite

Le modèle est caractérisé par les paramètres β et ϕ, dont il suppose l'existence, mais dont les valeurs exactes sont inconnues. Les prédictions du modèle sont obtenues en remplaçant, dans les expressions correspondantes, ces paramètres par leurs valeurs **estimées**.

En qui concerne l'espérance $\mathbb{E}(Y)$, sa valeur selon le modèle est :

$$\mu = g^{-1}\left(\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)}\right),$$

sa valeur estimée est :

$$\mu^{\text{est.}} = g^{-1}\left(\sum_{j=0}^{d} \beta_j^{\text{est.}} \mathrm{x}^{(j)}\right),$$

et l'estimateur de sa valeur (en remplaçant les y_i par les v.a. Y_i) est :

$$\hat{\mu} = g^{-1}\left(\sum_{j=0}^{d} \hat{\beta}_j \mathrm{x}^{(j)}\right).$$

Valeur prédite

Pour répercuter l'incertitude liée à l'estimation des coefficients sur la valeur estimée de $\mathbb{E}(Y)$, on peut construire un intervalle de confiance en utilisant l'approximation normale de $\hat{\beta}$ et la méthode Delta :

$$\text{Loi}(\hat{\mu} - \mu) \approx \mathcal{N}(0, \sigma^2),$$

la variance asymptotique σ^2 étant donnée par :

$$\sigma^2 = \frac{{}^t\mathrm{x} \times I(\beta)^{-1} \times \mathrm{x}}{g'\left(g^{-1}(\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)})\right)^2}, \text{ où } \mathrm{x} = \begin{pmatrix} \mathrm{x}^{(0)} \\ \vdots \\ \mathrm{x}^{(d)} \end{pmatrix}$$

En substituant $\hat{\beta}$ à β ci-dessus, on obtient $\hat{\sigma} \approx \sigma$ et des intervalles de confiance peuvent être effectivement calculés.

Pour une prédiction ne portant pas seulement sur l'espérance de Y, mais sur l'espérance d'une somme de réponses $\sum_{q=1}^{r} Y_q$ associées à des valeurs différentes des variables explicatives, des intervalles de confiance peuvent également être obtenus par une approche étendant ce qui précède.

Valeur prédite

Quelques points d'attention :

- Ne pas confondre intervalle de confiance pour $\mathbb{E}(Y)$ et intervalle de prédiction pour Y
- Ce qui précède montre comment obtenir un intervalle de confiance pour $\mathbb{E}(Y)$; de manière générale, il conviendrait de répercuter l'incertitude associée à l'estimation des paramètres sur les différentes prédictions tirées du modèle (variance, intervalles de prédiction, etc.)
- Lorsque ϕ est estimé, il y a également une incertitude sur ce paramètre, et pas seulement sur les coefficients β.
- La validité des prédictions obtenues repose bien entendu sur la validité du modèle GLM sous-jacent.

- Vraisemblance
- Prédiction avec un GLM
- **Exemple(s) sur des données d'assurance**
- Déviance
- Estimation du paramètre de dispersion
- Exemple(s) sur des données d'assurance

Avec R

Exemples, code et explications se trouvent dans les blocs-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-A.html`
`https://irma.math.unistra.fr/~jberard/nb-GLM-B.html`

- Vraisemblance
- Prédiction avec un GLM
- Exemple(s) sur des données d'assurance
- **Déviance**
- Estimation du paramètre de dispersion
- Exemple(s) sur des données d'assurance

Fonction de déviance

La fonction de déviance est une mesure d'écart adaptée à la famille exponentielle, qui généralise l'écart quadratique. Formellement, étant donnés une fonction de variance v et une valeur du paramètre de dispersion ϕ, on pose, pour tout couple[9] de nombres réels y, μ :

$$D(y,\mu) \overset{\text{déf.}}{=} 2\phi \cdot \left[\log f_{\mathscr{L}_v(y,\phi)}(y) - \log f_{\mathscr{L}_v(\mu,\phi)}(y)\right].$$

La valeur de $D(y,\mu)$ ne dépend en fait pas de ϕ comme on peut le constater en reprenant l'expression de la densité la famille exponentielle, qui fournit la formule explicite :

$$D(y,\mu) = 2\left(y\left[(b')^{-1}(y) - (b')^{-1}(\mu)\right] - \left[b((b')^{-1}(y)) - b((b')^{-1}(\mu))\right]\right).$$

On peut également exprimer la fonction de déviance en termes de divergence de Kullback-Leibler :

$$D(y,\mu) = 2\phi \cdot d_{\mathsf{KL}}\left(\mathscr{L}_v(y,\phi)||\mathscr{L}_v(\mu,\phi)\right).$$

9. La déviance n'est bien définie que lorsque y et μ appartiennent au support commun des lois caractérisées par la fonction de variance v et le paramètre de dispersion ϕ.

Déviance

On note les propriétés générales suivantes :

- $D(y, \mu) \geq 0$
- $D(y, \mu) = 0 \Leftrightarrow \mu = y$

En revanche, il n'est pas vrai en général que $D(y, \mu) = D(\mu, y)$.

Nom	Déviance $D(y, \mu)$
Normale	$(y - \mu)^2$
Poisson	$2\left[y \log(y/\mu) - (y - \mu)\right]$
Gamma	$-2\left[\log(y/\mu) - \frac{y-\mu}{\mu}\right]$
Binomiale	$2\left[y \log(y/\mu) + (1 - y) \log\left(\frac{1-y}{1-\mu}\right)\right]$

Déviance

Pour un modèle GLM, la déviance associée au jeu de données $(y_i, x_i)_{1 \leq i \leq N}$, aux poids $(w_i)_{1 \leq i \leq N}$, et aux valeurs $\mu_i = g^{-1}(\eta_i)$ est définie par :

$$D \overset{\text{déf.}}{=} \sum_{i=1}^{N} w_i \cdot D(y_i, \mu_i).$$

La log-vraisemblance associée se réécrit sous la forme :

$$\log L = -D/(2\phi) + C,$$

où C s'exprime uniquement en fonction de $(y_i)_{1 \leq i \leq N}$, $(w_i)_{1 \leq i \leq N}$ et ϕ. Par conséquent, l'estimation des coefficients du modèle par **maximum de vraisemblance** revient à une estimation par **minimum de déviance**. Le cas limite $D = 0$ correspond au cas où les valeurs μ_i sont **exactement** égales aux valeurs observées y_i.

Déviance

La commande `glm` de R calcule systématiquement :

- la déviance $D = \sum_{i=1}^{N} w_i \cdot D\left(y_i, \mu_i^{\text{est.}}\right)$ associée au modèle ajusté (appelée « residual deviance »)
- la déviance D_0 associée au modèle comportant uniquement un terme constant (appelée « null deviance »), c'est-à-dire $D_0 = \sum_{i=1}^{N} w_i \cdot D\left(y_i, \mu^{\text{est.}}\right)$, où $\mu^{\text{est.}} = \left(\sum_{i=1}^{N} w_i y_i\right) / \sum_{i=1}^{N} w_i$.

On peut définir[10] un « pseudo-R^2 » par :

$$R_D^2 \overset{\text{déf.}}{=} 1 - \frac{D}{D_0}.$$

- $0 \leq R_D^2 \leq 1$
- $R_D = 1$ correspond à un ajustement « parfait » ($\mu_i^{\text{est.}} = y_i$ pour tout i)
- $R_D = 0$ correspond à une absence totale de pertinence des variables explicatives utilisées par le modèle
- R_D^2 s'interprète comme une proportion de déviance « expliquée » par le modèle

10. Toutes les précautions d'usage relatives à l'interprétation du R^2 en modèle linéaire s'appliquent ici à plus forte raison !

Déviance normalisée

On note que la déviance D définie précédemment est uniquement fonction des $\mu_i^{\text{est.}}$, y_i et w_i, et ne fait donc pas intervenir le paramètre de dispersion ϕ. La déviance **normalisée** (« scaled deviance ») est définie quant à elle par :

$$\tilde{D} = \frac{D}{\phi}.$$

Lorsque la valeur de ϕ est déterminée *a priori* (par exemple, $\phi = 1$ en régression de Poisson ou binomiale), cette définition ne présente pas de difficulté. En revanche, lorsque la valeur de ϕ doit être estimée (p.ex. : en régression Gamma), on remplace souvent ϕ par une valeur estimée $\phi^{\text{est.}}$ dans l'expression de $\tilde{D}$, d'où

$$\tilde{D}' = \frac{D}{\phi^{\text{est.}}}.$$

 La terminologie concernant la déviance est quelque peu fluctuante, et la distinction entre déviance normalisée et non-normalisée (plus généralement entre D, $\tilde{D}$ et $\tilde{D}'$) n'est pas toujours bien claire dans les références disponibles. La « residual deviance » calculée par la fonction `glm` de R est la déviance non-normalisée D.

- Vraisemblance
- Prédiction avec un GLM
- Exemple(s) sur des données d'assurance
- Déviance
- **Estimation du paramètre de dispersion**
- Exemple(s) sur des données d'assurance

Estimation du paramètre de dispersion

Lorsque le paramètre ϕ doit être estimé (p.ex. : GLM basé sur une loi Gamma), plusieurs estimateurs sont classiquement utilisés :

- via la statistique de Pearson :

$$\hat{\phi} = \frac{\mathcal{X}^2}{N - (d+1)}, \text{ où } \mathcal{X}^2 = \sum_{i=1}^{N} w_i \cdot \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)};$$

- par maximum de vraisemblance : en utilisant les valeurs déjà estimées des coefficients, il s'agit d'une optimisation uni-dimensionnelle ;
- via la déviance[11] :

$$\hat{\phi} = \frac{D}{N - (d+1)}, \; D = \sum_{i=1}^{N} w_i \cdot D(Y_i, \hat{\mu}_i).$$

Voir l'annexe pour la méthode d'estimation utilisée par le paquet `mgcv`, basée sur [Fle12].

11. Méthode non-recommandée en général.

Estimation du paramètre de dispersion

Quelques points à noter :

- l'estimation par maximum de vraisemblance des coefficients β_j ne nécessite pas l'estimation de ϕ
- la valeur estimée de ϕ intervient dans les intervalles de confiance pour les β_j
- la valeur estimée de ϕ intervient dans les intervalles de prédiction pour Y
- l'estimation de ϕ n'est pas invariante par regroupement des données

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-C.html`

Un exemple simple et frappant [12] (modèles)

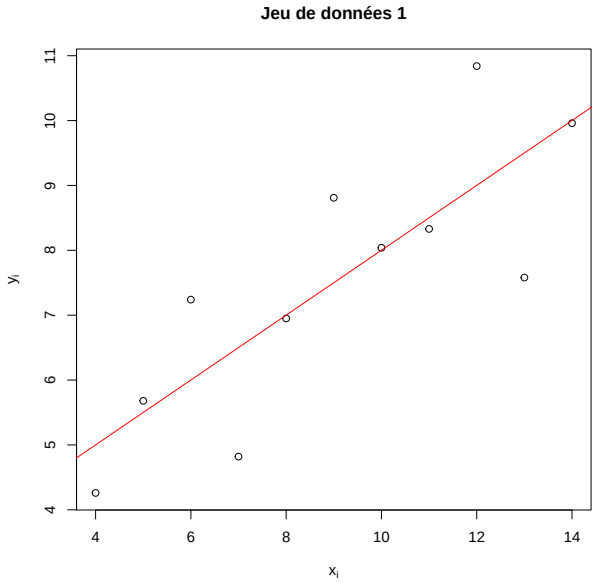
Ci-dessous, on présente les résultats de l'ajustement d'un modèle linéaire gaussien de la forme $Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \phi)$ sur quatre jeux de données distincts comportant chacun $N = 11$ couples de valeurs $(y_i, x_i)_{1 \leq i \leq N}$.

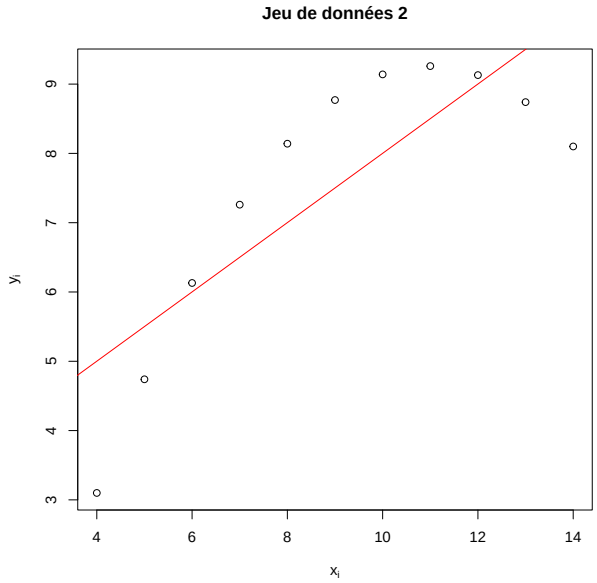
Jeu de données	1	2	3	4
$\beta_0^{\text{est.}}$	3.0001	3.001	3.0025	3.0017
$\sigma(\hat{\beta}_0)^{\text{est.}}$	1.1247	1.125	1.1245	1.1239
$\beta_1^{\text{est.}}$	0.5001	0.500	0.4997	0.4999
$\sigma(\hat{\beta}_1)^{\text{est.}}$	0.1179	0.118	0.1179	0.1178
$\phi^{\text{est.}}$	1.5292	1.5307	1.5285	1.5269

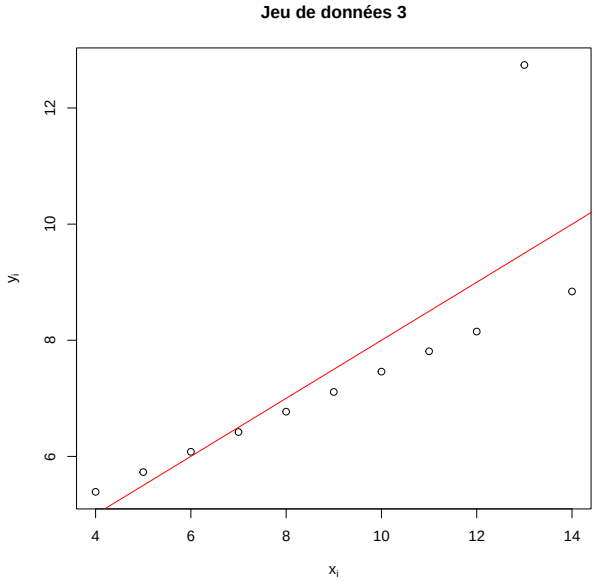
Les résultats ainsi obtenus sont donc essentiellement équivalents pour chacun des quatre jeux de données. Pourtant, un examen graphique très sommaire suffit à révéler des situations extrêmement différentes d'un jeu de données à l'autre, le modèle ajusté étant visiblement inadéquat dans tous les cas sauf celui du jeu de données 1.

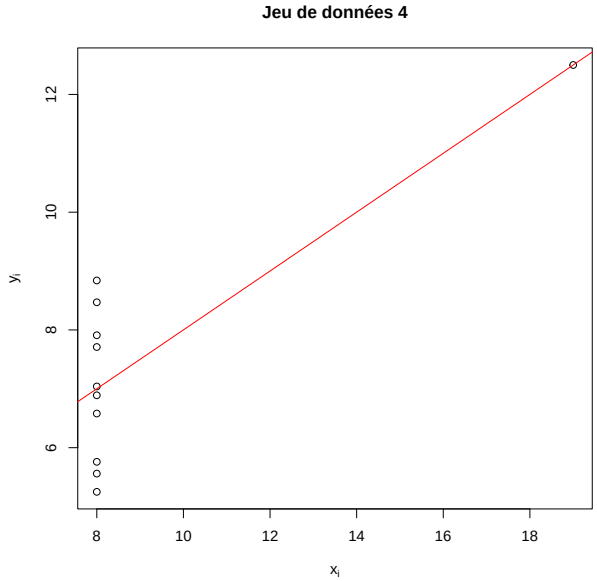
12. Exemple tiré de : Anscombe, Francis J. (1973). Graphs in statistical analysis. The American Statistician, 27, 17–21.

Un exemple simple et frappant (données)









Un exemple simple et frappant (morale de l'histoire)

Ce qui précède illustre dans un cadre très simple :

- la nécessité d'étudier la **validité** des modèles ajustés avant de les utiliser (quelle confiance pourrait-on accorder aux prédictions des modèles ajustés sur les jeux de données 2, 3, 4 ?)
- l'intérêt d'un examen **graphique** de l'ajustement effectué

Résidus

De manière générale, les résidus constituent une mesure de l'écart entre la valeur de l'espérance modélisée pour la réponse $\mu_i^{\text{est.}}$ et la valeur observée y_i de celle-ci, sur des données effectivement disponibles.
Différentes définitions et normalisations sont possibles pour l'écart entre y_i et $\mu_i^{\text{est.}}$, conduisant à autant de types distincts de résidus (voir plus loin).
Les résidus peuvent être calculés :

- sur les données ayant servi à ajuster le modèle, ou sur un jeu de données séparées (données de test)
- au niveau des données individuelles, ou à un niveau plus agrégé

L'examen graphique des résidus a notamment pour but de :

- vérifier l'adéquation entre modèle et données
- identifier d'éventuelles tendances/structures globalement présentes dans les données mais non prises en compte dans la modélisation
- repérer les données « aberrantes » (en désaccord avec le modèle)

Tracé des résidus

Classiquement, on visualise les résidus :

- en fonction de l'indice i (qui n'a en général pas de signification particulière)
- en fonction de la valeur moyenne prédite [13] $\mu_i^{\text{est.}}$ ou de la composante systématique $\eta_i^{\text{est.}}$
- en fonction des diverses variables explicatives disponibles (incluses ou non-incluses dans le modèle)
- du point de vue de leur distribution globale

Concrètement, de nombreux choix sont possibles pour le type de représentation graphique effectué, avec divers degrés de raffinement (voir plus loin).

13. Avec ou sans normalisation par l'exposition, selon les cas.

Le monde idéal pour les résidus

Dans l'idéal, on souhaiterait disposer de résidus $r_1, \ldots, r_N$ qui, **lorsque le modèle étudié est valide**[14], constituent des réalisations de variables aléatoires $R_1, \ldots, R_N$:

- centrées, c.à.d. $\mathbb{E}(R_i) = 0$
- de variance unité, c.à.d. $\mathbb{V}(R_i) = 1$
- de loi normale $\mathcal{N}(0, 1)$
- indépendantes

On peut alors éprouver la validité du modèle, en examinant la conformité des tracés produits vis-à-vis de ces hypothèses. Schématiquement :

- vérifier le caractère centré des résidus le long des divers tracés aide à vérifier l'absence de sur- ou de sous-estimation systématique de l'espérance ;
- pour des résidus centrés, vérifier la constance à 1 de la variance le long des divers tracés aide à vérifier l'absence de sur- ou de sous-estimation systématique de la variance ;
- pour des résidus centrés de variance unité, vérifier la normalité de la distribution le long des divers tracés aide à vérifier le caractère approprié de la loi utilisée par le modèle.

14. Dans notre contexte, cela signifie que les y_i peuvent effectivement être considérées comme des réalisations de v.a. $Y_1, \ldots, Y_N$ indépendantes et telles que, pour tout i, $Y_i \sim \mathscr{L}_v(\mu_i, \phi/w_i)$, où $\mu_i = g^{-1}(\sum_{j=0}^{d} \beta_j \mathrm{x}^{(j)})$.

Le monde idéal pour les résidus

On illustre ci-après le comportement des résidus dans une telle situation idéale, d'abord dans le cas d'un modèle valide, puis pour des modèles présentant divers types de défauts que l'examen graphique des résidus permet de déceler.
On a délibérément choisi de se placer dans le cas d'un nombre limité de résidus (400), de manière à illustrer un contexte classique en statistique (qui n'est pas forcément celui des applications les plus habituelles dans un contexte actuariel).
La construction, le tracé et l'examen des résidus dans des situations réelles et non pas idéales sera abordée ensuite.

Le monde idéal pour les résidus

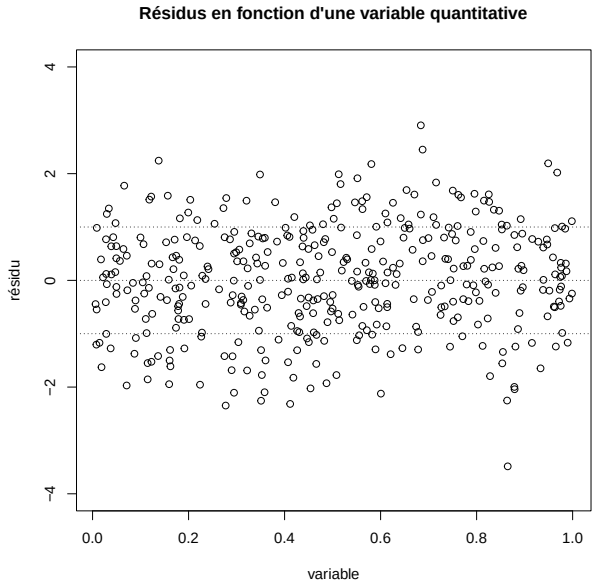
La première illustration présente le cas d'un modèle valide, pour lequel les résidus sont exactement conformes aux hypothèses précédentes. Il y a en tout 400 résidus, représentés :

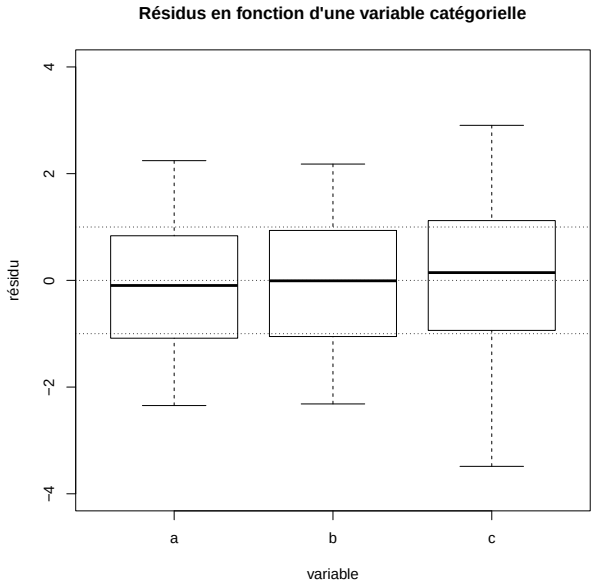
- en fonction de leur indice i
- en fonction d'une variable quantitative continue
- en fonction d'une variable catégorielle (un box-plot [15] par modalité)
- en comparaison avec une loi $\mathcal{N}(0,1)$ via un tracé quantile-quantile

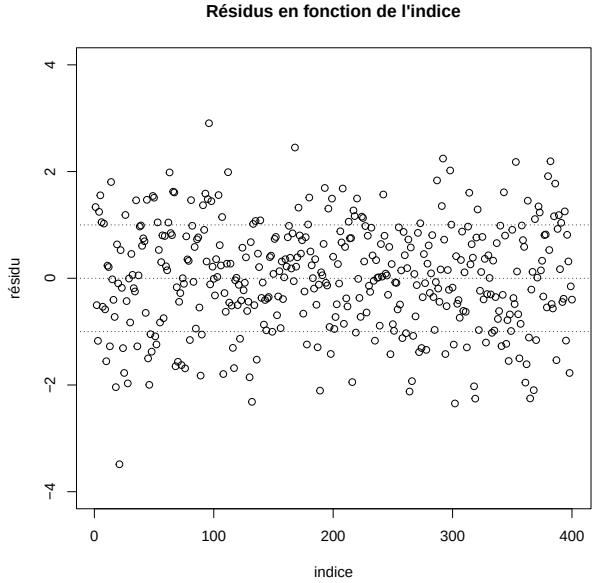
15. On utilise ici une version modifiée du box-plot classique : on représente la moyenne (à l'intérieur de la boîte), la moyenne moins la racine carrée des écarts quadratiques restreints aux valeurs inférieures à la moyenne, la moyenne plus la racine carrée des écarts quadratiques restreints aux valeurs supérieures à la moyenne (les bords de la boîte), et les valeurs minimales et maximales (les moustaches).

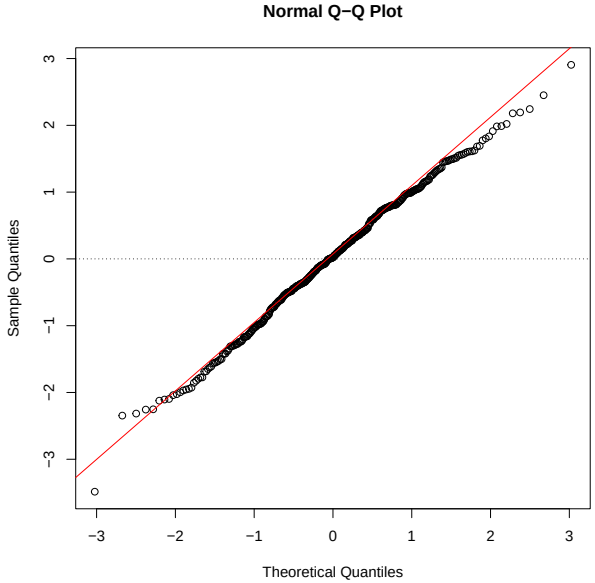
Le monde idéal pour les résidus

Illustration : modèle valide.









Le monde idéal pour les résidus

Pour définir des seuils de valeurs anormalement élevées (positives ou négatives) pour les résidus, on peut s'appuyer sur la loi normale, en définissant t_α par l'identité

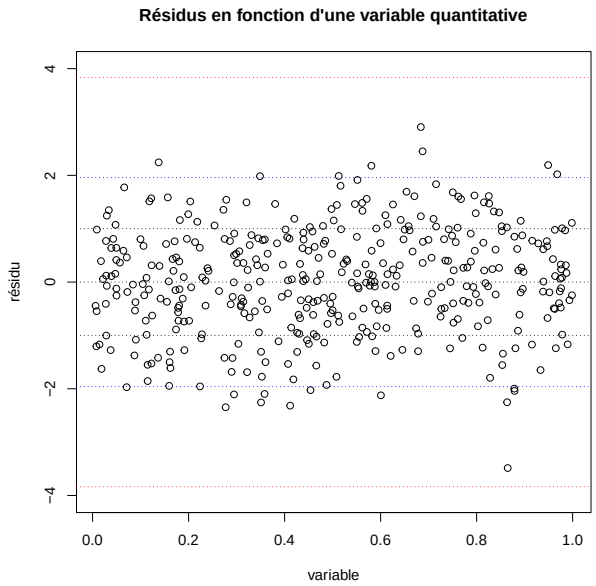
$$\mathbb{P}(Z \notin [-t_\alpha, t_\alpha]) = \alpha, \ Z \sim \mathcal{N}(0,1).$$

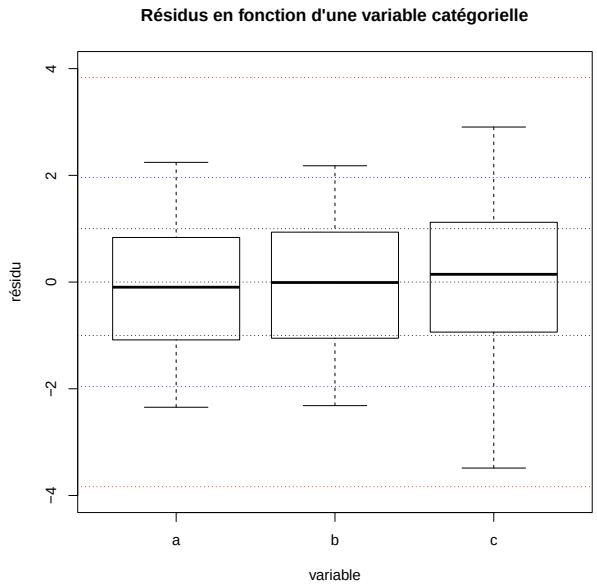
Ainsi :

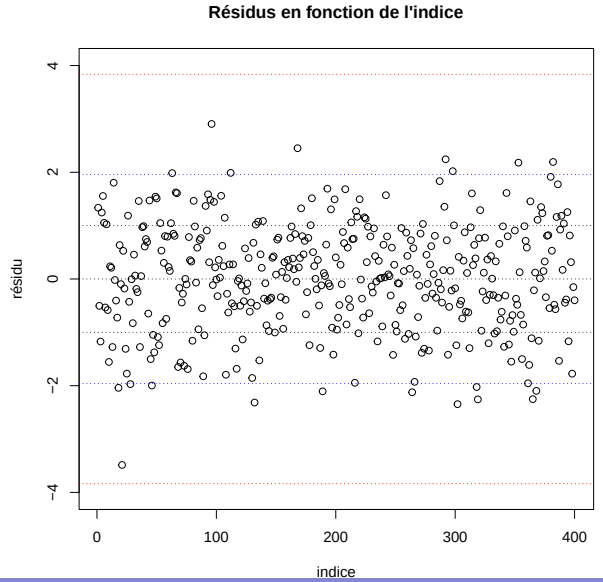
- chaque résidu a une probabilité α de se trouver hors de $[-t_\alpha, t_\alpha]$
- la probabilité qu'au moins un résidu sur les N se trouve hors de $[-t_{\alpha/N}, t_{\alpha/N}]$ est $\lesssim \alpha$

Le monde idéal pour les résidus

Illustration : modèle valide, avec les deux niveaux de seuils pour $\alpha = 0.05$.





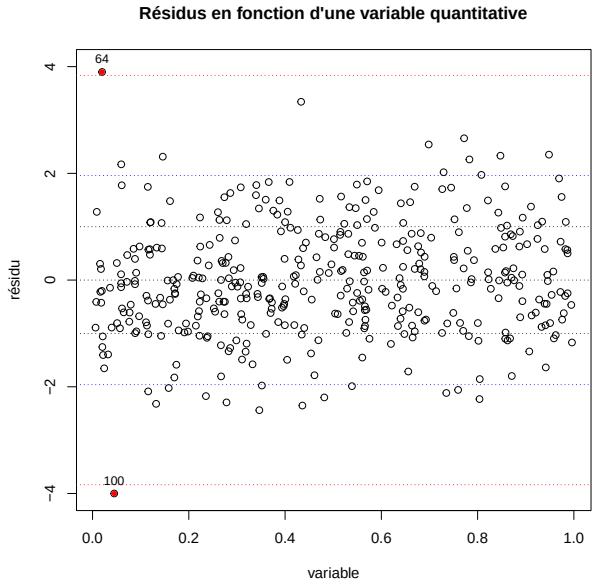


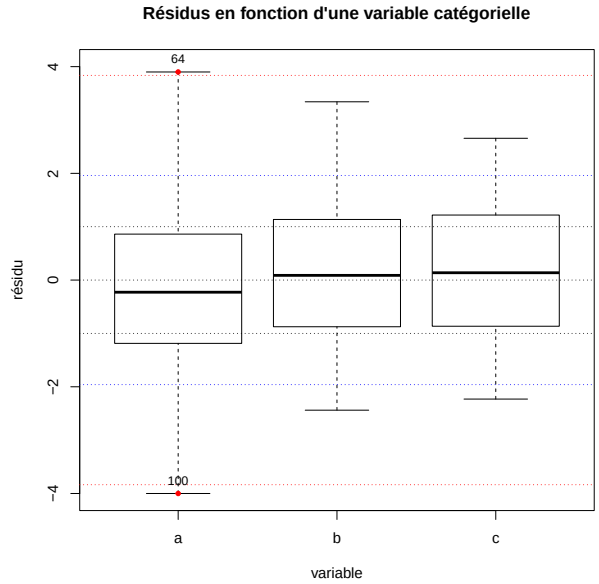
Le monde idéal pour les résidus

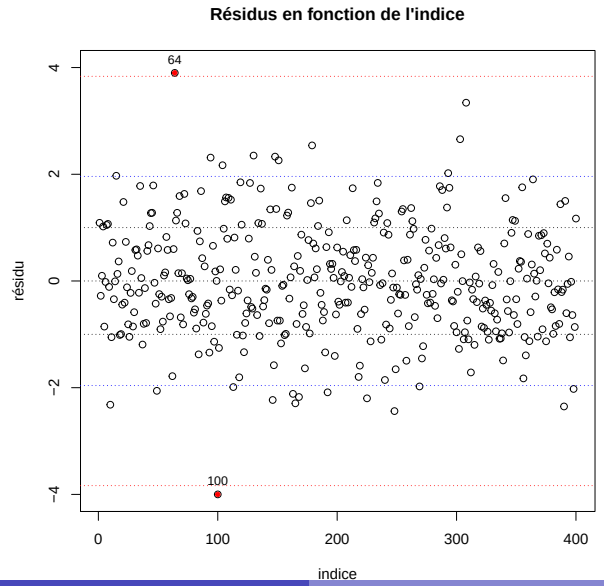
La deuxième illustration présente le cas d'un modèle globalement valide, mais avec la présence de deux points « aberrants », donnant lieu à des résidus de taille anormalement élevée, repérés en rouge sur les divers tracés. Les indices correspondants sont également indiqués : $i = 64$ et $i = 100$.

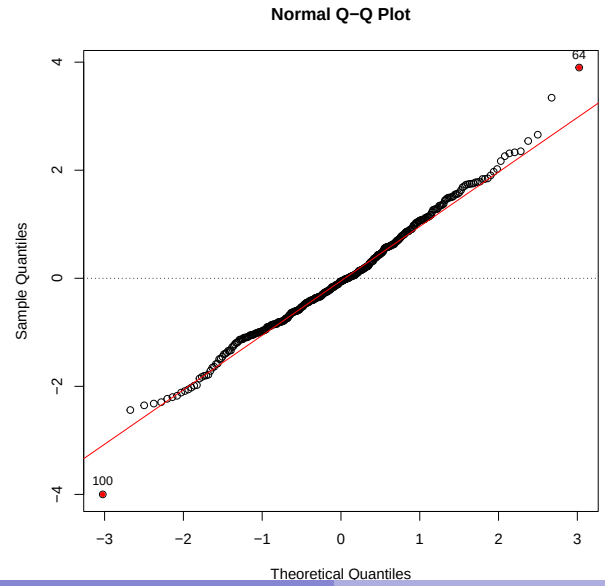
Le monde idéal pour les résidus

Illustration : présence de points aberrants.







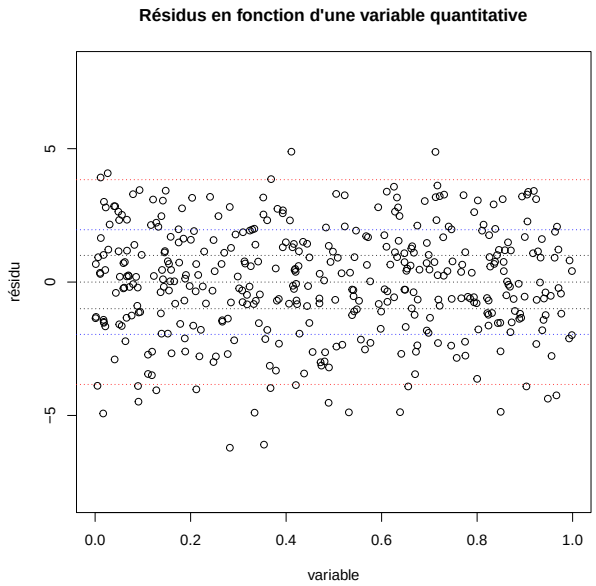


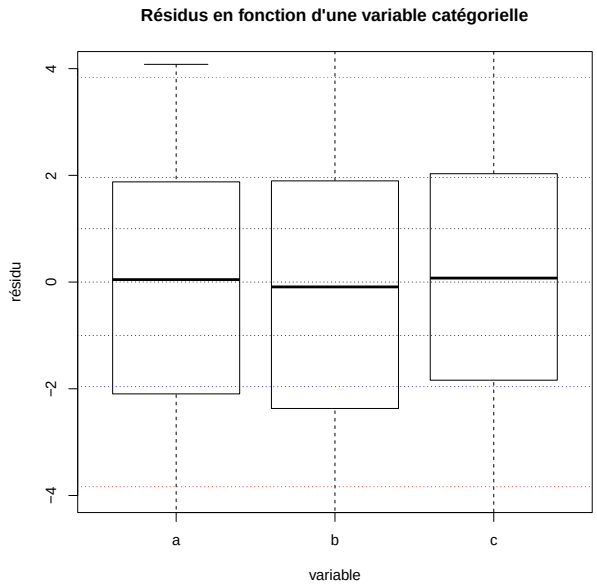
Le monde idéal pour les résidus

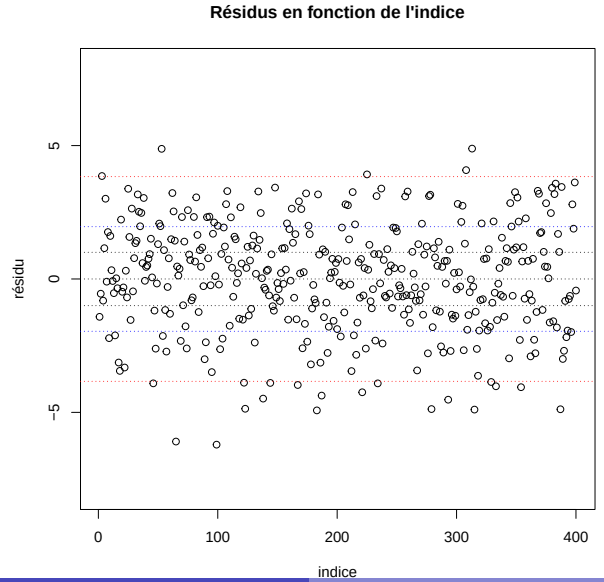
La troisième illustration présente le cas d'un modèle invalide en raison d'une sur-dispersion globale : la variance des résidus est globalement supérieure à celle attendue, les autres propriétés étant satisfaites.

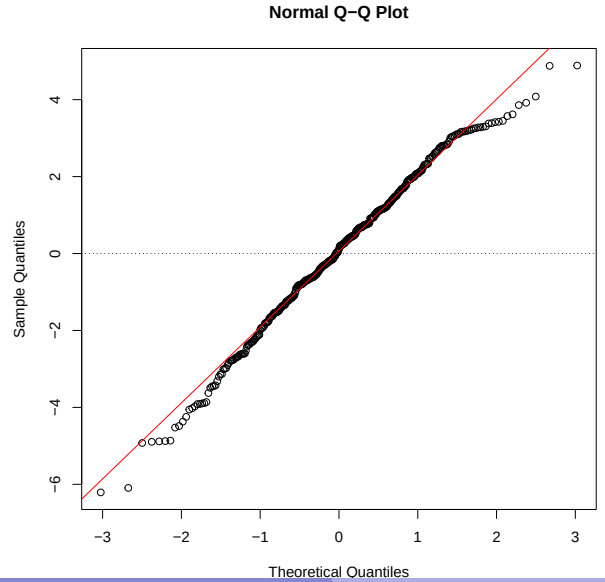
Le monde idéal pour les résidus

Illustration : surdispersion globale







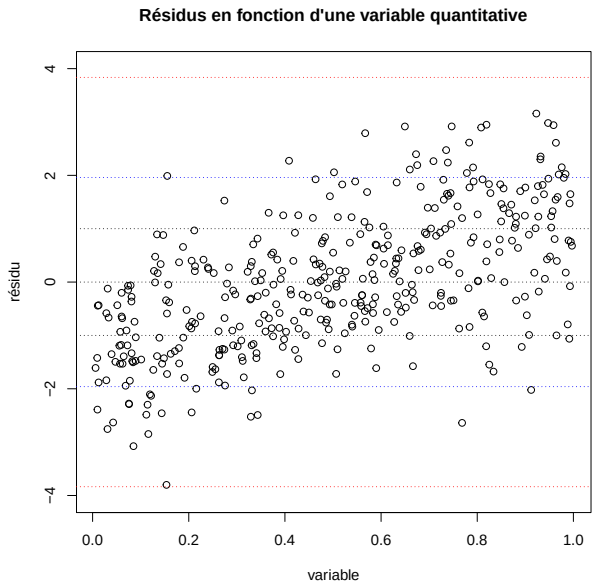


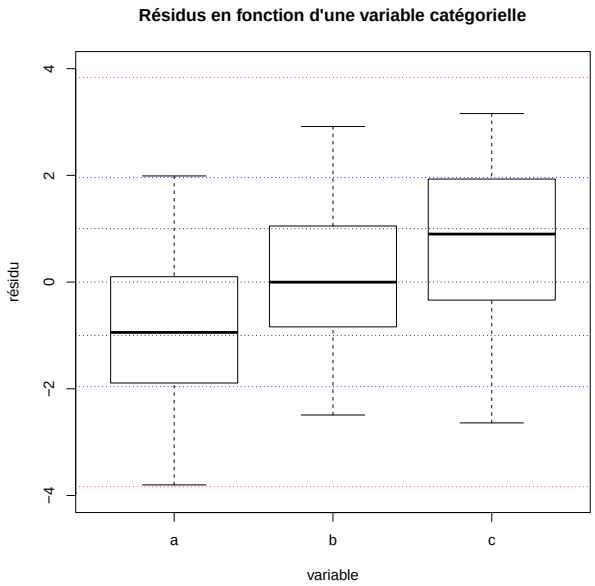
Le monde idéal pour les résidus

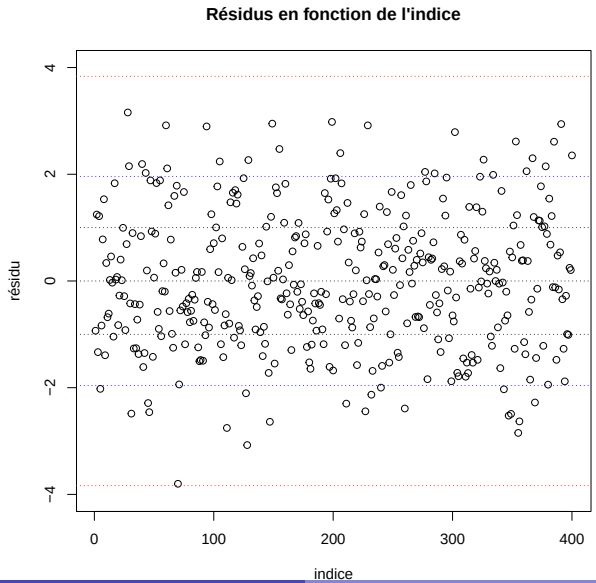
La quatrième illustration présente le cas d'un modèle invalide en raison d'une tendance sur l'espérance : l'espérance des résidus varie en fonction des variables explicatives.

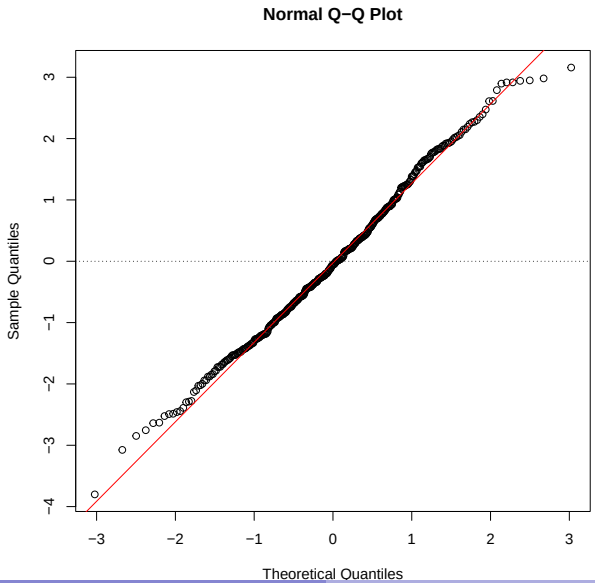
Le monde idéal pour les résidus

Illustration : tendance sur l'espérance







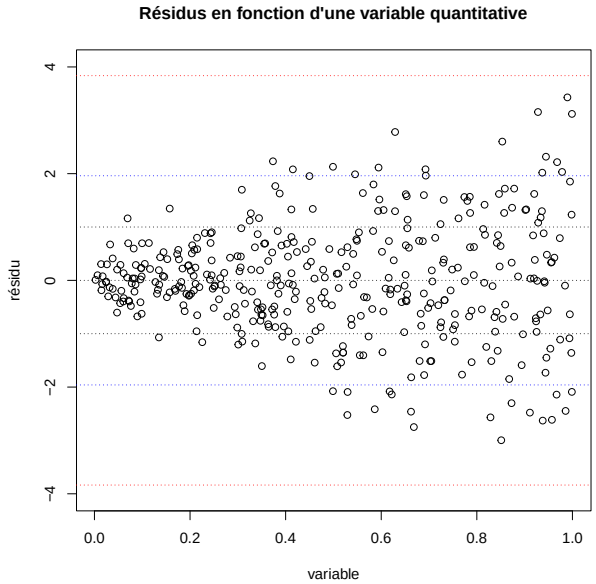


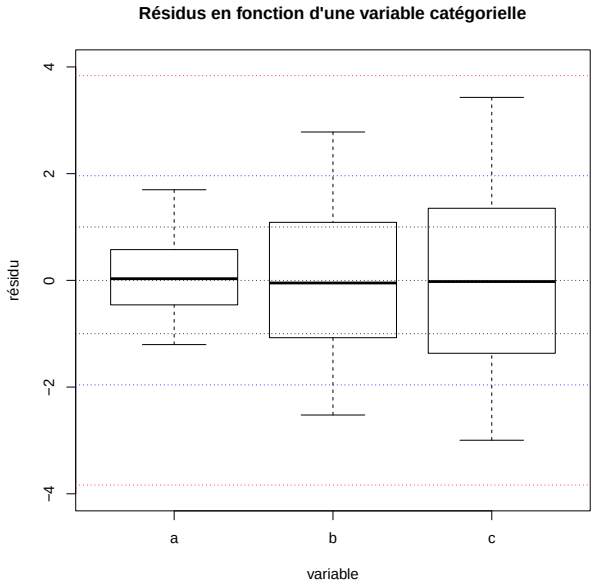
Le monde idéal pour les résidus

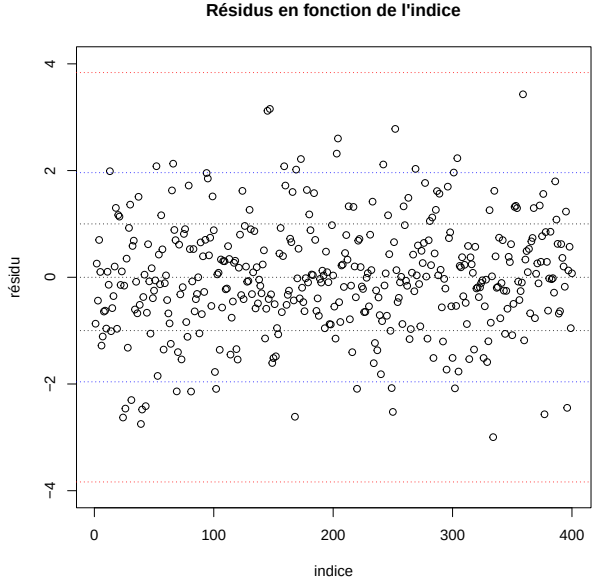
La cinquième illustration présente le cas d'un modèle invalide en raison d'une tendance sur la variance : la variance des résidus varie en fonction des variables explicatives.

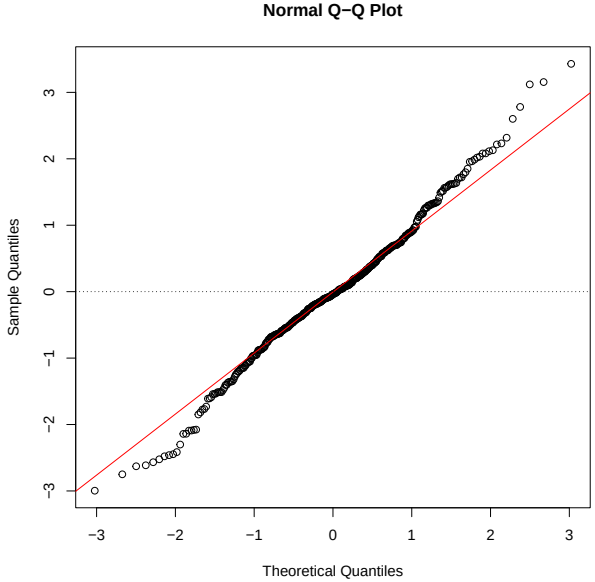
Le monde idéal pour les résidus

Illustration : tendance sur la variance







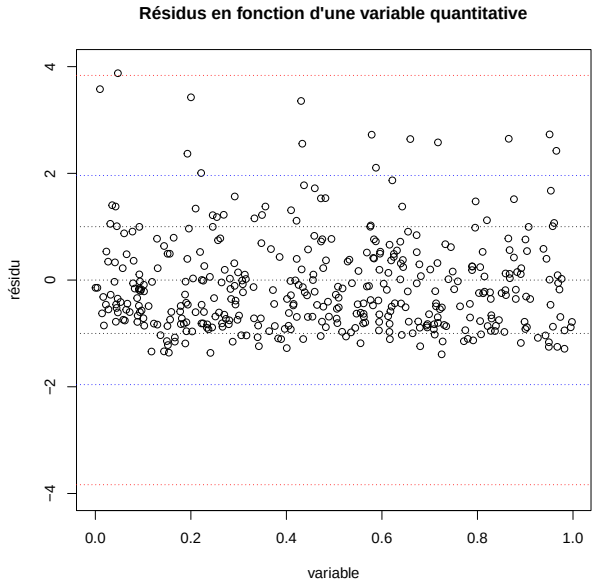


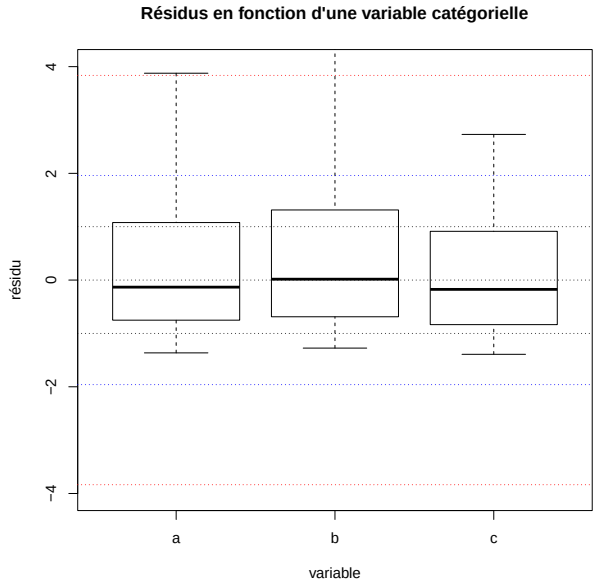
Le monde idéal pour les résidus

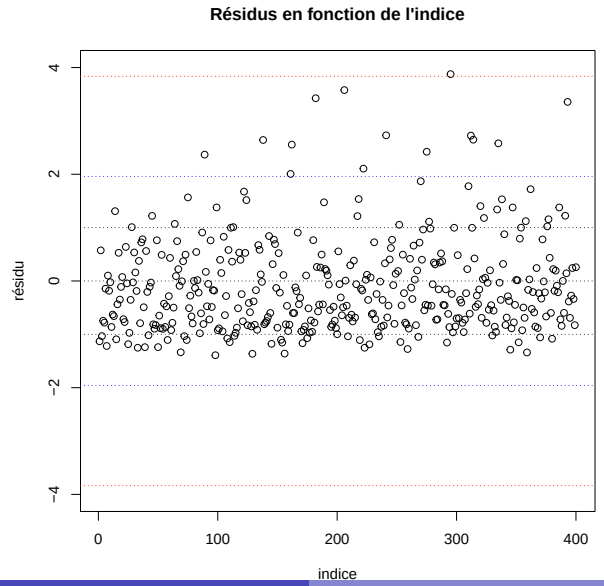
La sixième illustration présente le cas d'un modèle invalide en raison d'une dissymétrie : la loi des résidus n'est pas symétrique.

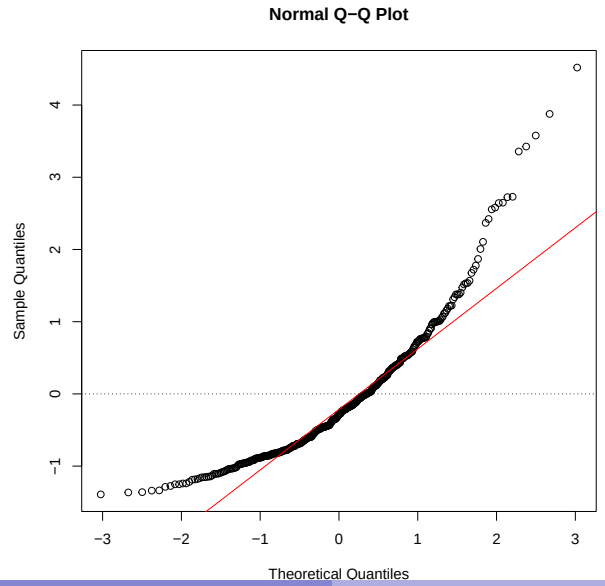
Le monde idéal pour les résidus

Illustration : dissymétrie









Le monde réel

Malheureusement, dans le cadre des modèles linéaires généralisés, on ne dispose pas d'une définition des résidus permettant, lorsque le modèle est valide, de satisfaire exactement les propriétés précédentes (résidus centrés, de variance unité, normaux, indépendants). Ce cadre idéalisé reste néanmoins souvent utilisé comme une référence approximative.

Dans ce qui suit, on va définir plusieurs types de résidus, présentant une proximité variable avec le cadre idéalisé précédent.

Afin de pouvoir examiner ces résidus de manière pertinente, il est important de savoir quelles propriétés sont (approximativement) attendues, afin de ne pas sur-interpréter des écarts acceptables et de ne pas sous-estimer des écarts problématiques !

Ecarts entre réponse observée et espérance modélisée

De manière générale, en supposant le modèle valide, on peut écrire :

$$Y_i - \hat{\mu}_i = \underbrace{Y_i - \mu_i}_{\text{partie aléatoire}} + \underbrace{\mu_i - \hat{\mu}_i}_{\text{erreur d'estimation}},$$

et les deux termes de la décomposition sont importants.

- La partie aléatoire provient de l'aléa résiduel sur la réponse une fois les variables explicatives prises en compte
- L'erreur d'estimation provient du fait que, même en supposant la validité du modèle GLM utilisé et un jeu de données suffisamment grand, les paramètres du modèle ne peuvent pas être déterminés avec certitude
- Selon les situations, l'un ou l'autre terme peut dominer

Résidus « bruts » (« raw » residuals)

On commence par deux types de résidus très simples :

- Résidu « brut » : $R_i^B \overset{\text{déf.}}{=} Y_i - \hat{\mu}_i$
- Résidu « O/P » : $R_i^{O/P} \overset{\text{déf.}}{=} \left(\frac{Y_i}{\hat{\mu}_i} - 1\right) = r_i^B / \hat{\mu}_i$ (typiquement pour des modèles où on doit avoir $\hat{\mu}_i > 0$)

Ces résidus fournissent une première quantification précise des écarts entre valeur moyenne prédite et valeur observée. Si le modèle est valide, dans la limite d'un grand nombre de données (et sous certaines hypothèses de régularité), on s'attend à ce qu'ils se comportent **approximativement** comme des v.a. **centrées** et **indépendantes**.
Néanmoins, même en supposant le modèle parfaitement valide, la variance attendue de ces résidus est susceptible de varier fortement d'un résidu à l'autre, selon la valeur de μ_i.

Résidus de Pearson

Afin de disposer de résidus qui, si le modèle est valide, dans la limite d'un grand nombre de données (et sous certaines hypothèses de régularité), se comportent **approximativement** comme des v.a. **centrées**, **indépendantes** et de **variance constante**, on introduit les résidus de Pearson :

$$R_i^P \overset{\text{déf.}}{=} \frac{Y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)/w_i}}.$$

La terminologie concernant les résidus est assez peu homogène d'une référence à l'autre.
Lorsque ϕ n'est pas *a priori* égal à 1, on peut aussi considérer $\widetilde{R_i^P} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\phi v(\hat{\mu}_i)/w_i}}$, et (lorsque ϕ doit être estimé) $\widetilde{R_i^P}' = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} v(\hat{\mu}_i)/w_i}}$, qui sont également appelés résidus de Pearson.
Dans tous les cas, l'idée est de diviser $Y_i - \hat{\mu}_i$ par l'écart-type estimé de $Y_i - \mu_i$ (éventuellement à une constante multiplicative près). En supposant la validité du modèle, dans la limite d'un grand nombre de données, et sous des hypothèses de régularité, les résidus de Pearson doivent se comporter approximativement comme des v.a. centrées, de variance ϕ (pour R_i^P) ou 1 (pour $\widetilde{R_i^P}$ et $\widetilde{R_i^P}'$), et indépendantes.
En multipliant les résidus précédents par $\frac{1}{\sqrt{1-\hat{h}_i}}$, où $\hat{h}_i$ provient de la « hat-matrix » définie plus loin, on obtient des résidus de Pearson dits « standardisés », qui incluent un facteur correctif destiné à rapprocher la variance de sa valeur constante supposée.

Résidus de déviance

On définit les résidus de déviance par :

$$R_i^D \overset{\text{déf.}}{=} \text{signe}(Y_i - \hat{\mu}_i) \cdot \sqrt{w_i \cdot D(Y_i, \hat{\mu}_i)}.$$

On a alors la formule :

$$D = \sum_{i=1}^{N} \left(R_i^D\right)^2,$$

dans laquelle les résidus de déviance jouent le même rôle que les résidus de Pearson pour la formule

$$\mathcal{X}^2 = \sum_{i=1}^{N} \left(R_i^P\right)^2.$$

La commande `glm` de R renvoie par défaut quelques indicateurs (min., max., médiane, 1er et 3ème quartiles) de la distribution empirique des résidus de déviance calculés sur les données ayant servi à ajuster le modèle.

Comme pour les résidus de Pearson, on peut considérer des résidus de déviance normalisés par ϕ ou $\hat{\phi}$, éventuellement standardisés en multipliant par le facteur correctif $\frac{1}{\sqrt{1-\hat{h}_i}}$.

Résidus d'Anscombe

On définit les résidus d'Anscombe par

$$R_i^A \overset{\text{déf.}}{=} \frac{A(Y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{v(\hat{\mu}_i)/w_i}},$$

où la fonction A vérifie

$$A'(y) = v(y)^{-1/3}.$$

Résidus agrégés

De manière générale, aucune des définitions précédentes ne conduit à des résidus censés – lorsque le modèle est valide, et dans la limite d'un grand nombre de données – suivre approximativement une loi normale [16].
Cette propriété ne se manifeste pour le résidu R_i que dans la limite où l'exposition totale associée à la ligne i du jeu de données tend vers $+\infty$.
Pour se rapprocher de cette situation, on peut considérer des **résidus agrégés** (« crunched residuals ») dans lesquels on mesure l'écart entre l'espérance modélisée et la réponse totale observée pour un groupe de données G représentant une exposition totale suffisante. On aura par exemple le résidu de Pearson agrégé sur le groupe G :

$$R_G^P \overset{\text{déf.}}{=} \frac{\sum_{i\in G} w_i Y_i - \sum_{i\in G} w_i \hat{\mu}_i}{\sqrt{\sum_{i\in G} w_i v(\hat{\mu}_i)}}.$$

16. Penser par exemple au cas où Y_i suit une loi de Poisson de paramètre $\mu \lessapprox 1$: le caractère discret des valeurs de Y_i est incontournable.

Résidus quantiles randomisés normalisés

Une autre manière d'obtenir des résidus censés, lorsque le modèle est valide, et dans la limite d'un grand nombre de données, suivre approximativement une loi normale, est de considérer les résidus quantile randomisés normalisés ([DS96])

$$R_i^{QR} \overset{\text{déf.}}{=} F^{-1}_{\mathcal{N}(0,1)}(\hat{F}_i(Y_i)),$$

où $\hat{F}_i$ est la fonction de répartition estimée pour la réponse (paramètres $\hat{\mu}_i$ et ϕ/w_i ou $\hat{\phi}/w_i$), randomisée dans le cas d'une loi discrète.
Lorsque le modèle est valide :

- à l'approximation de μ_i par $\hat{\mu}_i$ (et de ϕ par $\hat{\phi}$ le cas échéant) près, les résidus quantile randomisés normalisés doivent suivre une loi $\mathcal{N}(0,1)$
- pour une réponse de loi discrète, les résidus quantile randomisés normalisés introduisent un bruit aléatoire supplémentaire, non présent dans les données

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-D.html`

Données influentes

Une donnée (y_i, x_i) est dite **influente** quand elle est susceptible de produire à elle seule un impact non-négligeable sur :

- la qualité globale de l'ajustement
- la valeur des paramètres estimés
- les valeurs prédites par le modèle

L'un des outils techniques permettant de mesurer l'influence des données dans le cadre GLM est l'extension à ce cadre de la « hat-matrix » définie en modèle linéaire.

La « hat-matrix »

La « hat-matrix » est définie dans le cadre GLM par :

$$H \overset{\text{déf.}}{=} W^{1/2}\mathfrak{X} \underbrace{\left({}^t\mathfrak{X}W\mathfrak{X}\right)^{-1}}_{=I(\beta)^{-1}\approx\Sigma(\hat{\beta})} {}^t\mathfrak{X}W^{1/2}$$

avec (rappel) :

- $\mathfrak{X} = (\mathrm{x}_i^{(j)})_{\substack{1\leq i\leq N\\ 0\leq j\leq d}}$ vu comme une matrice $N \times (d+1)$
- $W = \operatorname{diag}\left[\frac{w_i}{\phi v(\mu_i)}\left(\frac{1}{g'(\mu_i)}\right)^2,\ 1 \leq i \leq N\right]$

En remplaçant [17] μ_i par $\hat{\mu}_i$, on obtient un estimateur de la « hat-matrix » $\hat{H} \approx H$, et l'on définit, pour $i = 1, \ldots, N$, les **coefficients de levier** :

$$\hat{h}_i \overset{\text{déf.}}{=} \hat{H}_{ii}.$$

On note la relation (due au fait que $\hat{H}$ est une matrice de projection)

$$\sum_{i=1}^{N} \hat{h}_i = d+1.$$

17. Quant à ϕ, sa valeur se simplifie dans l'expression de H donnée ci-dessus, si bien que H ne dépend donc pas de ϕ.

Influence et hat-matrix

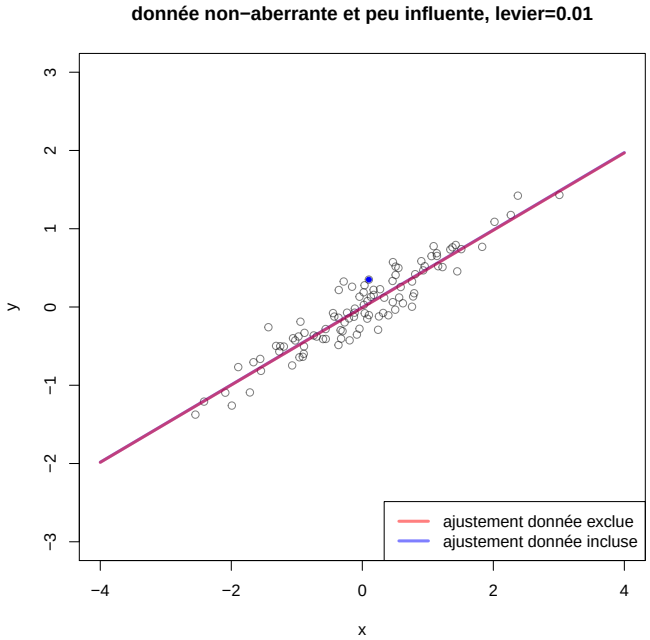
Lorsque le modèle est valide, dans la limite d'un grand nombre de données (et sous des hypothèses de régularité), on a l'approximation :

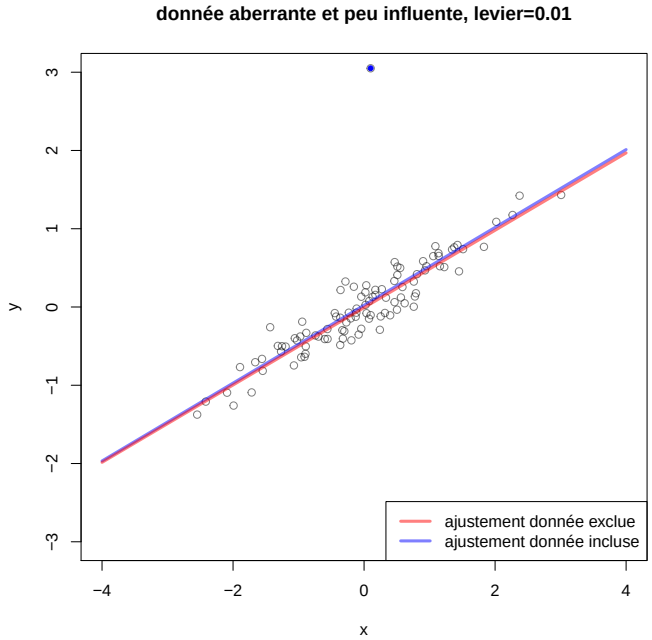
$$\left[\frac{\hat{\mu}_i - \mu_i}{\sqrt{v(\mu_i)}}\right]_{1\leq i\leq N} \approx \hat{H} \times \left[\frac{Y_i - \mu_i}{\sqrt{v(\mu_i)}}\right]_{1\leq i\leq N}.$$

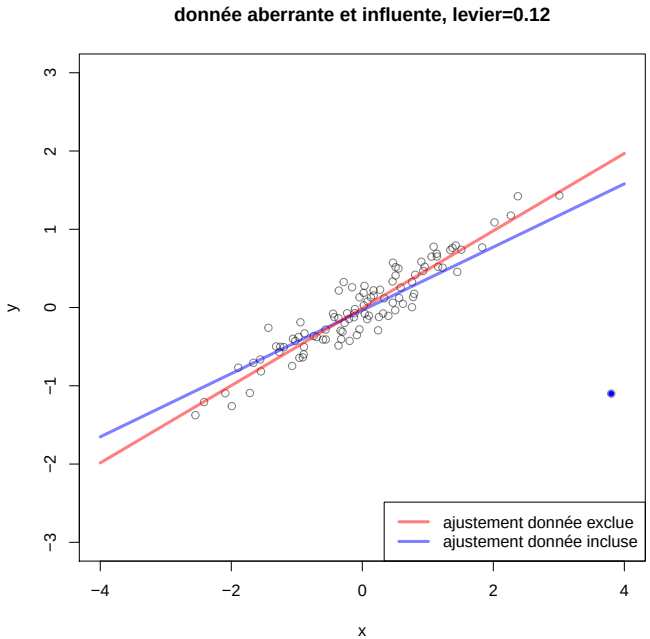
Cette relation étend au cadre GLM la relation analogue valable en modèle linéaire. La matrice $\hat{H}$ permet ainsi en un certain sens de mesurer l'influence des données individuelles (y_i, x_i) sur les valeurs prédites par le modèle. Spécifiquement, le coefficient de levier numéro i, $\hat{h}_i$ fournit une mesure de l'influence de (y_i, x_i) sur $\mu_i^{\text{est.}}$, pour un i donné.

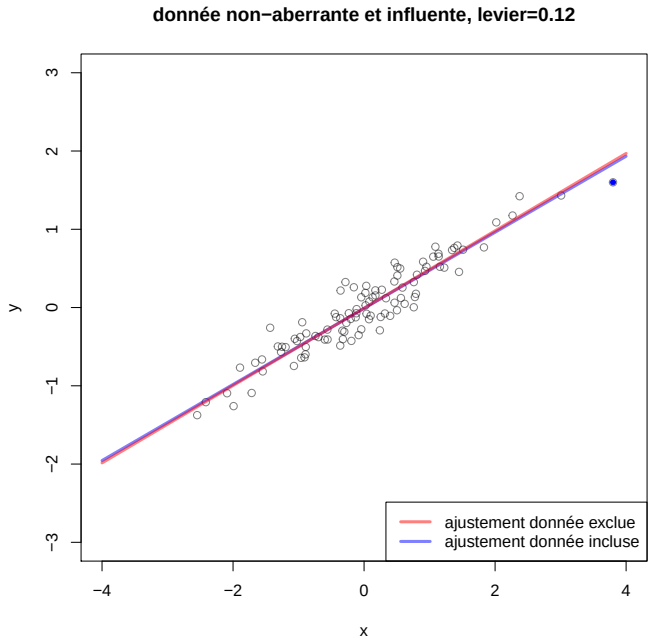
Néanmoins, contrairement au cas du modèle linéaire, il ne s'agit que d'une identité approchée, valable uniquement lorsque certaines hypothèses sont satisfaites, et de plus la matrice $\hat{H}$ dépend des Y_i et non pas uniquement des variables explicatives.

Illustration (en modèle linéaire simple)









Distance de Cook et DFBETA(S)

Idée : mesurer l'impact individuel d'une donnée sur l'estimation des coefficients

$\hat{\beta}^{[-i]} \overset{\text{déf.}}{=}$ estimateur de $\beta = (\beta_0, \ldots, \beta_d)$ après suppression de la ligne i.

On introduit :

- $\text{DFBETA}_i \overset{\text{déf.}}{=} \hat{\beta}^{[-i]} - \hat{\beta}$
- $\text{DFBETAS}_i \overset{\text{déf.}}{=} \left((\hat{\beta}_j^{[-i]} - \hat{\beta}_j) / \hat{\sigma}(\hat{\beta}_j^{[-i]}) \right)_{0 \leq j \leq d+1}$

Plutôt que de réajuster N fois le modèle, ce qui peut être numériquement impraticable, on utilise couramment l'approximation :

$$\hat{\beta}^{[-i]} \approx \hat{\beta} - (\sqrt{\widehat{W}_{ii}} \cdot R_i^P) \hat{\Sigma}(\hat{\beta}) \times \mathrm{x}_i / (1 - \hat{h}_i)$$

Distance de Cook et DFBETA(S)

La distance de Cook associée à la ligne i du jeu de données est définie par :

$$C_i \overset{\text{déf.}}{=} \frac{1}{d+1} {}^t(\hat{\beta} - \hat{\beta}^{[-i]}) \times \hat{\Sigma}(\hat{\beta})^{-1} \times (\hat{\beta} - \hat{\beta}^{[-i]})$$

Cette formule est une version normalisée de la distance $||\hat{\beta} - \hat{\beta}^{[-i]}||^2$ adaptée à la structure de covariance de $\hat{\beta}$.

On utilise couramment l'approximation :

$$C_i \approx \frac{1}{d+1} \frac{\left(R_i^P\right)^2}{\hat{\phi}} \frac{\hat{h}_i}{(1-\hat{h}_i)^2}$$

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-D-bis.html`

Triade Wald/LRT/Score

Dans le cadre d'une modélisation GLM, il est courant de tester diverses hypothèses portant sur les coefficients au moyen des tests suivants :

- Test de Wald
- Test du rapport de vraisemblance
- Test du score

Ces trois tests reposent sur des propriétés asymptotiques, valables dans la limite d'un grand nombre de données, et sous des hypothèses de régularité.

Triade Wald/LRT/Score

De manière générale, on suppose *a priori* la validité d'un modèle GLM associé au jeu de coefficients $\beta = (\beta_0, \ldots, \beta_d)$, et l'on cherche dans ce cadre à tester une hypothèse (H_0) sur les coefficients, de la forme

$$(H_0) \ : \ h(\beta) = 0,$$

où :

- $h(\beta) = \begin{pmatrix} h_1(\beta) \\ \vdots \\ h_r(\beta) \end{pmatrix}$ est une fonction « régulière » de β
- l'hypothèse alternative est $(H_1) \ : \ h(\beta) \neq 0$

Exemples :

- test de $\beta_k = 0$: $h(\beta) = \beta_k$ $(r = 1)$
- test de $\beta_{\ell_1} = \cdots = \beta_{\ell_b} = 0$: $h(\beta) = (\beta_{\ell_1}, \ldots, \beta_{\ell_b})$ $(r = b)$
- test de $\beta_1 = \beta_2$: $h(\beta) = \beta_1 - \beta_2$ $(r = 1)$

Test de Wald

Rappel : sous des hypothèses de régularité, dans la limite d'un grand nombre de données, on a la propriété :

$$\text{Loi}(\hat{\beta} - \beta) \approx \mathcal{N}(0, \Sigma = I(\beta)^{-1}) \approx \mathcal{N}(0, \Sigma = I(\hat{\beta})^{-1})$$

Principe du test de Wald pour $h(\beta) = 0$

$$\text{Sous } (H_0), \text{ on a : Loi } \left({}^t h(\hat{\beta}) \times Q(\hat{\beta})^{-1} \times h(\hat{\beta}) \right) \approx \chi^2(r)$$

avec :

- $Q(\beta) = B(\beta) \times I(\beta)^{-1} \times {}^t B(\beta)$
- $B(\beta) := \left[\frac{\partial h_\ell}{\partial \beta_j} \right]_{\substack{1 \leq \ell \leq r \\ 0 \leq j \leq d}}$

Le test consiste à situer la valeur observée de $T_{\text{Wald}} = {}^t h(\hat{\beta}) \times Q(\hat{\beta})^{-1} \times h(\hat{\beta})$ vis-à-vis de la loi $\chi^2(r)$. La $p-$valeur du test est donnée par $1 - F_{\chi^2(r)}(T_{\text{Wald}}^{\text{obs.}})$.

Test de Wald pour la nullité des coefficients

La commande `glm` de R effectue systématiquement un test de Wald de l'hypothèse $\beta_j = 0$ pour chacun des coefficients $\beta_j, j = 0, \ldots, d$ estimés. Dans ce cas particulier, la $p-$valeur du test est simplement donnée par $\mathbb{P}(|Z| \geq |z_j|)$, où $Z \sim \mathcal{N}(0,1)$, et $z_j = \hat{\beta}_j^{\text{obs.}} / \hat{\sigma}(\hat{\beta}_j)^{\text{obs.}}$.

L'obtention d'un résultat « statistiquement significatif » pour le coefficient β_j correspond simplement au fait que l'hypothèse $\beta_j = 0$ produit une $p-$valeur de test suffisamment faible pour conduire à un rejet de cette hypothèse, un niveau de risque (de première espèce) étant fixé.

Attention alors au mirage des $***$: le fait qu'un modèle GLM ajusté donne lieu à des résultats de test de Wald fortement significatifs pour chacun des coefficients **ne garantit en aucun cas** la validité ou la qualité du modèle. Le travail de validation, d'amélioration et d'évaluation du modèle reste à mener !

Dans le cas où, pour certains coefficients, les résultats de tests de Wald sont non-significatifs, le travail de validation, d'amélioration et d'évaluation du modèle n'en est pas moins à mener, mais il doit également tenir compte du possible problème posé par ces résultats.

Test du rapport de vraisemblance

Principe du test du rapport de vraisemblance pour $h(\beta) = 0$

$$\text{Sous } (H_0), \text{ on a : Loi } \left(2\left[\log L(\hat{\beta}) - \log L(\tilde{\beta})\right]\right) \approx \chi^2(r)$$

avec :

- $\hat{\beta}$: le jeu de paramètres estimé par max. de vraisemblance
- $\tilde{\beta}$ le jeu de paramètres estimé par max. de vraisemblance **sous la contrainte** $h(\beta) = 0$

Le test consiste à situer la valeur obtenue pour $T_{\text{LRT}} = 2\left[\log L(\hat{\beta}) - \log L(\tilde{\beta})\right]$ vis-à-vis de la loi $\chi^2(r)$. La $p-$valeur du test est donnée par $1 - F_{\chi^2(r)}(T_{\text{LRT}}^{\text{obs.}})$.

Lorsque la valeur de ϕ est fixée *a priori*, le calcul de L dépend simplement de la valeur des coefficients estimés $\hat{\beta}$ et $\tilde{\beta}$. Lorsque ϕ doit être estimé, une possibilité est d'utiliser la valeur $\hat{\phi}$ et l'approximation $T_{\text{LRT}} \approx 2\left[\log L(\hat{\beta}, \hat{\phi}) - \log L(\tilde{\beta}, \hat{\phi})\right]$ (plutôt que $2\left[\log L(\hat{\beta}, \hat{\phi}) - \log L(\tilde{\beta}, \tilde{\phi})\right]$). Noter le lien avec la déviance : $L(\hat{\beta}, \phi) - L(\tilde{\beta}, \phi) = D(\hat{\beta})/\phi - D(\tilde{\beta})/\phi$, et la **différence des log-vraisemblances** est donc égale à la **différence des déviances normalisées**.

Test du score

Principe du test du score pour $h(\beta) = 0$

$$\text{Sous } (H_0), \text{ on a : Loi } \left({}^t\nabla_{\tilde{\beta}}(\log L) \times I(\tilde{\beta})^{-1} \times \nabla_{\tilde{\beta}}(\log L) \right) \approx \chi^2(r)$$

avec :

- $\tilde{\beta}$ le jeu de paramètres obtenu par max. de vraisemblance **sous la contrainte** $h(\beta) = 0$

Le test consiste à situer la valeur de $T_{\text{score}} = {}^t\nabla_{\tilde{\beta}}(\log L) \times I(\tilde{\beta})^{-1} \times \nabla_{\tilde{\beta}}(\log L)$ vis-à-vis d'une loi $\chi^2(r)$. La $p-$valeur du test est donnée par :

$$1 - F_{\chi^2(r)}(T_{\text{score}}^{\text{obs.}}).$$

Remarques

- Les trois tests Wald/LRT/Score donnent en général des résultats proches sur des jeux de données volumineux (asymptotiquement équivalents lorsque $N \to +\infty$)
- Les trois tests reposent sur des approximations asymptotiques, et supposent certaines hypothèses de régularité (p. ex. : hypothèse de point intérieur pour le LRT)
- le test du score suppose uniquement l'ajustement du modèle avec contrainte, le test du rapport de vraisemblance suppose l'ajustement du modèle avec contrainte et du modèle sans contrainte, le test de Wald suppose uniquement l'ajustement du modèle sans contrainte
- Ils peuvent s'étendre en des tests portant sur d'autres paramètres que les coefficients β (dispersion ϕ, paramètres de la fonction de variance ou de lien, etc.)
- Leur utilisation directe est limitée à la comparaison de modèles **emboîtés**  (l'approche de Vuong [Vuo89] permet en principe des tests de comparaison entre modèles non emboîtés)

A propos des tests

De manière générale :

- une $p-$valeur supérieure au seuil de rejet **ne signifie pas** que (H_0) peut être inconditionnellement acceptée comme « vraie » ($\rightarrow$ quelle est la puissance du test c.à.d. sa capacité à distinguer (H_0) des alternatives ?).
- les seuils pour les $p-$valeurs (p. ex. 0.05) sont – dans une certaine mesure – conventionnels
- une $p-$valeur inférieure au seuil de rejet montre un écart *statistiquement* significatif des données par rapport à (H_0), mais pas nécessairement *pratiquement* significatif (ex. : coefficient très petit mais néanmoins statistiquement significatif dans un grand jeu de données).
- la multiplication des tests et l'orientation des tests par l'examen des données affectent les propriétés nominales du test
- les lois utilisées pour calculer les $p-$valeurs sont le plus souvent approximatives
- l'opposition entre (H_0) et (H_1) est le plus souvent testée par rapport à un ensemble d'hypothèses de référence qui ne sont pas explicitement mises en question par le test (indépendance, forme du modèle, etc.).

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-E.html`

Tests d'adéquation globale

But : vérifier par une procédure de test formelle l'adéquation globale entre modèle et données.

Typiquement, on considère une mesure quantitative d'ajustement entre modèle et données (D, $\mathcal{X}^2$,...) et l'on cherche à vérifier que la valeur observée de cette quantité est compatible avec le modèle, en la situant par rapport à sa loi de probabilité sous l'hypothèse que le modèle est valide.

Il est alors nécessaire de disposer (au moins d'une approximation) de cette loi, afin de pouvoir calculer la $p-$valeur du test.

Tests d'adéquation basé sur la déviance

- Sous certaines hypothèses  , on a, **en supposant la validité du modèle GLM**, l'approximation :

$$\text{Loi}(D/\phi) \approx \chi^2(N-(d+1))$$

- Dans ce cas, on peut effectuer un test global d'adéquation du modèle basé sur la déviance en calculant une $p-$valeur approchée

$$1 - F_{\chi^2(N-(d+1))}((D^{\text{obs.}}/\phi)$$

- L'approximation par une loi $\chi^2(N-(d+1))$ n'est **pas valable** de manière générale pour un grand jeu de données (voir annexe)
- D'autres approximations ou des méthodes de boostrap peuvent éventuellement être employées pour calculer des $p-$valeurs

Le cas de données binomiales non-groupées est un cas dégénéré pour ce type de test.

Tests d'adéquation basé sur la statistique de Pearson

- Sous certaines hypothèses  , on a, **en supposant la validité du modèle GLM**, l'approximation :

$$\text{Loi}(\mathcal{X}^2/\phi) \approx \chi^2(N-(d+1))$$

- Dans ce cas, on peut effectuer un test global d'adéquation du modèle basé sur la déviance en calculant une $p-$valeur approchée

$$1 - F_{\chi^2(N-(d+1))}((\mathcal{X}^2)^{\text{obs.}}/\phi)$$

- L'approximation par une loi $\chi^2(N-(d+1))$ n'est **pas valable** de manière générale pour un grand jeu de données (voir annexe)
- D'autres approximations ou des méthodes de boostrap peuvent éventuellement être employées pour calculer des $p-$valeurs

Tests d'adéquation basé sur des regroupements de données

Une autre catégorie de tests d'adéquation globale s'appuie sur un regroupement préalable des données en classes de taille comparable, après tri par ordre croissant des espérances modélisées.
Des $p-$valeurs approchées (approximation asymptotique, ou bootstrap) peuvent alors être calculées.
Un exemple classique est le test de Hosmer-Lemeshow en régression logistique binaire :

$$T_{\mathsf{HL}} = \sum_{j=1}^{K} \frac{\left(\sum_{i \in G_j} y_{(i)} - \hat{\mu}_{(i)}\right)^2}{\sum_{i \in G_j} \hat{\mu}_{(i)}(1 - \hat{\mu}_{(i)})},$$

où les $G_1, \ldots, G_K$ forment une partition de $\{1, \ldots, N\}$ en K sous-intervalles consécutifs de tailles approximativement égales, et les indices (i) désignent le tri des données par ordre croissant des valeurs prédites : $\hat{\mu}_{(1)} \leq \cdots \leq \hat{\mu}_{(N)}$.

La $p-$valeur du test de Hosmer-Lemeshow est calculée en utilisant l'approximation selon laquelle, en supposant la validité du modèle, $\text{Loi}(T_{\mathsf{HL}}) \approx \chi^2(K-2)$.

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-F.html`

Adéquation et performance prédictive

On a présenté auparavant différentes approches (visualisation des résidus, tests statistiques) visant à contrôler sur les données l'absence de violations (facilement) détectables des hypothèses du modèle.

Dans cette partie, on s'intéresse à des méthodes visant à estimer directement les performances du modèle en matière de prédiction, afin de pouvoir comparer entre elles les performances de différents modèles.

Performance prédictive

Concrètement, on considère une fonction d'écart ρ, par exemple :

$$\rho(y, \mu, w) = \begin{cases} w \cdot D(y, \mu) & (\rho^D) \\ w \cdot (y - \mu)^2 & (\rho^{SS}) \\ w \cdot \frac{(y-\mu)^2}{v(\mu)} & (\rho^P) \end{cases},$$

et, dans l'idéal, on souhaiterait évaluer une quantité telle que :

$$\rho_{\text{préd.}} = \sum_{i=1}^{N} \mathbb{E}\left[\rho(Y_i', \hat{\mu}_i^{\text{obs.}}, w_i)\right],$$

où $(Y_i')_{1 \leq i \leq N}$ désigne une suite de v.a. de même loi que $(Y_i)_{1 \leq i \leq N}$ et indépendante de celle-ci.
Les $(Y_i')_{1 \leq i \leq N}$ constituent un jeu de « nouvelles données » possédant les mêmes propriétés statistiques que les données utilisées pour l'ajustement.
La quantité $\rho_{\text{préd.}}$ représente donc une mesure de la **performance prédictive** du modèle.

Performance prédictive

Telle quelle, la valeur exacte de

$$\rho_{\text{préd.}} = \sum_{i=1}^{N} \mathbb{E}\left[\rho(Y_i', \hat{\mu}_i^{\text{obs.}}, w_i)\right]$$

est inaccessible, mais on pourrait envisager de l'approcher par :

$$\rho_{\text{ajust.}} = \sum_{i=1}^{N} \rho(y_i, \hat{\mu}_i^{\text{obs.}}, w_i).$$

Du fait que $\rho_{\text{ajust.}}$ mesure un écart entre valeurs observées y_i et valeurs prédites par le modèle $\hat{\mu}_i$ sur les **mêmes données** que celles ayant servi à ajuster le modèle, $\rho_{\text{ajust.}}$ a tendance à sous-estimer $\rho_{\text{préd.}}$, et ce d'autant plus que le modèle ajusté est plus flexible [more on this later].
D'autres approches sont donc nécessaires pour comparer correctement les performances prédictives de modèles présentant des degrés variables de flexibilité.

Utilisation d'un échantillon de test

Idée : séparer le jeu de données en deux parties, l'une pour effectuer l'ajustement du modèle, et l'autre pour mesurer la proximité entre valeurs prédites et observées.

- Avantages : simplicité ! L'indépendance (supposée) entre échantillon d'ajustement et échantillon de test assure une évaluation fiable des performances prédictives (sur des cas futurs ayant des propriétés statistiques identiques aux données utilisées)
- Inconvénients : la perte de précision dans l'ajustement **et** dans l'évaluation des performances, liée au fait que l'on n'utilise qu'une partie des données pour chaque tâche (à apprécier selon la taille et la complexité des données et du modèle)

 Sur des données fortement groupées et/ou insuffisamment nombreuses, les inconvénients évoqués ci-dessus peuvent s'avérer rédhibitoires.

Utilisation d'un échantillon de test

Concrètement, on considère une fonction d'écart $\rho(y, \mu, w)$, et l'on calcule

$$\rho_{\text{Test}} = \sum_{i \in I_{\text{Test}}} \rho(y_i, \hat{\mu}_i^{[\text{Aj.}]}, w_i),$$

où :

- l'ensemble des indices $\{1, \ldots, N\}$ est partitionné en deux sous-ensembles $I_{\text{Ajust.}}$ et I_{Test}, le partitionnement étant généralement produit par tirage aléatoire ;
- $\hat{\mu}_i^{[\text{Aj.}]}$ est la valeur moyenne prédite pour Y_i, par le modèle ajusté **uniquement** sur les données $(y_i, x_i)_{i \in I_{\text{Ajust.}}}$

Validation croisée

Idée : répéter plusieurs fois la séparation du jeu de données en échantillon d'ajustement et de test, et sommer les performances ainsi obtenues.

- Avantages : on utilise l'intégralité des données pour l'ajustement et l'évaluation des performances prédictives.
- Inconvénients : l'interprétation exacte du critère calculé est moins immédiate

Validation croisée « leave-one-out »

Etant donné une fonction d'écart $\rho(y, \mu, w)$,

$$\rho_{\text{LOOCV}} = \sum_{i=1}^{N} \rho(y_i, \hat{\mu}_i^{[-i]}, w_i),$$

où $\hat{\mu}_i^{[-i]}$ est la valeur moyenne prédite pour Y_i, par le modèle ajusté **uniquement** sur les données $(y_\ell, x_\ell)_{\ell \in \{1,\dots,N\} \setminus \{i\}}$

Validation croisée

Pour un jeu de données volumineux, il n'est pas forcément réalisable de calculer ρ_{LOOCV} en ré-ajustant N fois le modèle.

Pour $\rho = \rho^D$, on peut à la place utiliser l'approximation suivante utilisant les coefficients de levier (voir [Woo17], p. 262) :

$$\rho^{\text{approx.}}_{\text{LOOCV}} \overset{\text{déf.}}{=} \sum_{i=1}^{N} \rho^D(y_i, \hat{\mu}_i, w_i) + \frac{2\hat{h}_i}{1-\hat{h}_i}\rho^P(y_i, \hat{\mu}_i, w_i).$$

On peut aussi utiliser l'approximation supplémentaire consistant à remplacer les $\hat{h}_i$ par leur valeur moyenne : $\text{tr}(\hat{H})/N = (d+1)/N$, d'où le critère de **validation croisée généralisée** :

$$\rho^*_{\text{GCV}} = D + 2\frac{\text{tr}(\hat{H})}{N - \text{tr}(\hat{H})}\mathcal{X}^2.$$

Validation croisée

Une autre possibilité est de faire appel à la validation croisée de type « $K-$fold », qui ne nécessite qu'un nombre limité de ré-estimations du modèle (et est applicable avec tout modèle de régression, indépendamment d'une structure linéaire sous-jacente).

Validation croisée « $K-$fold »

On crée une partition (en général par tirage aléatoire) de $\{1, \ldots, N\}$ en K sous-ensembles $I_1, \ldots, I_K$ de tailles approximativement égales, et l'on considère

$$\rho_{K\text{-fold}} = \sum_{k=1}^{K} \sum_{i \in I_k} \rho(y_i, \hat{\mu}_i^{[-k]_K}, w_i),$$

où $\hat{\mu}_i^{[-k]_K}$ est la valeur moyenne prédite pour Y_i, par le modèle ajusté **uniquement** sur les données $(y_\ell, x_\ell)_{\ell \in \{1,\ldots,N\} \setminus I_k}$.

On utilise typiquement $K \in \{5, \ldots, 10\}$.Noter que de nombreuses extensions et variantes existent, notamment combinées à des techniques de type bootstrap.

Critère AIC (= « Akaike Information Criterion »)

De manière générale, étant donnée une variable aléatoire Z possédant une fonction de vraisemblance L^*, et une autre fonction de vraisemblance L, l'écart entre les lois de probabilité définies par L et L^* peut être mesuré via la divergence de Kullback-Leibler :

$$d_{\mathrm{KL}}(L^* || L\ ) = \mathbb{E}\left(\log\left(\frac{L^*(Z)}{L(Z)}\right)\right).$$

Dans notre contexte, la vraie fonction de vraisemblance L_i^* de Y_i est inconnue, et l'on cherche à l'approcher par la fonction de vraisemblance estimée $\hat{L}_i^{\text{obs.}} = L(\cdot, \hat{\mu}_i^{\text{obs.}}, \hat{\phi}_i^{\text{obs.}})$, avec $\hat{\phi}_i = \hat{\phi}/w_i$. En vue de mesurer la proximité entre la loi estimée et la loi de probabilité de nouvelles données (pour les distinguer des données utilisées pour produire $\hat{\mu}_i$ et $\hat{\phi}_i$), on s'intéresse à :

$$d_{\mathrm{KL}}(L^* || \hat{L}^{\text{obs.}}\ ) = \mathbb{E}\left(\sum_{i=1}^{N} \log\left(\frac{L_i^*(Y_i')}{L(Y_i', \hat{\mu}_i^{\text{obs.}}, \hat{\phi}_i^{\text{obs.}})}\right)\right)$$

Critère AIC

On écrit alors :

$$d_{\mathrm{KL}}(L^* || \hat{L}^{\text{obs.}}) = \mathbb{E}\left(\sum_{i=1}^{N} \log L_i^*(Y_i')\right) - \mathbb{E}\left(\sum_{i=1}^{N} \log L(Y_i', \hat{\mu}_i^{\text{obs.}}, \hat{\phi}_i^{\text{obs.}})\right)$$

Pour des valeurs ν_i fixées, on s'attend à avoir l'approximation

$$\mathbb{E}\left(\sum_{i=1}^{N} \log L(Y_i', \nu_i)\right) \approx \sum_{i=1}^{N} \log L(Y_i, \nu_i).$$

Le fait que les $\hat{\mu}_i$ et $\hat{\phi}_i$ soient estimés à partir des v.a. Y_i introduit un biais dans cette approximation si l'on prend $\nu_i = (\hat{\mu}_i, \hat{\phi}_i)$, (qui constituent ici des variables aléatoires et non pas des valeurs fixées). Ce biais peut être approximativement corrigé au moyen du terme correctif suivant :

$$\mathbb{E}\left(\sum_{i=1}^{N} \log L(Y_i', \hat{\mu}_i, \hat{\phi}_i) - \sum_{i=1}^{N} \log L(Y_i, \hat{\mu}_i, \hat{\phi}_i)\right) \approx -p,$$

où p est le nombre de paramètres estimés du modèle, ceux-ci étant estimés par maximum de vraisemblance. (Pour un GLM avec ϕ connu, $p = d + 1$).

Critère AIC

On aboutit finalement à l'approximation :

$$d_{\mathrm{KL}}(L^* || \hat{L}^{\text{obs.}}) \approx p - \sum_{i=1}^{N} \log L(Y_i, \hat{\mu}_i^{\text{obs.}}, \hat{\phi}_i^{\text{obs.}}) + \mathbb{E}\left(\sum_{i=1}^{N} \log L_i^*(Y_i') \right)$$

En posant

$$\mathrm{AIC} \overset{\text{déf.}}{=} -2 \sum_{i=1}^{N} \log L(Y_i, \hat{\mu}_i^{\text{obs.}}, \hat{\phi}_i^{\text{obs.}}) + 2p,$$

on aboutit donc à :

$$d_{\mathrm{KL}}(L^* || \hat{L}^{\text{obs.}}) \approx \tfrac{1}{2}\mathrm{AIC} + \mathbb{E}\left(\textstyle\sum_{i=1}^{N} \log L_i^*(Y_i') \right).$$

Critère AIC

Dans l'approximation

$$d_{\mathrm{KL}}(L^*||\hat{L}^{\text{obs.}}) \approx \tfrac{1}{2}\mathrm{AIC} + \mathbb{E}\left(\sum_{i=1}^{N} \log L_i^*(Y_i')\right).$$

le terme de droite ne dépend que de la vraie fonction de vraisemblance des données L_i^*. Etant donnés deux modèles concurrents donnant lieu à $\hat{L}_1^{\text{obs.}}$ et $\hat{L}_2^{\text{obs.}}$, on a :

$$d_{\mathrm{KL}}(L^*||\hat{L}_2^{\text{obs.}}) - d_{\mathrm{KL}}(L^*||\hat{L}_1^{\text{obs.}}) \approx \tfrac{1}{2}(\mathrm{AIC}_2 - \mathrm{AIC}_1).$$

On peut donc comparer les modèles sur la base de leurs AIC pour en déduire une comparaison sur leurs d_{KL} respectives. Un AIC plus faible s'interprétera comme le signe d'un modèle plus proche de la véritable loi engendrant les données.

Critère AIC

Quelques points d'attention :

- l'approximation du biais par p suppose que $\hat{L} \approx L^*$, et peut donc être pertinente pour départager entre eux des modèles « proches » de la véritable loi engendrant les données ;
- on peut simplement voir AIC comme un critère de vraisemblance incluant une pénalité en fonction du nombre de paramètres
- lorsque ϕ est connue *a priori*, le calcul de l'AIC d'un modèle GLM fait uniquement intervenir les coefficients estimés $\hat{\beta}$; lorsque ϕ doit être estimé, le critère AIC *stricto sensu* suppose que ϕ est également estimé par maximum de vraisemblance, et ϕ doit être comptabilisé parmi les paramètres. En toute rigueur, il n'est donc pas correct de calculer un tel AIC avec une valeur de ϕ estimée par une méthode de moments, comme R le fait en standard...
- AIC inclut une correction approximative du biais dans l'estimation de $d_{\mathrm{KL}}(L^* || \hat{L}^{\text{obs.}})$, mais quid de la variance...

Critère(s)

Récapitulatif des différents critères introduits précédemment, ainsi que quelques variantes. **Les divers ρ sont très souvent normalisés en les divisant par N** (ou par $|I_{\text{Test}}|$ pour ρ_{Test}).

- ρ_{Test} (suppose un échantillon d'ajustement et un échantillon de test)
- ρ_{LOOCV} (suppose N réajustements du modèle en supprimant chaque donnée tour-à-tour)
- $\rho^{\text{approx.}}_{\text{LOOCV}}$ (à l'aide des leviers du modèle ajusté)
- ρ^{*}_{GCV} (leviers remplacés par la valeur unique $\text{tr}(\hat{H})/N$)
- $\rho^{\text{approx.}}_{K-\text{fold}}$ (suppose seulement K réajustements du modèle)
- AIC (typiquement : ϕ connu et $d+1$ coefficients estimés par max. de vraisemblance)
- $\rho_{\text{UBRE}} \overset{\text{déf.}}{=} D + 2\phi\text{tr}(\hat{H})$ (utilisé notamment par le paquet `mgcv` pour les GAM)
- $\rho_{\text{GCV}} \overset{\text{déf.}}{=} \frac{D}{\left(1-\text{tr}(\hat{H})/N\right)^2}$ (utilisé notamment par le paquet `mgcv` pour les GAM)

Si (dans le cas où ϕ est connu *a priori*) on calcule les quantités $\widetilde{\text{AIC}}$ et $\widetilde{\rho_{\text{UBRE}}}$ en remplaçant ϕ par sa valeur estimée via la statistique de Pearson, on a $\widetilde{\text{AIC}} = \text{cte} + \frac{\rho^{*}_{\text{GCV}}}{\hat{\phi}}$ et $\widetilde{\rho_{\text{UBRE}}} = \rho^{*}_{\text{GCV}}$.
Si ϕ est remplacé par sa valeur estimée via la déviance, on a $\widetilde{\rho_{\text{UBRE}}} \approx \rho_{\text{GCV}}$ (pour $\text{tr}(\hat{H})/N << 1$).

Courbe de Lorenz

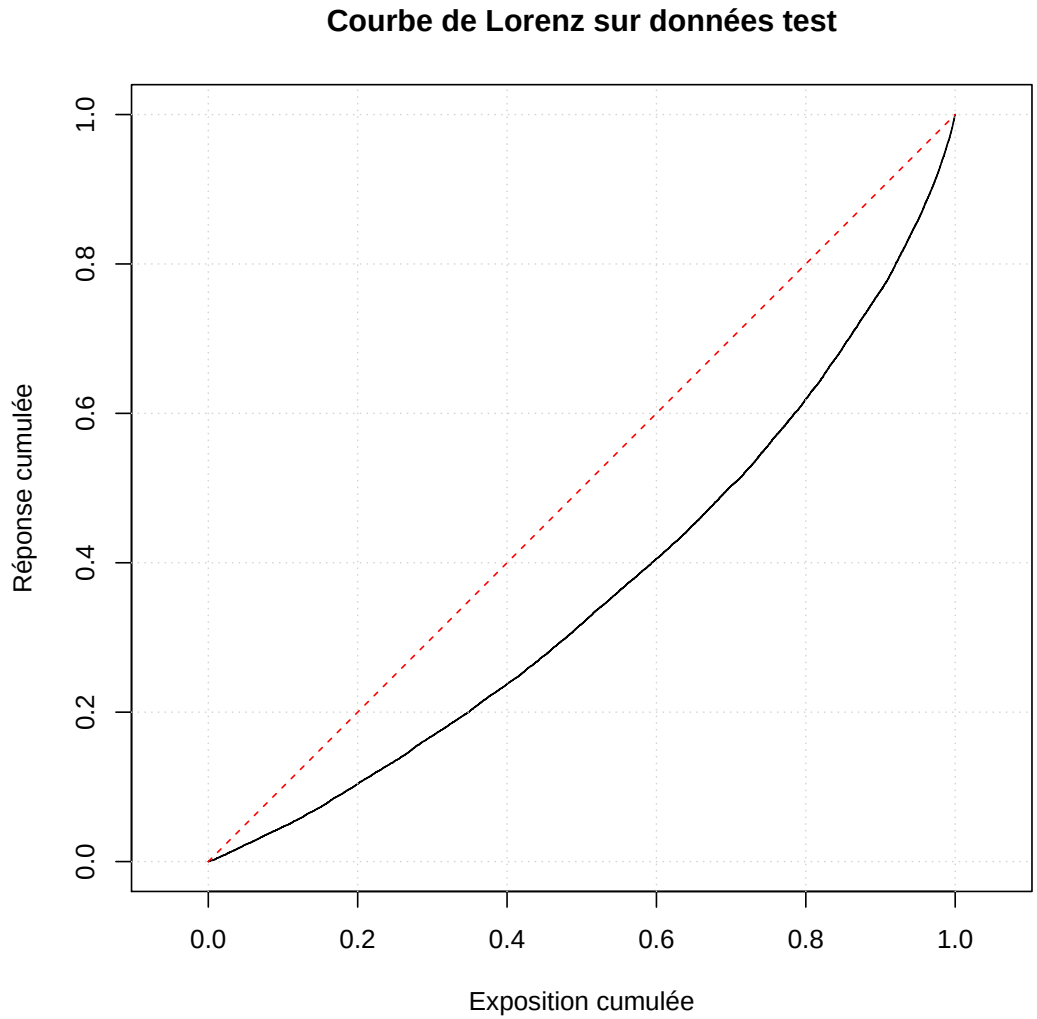
La courbe de Lorenz fournit une visualisation de la capacité du modèle à discriminer efficacement entre les valeurs moyennes plus ou moins élevées de la réponse (on se place dans le cas de réponses à valeurs dans $[0, +\infty[$).

Classiquement, on considère un échantillon de test $(y_i, x_i)_{i \in I_{\text{Test}}}$ indépendant de l'échantillon utilisé pour l'ajustement du modèle, et la courbe est obtenue de la manière suivante :

- tri des données de test par ordre croissant de la valeur moyenne (normalisée) prédite pour obtenir l'échantillon trié $(y_{(i)}, x_{(i)})_{i \in I_{\text{Test}}}$
- tracé des points de coordonnées $(u_{(i)}, v_{(i)})_{i \in I_{\text{Test}}}$ définies par :
  - $u_{(i)}$ = somme des expositions associées aux données $(y_{(\ell)}, x_{(\ell)})_{(\ell) \leq (i)}$
  - $v_{(i)}$ = somme des réponses $(y_{(\ell)})_{(\ell) \leq (i)}$

On normalise en général les coordonnées u et v, respectivement par la somme des expositions et des réponses.

Courbe de Lorenz



Courbe de Lorenz

Dans le cas normalisé, on obtient (normalement !) une courbe croissante et convexe joignant les deux extrémités $(0, 0)$ et $(1, 1)$.

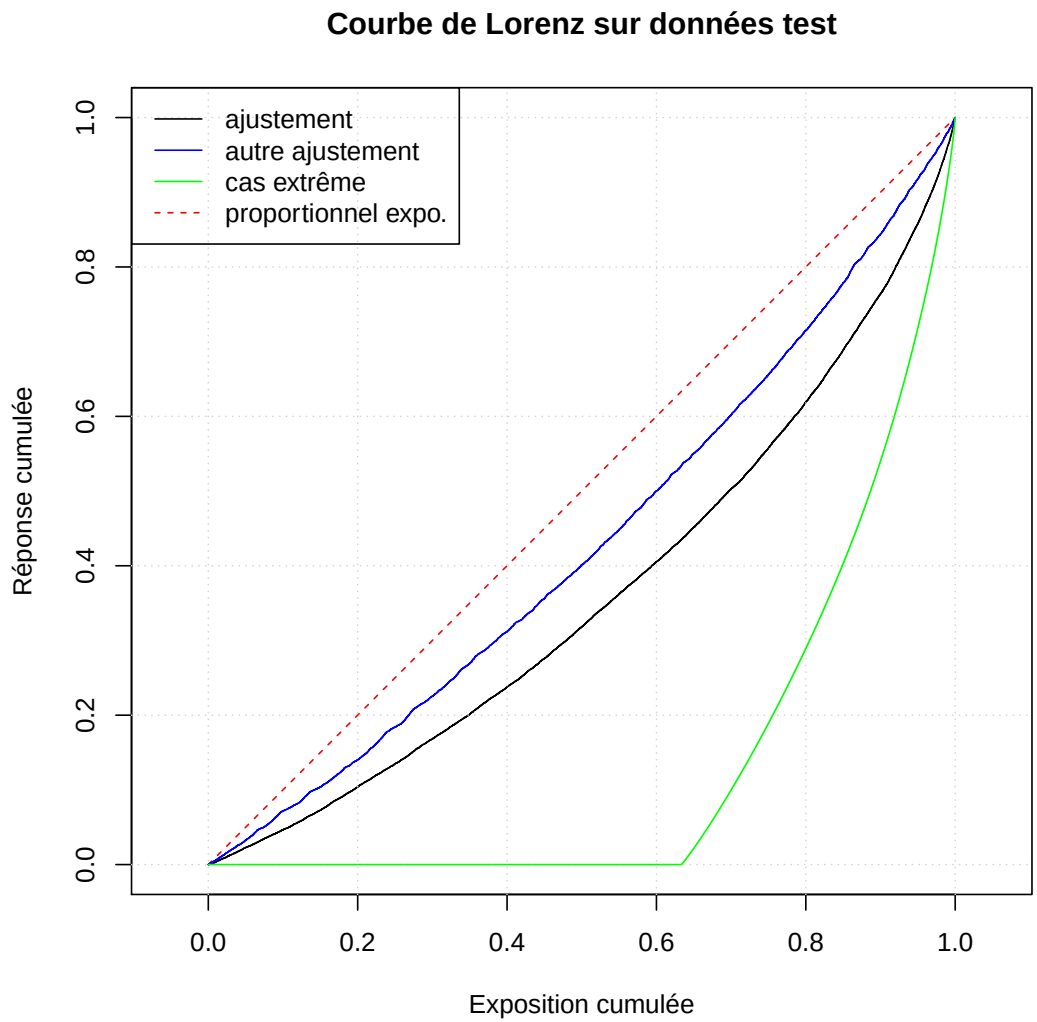
Plus la convexité est forte, plus le pouvoir discriminant du modèle est élevé.

Cas limites :

- la droite reliant les deux extrémités correspond à un modèle prédisant systématiquement la réponse moyenne du jeu de données de test, indépendamment des variables explicatives ;
- la courbe obtenue en prenant comme réponse moyenne prédite la valeur observée de la réponse correspond à la discrimination parfaite.

Voir [DST19] pour une discussion théorique approfondie.

Courbe de Lorenz



Courbe ROC

En régression logistique binaire, on utilise plutôt la courbe ROC (« Receiver Operating Characteristic »).
Une valeur de référence π étant fixée, on peut utiliser le modèle comme un classificateur fournissant une prédiction de type 0/1 pour la réponse :

- on prédit 1 lorsque $\hat{\mu} > \pi$
- on prédit 0 lorsque $\hat{\mu} \leq \pi$

Deux types d'erreur de classification :

- prédire 0 lorsque la réponse est 1
- prédire 1 lorsque la réponse est 0

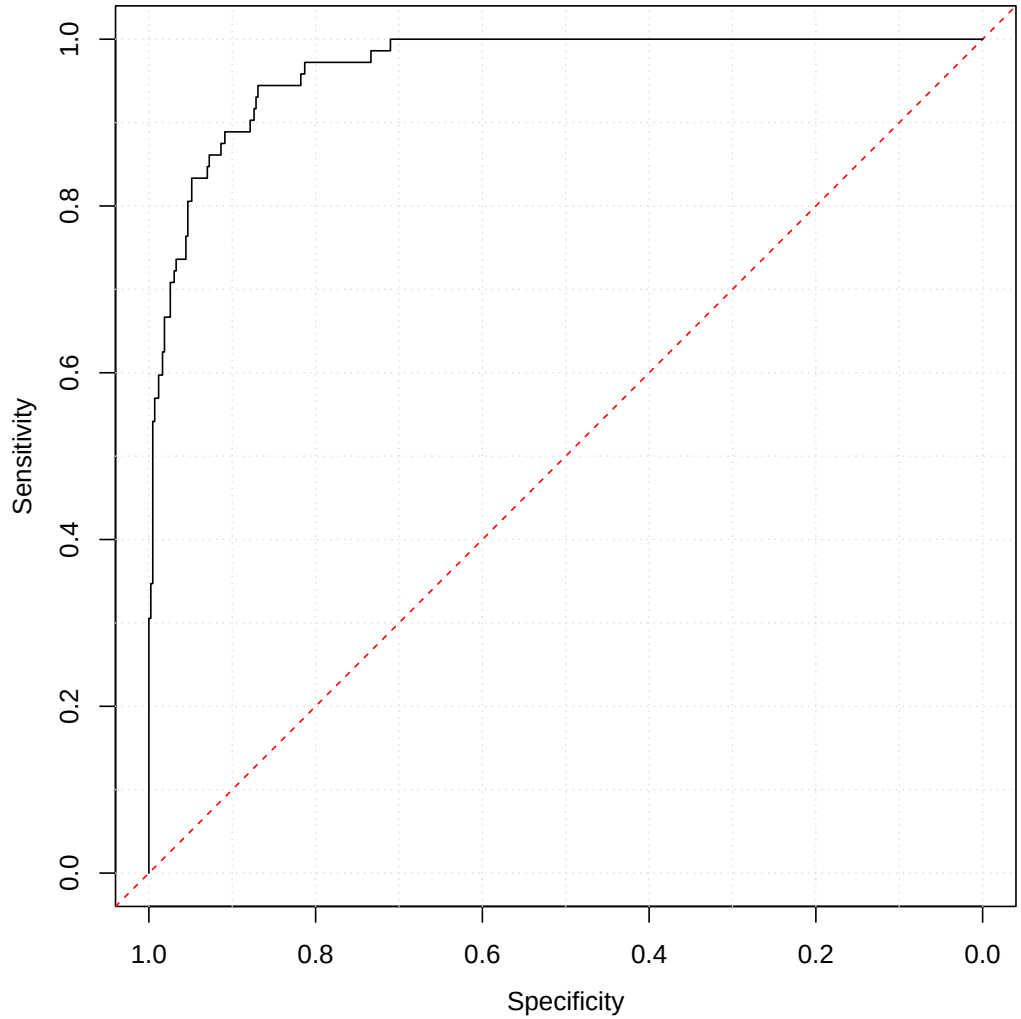
Sur un jeu de données de test (de préférence), on calcule, pour une valeur donnée de π :

- sensibilité : proportion de réponses 1 correctement prédites ;
- spécificité : proportion de réponses 0 correctement prédites.

Lorsque π augmente, la sensibilité décroît tandis que la spécificité croît.

Courbe ROC

La courbe ROC est obtenue en représentant, lorsque π parcourt l'intervalle $[0, 1]$, les points de coordonnées : $(1 - \text{spécificité}, \text{sensibilité})$.



Courbe ROC

On obtient (normalement !) une courbe concave joignant $(0,0)$ et $(1,1)$.
Plus la concavité est forte, plus le pouvoir discriminant du modèle est élevé.
Cas limites :

- la droite reliant les deux extrémités correspond à un modèle prédisant systématiquement la réponse moyenne du jeu de données de test, indépendamment des variables explicatives ;
- la courbe obtenue en prenant comme réponse moyenne prédite la valeur observée de la réponse correspond à la discrimination parfaite et se limite aux points $(0,0),(0,1),(1,1)$.

Courbe ROC

On utilise souvent l'aire sous la courbe ROC (AUC = « Area Under Curve ») pour mesurer la proximité de la courbe avec le cas d'une discrimination parfaite.

$$\text{aucune discrimination } = 1/2 \leq \text{AUC} \leq 1 = \text{ discrimination parfaite.}$$

La valeur de l'AUC est liée au test de Wilcoxon-Mann-Whitney de comparaison de populations.


Ne pas confondre courbe de Lorenz et courbe ROC en régression logistique (voir annexe).

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-G.html`

Termes lisses

Pour représenter correctement l'effet d'une variable explicative numérique dans un modèle GLM, l'une des principales approches consiste à intégrer la variable sous forme d'une fonction lisse écrite comme combinaison linéaire de fonctions de base.
On commence par présenter deux familles classiques de fonctions lisses :

- polynômes
- splines cubiques

Polynômes

Une fonction polynôme de degré $\leq m$ se met par définition sous la forme :

$$f(x) = \sum_{k=0}^{m} c_k x^k.$$

Pour l'utilisation en régression GLM, on choisira plutôt l'écriture dans une autre base [18] :

$$f(x) = \sum_{k=0}^{m} a_k b_k(x),$$

les polynômes $b_0, \ldots, b_m$ formant également une famille échelonnée en degré, mais plus commode d'un point de vue numérique que la base classique des monômes $1, x, \ldots, x^m$.

Lorsque l'on inclut un terme utilisant ce type de base dans un modèle GLM comportant par ailleurs un coefficient constant, on enlève de la base la fonction constante b_0.

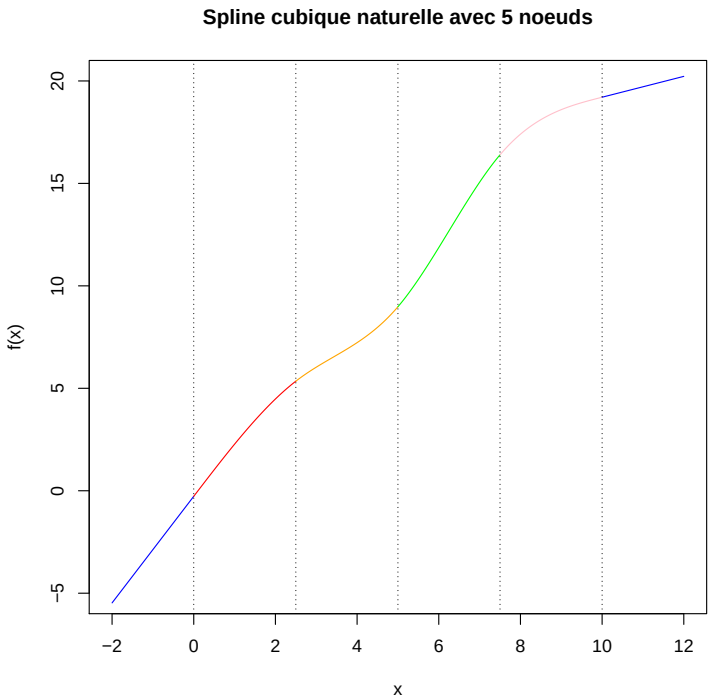
18. Par exemple une base orthogonalisée sur les données, le choix par défaut avec R.

Splines

Définition

Les splines (en français : « cerces ») sont des fonctions polynomiales par morceaux, auxquelles on impose diverses contraintes de raccordement.

Une **spline cubique naturelle** associée à une suite $t_1 < \cdots < t_m$ de « noeuds » est une fonction de $\mathbb{R}$ dans $\mathbb{R}$ qui s'écrit comme un polynôme de degré 3 sur chaque intervalle $[t_\ell, t_{\ell+1}]$, qui est globalement de classe C^2, et s'écrit comme une fonction affine sur $]-\infty, t_1]$ et $[t_m, +\infty[$.



Splines

Les noeuds $t_1, \ldots, t_m$ étant fixés, une telle fonction peut s'écrire comme combinaison linéaire de m splines cubiques naturelles de base [19] :

$$f(x) = \sum_{k=0}^{m-1} a_k b_k(x).$$

En effet, une telle fonction est caractérisée par $2 + 4(m-1) + 2 = 4m$ coefficients de polynômes (4 pour chaque $[t_\ell, t_{\ell+1}]$, 2 pour $]-\infty, t_1]$ et 2 pour $[t_m, +\infty[$) devant satisfaire $3m$ relations (linéaires) entre les coefficients provenant des contraintes de raccordement, d'où une dimension de $4m - 3m = m$.

Lorsque l'on inclut un terme utilisant ce type de base dans un modèle GLM comportant par ailleurs un coefficient constant, on doit enlever une fonction de la base.

Si $\{t_1, \ldots, t_m\}$ n'est pas un sous-ensemble de $\{s_1, \ldots, s_{m+1}\}$, l'ensemble des splines cubiques naturelles construites sur les noeuds $(t_1, \ldots, t_m)$ n'est pas inclus dans l'ensemble des splines cubiques naturelles construites sur les noeuds $(s_1, \ldots, s_{m+1})$.

19. Voir l'appendice pour une discussion de diverses bases de fonctions splines.

Degrés de liberté

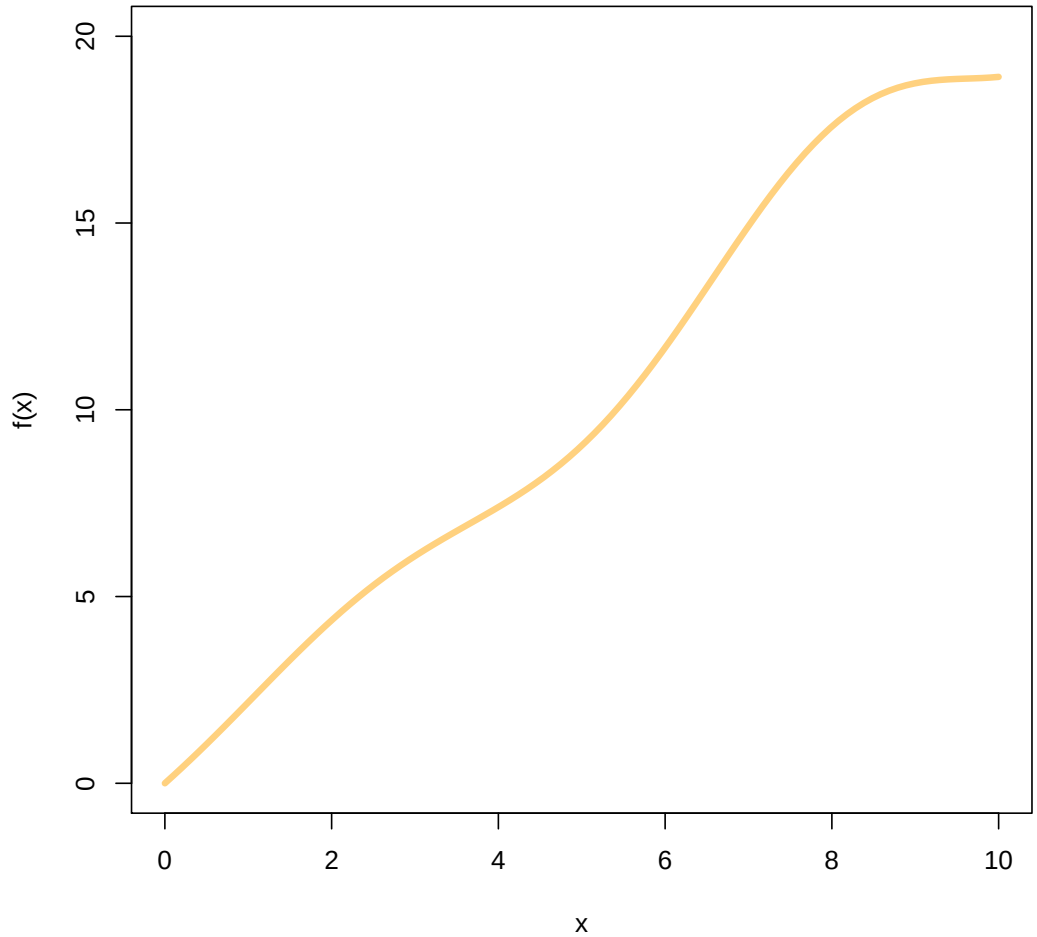
Pour les deux familles de fonctions précédentes, un paramètre (le degré m pour les polynômes, et $(m-1)$ pour les splines cubiques naturelles, où m est le nombre de noeuds) compte le **nombre de degrés de liberté** introduits dans le modèle par l'utilisation de cette famille. Dans ce cadre simple, le nombre de degrés de liberté s'identifie au nombre de coefficients à estimer dans la composante systématique.

Plus ce nombre est élevé, et plus la famille de fonctions employée est flexible, et susceptible d'approcher avec précision une vaste gamme de fonctions.

On pourrait donc *a priori* penser que les performances prédictives du modèle seront d'autant meilleures que le modèle est plus flexible, et donc que le nombre de degrés de liberté est plus élevé...

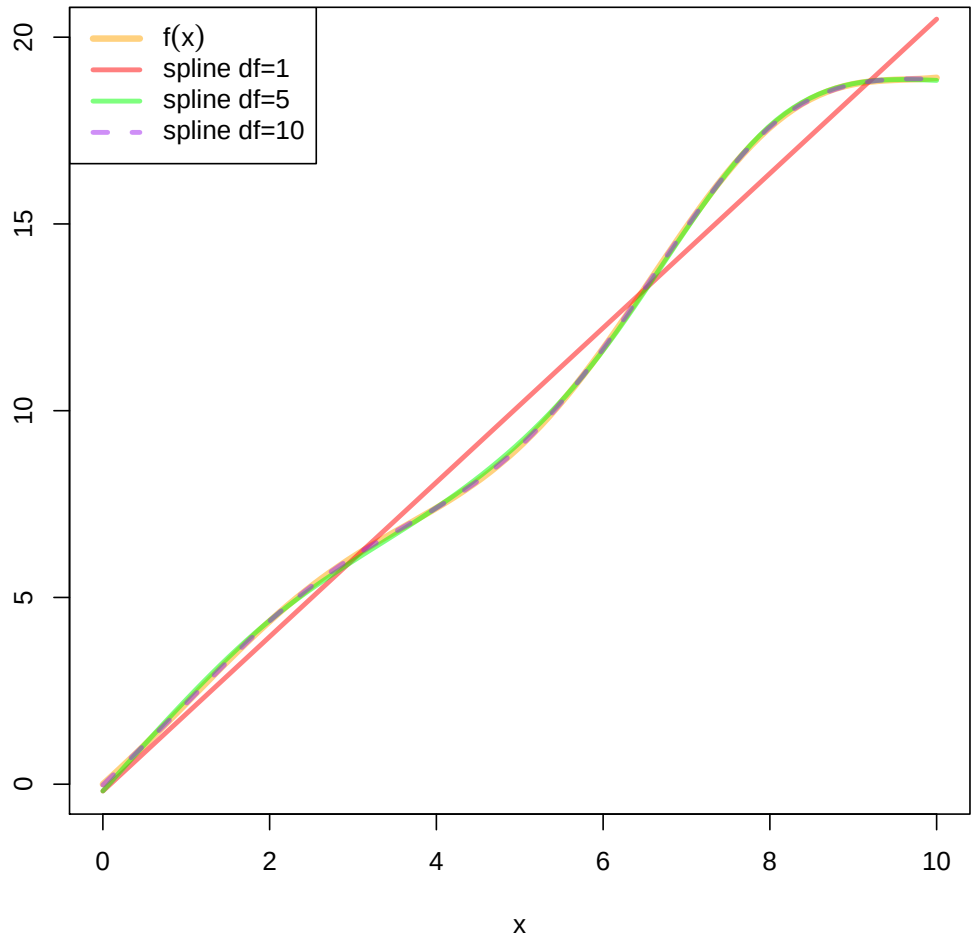
Une petite expérience

On montre ci-après un tracé de la fonction $f : x \mapsto 2x + 0,2x\sin(x)$ sur l'intervalle $[0, 10]$.



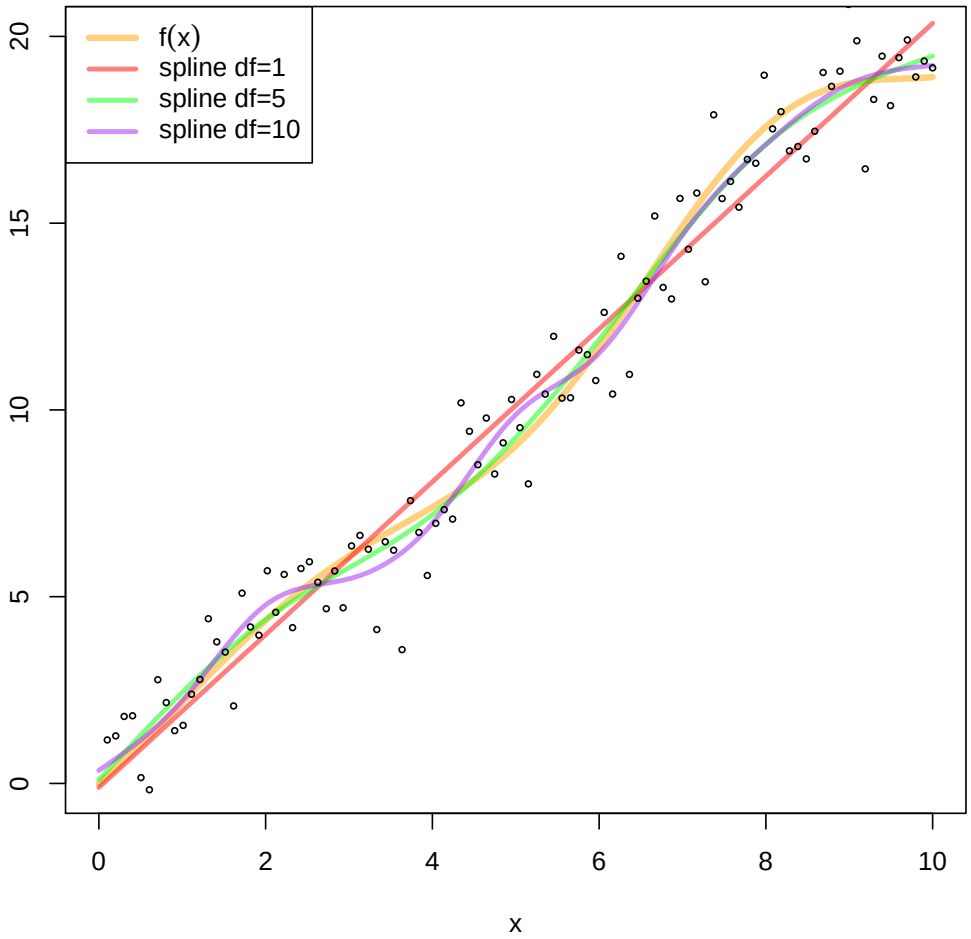
Une petite expérience

On montre ci-après un tracé de la fonction $f : x \mapsto 2x + 0,2x\sin(x)$ sur l'intervalle $[0, 10]$, et trois approximations de f par constante + spline cubique naturelle de degré de liberté : 1 (affine), 5 et 10.



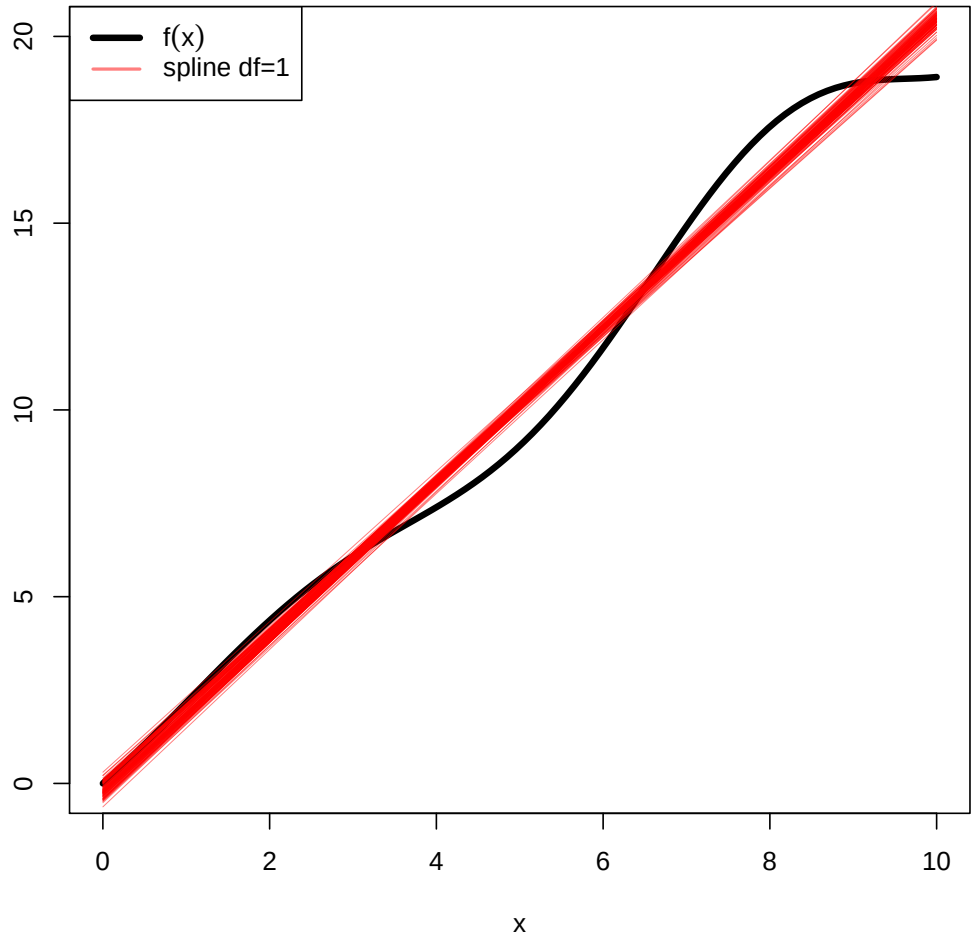
Une petite expérience

On fabrique un jeu de $N = 100$ données simulées (y_i, x_i) avec les x_i formant un découpage régulier de $[0, 10]$, et $y_i = f(x_i) + \varepsilon_i$, où $\varepsilon_i \sim \mathcal{N}(0, 1)$, et l'on ajuste sur ces données un modèle linéaire gaussien $Y = c + s_q(x) + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, où s_q est une spline cubique naturelle à q degrés de liberté.



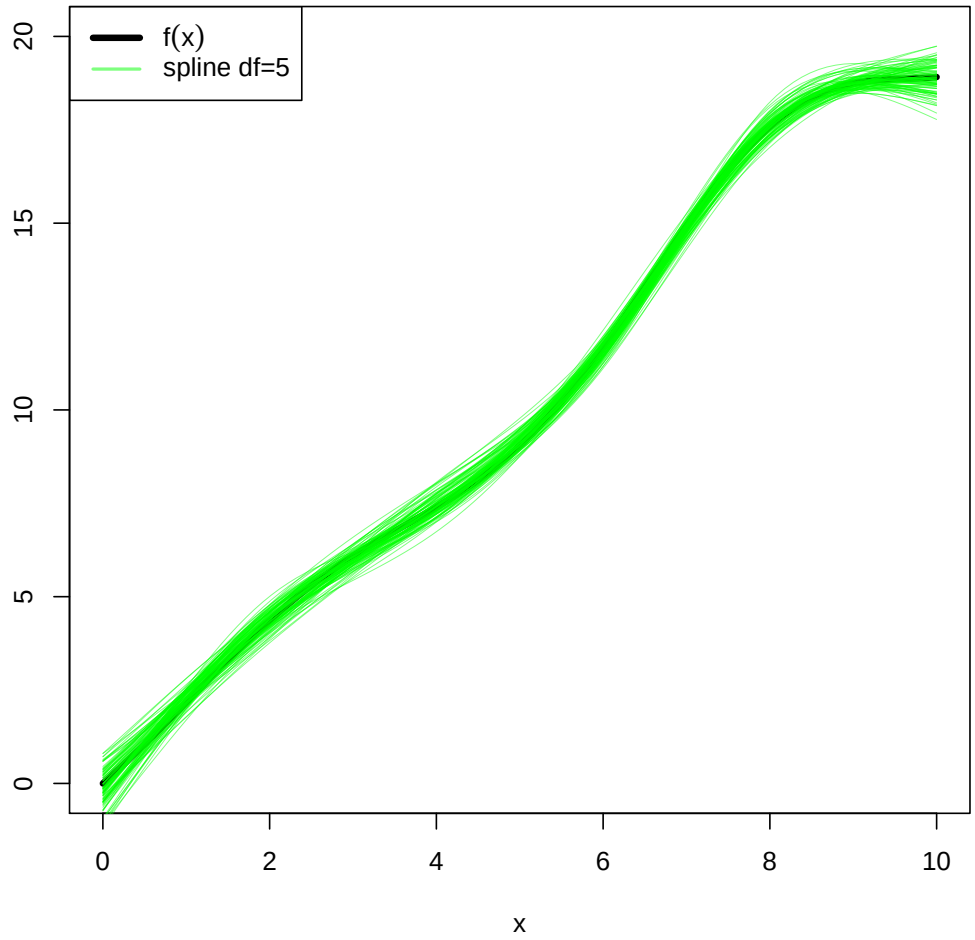
Une petite expérience

Résultats pour 100 répétitions et $q = 1$



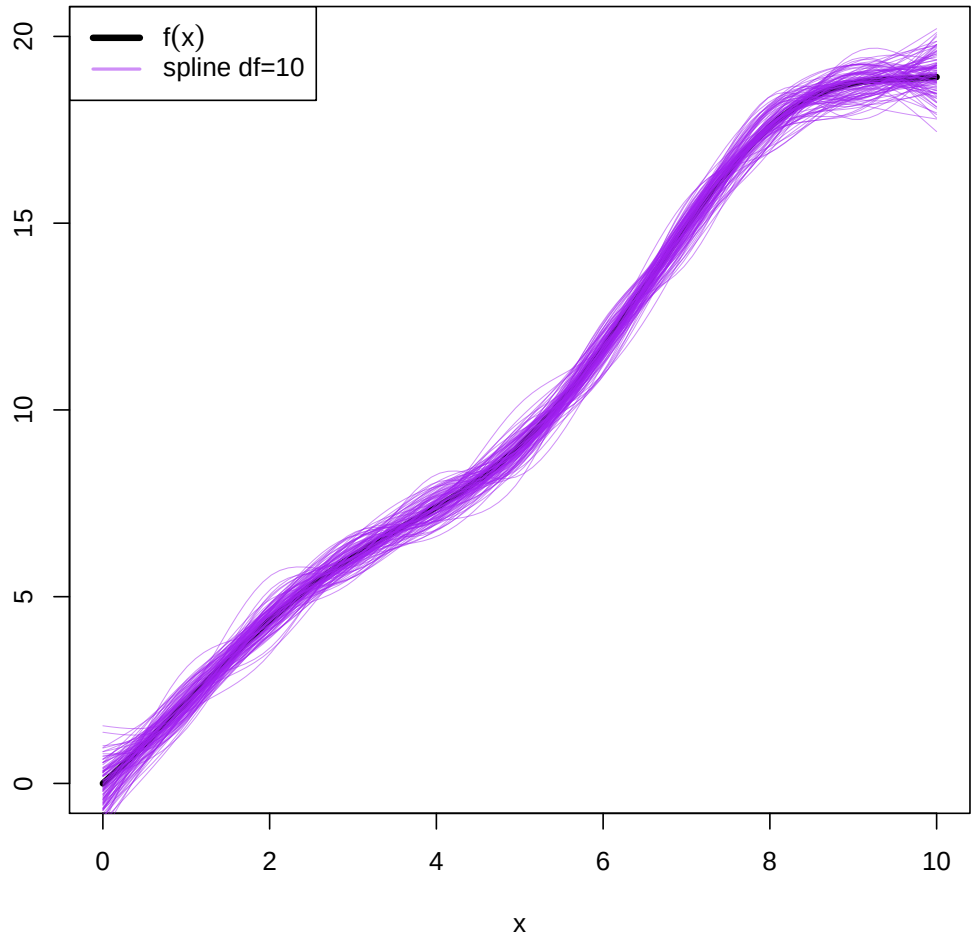
Une petite expérience

Résultats pour 100 répétitions et $q = 5$



Une petite expérience

Résultats pour 100 répétitions et $q = 10$



Une petite expérience

Pour une valeur de x donnée, l'erreur quadratique de prédiction de la moyenne s'écrit [20] :

$$\underbrace{\mathbb{E}\left[(\hat{f}_q(x) - f(x))^2\right]}_{\text{err. quad.}(x)} = \underbrace{\left(\mathbb{E}\left[\hat{f}_q(x)\right] - f(x)\right)^2}_{\text{biais}(x)^2} + \underbrace{\mathbb{E}\left[\left(\hat{f}_q(x) - \mathbb{E}\left[\hat{f}_q(x)\right]\right)^2\right]}_{\text{variance}(x)}$$

Avec $n = 1000$ répétitions, et pour $q = 1, \ldots, 20$, on calcule des estimations de :

- err. quad. $\overset{\text{déf.}}{=} \frac{1}{N}\sum_{i=1}^{N} \text{err. quad.}(x_i)$
- biais^2 $\overset{\text{déf.}}{=} \frac{1}{N}\sum_{i=1}^{N} \text{biais}(x_i)^2$
- variance $\overset{\text{déf.}}{=} \frac{1}{N}\sum_{i=1}^{N} \text{variance}(x_i)$

20. L'espérance $\mathbb{E}$ se réfère ici à l'aléa présent dans les données y_i, qui se répercute sur l'estimateur $\hat{f}_q(x)$ de $f(x)$.

Une petite expérience

De manière précise, avec $n = 1000$ répétitions, et pour $q = 1, \ldots, 20$, on calcule :

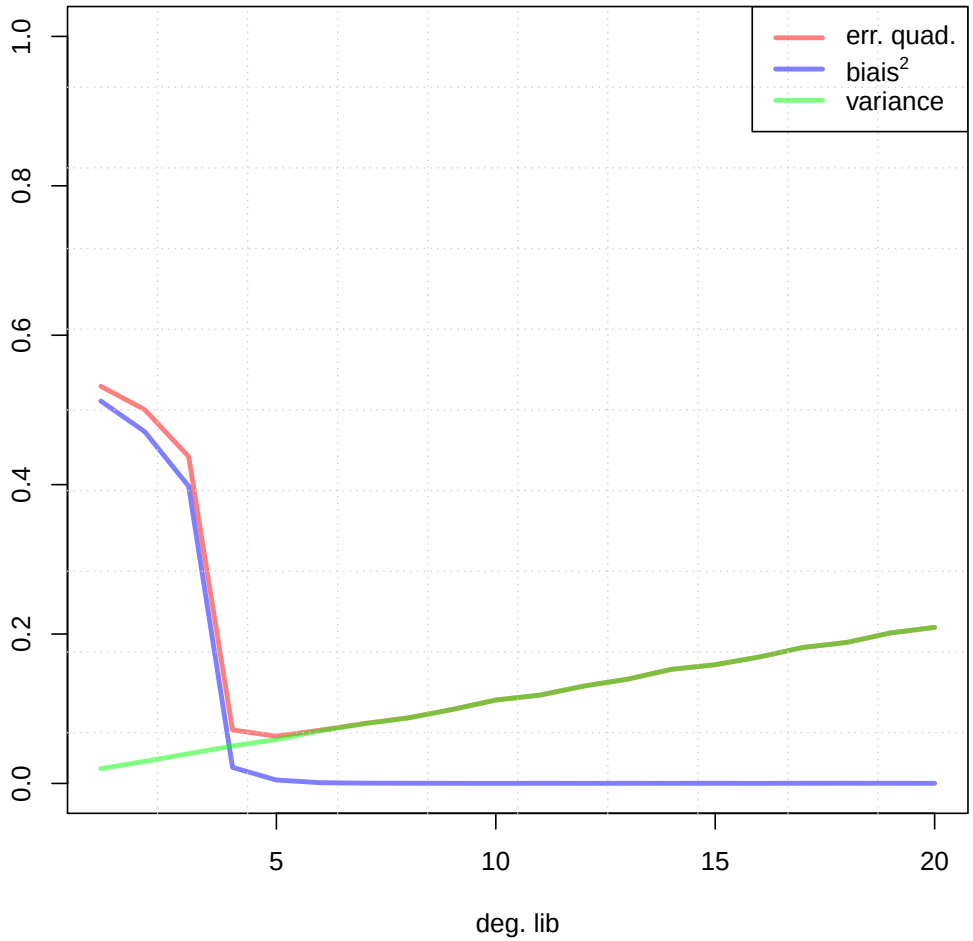
$$\overline{\text{err. quad.}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n} \sum_{b=1}^{n} \left(\hat{f}_q^{(b)}(x_i) - f(x_i) \right)^2$$

$$\overline{\text{biais}^2} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{n} \sum_{b=1}^{n} \hat{f}_q^{(b)}(x_i) - f(x_i) \right)^2$$

$$\overline{\text{var}} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{n} \sum_{b=1}^{n} \left(\hat{f}_q^{(b)}(x_i) - \frac{1}{n} \left(\sum_{b=1}^{n} \hat{f}_q^{(b)}(x_i) \right) \right)^2$$

Une petite expérience

Résultats pour $n = 1000$ répétitions, et pour $q = 1, \ldots, 20$:



Une petite expérience

Observations :

- pour un modèle **trop peu** flexible, le biais est élevé et la variance est faible, le modèle n'est pas capable de s'ajuster assez finement aux données et l'erreur quadratique est principalement due au biais, on parle de « **sous-ajustement** »
- pour un modèle **trop** flexible, le biais est faible et la variance est élevée, le modèle s'ajuste en partie au bruit aléatoire présent dans les données, et l'erreur quadratique est principalement due à la variance, on parle de « **sur-ajustement** »
- le modèle le plus précis en prédiction réalise un **compromis** entre les deux effets

Une petite expérience

Remarques :

- l'exemple précédent est un cas d'école, dans lequel on connaît à l'avance le modèle ayant engendré les données ; sur de véritables données, on ne peut pas accéder aussi facilement à l'erreur de prédiction sur la moyenne commise par le modèle ;
- le phénomène mis en évidence reste (qualitativement) valable dans un cadre très général
- dans un cadre GLM, la flexibilité du modèle ne se limite pas à l'utilisation de bases de fonctions plus ou moins riches pour intégrer les effets des variables quantitatives, on peut citer en outre :
  - le choix d'incorporer ou non certaines variables explicatives (lorsque celles-ci sont nombreuses)
  - le choix de fusionner ou non certaines catégories pour des variables catégorielles
  - le choix d'incorporer plus ou moins de termes d'interaction
- du point de vue inférentiel classique, un manque de flexibilité se traduira par une mauvaise adéquation, et un excès de flexibilité par une bonne adéquation mais une incertitude élevée sur les coefficients

Choix manuel du nombre de degrés de liberté

Ce qui précède illustre l'importance du choix du nombre de degrés de liberté. Pour les deux familles de fonctions lisses vue auparavant (polynômes et splines cubiques), on peut envisager d'ajuster des modèles pour différents degrés de liberté, puis de les comparer en utilisant par exemple :

- un examen graphique de l'ajustement obtenu
- des tests statistiques comme le test du rapport de vraisemblance
  (lorsque les modèles sont emboîtés, ce qui est le cas pour les polynômes)
- la comparaison d'un critère de performance prédictive (AIC, CV, etc.)

Approche par pénalisation

Une autre approche consiste à partir d'une famille de fonctions très flexible, et à introduire un terme de **pénalité** dans l'ajustement des coefficients.
Par exemple, lorsque la composante systématique du modèle inclut une fonction de la variable quantitative $x^{(j)}$ décomposée sous la forme :

$$f_j(x^{(j)}) = \sum_{k=1}^{m_j} \beta_k^{(j)} b_k^{(j)}(x^{(j)}),$$

on estimera l'ensemble des coefficients β du modèle en maximisant la fonction :

$$\log L(\beta) - \lambda^{(j)} \int_{-\infty}^{+\infty} \left(f_j''(x^{(j)}) \right)^2 dx^{(j)},$$

où $\lambda^{(j)} \geq 0$ est un paramètre de contrôle fixé.
L'ajustement réalise ainsi automatiquement un compromis entre :

- la fidélité aux données (mesurée par $L(\beta)$)
- la régularité de f_j (mesurée par $\int_{-\infty}^{+\infty} \left(f_j''(x^{(j)}) \right)^2 dx^{(j)}$)

dont les poids respectifs sont contrôlés par la valeur de $\lambda^{(j)}$.

Approche par pénalisation

En ajustant les coefficients par maximisation de la vraisemblance pénalisée

$$\underbrace{\log L(\beta)}_{\text{fidélité}} - \lambda^{(j)} \underbrace{\int_{-\infty}^{+\infty} \left(f_j''(x^{(j)}) \right)^2 dx^{(j)}}_{\text{pénalité}},$$

on a les deux situations extrêmes suivantes :

- lorsque $\lambda^{(j)} \to +\infty$, le terme de pénalité domine et force la fonction f_j estimée à être proche d'une fonction affine, ce qui correspond à un nombre de degrés de liberté égal à 2 pour f_j ;
- lorsque $\lambda^{(j)} \to 0$, le terme de fidélité domine et l'on retrouve l'estimation sans pénalisation, ce qui correspond à un nombre de degrés de liberté égal à m_j pour f_j.

En faisant varier $\lambda^{(j)}$, on fait ainsi varier le degré de flexibilité autorisé pour l'ajustement.

Approche par pénalisation

De manière précise, l'ajustement par maximum de vraisemblance pénalisée donne lieu à une matrice $\hat{H}(\lambda^{(j)})$ analogue à celle obtenue pour un GLM estimé par maximum de vraisemblance (sans pénalisation).

On peut ainsi utiliser la quantité $\text{tr}(\hat{H}(\lambda^{(j)}))$ pour définir le **nombre de degrés de liberté effectif** de l'ajustement, reflétant la réduction du nombre de degrés de liberté total (= nombre de coefficients à estimer) sous l'effet de la pénalisation.

 Contrairement à la méthode du maximum de vraisemblance non-pénalisée, l'utilisation d'une pénalité introduit une certaine dose de biais dans l'estimation. On ne peut donc pas utiliser cette approche impunément.

Approche par pénalisation

Plus généralement, lorsque la composante systématique du modèle fait intervenir q variables quantitatives $x^{(j_1)}, \ldots, x^{(j_q)}$ dont les effets respectifs sont, pour $j \in \{j_1, \ldots, j_q\}$, représentés sous la forme :

$$f_j(x^{(j)}) = \sum_{k=1}^{m_j} \beta_k^{(j)} b_k^{(j)}(x^{(j)}),$$

on estimera l'ensemble des coefficients β du modèle en maximisant la fonction :

$$\log L(\beta) - \sum_{j \in \{j_1, \cdots, j_q\}} \lambda^{(j)} \int_{-\infty}^{+\infty} \left(f_j''(x^{(j)}) \right)^2 dx^{(j)},$$

avec un jeu de q paramètres de contrôle positifs $\lambda = (\lambda^{(j_1)}, \ldots, \lambda^{(j_q)})$.

Approche par pénalisation

L'ajustement du modèle comporte deux niveaux imbriqués :

- les valeurs des paramètres de contrôle $\lambda = (\lambda^{(j_1)}, \ldots, \lambda^{(j_q)})$ étant fixés, estimer les coefficients β en maximisant la vraisemblance pénalisée

  La maximisation de la vraisemblance pénalisée s'effectue de manière similaire à la maximisation de la vraisemblance non-pénalisée pour un GLM classique. On obtient en particulier une matrice $\hat{H}(\lambda)$ adaptée à ce cas, et la valeur $\mathrm{tr}(\hat{H}(\lambda))$ joue alors le rôle du **nombre de degrés de liberté effectifs** du modèle, mesurant la réduction du nombre de degrés de liberté total (= nombre de coefficients) sous l'effet de la pénalisation. La valeur $\mathrm{tr}(\hat{H}(\lambda))$ se décompose en la somme d'un nombre de degrés de liberté effectifs pour la représentation de chacune des variables $x^{(j_i)}$.

- optimiser la valeur des paramètres de contrôle $\lambda^{(j_1)}, \ldots, \lambda^{(j_q)}$ de manière à maximiser la performance prédictive du modèle ajusté

  Par exemple, le paquet `mgcv` utilise par défaut le critère UBRE lorsque ϕ est connue *a priori*, et le critère GCV sinon. Ces deux critères font intervenir $\mathrm{tr}(\hat{H}(\lambda))$.

Splines pénalisées

Plusieurs choix sont possibles pour la famille de fonctions utilisée dans cette approche (rappel : cette famille est supposée très flexible au départ, la pénalisation servant ensuite à régler la flexibilité effective du modèle obtenu)

- splines cubiques naturelles ayant pour noeuds la totalité des valeurs de $x^{(j)}$ présentes dans le jeu de données

  Pour un jeu de données même modérément volumineux, un tel choix conduit à un nombre excessif de paramètres d'un point de vue numérique, et également du point de vue de la modélisation statistique. On a donc plutôt recours aux deux autres options décrites ici, qui consistent en des approximations de dimension plus réduite :

- « penalized regression splines » : splines cubiques naturelles définies sur un jeu de noeuds de taille « raisonnable » construit à partir des données
- « thin plate regression splines » uni-dimensionnelles : fonctions lisses spécifiées à partir d'une approximation de (en fait, d'une sous-matrice de taille gérable de) la matrice $\left(|x_i^{(j)} - x_\ell^{(j)}|^3\right)_{1 \leq i,\ell \leq N}$ ne conservant que les plus grandes valeurs propres

Modèles additifs généralisés

L'approche d'estimation pénalisée décrite précédemment se rattache aux **modèles additifs généralisés** (« Generalized Additive Models »). La structure est analogue à celle d'un GLM, avec une composante systématique se décomposant sous la forme additive suivante :

$$\eta = \beta_0 + \underbrace{\sum_{j=1}^{q} f_j(x^{(j)})}_{\text{effet des var. quantitatives}} + \underbrace{\sum_{j=q+1}^{m} g_j(x^{(j)})}_{\text{effet des var. catégorielles}},$$

les fonctions f_j et g_j s'écrivant comme combinaisons linéaires de fonctions de base. Typiquement : base de splines pour les variables quantitatives, et utilisation de matrices d'encodage pour les variables catégorielles.
De nombreuses extensions (p. ex. avec des fonctions lisses de plusieurs variables) existent. Nous renvoyons à l'ouvrage [Woo17] pour un traitement détaillé de ces modèles.

Avec R

On utilisera le paquet `mgcv` pour ajuster les modèles additifs généralisés via l'approche de vraisemblance pénalisée pour les bases de fonctions splines. (Voir par exemple le paquet `gam` pour une approche alternative, plus conforme à la méthodologie proposée lorsque les modèles additifs généralisés ont été introduits.)

 D'un point de vue inférentiel, la prise en compte correcte de la pénalisation et de l'optimisation de ses paramètres est un problème complexe. Les intervalles de confiance et $p-$valeurs renvoyés par `mgcv` utilisent une approche spécifique décrite dans [Woo17].

Régression locale

Une autre approche pour intégrer des effets lisses est celle de la régression locale, décrite ci-après dans le cas d'une seule variable (quantitative) x.

Régression locale (uni-dimensionnelle)

La valeur moyenne prédite $\hat{\mu}(x)$ est obtenue en ajustant un modèle GLM basé sur un polynôme de petit degré en x (typiquement 0, 1, ou 2) dans lequel les données sont **pondérées** en fonction de leur proximité avec x.

Exemple :

- on identifie l'ensemble $\mathcal{V}(x)$ formé par les $\lceil \alpha N \rceil$ points x_i les plus proches de x dans le jeu de données $(y_i, x_i)_{1 \leq i \leq N}$
- $d \overset{\text{déf.}}{=}$ distance à x du plus éloigné des points de $\mathcal{V}(x)$, avec $\alpha = 70\%$
- $W(x_i, x) := K\left(\min\left[\frac{|x - x_i|}{d}, 1\right]\right)$ où $K(u) = (1 - u^3)^3$
- $\hat{\mu}(x) =$ valeur moyenne prédite par un ajustement GLM avec les caractéristiques :
  - $\eta = \beta_0 + \beta_1 x + \beta_2 x^2$
  - données $(x_i, y_i)_{1 \leq i \leq N}$ avec poids $w_i \cdot W(x_i, x)$

Régression locale

Quelques points d'attention :

- le paramètre α contrôle la flexibilité de l'ajustement
- on peut ici encore définir une matrice $\hat{H}$ et un nombre de degré de liberté effectif
- en général, une approximation du résultat est effectuée, à partir du calcul pour une famille restreinte de valeurs de x
- ces approches sont plutôt utilisées dans un contexte exploratoire (ex. : visualisation de tendances)
- leur efficacité est souvent limitée au cas uni- ou bi-dimensionnel.
- diverses approches pour le calcul de la variance (local ou non) sont disponibles

Nous renvoyons à l'ouvrage [Loa99] pour un traitement détaillé de ces modèles.

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-H.html`

Régression avec un arbre

Avec m variables explicatives $x_1, \ldots, x_m$, on considère l'ensemble S formé par toutes les combinaisons possibles de valeurs de $x = (x_1, \ldots, x_m)$.

Une **partition arborescente** de S associée à un arbre $\mathscr{A}$ est la donnée, pour chaque sommet s de $\mathscr{A}$, d'un sous-ensemble $I(s)$ de S, avec les propriétés suivantes :

- $I(r) = S$, où r est la racine de $\mathscr{A}$,
- pour tout sommet non-terminal s de $\mathscr{A}$, les ensembles $I(e)$, où e parcourt l'ensemble des enfants de s, forment une partition de $I(s)$
- cette partition est obtenue en imposant aux éléments de $I(s)$ une condition supplémentaire portant sur une variable explicative $x_{v(s)}$

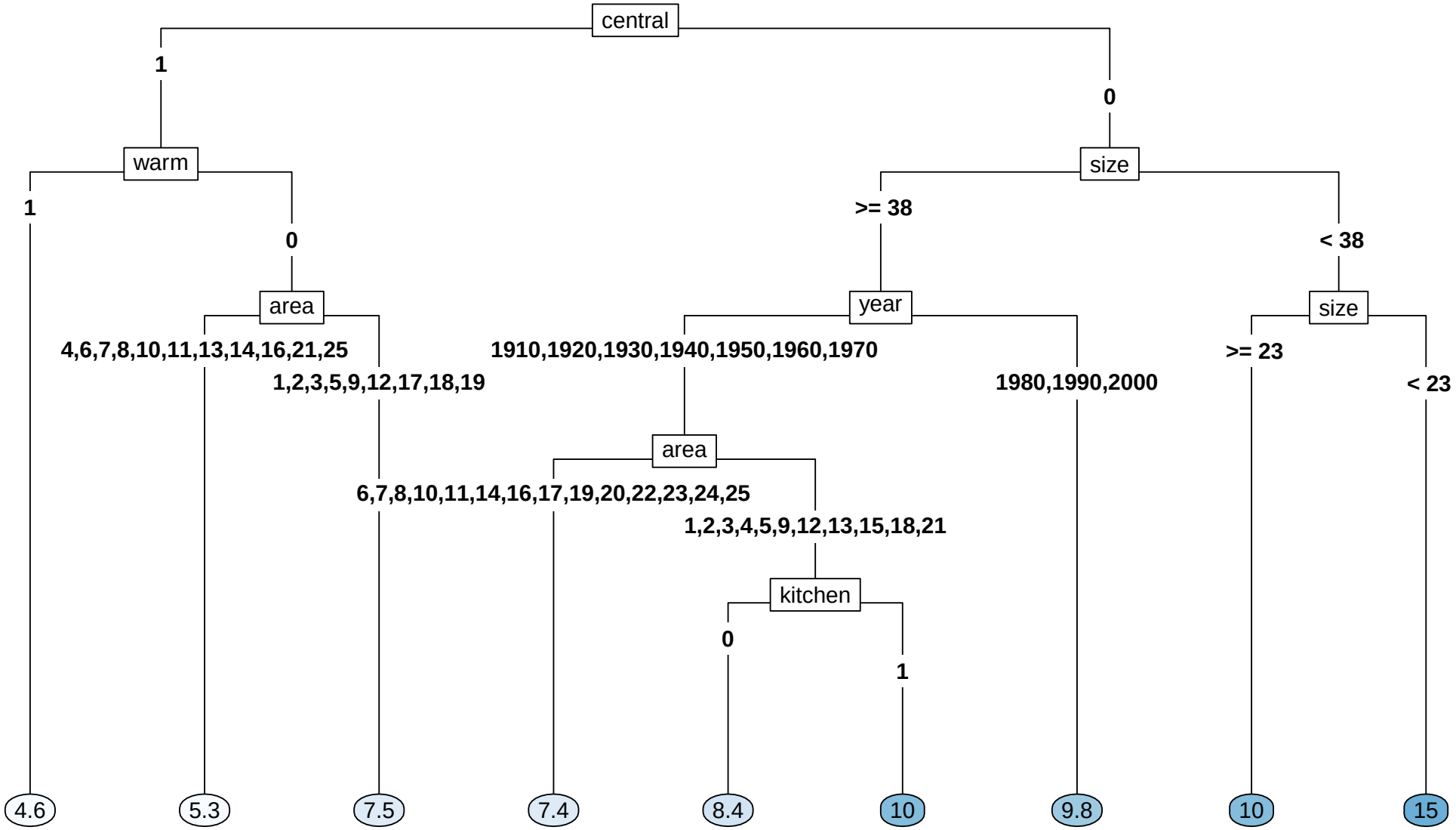
La **fonction de régression** associée s'écrit :

$$f(x) = \sum_{s \in F(\mathscr{A})} f_s \cdot \mathbf{1}(x \in I(s)),$$

où $F(\mathscr{A})$ désigne l'ensemble des feuilles (sommets terminaux) de l'arbre.

Régression avec un arbre

Exemple : modélisation du loyer mensuel au m^2 (Munich, 2003)



Régression avec un arbre

Partant d'un jeu de données d'ajustement $(y_i, x_i)_{1 \leq i \leq N}$ et des poids $(w_i)_{1 \leq i \leq N}$, on cherche à identifier la structure (arbre et partition) et les valeurs f_s qui minimisent l'écart :

$$\sum_{i=1}^{N} w_i \cdot D(y_i, f(x_i)) + \alpha |F(\mathscr{A})|,$$

où le terme $\alpha|F(\mathscr{A})|$ vient pénaliser la complexité de l'arbre utilisé, afin d'éviter le surajustement du modèle, le paramètre α étant lui-même optimisé afin de maximiser la performance prédictive du modèle ajusté.

La structure (arbre et partition) étant donnée, les valeurs attribuées aux sommets terminaux s'expriment alors comme la moyenne

$$f_s = \frac{\sum_{i \in J(s)} w_i y_i}{\sum_{i \in J(s)} w_i},$$

où $J(s)$ est l'ensemble des indices $1 \leq i \leq N$ tels que $x_i \in I(s)$.

Régression avec un arbre

En général, l'ensemble des structures possibles est trop vaste pour pouvoir être exploré de manière exhaustive, et l'on a recours à des approches simplifiées.
L'approche CART consiste *grosso modo* à :

1. Faire croître un arbre par étapes à partir de la racine, en ajoutant à chaque étape la ramification produisant le meilleur gain de déviance par rapport à l'arbre précédent, jusqu'à satisfaire un critère d'arrêt [21].
2. Elaguer progressivement l'arbre ainsi obtenu en optimisant l'écart pénalisé pour des valeurs croissantes de α.
3. Optimiser la valeur de α par validation croisée $K-$fold.

L'utilisation de la validation croisée à l'étape 3 présente une subtilité : il convient de répéter les étapes 1. et 2. K fois, en excluant à chaque fois l'un des groupes des données utilisées pour l'ajustement, ce groupe étant utilisé pour mesurer l'écart de déviance ainsi obtenu.

21. Par exemple : une profondeur maximale pour l'arbre, un seuil d'amélioration pour les nouvelles ramifications, un volume de données minimal dans chaque partition, etc.

Régression avec un arbre

L'utilisation d'un arbre de régression conduit en général à un modèle dont l'**interprétabilité** est excellente, mais dont la **performance prédictive** est limitée, notamment du fait de la variabilité importante de l'arbre obtenu en fonction du jeu de données [22].

Afin d'améliorer la performance prédictive (et au prix d'une diminution de l'interprétabilité), des méthodes de régression reposant sur des familles d'arbres ont été développées :

- « bagging »
- « boosting »

22. Par exemple, si l'on sépare aléatoirement le jeu de données en deux parties de volume équivalent, les arbres ajustés sur chaque partie pourront différer sensiblement.

Régression avec plusieurs arbres : « bagging »

L'idée générale est d'ajuster B arbres de régression sur B jeux de données produits par ré-échantillonnage aléatoire du jeu de données initial, et d'utiliser la moyenne des B fonctions de régression ainsi obtenues :

$$\hat{f}(x) = \frac{1}{B} \sum_{i=1}^{B} \hat{f}^{(i)}(x).$$

On tente ainsi de se rapprocher de la situation théorique idéale où l'on aurait B réalisations i.i.d. d'un estimateur sans biais de $\mathbb{E}(Y | X = x)$, dont on prendrait la moyenne pour réduire la variance.

Dans ce contexte, pénaliser la complexité des arbres ajustés ne revêt plus une importance cruciale, et l'on peut se limiter à utiliser l'approche « gloutonne » de l'étape 1 de l'approche CART pour ajuster chaque arbre.

Régression avec plusieurs arbres : « bagging »

Une possibilité (parmi d'autres)[23] pour estimer les performances prédictives du modèle obtenu dans le contexte du « bagging » est de faire appel à l'erreur dite « out-of-bag ».

Pour chaque $1 \leq i \leq N$, on a en moyenne $B(1 - \frac{1}{N})^B \approx B/e$ jeux de données ré-échantillonnés dans lesquels la donnée (x_i, y_i) ne figure pas.

En notant $\hat{f}_{OOB}^{[-i]}$ la moyenne des fonctions de régression obtenus sur ces jeux de données, on considérera

$$\sum_{i=1}^{N} w_i \cdot D(y_i, \hat{f}_{OOB}^{[-i]}(x_i)).$$

23. On peut voir cette approche comme une variante de la validation croisée « leave-one-out » adaptée au cas du « bagging ».

Régression avec plusieurs arbres : « random forest »

L'approche dite « random forest » est essentiellement une variante du « bagging » dans laquelle on impose une limitation *a priori* sur le jeu de variables pouvant être utilisé à chaque ramification.

Par exemple, on se limitera à un tirage aléatoire uniforme de $\sqrt{m}$ variables parmi les m en possibles.

L'idée est que cette limitation contribue à diminuer la corrélation entre les estimateurs $\hat{f}^i$, et donc la variance de leur moyenne.

Régression avec plusieurs arbres : « boosting »

L'idée est de faire évoluer de manière **graduelle** la fonction de régression vers son objectif, en ajustant à chaque étape un nouvel arbre permettant d'effectuer un petit pas dans la bonne direction (en moyenne).

De manière générale, on aura le schéma suivant :

- Initialisation :
  - $f^{[0]} \equiv 0$
- Itération : pour $b = 1, \ldots, B$
  - ajuster un arbre de régression $\hat{f}^b$ sur les données
    $$\left(y_i - f^{[b-1]}(x), x_i\right)_{1 \leq i \leq N}$$
  - $f^{[b]}(x) = f^{[b-1]}(x) + \lambda \hat{f}^b(x)$

Le paramètre de rétrécissement (« shrinkage ») λ contrôle la taille des pas effectués, avec par exemple des valeurs de l'ordre de 10^{-1} à 10^{-3}.

La valeur de B doit être contrôlée pour éviter le surajustement du modèle.

On se limite souvent à ajuster à chaque étape un arbre peu profond.

Régression avec plusieurs arbres : « boosting »

Divers outils logiciels permettent de mettre en œuvre cette approche parmi lesquels :

- XGBoost [24]
- LightGBM [25]
- CatBoost [26]

Par rapport à la présentation générale précédente, ces outils présentent un certain nombre de spécificités et d'ajouts [27], notamment diverses approximations afin de réduire le coût global (en temps de calcul et espace mémoire) de l'ajustement, en particulier en présence de données massives.

---

24. Voir `https://xgboost.readthedocs.io` et [CG16].
25. Voir `https://lightgbm.readthedocs.io` et [KMF+17].
26. Voir `https://catboost.ai/en/docs/` et [PGV+18].
27. Par exemple, XGBoost ajoute systématiquement un terme de pénalisation proportionnel à $\sum_{s \in F(\mathscr{A})} f_s^2$ dans l'ajustement des arbres de régression.

Régression avec plusieurs arbres : mesure de l'importance des variables

Plusieurs notions bien distinctes d'**importance des variables** apparaissent naturellement dans le cadre de la régression avec plusieurs arbres.

Pour une variable explicative x_k, on pourra considérer :

- la réduction d'écart entre réponses observées et valeurs modélisées produite par les ramifications faisant intervenir x_k
- le volume de données pour lesquelles une ramification faisant intervenir x_k est effectivement utilisée par le modèle
- la fréquence avec laquelle x_k intervient dans les ramifications des différents arbres du modèle

D'autres approches (non-spécifiques des modèles de régression par arbres) peuvent également être utilisées, telles que SHAP (« SHapley Additive exPlanations »).

Encodage des variables

Lors de l'utilisation d'un modèle de régression par arbres, il est indispensable de vérifier la façon dont les variables explicatives sont encodées (notamment les variables catégorielles)[28], car divers choix sont possibles, non équivalents du point de vue du modèle obtenu, et conduisant potentiellement à des performances différentes.

28. Par exemple, encoder une variable catégorielle par les indicatrices des différentes modalités (« one-hot ») n'autorise pas les mêmes ramifications qu'un encodage qui la traite comme une seule variable.

Pour aller plus loin

[JWHT13] (chapitre 8), [DHT20]

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-I.html`

Termes d'interaction dans un GLM

L'utilisation de termes d'interaction dans un modèle GLM permet d'affiner la modélisation, de manière parfois considérable, en allant au-delà de la structure purement additive des effets obtenue lorsque les variables sont incorporées individuellement dans la composante systématique du modèle.

En considérant des interactions entre paires de variables, on autorise le modèle GLM à inclure une **modulation** des effets d'une variable en fonction des valeurs d'une autre, dans la composante systématique.

Termes d'interaction dans un GLM

On commence par discuter le cas d'une interaction entre deux **variables catégorielles**.
Dans un GLM de base, les variables explicatives catégorielles $x^{(\ell_1)}$ et $x^{(\ell_2)}$ interviennent dans la composante systématique du modèle via les coefficients $\beta_{(\ell_1,1)}, \ldots, \beta_{(\ell_1,K_1-1)}$ et $\beta_{(\ell_2,1)}, \ldots, \beta_{(\ell_2,K_2-1)}$:

$$\eta = \underbrace{\sum_{k_1=1}^{K_1-1} \beta_{(\ell_1,k_1)} \mathbf{1}(x^{(\ell_1)} = a_{k_1}^{(\ell_1)})}_{\text{effet de } x^{(\ell_1)}} + \underbrace{\sum_{k_2=1}^{K_2-1} \beta_{(\ell_2,k_2)} \mathbf{1}(x^{(\ell_2)} = a_{k_2}^{(\ell_2)})}_{\text{effet de } x^{(\ell_2)}} + \underbrace{\cdots}_{\text{effet des autres var.}}$$

L'effet des variables explicatives sur η est donc **purement additif** : le coefficient $\beta_{(\ell_1,k_1)}$, pour $1 \leq k_1 \leq K_1 - 1$, traduit le changement dans la valeur de η produit par le passage de la modalité de référence $x^{(\ell_1)} = a_0^{(\ell_1)}$ à la modalité $x^{(\ell_1)} = a_{k_1}^{(\ell_1)}$, *quelle que soit la valeur de* $x^{(\ell_2)}$ (et des autres variables explicatives).

Termes d'interaction dans un GLM

L'utilisation de termes d'interaction entre $x^{(\ell_1)}$ et $x^{(\ell_2)}$ dans le modèle permet de modéliser un **effet conjoint** de ces deux variables sur η ne se réduisant pas à une somme d'effets individuels.
Deux manières classiques (mathématiquement équivalentes, mais distinctes du point de vue de l'interprétation et de la mise en pratique) de faire apparaître ce type d'effet :

- interaction « complète » : on crée une nouvelle variable catégorielle $x^{(\ell_1,\ell_2)} = (x^{(\ell_1)}, x^{(\ell_2)})$ qui décrit les valeurs conjointes de $x^{(\ell_1)}$ et $x^{(\ell_2)}$
- interaction « marginale » : on conserve les coefficients individuels associés à $x^{(\ell_1)}$ et $x^{(\ell_2)}$, en ajoutant des coefficients supplémentaires liés aux valeurs conjointes de $x^{(\ell_1)}$ et $x^{(\ell_2)}$, qui viennent ainsi « corriger » la partie purement additive du modèle

Termes d'interaction dans un GLM

Dans le cas d'une **interaction complète**, on aura, en prenant comme modalité de référence pour $x^{(\ell_1,\ell_2)}$ la valeur $(a_0^{(\ell_1)}, a_0^{(\ell_2)})$:

$$\eta = \underbrace{\beta_0 + \sum_{(k_1,k_2)\neq(0,0)} \beta_{((\ell_1,\ell_2),(k_1,k_2))} \mathbf{1}(x^{(\ell_1)} = a_{k_1}^{(\ell_1)}) \mathbf{1}(x^{(\ell_2)} = a_{k_2}^{(\ell_2)})}_{\text{effet de } (x^{(\ell_1)}, x^{(\ell_2)})} + \underbrace{\cdots}_{\text{effet des autres var.}}$$

L'effet global de $x^{(\ell_1)}$ et $x^{(\ell_2)}$ est reflété par (en tout) $K_1 \times K_2 - 1$ coefficients. Il n'y a pas de coefficients correspondant à un « effet individuel » de $x^{(\ell_1)}$ ou de $x^{(\ell_2)}$.

Termes d'interaction dans un GLM

Dans le cas d'une **interaction marginale**, on aura :

$$\eta = \beta_0 + \underbrace{\sum_{k_1=1}^{K_1-1} \beta_{(\ell_1,k_1)} \mathbf{1}(x^{(\ell_1)} = a_{k_1}^{(\ell_1)})}_{\text{effet principal de } x^{(\ell_1)}} + \underbrace{\sum_{k_2=1}^{K_2-1} \beta_{(\ell_2,k_2)} \mathbf{1}(x^{(\ell_2)} = a_{k_2}^{(\ell_2)})}_{\text{effet principal de } x^{(\ell_2)}}$$

$$+ \underbrace{\sum_{k_1=1}^{K_1-1} \sum_{k_2=1}^{K_2-1} \beta_{((\ell_1,\ell_2),(k_1,k_2))} \mathbf{1}(x^{(\ell_1)} = a_{k_1}^{(\ell_1)}) \mathbf{1}(x^{(\ell_2)} = a_{k_2}^{(\ell_2)})}_{\text{interaction marginale de } x^{(\ell_1)} \text{ et } x^{(\ell_2)}} + \underbrace{\cdots}_{\text{effet des autres var.}}$$

L'effet de $x^{(\ell_1)}$ et $x^{(\ell_2)}$ est reflété par (en tout)
$(K_1 - 1) + (K_2 - 1) + (K_1 - 1) \times (K_2 - 1) = K_1 \times K_2 - 1$ coefficients,
comme pour une interaction complète.

Termes d'interaction dans un GLM

L'introduction de termes d'interaction entre deux variables catégorielles comportant des modalités nombreuses entraîne une prolifération de coefficients qui n'est pas souhaitable en général.

Une solution intermédiaire peut consister à introduire un terme d'interaction basé sur une version de $(x^{(1)}, x^{(2)})$ dont certaines modalités ont été fusionnées, afin de conserver l'interaction tout en limitant la prolifération des coefficients.

Termes d'interaction dans un GLM

On peut ajouter des interactions entre plusieurs paires de variables catégorielles Cependant, l'interprétation en termes d'effets peut devenir délicate lorsque plusieurs paires ayant des variables en commun interviennent (p.ex. $(x^{(1)}, x^{(2)})$, $(x^{(2)}, x^{(3)})$, $(x^{(1)}, x^{(3)})$).

On peut *a priori* considérer des interactions entre des $k-$uplets de variables pour $k \geq 3$.  Problèmes éventuels de sur-paramétrisation et d'interprétation.

Termes d'interaction dans un GLM

On discute maintenant des interactions entre **variables quantitatives**, dans le cadre des modèles additifs généralisés. L'interaction entre deux variables pourra se représenter naturellement dans une base tensorielle de fonctions lisses.

- Dans le cas d'une interaction complète :

$$\eta = \beta_0 + \underbrace{\sum_{k_1,k_2} \beta_{k_1,k_2}^{(\ell_1,\ell_2)} b_{k_1}^{(\ell_1)}(x^{(\ell_1)}) \cdot b_{k_2}^{(\ell_2)}(x^{(\ell_2)})}_{\text{effet de } (x^{(\ell_1)}, x^{(\ell_2)})} + \underbrace{\cdots}_{\text{effet des autres var.}} ,$$

- Dans les cas d'une interaction marginale :

$$\eta = \beta_0 + \underbrace{f_{\ell_1}(x^{(\ell_1)})}_{\text{effet princ. de } x^{(\ell_1)}} + \underbrace{f_{\ell_2}(x^{(\ell_2)})}_{\text{effet princ. de } x^{(\ell_2)}} + \underbrace{\sum_{k_1,k_2} \beta_{k_1,k_2}^{(\ell_1,\ell_2)} b_{k_1}^{(\ell_1)}(x^{(\ell_1)}) \cdot b_{k_2}^{(\ell_2)}(x^{(\ell_2)})}_{\text{interaction marginale de } (x^{(\ell_1)}, x^{(\ell_2)})} + \underbrace{\cdots}_{\text{effet des autres var.}} ,$$

  les fonctions f_{ℓ_1}, f_{ℓ_2} étant elles-mêmes décomposées dans des bases de fonctions lisses appropriées.

Termes d'interaction dans un GLM

On renvoie à [Woo17] pour une discussion approfondie de cette approche, à la fois sur le plan théorique et sur celui de sa mise en pratique avec le paquet `mgcv`. Voir notamment :

- `te()` pour les interactions complètes,
- `ti()` pour les interactions marginales,
- `s()` avec un couple de variables pour obtenir une décomposition isotrope adaptée par exemple à un couple de coordonnées spatiales.

Termes d'interaction dans un GLM

Toujours dans le cadre des modèles additifs généralisés, on peut considérer des interactions mixtes, entre une variable **quantitative** et une variable **catégorielle** :

$$\eta = \beta_0 + \underbrace{\sum_{k_1=1}^{K_1-1} \beta_{(\ell_1,k_1)} \mathbf{1}(x^{(\ell_1)} = a_{k_1}^{(\ell_1)})}_{\text{effet principal de } x^{(\ell_1)}} + \underbrace{f_{\ell_2}(x^{(\ell_2)})}_{\text{effet principal de } x^{(\ell_2)}}$$
$$+ \underbrace{\sum_{k_1=0}^{K_1-1} \mathbf{1}(x^{(\ell_1)} = a_{k_1}^{(\ell_1)}) g_{\ell_2}^{(a_{k_1})}(x^{(\ell_2)})}_{\text{interaction marginale de } x^{(\ell_1)} \text{ et } x^{(\ell_2)}}$$

les fonctions f_{ℓ_2} et $g_{\ell_2}^{(a_{k_1})}$ étant décomposées dans des bases appropriées.

Voir le paramètre `by` de la fonction `s()` du paquet `mgcv` (ainsi que le paramètre `id` pour contrôler plus précisément la pénalisation).

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-J.html`

Sélection de variables

D'un point de vue purement statistique, une prolifération excessive de termes dans un modèle de régression est problématique :

- dégradation des performances prédictives (sur-ajustement) ;
- perte de précision dans l'estimation des effets (intervalles de confiance larges et niveaux de significativité faibles).

C'est l'une des raisons (parmi d'autres : robustesse, interprétabilité, etc.) qui peuvent pousser à réduire le jeu de variables utilisé, en ne conservant que les variables qui apparaissent comme pertinentes vis-à-vis de la régression :

- suppression pure et simple d'une variable
- prise en compte adaptée (fusion de catégories, catégorisation de variables quantitatives, approximation via des bases de fonctions appropriées, regroupement de variables, utilisation limitée de termes d'interaction,...)

Sélection de variables

De manière générale, la **sélection de variables** désigne l'ensemble du processus qui, partant des données initiales, aboutit à la façon précise dont celles-ci sont intégrées dans le modèle retenu *in fine*.

De nombreuses approches existent pour mener à bien ce processus, et l'on s'appuie classiquement sur une succession d'étapes du type suivant :

- Etape 1 : examen et retraitement préalables des variables disponibles
- Etape 2 : pré-sélection des variables
- Etape 3 : sélection aboutissant à un modèle avec effets principaux
- Etape 4 : ajout de termes d'interaction

On procède ensuite à la validation/vérification/évaluation globale du modèle obtenu.

Etape 1 : examen et retraitement préalables des variables disponibles

Cette étape consiste d'abord à analyser de manière individuelle chacune des variables disponibles :

- définition opérationnelle de la variable ;
- type et encodage de la variable dans le jeu de données ;
- présence éventuelle de valeurs extrêmes, incohérentes ou manquantes ;
- caractéristiques globales de la distribution des valeurs.

Selon les cas, un certain nombre de retraitements peuvent être effectués :

- suppression / imputation des valeurs incohérentes ou manquantes [29] ;
- catégorisation / changement d'échelle / choix d'une base de fonctions, pour des variables quantitatives ;
- recatégorisation de variables catégorielles ;
- élimination *a priori* de variables non-pertinentes.

29. Attention : supprimer une fraction non-négligeable des données est problématique ! De plus, même une faible fraction des données est susceptible de représenter une part essentielle de l'information qu'elles contiennent !

Etape 2 : pré-sélection des variables

Cette étape repose sur l'étude de :

(A) la dépendance entre la variable réponse et chacune des variables explicatives, prise individuellement ;

(B) la dépendance entre les variables explicatives.

En s'appuyant sur (A), on peut ensuite éventuellement [30] :

- éliminer les variables explicatives dont la liaison individuelle avec la variable réponse est jugée insuffisante ;
- (re-)catégoriser certaines variables en tenant compte de leur liaison individuelle avec la variable réponse pour définir les catégories.

  (idem pour le choix d'une base de fonctions)

En s'appuyant sur (B), on peut d'autre part :

- éliminer certaines variables explicatives dont la liaison avec les autres est problématique (notamment, en présence de multi-colinéarité).

30. Prudence ! On ne prend pas ici en compte la dépendance conjointe de la réponse vis-à-vis de l'ensemble des variables explicatives, qui peut conduire à des résultats différents. Une vérification *a posteriori* de la pertinence des choix effectués s'impose.

Etape 3 : sélection aboutissant à un modèle avec effets principaux

Cette étape repose sur l'étude de la dépendance **conjointe** de la variable réponse vis-à-vis de l'ensemble des variables explicatives retenues.

Une exploration itérative des différentes possibilités d'inclusion/exclusion, de (re-)catégorisation, de choix d'une base de fonctions, pour les différentes variables, est effectuée, afin d'aboutir à un modèle intégrant, en principe, un effet principal pour chaque variable pertinente, encodée de manière appropriée.

On peut ensuite vérifier *a posteriori* en prenant ce modèle comme référence, la pertinence d'une partie des décisions (inclusion/exclusion, catégorisation, etc.) prises lors des étapes 1 et 2.

Etape 4 : ajout de termes d'interaction

En prenant comme référence le modèle avec effets principaux issu de l'étape précédente, on envisage l'ajout de termes d'interaction susceptibles d'améliorer la qualité du modèle.

Pour limiter une prolifération excessive de termes, on limite parfois les termes d'interaction considérés à un sous-ensemble d'interactions jugées pertinentes *a priori*.

(Re)voir le chapitre sur les interactions !

Divers outils pour la sélection de variables

Dans ce qui suit, on présente (ou rappelle !) quelques outils techniques pouvant être utilisés dans les différentes étapes décrites précédemment. Il ne s'agit que d'une sélection d'exemples, au sein d'un grand nombre d'approches et de méthodes envisageables !

Mesures de dépendance entre variables

Il existe une multitude de mesures quantitatives de dépendance entre deux variables, susceptibles dans certains cas de donner lieu à des tests statistiques formels de l'existence d'une dépendance entre celles-ci.
Parmi les plus classiques, selon la nature des variables considérées :

- **catégorielle-catégorielle** : $\mathscr{V}$ de Cramér ;
- **catégorielle-quantitative** : η^2 basé sur les valeurs ou sur les rangs ;
- **quantitative-quantitative** : ρ de Pearson, τ de Kendall, ρ de Spearman.

Mesures de dépendance entre variables : $\mathscr{V}$ de Cramér

Etant donné deux variables u et v prenant respectivement leurs valeurs dans les ensembles $\{a_1, \ldots, a_K\}$ et $\{b_1, \ldots, b_L\}$, et pour lesquelles on dispose de N observations conjointes $(u_i, v_i)_{1 \leq i \leq N}$, le $\mathscr{V}$ de Cramér est défini par :

$$\mathscr{V} \overset{\text{déf.}}{=} \sqrt{\frac{\sum_{k=1}^{K} \sum_{\ell=1}^{L} \frac{\left(\frac{n_{k\ell}}{N} - \frac{n_{k\cdot}}{N} \times \frac{n_{\cdot\ell}}{N}\right)^2}{\frac{n_{k\cdot}}{N} \times \frac{n_{\cdot\ell}}{N}}}{\min(K-1, L-1)}},$$

avec les notations

$$n_{k\ell} \overset{\text{déf.}}{=} \sum_{i=1}^{N} \mathbf{1}_{\{u_i = a_k \text{ et } v_i = b_\ell\}}, \; n_{k\cdot} \overset{\text{déf.}}{=} \sum_{i=1}^{N} \mathbf{1}_{\{u_i = a_k\}} \text{ et } n_{\cdot\ell} \overset{\text{déf.}}{=} \sum_{i=1}^{N} \mathbf{1}_{\{v_i = b_\ell\}}.$$

Mesures de dépendance entre variables : $\mathscr{V}$ de Cramér

- La valeur de $\mathscr{V}$ est toujours comprise entre 0 et 1.
- La valeur limite 1 correspond au cas où l'une des variables peut s'écrire comme une fonction de l'autre.
- Lorsque les valeurs (u_i, v_i) sont issues d'un échantillon i.i.d. de la loi d'un couple de v.a. (U, V), où U et V sont **indépendantes**, $\mathscr{V}$ est proche de 0.
- Dans ce dernier cas, lorsque $N \to +\infty$, on a la convergence $\text{Loi}(N\mathscr{V}^2 \min(K-1, L-1)) \to \chi^2((K-1)(L-1))$, ce qui permet de pratiquer le test du χ^2 d'indépendance à partir de la valeur de $\mathscr{V}$.

Mesures de dépendance entre variables : η^2

Etant donné une variable u (catégorielle) prenant ses valeurs dans $\{a_1, \ldots, a_K\}$ et une variable quantitative v, pour lesquelles on dispose de N observations conjointes $(u_i, v_i)_{1 \leq i \leq N}$, le η^2 est défini par :

$$\eta^2 \overset{\text{déf.}}{=} 1 - \frac{\sum_{k=1}^{K} \sum_{i \in A_k} (v_i - \bar{v}_k)^2}{\sum_{i=1}^{N} (v_i - \bar{v})^2},$$

avec les notations

$$\bar{v} = \frac{\sum_{i=1}^{N} v_i}{N}, A_k = \{i;\ u_i = a_k\}, n_k = \#A_k \text{ et } \bar{v}_k = \frac{\sum_{i \in A_k}^{n} v_i}{n_k}.$$

Mesures de dépendance entre variables : η^2

Le η^2 n'est autre que le R^2 obtenu avec un modèle linéaire avec v pour variable réponse et u pour variable explicative, ce qui clarifie son interprétation.

- La valeur de η^2 est toujours comprise entre 0 et 1.
- La valeur 1 correspond au cas où v s'exprime comme une fonction de u
- Lorsque les valeurs (u_i, v_i) sont issues d'un échantillon i.i.d. de la loi d'un couple de v.a. (U, V), où U et V sont **indépendantes**, η^2 est proche de 0.
- Dans la limite où $N \to +\infty$, les tests d'ANOVA à un facteur restent valables même en l'absence de normalité, ce qui permet de pratiquer un test d'indépendance.

  Dans les faits, il s'agit d'un test d'égalité des moyennes entre les différents groupes, valable asymptotiquement sans l'hypothèse de normalité, mais sensible à une éventuelle hétéroscédasticité entre les groupes.

Mesures de dépendance entre variables : η^2 sur les rangs

Dans le même contexte que celui pour lequel est défini le η^2, on considère les valeurs des rangs [31] :

$$\tilde{v}_i \overset{\text{déf.}}{=} \text{rang de } v_i \text{ parmi } v_1, \ldots, v_N.$$

On définit ensuite le $\tilde{\eta}^2$ comme étant le η^2 associé aux couples $(u_i, \tilde{v}_i)_{1 \leq i \leq N}$.

- Les propriétés générales précédentes du η^2 sont conservées.
- Le fait de travailler avec des rangs permet d'utiliser le $\tilde{\eta}^2$ pour pratiquer un test d'indépendance non-paramétrique : le test de Kruskal-Wallis.

31. Lorsque plusieurs valeurs sont égales, on attribue à chacune d'entre elles la valeur moyenne des rangs correspondants.

Mesures de dépendance entre variables : ρ, τ, ρ_S

Etant donné deux variables quantitatives u et v, et pour lesquelles on dispose de N observations conjointes $(u_i, v_i)_{1 \leq i \leq N}$, on considère :

- le coefficient de corrélation ρ de Pearson :

$$\rho = \frac{\sum_{i=1}^{N}(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^{N}(u_i - \bar{u})^2}\sqrt{\sum_{i=1}^{N}(v_i - \bar{v})^2}}$$

- le τ de Kendall :

$$\tau = \frac{\sum_{1 \leq i < j \leq N} \text{signe}((u_i - v_i)(u_j - v_j))}{N(N-1)/2}$$

- le ρ_S de Spearman : c'est le ρ de Pearson calculé sur les rangs $\tilde{u}_i$ et $\tilde{v}_i$

Mesures de dépendance entre variables : ρ, τ, ρ_S

Pour chacune des trois mesures ainsi définies :

- La valeur est toujours comprise entre -1 et 1.
- Les valeurs ± 1 correspondent au cas où une variable s'exprime comme une fonction (affine pour ρ, monotone pour τ et ρ_S) de l'autre
- Lorsque les valeurs (u_i, v_i) sont issues d'un échantillon i.i.d. de la loi d'un couple de v.a. (U, V), où U et V sont **indépendantes**, la valeur est proche de 0.
- Des tests d'indépendance peuvent être pratiqués à partir de ces valeurs (asymptotique ou sous l'hypothèse de normalité pour ρ, non-paramétriques pour τ et ρ_S)

Mesures de dépendance entre variables

Quelques remarques et points d'attention :

- L'utilisation de versions catégorisées des variables quantitatives  permet de traiter celles-ci comme des variables catégorielles dans les définitions des diverses mesures précédentes.
- En cas d'utilisation avec des données groupées, il faut intégrer correctement les effectifs des groupes dans les calculs.
- La présence de groupes de taille insuffisante, ou d'égalités entre valeurs, peut fausser la fiabilité des lois de référence utilisées pour les tests.
- En cas d'utilisation avec une variable V (typiquement, la réponse) dont la loi est *a priori* modulée par une exposition elle-même variable, l'utilisation directe des mesures précédentes peut se révéler inadaptée.

  Faute d'approche plus adaptée, il convient par exemple de s'assurer que l'exposition peut bien être considérée comme indépendante de la variable U.

Utilisation pour la sélection de variables

Dans une optique de sélection de variables, de telles mesures de dépendance peuvent être utilisées :

- pour quantifier l'intensité de l'association individuelle entre une variable explicative et la réponse, afin de pré-sélectionner les variables explicatives présentant l'association la plus forte avec la réponse ;
- pour quantifier l'intensité de l'association entre paires de variables explicatives, afin d'identifier des dépendances susceptibles de poser problème et/ou pouvant justifier l'élimination de certaines variables.

Association individuelle entre variable explicative et réponse

L'examen de l'association individuelle de chaque variable explicative avec la réponse fournit une information intéressante, mais ne prend pas en compte la dépendance **conjointe** de la réponse vis-à-vis de l'ensemble des variables explicatives, qui peut conduire à des résultats différents :

- association individuelle observée, disparaissant une fois que la dépendance conjointe est modélisée ;
- association individuelle non-observée, apparaissant une fois que la dépendance conjointe est modélisée.

Par conséquent, les résultats d'une pré-sélection basée sur l'association individuelle doivent être vérifiés *a posteriori*.

Parallèlement aux mesures de dépendance précédentes, on utilisera souvent avec profit un tracé de la valeur moyenne de la réponse en fonction de la variable explicative considérée.

Dépendance entre variables explicatives

L'examen des dépendances entre paires de variables explicatives permet déjà d'identifier certaines redondances *a priori* problématiques dans la composante systématique du modèle :

- multi-colinéarité exacte (impossibilité d'estimer les coefficients) ;
- multi-colinéarité approchée (forte imprécision sur l'estimation des coefficients).

Dans ces cas, on choisit souvent d'éliminer purement et simplement l'une des deux variables.

 Cette approche ne permet pas directement d'identifier des redondances problématiques au sein de groupes de plus de deux variables.

Dépendance entre variables explicatives

Plus globalement, on pourra chercher à construire des groupes de variables explicatives liées :

- pour identifier d'éventuelles redondances,
- dans une optique de limitation délibérée du nombre de variables explicatives,
- dans un but purement exploratoire.

Voir par exemple le paquet `ClustOfVar` et [CKSLS12] pour un exemple d'approche permettant de mener cela à bien, autorisant une visualisation sous forme de dendrogramme. Plus généralement, voir le cours d'analyse de données !

Dépendance entre variables explicatives : VIF

Les facteurs d'inflation de la variance (« Variance Inflation Factor » ou plus brièvement VIF en anglais) sont un outil permettant d'évaluer *a posteriori* l'impact de la dépendance entre variables explicatives sur la précision de l'estimation des coefficients.

Une définition adaptée aux GLM et à l'encodage de l'effet d'une variable explicative via plusieurs coefficients est proposée dans [Fox15, FM92], et disponible via la fonction `vif` du paquet `car`.

L'approche consiste (en gros) à comparer le volume de l'ellipsoïde traduisant l'incertitude d'estimation sur les coefficients d'une variable, au volume que l'on obtiendrait en l'absence de corrélation avec les autres variables, toutes choses égales par ailleurs.

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-K.html`

Exploration des sous-ensembles de variables

Selon la taille du jeu de données (nombre de variables ET nombre d'observations), diverses possibilités d'exploration automatique des sous-ensembles de variables peuvent être envisagées.

- exploration « stepwise » : exploration par ajout et/ou suppression pas-à-pas des différentes variables ;
- exploration exhaustive : toutes les combinaisons de variables sont examinées ;
- exploration guidée par un algorithme d'optimisation élaboré (p. ex. : algo. génétique)

Ces différentes approches supposent le choix d'un critère de comparaison des modèles (p. ex. : AIC, performance prédictive sur données de test, tests statistiques, etc.)

Exploration des sous-ensembles de variables

 Sélectionner les variables de manière purement automatique n'est pas forcément conseillé, et une part plus ou moins grande peut être accordée à l'**appréciation individuelle** pour guider l'exploration (voir par exemple l'approche « purposeful selection of covariates » présentée dans [HLS13]).

Même dans une approche automatisée, diverses considérations peuvent amener à imposer des contraintes supplémentaires (p.ex. : variables ou blocs de variables à inclure obligatoirement, limitation volontaire du nombre de variables, etc.).

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-L.html`

Fusion de catégories

Au cours des diverses étapes de sélection des variables (et notamment l'étape 3), le problème de la fusion éventuelle de certaines catégories d'une même variable (catégorielle) se pose fréquemment, en particulier en présence d'un nombre important de catégories différentes. Les motivations pour fusionner des catégories peuvent notamment être les suivantes :

- limiter l'incertitude sur l'estimation des effets ;
- éviter la présence dans le modèle d'effets distincts mais dont la différence possède une faible significativité statistique ;
- améliorer les performances prédictives.

Ces objectifs sont à mettre en regard avec la perte d'information que représente l'abandon de la distinction entre des catégories fusionnées.

Le type de fusion envisageable dépend du type de variable considéré (catégorielle pure, catégorielle ordinale, catégorielle avec une structure géographique, etc.)

Fusion de catégories guidée par des tests statistiques

Les tests statistiques d'égalité entre coefficients fournissent une manière d'aborder la fusion de catégories. On peut les utiliser dans le cadre d'une approche « manuelle », par exemple : tracé des coefficients avec intervalles de confiance, proposition de fusion validée ensuite par un test statistique, ou partiellement automatique, par exemple, fusion des deux catégories pour lesquelles la $p-$valeur d'un test d'égalité est la plus élevée, cette procédure étant répétée jusqu'à ce qu'aucun test d'égalité ne donne une $p-$valeur supérieure à un seuil pré-défini.

Ces approches sont simples à mettre en œuvre et à comprendre, mais ne sont pas exemptes de défauts : la multiplicité des tests pratiqués, et leur orientation par l'examen préalable des données, rendent peu claire la significativité statistique globale des résultats obtenus. Elles peuvent par exemple être complétées par des indicateurs de performance prédictive comparant les modèles ainsi recatégorisés.

Il s'agit d'une approche « descendante » au sens où l'on élimine progressivement la distinction entre des catégories.

Fusion de catégories par ajustement itératif d'arbres

La fusion de catégories par ajustement itératif d'arbres consiste à construire itérativement, pour chacune des variables catégorielles concernées, un partitionnement arborescent des modalités utilisées par le modèle GLM. Par exemple, à chaque étape, un découpage supplémentaire est ajouté, en choisissant le découpage provoquant la meilleure amélioration globale du modèle au sens de la déviance, voir [TB18]. Il s'agit alors d'une version « ascendante » de l'approche précédente : on raffine progressivement une partition par fission, au lieu de la rendre progressivement plus grossière par fusion.

 Il ne s'agit pas de construire un arbre de régression : on construit un modèle GLM dans lequel certaines variables catégorielles sont recatégorisées via un découpage arborescent de leurs modalités.

Fusion de catégories par pénalisation de type LASSO

La fusion de catégories par pénalisation de type LASSO (« Least Absolute Shrinkage and Selection Operator ») consiste à ajouter à la vraisemblance un terme qui pénalise les différences d'effets entre catégories, la forme spécifique de la pénalité choisie permettant de forcer certaines différences pénalisées à prendre une valeur nulle.

Les catégories dont les effets ainsi ajustés sont identiques sont ensuite fusionnées dans un GLM dont les coefficients sont estimés par maximum de vraisemblance (non-pénalisée).

En faisant varier l'intensité de la pénalisation (contrôlée par un unique paramètre global), on parcourt un ensemble de fusions de catégories possibles, parmi lesquelles on peut choisir celle donnant lieu au modèle qui maximise un critère de performance prédictive.

Fusion de catégories par pénalisation de type LASSO

Pour une variable catégorielle $x^{(j)}$ possédant les modalités $a_0, \ldots, a_{K-1}$, avec comme modalité de référence a_0, les coefficients correspondant étant notés $\beta_{(j,0)}, \ldots, \beta_{(j,K-1)}$ (avec $\beta_{(j,0)} \equiv 0$) on rencontrera les pénalités [32] :

- *fused* LASSO pour une variable catégorielle ordinale dont les modalités $a_0, \ldots, a_{K-1}$ sont ordonnées (avec $K \geq 2$) :

$$-\sum_{k=0}^{K-2} \lambda_{j,k} \left|\beta_{(j,k+1)} - \beta_{(j,k)}\right|$$

- *generalized fused* LASSO pour une variable catégorielle pure (avec $K \geq 2$) :

$$-\sum_{0 \leq k < \ell < K-1} \lambda_{j,k,\ell} \left|\beta_{(j,\ell)} - \beta_{(j,k)}\right|$$

32. Mentionnons encore la possibilité plus générale consistant à pénaliser les différences entre coefficients associés à des modalités « adjacentes » au sens d'une matrice d'adjacence donnée (*graph guided fused* lasso).

Fusion de catégories par pénalisation de type LASSO

Une forme possible pour les coefficients de pénalisation précédents est :

$$\lambda_{j,k} = \lambda w_{j,k},\ \lambda_{j,k,\ell} = \lambda w_{j,k,\ell},$$

où λ est un unique coefficient global de pénalisation, et les facteurs $w_{j,k}$, $w_{j,k,\ell}$ sont des poids spécifiés *a priori*.
Pour définir les poids, on peut par exemple partir des coefficients $\hat{\beta}$ estimés par un modèle GLM non-pénalisé, et prendre :

$$w_{j,k} = 1/|\hat{\beta}_{j,k+1} - \hat{\beta}_{j,k}|,\ w_{j,k,\ell} = 1/|\hat{\beta}_{j,\ell} - \hat{\beta}_{j,k}|.$$

Il reste ensuite à parcourir un jeu de valeurs pour λ, en optimisant un critère de performance pour le modèle ajusté (GCV, etc.), comme dans l'approche GAM.
On renvoie à [DARV21] et à la documentation du paquet `smurf` pour une description détaillée de cette approche et de sa mise en œuvre.

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-M.html`

Crédibilité

En présence d'une variable catégorielle (disons $x^{(\ell)}$) présentant des modalités trop nombreuses pour autoriser une estimation suffisamment précise des effets de chaque catégorie, une alternative (parmi d'autres) à la fusion de catégories est l'utilisation d'un **modèle à effet mixtes**, traitant l'effet de $x^{(\ell)}$ comme une variable aléatoire, dont l'estimation peut par exemple s'appuyer sur une approche de type **crédibilité**.

Pour un GLM log-Poisson, l'approche proposée par [OJ10, Ohl08] conduit à écrire, lorsque $x^{(\ell)} = a$, les variables $x^{(k)}$ pour $k \neq \ell$ étant encodées par les $\mathrm{x}^{(j)}$, $1 \leq j \leq d_{-\ell}$:

$$\mathbb{E}\left(Y|X = x, U_a\right) = U_a \cdot \exp\left(\sum_{j=0}^{d_{-\ell}} \beta_j \mathrm{x}^{(j)}\right), \tag{1}$$

où les effets aléatoires $(U_a)_a$ sont vus *a priori* comme des v.a. i.i.d. vérifiant $\mathbb{E}(U_a) = 1$.

Crédibilité

Pour estimer les différents paramètres (les effets fixes, donnés par les β_j, et les effets aléatoires, donnés par les U_j), on note que :

- les valeurs des U_a étant supposées connues, l'équation (1) conduit à estimer les β_j par maximum de vraisemblance, en traitant $\log(U_a)$ comme un offset ;
- les valeurs des β_j étant supposées connues, l'équation (1) conduit à estimer les U_a par l'approche de Bühlmann-Straub : $\hat{U}_a = z_a \bar{\tilde{y}}_a + (1 - z_a)$, où les z_a sont des poids de crédibilité, et $\bar{\tilde{y}}_a$ est une moyenne pondérée des valeurs normalisées des réponses y_i lorsque $x_i^{(\ell)} = a$.

L'ensemble des paramètres du modèle est alors estimé par une approche itérative, en alternant les estimations des β_j et des U_a à partir des valeurs précédemment estimées, jusqu'à la convergence apparente des valeurs obtenues.
Voir [DHT19] pour une discussion plus générale des modèles à effets mixtes dans un contexte actuariel.

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-N.html`

Catégorisation d'effets lisses

Il arrive fréquemment que l'on souhaite, *in fine*, utiliser un modèle dans lequel l'intégralité des variables explicatives sont catégorielles, même si l'on dispose des valeurs numériques des variables quantitatives.
Une alternative à la catégorisation *a priori* des variables quantitatives consiste à procéder en deux étapes :

- ajustement des effets sous forme de termes lisses à l'aide d'un modèle GAM ;
- catégorisation *a posteriori* des effets ainsi modélisés.

La catégorisation *a posteriori* peut être effectuée manuellement (p. ex. à l'aide d'un examen graphique des effets ajustés), soit en s'appuyant sur une approche plus systématique, par exemple en ajustant un arbre de régression sur les effets modélisés (voir par exemple [HACV18]).

Avec R

Exemples, code et explications se trouvent dans le bloc-notes :
`https://irma.math.unistra.fr/~jberard/nb-GLM-0.html`

Spécification des modèles

Les modèles GLM étudiés dans ce cours reposent *a priori* sur un ensemble d'hypothèses fortes sur la loi de Y_i, qui peuvent se révéler inadaptées aux données que l'on manipule. Dès lors, la question se pose de savoir comment se comportent les méthodes d'inférence décrites précédemment lorsque ces hypothèses ne sont pas – ou du moins pas totalement – satisfaites.

Dans la suite, on continue de supposer que les réponses observées $y_1, \ldots, y_N$ peuvent être modélisées par des variables aléatoires indépendantes $Y_1, \ldots, Y_N$, et l'on considère plusieurs situations où une partie seulement des caractéristiques des v.a. $Y_1, \ldots, Y_N$ est spécifiée de manière correcte par le modèle.

Spécification des modèles

- On dit que l'espérance est correctement spécifiée par le modèle lorsque l'égalité : $\mathbb{E}(Y_i) = g^{-1}(\eta_i)$ est effectivement vérifiée où $\eta_i = \sum_{j=0}^{d} \beta_j x_i^{(j)}$, g est la fonction de lien du modèle, et $(\beta_j)_{0 \leq j \leq d}$ une famille de coefficients.
- On dit que la variance est correctement spécifiée par le modèle lorsque l'égalité : $\mathbb{V}(Y_i) = (\phi / w_i) \cdot v\left(g^{-1}(\eta_i)\right)$ est effectivement vérifiée, où v est la fonction de variance du modèle, et $\phi > 0$ une valeur possible du paramètre de dispersion.
- On dit que la loi de Y_i est correctement spécifiée, lorsque $\text{Loi}(Y_i) = \mathscr{L}_v(g^{-1}(\eta_i), \phi / w_i)$; le modèle est alors bien spécifié dans son ensemble.

Espérance correctement spécifiée

Lorsque l'espérance est correctement spécifiée, l'ajustement du modèle GLM par maximum de vraisemblance produit, sous des hypothèses de régularité et dans la limite d'un grand nombre de données, une estimation consistante et asymptotiquement normale des coefficients β :

$$\hat{\beta} \approx \beta \text{ et } \text{Loi}(\hat{\beta} - \beta) \approx \mathcal{N}(0, \Sigma(\hat{\beta})).$$

Ainsi, bien que l'on ne suppose pas que la loi, ni même la variance, soit correctement spécifiée, on obtient encore une estimation pertinente des coefficients. En revanche, sans hypothèse supplémentaire, la variance n'est pas correctement estimée en général, et la matrice de variance-covariance des coefficients non plus ; en particulier, les intervalles de confiance et tests pratiqués sur les coefficients produisent en général des résultats **incorrects**.

On parle de **pseudo-vraisemblance** pour qualifier cette approche : la fonction de vraisemblance que l'on maximise n'est plus la fonction de vraisemblance de la loi qui gouverne les données, mais peut néanmoins être utilisée pour produire une estimation pertinente d'une partie des paramètres du modèle.

Espérance correctement spécifiée

Pour obtenir une estimation consistante de la matrice $\Sigma(\hat{\beta})$ de $\hat{\beta}$ on peut utiliser l'estimateur « sandwich » :

$$\hat{V}^{-1} \times \hat{B} \times \hat{V}^{-1},$$

où les matrices de taille $(d+1) \times (d+1)$ $\hat{V}$ et $\hat{B}$ sont définies par :

$$\hat{V}_{j,k} = \sum_{i=1}^{N} w_i \frac{\mathrm{x}_i^{(j)} \mathrm{x}_i^{(k)}}{v(\hat{\mu}_i)} \left(\frac{1}{g'(\hat{\mu}_i)} \right)^2$$

$$\hat{B}_{j,k} = \sum_{i=1}^{N} \left(w_i \frac{y_i - \hat{\mu}_i}{v(\hat{\mu}_i) g'(\hat{\mu}_i)} \right)^2 \mathrm{x}_i^{(j)} \mathrm{x}_i^{(k)}$$

Une variante consiste à prendre pour $\hat{V}$ la matrice définie par

$$\hat{V}_{j,k} = \sum_{i=1}^{N} \frac{\partial(\phi \log L_i)}{\partial \beta_j}(\hat{\beta}) \frac{\partial(\phi \log L_i)}{\partial \beta_k}(\hat{\beta})$$

Espérance et variance correctement spécifiées

Lorsque l'on suppose l'espérance et la variance sont bien spécifiées, on peut voir les équations habituelles d'estimation des coefficients :

$$\sum_{i=1}^{N} \frac{w_i(y_i - \mu_i)x_i^{(j)}}{v(\mu_i)g'(\mu_i)} = 0, \text{ pour } j = 0, \ldots, d, \tag{2}$$

comme une approche d'estimation à part entière, sans référence à la véritable fonction de vraisemblance qui gouverne les données. Cette approche demeure pertinente pour une fonction v et des valeurs de ϕ arbitraires, sans référence à la famille exponentielle (qui impose des restrictions sur v et ϕ).
Les équations (2) se résolvent numériquement comme pour un GLM classique, et l'on peut utiliser une méthode de moments pour l'estimation de ϕ. On parle alors de méthode de **quasi-vraisemblance**, la fonction que l'on maximise pour estimer les coefficients n'étant plus supposée être une fonction de vraisemblance, même incorrectement spécifiée.

Espérance et variance correctement spécifiées

Si l'espérance et la variance sont bien spécifiées, l'approche de quasi-vraisemblance fournit une estimation consistante des paramètres, même lorsqu'il n'existe pas de loi de la famille exponentielle correspondant aux valeurs de v et ϕ.
De plus, les coefficients possèdent encore une loi asymptotiquement normale dont la matrice de variance-covariance est estimée de manière consistante par les formules utilisées dans le cadre GLM classique.

Un cadre plus général, non-développé ici, est celui des équations d'estimation généralisées (« GEE » pour Generalized Estimating Equations en anglais), qui permettent notamment de tenir compte de dépendances entre les variables Y_i.

Quelques limites courantes du modèle poissonien

Le modèle de comptage fondé sur la loi de Poisson (souvent avec une fonction de lien logarithmique) est fréquemment utilisé, au moins comme première approximation.
Il arrive couramment de rencontrer, de manière plus ou moins prononcée, les problèmes suivants :

- la variance observée dans les données est supérieure à celle prédite par le modèle (« sur-dispersion ») ;
- la fréquence des zéros est supérieure de celle prédite par le modèle (« excès de zéros »).

Diverses extensions du modèle visent, entre autres, à remédier à ces problèmes.

Sur-dispersion

Deux manières classiques (et relativement simples) d'incorporer une sur-dispersion par rapport au modèle poissonien :

- GLM basé sur la loi binomiale négative. La fonction de variance associée est alors (voir annexe) $v(\mu) = \mu(1 + \mu/k)$, le paramètre k devant être estimé, par exemple par maximum de vraisemblance (la valeur de $k > 0$ étant fixée, le modèle est alors un GLM classique).
- approche de quasi-vraisemblance avec $v(\mu) = \mu$. La valeur de ϕ n'étant plus contrainte à valoir 1, on peut ainsi modéliser des situations où $\mathbb{V}(Y) = \phi\mathbb{E}(Y)$, avec $\phi \neq 1$. Le cas $\phi > 1$ traduit une sur-dispersion, tandis que le cas $\phi < 1$ traduit une sous-dispersion. On note que, si $Z \sim \mathscr{P}(\mu/\phi)$, $Y = \phi Z$ a précisément pour espérance μ et pour variance $\phi\mu$, donc on a au moins un exemple simple de loi possédant explicitement ces propriétés, mais les valeurs de Y ne sont pas en général des entiers...

Excès de zéros

Deux manières assez classiques d'incorporer un excès de zéros par rapport au modèle poissonien :

- modèle à inflation de zéros (« zero inflated ») :
  $Y \sim p\delta_0 + (1-p)\mathscr{P}(\lambda)$,
- modèle avec loi de Poisson tronquée :
  $Y \sim p\delta_0 + (1-p)\text{Loi}(Z|Z \geq 1)$, où $Z \sim \mathscr{P}(\lambda)$;

où λ et p sont reliés aux variables explicatives par leur propre fonction de lien et leurs propres coefficients.
Ces modèles sont classiquement estimés par maximum de vraisemblance.
On peut également utiliser une loi binomiale négative au lieu d'une loi de Poisson, pour augmenter la flexibilité du modèle.

Quelques remarques

- Il n'est pas recommandé d'avoir recours aux modèles présentés précédemment sans avoir au préalable éliminé d'autres causes possibles d'inadéquation du modèle utilisé (mauvaise prise en compte des effets individuels de certaines variables, omission de termes d'interaction importants, voire choix inadapté de la fonction de lien, etc.)
- Il se peut bien entendu qu'aucun des modèles présentés précédemment ne permette une modélisation adéquate de la sur-dispersion et/ou de l'excès de zéros observé. On rappelle que, de manière générale, et particulièrement si l'objectif principal est la modélisation de l'espérance de Y, l'approche de pseudo-vraisemblance demeure un outil valable pour travailler avec des modèles incorrectement spécifiés.

Bibliographie

Quelques références utilisées dans ce cours :

- Régression (de manière générale) : [FKLM13]
- Modèles linéaires généralisés (de manière générale) : [FT01, DS18]
- Applications actuarielles : [DJH+08, DC+05, DHT19, DHT20, DMPW07, Fre10, GFKT16, OJ10, AFM+04]
- Modèles spécifiques : [HLS13, Hil11]
- Modélisation prédictive : [JWHT13, Tuf15, Woo17]
- Points plus spécifiques : [Fox15, FM92, DS96, Jör97, Son07]

[AFM+04] Duncan Anderson, Sholom Feldblum, Claudine Modlin, Doris Schirmacher, Ernesto Schirmacher, and Neeza Thandi, A practitioner's guide to generalized linear models, Casualty Actuarial Society Discussion Paper Program (2004), 1–116.

[CG16] Tianqi Chen and Carlos Guestrin, Xgboost : A scalable tree boosting system, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '16, Association for Computing Machinery, 2016, p. 785–794.

[CKSLS12] Marie Chavent, Vanessa Kuentz-Simonet, Benoît Liquet, and Jérôme Saracco, ClustOfVar : An R Package for the Clustering of Variables, Journal of Statistical Software **50** (2012), no. i13.

[DARV21] Sander Devriendt, Katrien Antonio, Tom Reynkens, and Roel Verbelen, Sparse regression with multi-type regularized feature modeling, Insurance : Mathematics and Economics **96** (2021), 248–261.

[DC+05] Michel Denuit, Arthur Charpentier, et al., Mathématiques de l'assurance non-vie. Tome II : tarification et provisionnement, Economica, 2005.

[DHT19] Michel Denuit, Donatien Hainaut, and Julien Trufin, Effective statistical learning methods for actuaries I, Springer Actuarial, Springer, 2019, GLMs and extensions.

[DHT20] _____, Effective statistical learning methods for actuaries II, Springer Actuarial, Springer, 2020, Tree-Based Methods and Extensions.

[DJH+08] Piet De Jong, Gillian Z Heller, et al., Generalized linear models for insurance data, Cambridge University Press, 2008.

[DMPW07] Michel Denuit, Xavier Maréchal, Sandra Pitrebois, and Jean-François Walhin, Actuarial modelling of claim counts, John Wiley & Sons, Ltd., Chichester, 2007, Risk classification, credibility and bonus-malus systems, With a foreword by Ragnar Norberg.

[DS96] Peter K. Dunn and Gordon K. Smyth, Randomized quantile residuals, Journal of Computational and Graphical Statistics **5** (1996), no. 3, 236–244.

[DS18] _____, Generalized linear models with examples in R, Springer Texts in Statistics, Springer, New York, 2018.

[DST19] Michel Denuit, Dominik Sznajder, and Julien Trufin, Model selection based on Lorenz and concentration curves, Gini indices and convex order, Insurance : Mathematics and Economics **89** (2019), 128 – 139.

[FKLM13] Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian Marx, Regression, Springer, Heidelberg, 2013, Models, methods and applications.

[Fle12] D. J. Fletcher, Estimating overdispersion when fitting a generalized linear model to sparse data, Biometrika **99** (2012), no. 1, 230–237.

[FM92] John Fox and Georges Monette, Generalized collinearity diagnostics, Journal of the American Statistical Association **87** (1992), no. 417, 178–183.

[Fox15] John Fox, Applied Regression Analysis and Generalized Linear Models, Sage Publications, 2015.

[Fre10] Edward W. Frees, Regression modeling with actuarial and financial applications, International Series on Actuarial Science, Cambridge University Press, Cambridge, 2010.

[FT01] Ludwig Fahrmeir and Gerhard Tutz, Multivariate statistical modelling based on generalized linear models, second ed., Springer Series in Statistics, Springer-Verlag, New York, 2001, With contributions by Wolfgang Hennevogl.

[GFKT16] Mark Goldburd, Sholom Feldblum, Anand Khare, and Dan Tevet, Generalized linear models for insurance rating, Casualty Actuarial Society, 2016.

[HACV18] Roel Henckaerts, Katrien Antonio, Maxime Clijsters, and Roel Verbelen, A data driven binning strategy for the construction of insurance tariff classes, Scand. Actuar. J. (2018), no. 8, 681–705.

[Hil11] Joseph M. Hilbe, Negative binomial regression, second ed., Cambridge University Press, Cambridge, 2011.

[HLS13] David Hosmer, Stanley Lemeshow, and Rodney Sturdivant, Applied logistic regression, John Wiley & Sons, 2013.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Springer New York, New York, NY, 2009.

[Jör97] Bent Jörgensen, The theory of dispersion models, Monographs on Statistics and Applied Probability, vol. 76, Chapman & Hall, London, 1997.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An introduction to statistical learning, Springer Texts in Statistics, vol. 103, Springer, New York, 2013, With applications in R.

[KMF+17] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu, Lightgbm : A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017).

[Loa99] Clive Loader, Local regression and likelihood, Statistics and Computing, Springer-Verlag, New York, 1999.

[Ohl08] Esbjörn Ohlsson, Combining generalized linear models and credibility models in practice, Scandinavian Actuarial Journal 2008 (2008), no. 4, 301–314.

[OJ10] Esbjörn Ohlsson and Björn Johansson, Non-life insurance pricing with generalized linear models, Springer, 2010.

[PGV+18] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin, Catboost : unbiased boosting with categorical features, Advances in neural information processing systems 31 (2018).

[Son07] Peter X.-K. Song, Correlated data analysis : modeling, analytics, and applications, Springer Series in Statistics, Springer, New York, 2007.

[TB18] Gerhard Tutz and Moritz Berger, Tree-structured modelling of categorical predictors in generalized additive regression, Advances in Data Analysis and Classification 12 (2018), no. 3, 737–758.

[Tuf15] Stéphane Tufféry, Modélisation prédictive et apprentissage statistique avec R, Éditions Technip, 2015.

[Vuo89] Quang H. Vuong, Likelihood ratio tests for model selection and nonnested hypotheses, Econometrica 57 (1989), no. 2, 307–333. MR 996939

[Woo17] Simon N. Wood, Generalized additive models, Texts in Statistical Science Series, CRC Press, Boca Raton, FL, 2017, An introduction with R, Second edition.

On n'en a pas parlé (ou à peine)

- modèles à effets mixtes
- pénalisation Ridge et lissage manuel des effets pour des variables catégorielles
- modèles additifs généralisés pour position, échelle et forme (GAMLSS)
- modèles « single-index »
- ...

Famille exponentielle : expression de $\mathbb{E}(Y)$ et $\mathbb{V}(Y)$

On se place dans le cas discret et l'on raisonne de manière formelle (sans chercher à justifier les interversions entre dérivation et sommation). On part de l'identité

$$\sum_y f(y) = 1, \text{ où } f(y) = c(y,\phi) \cdot \exp\left(\frac{y\theta - b(\theta)}{\phi}\right).$$

En dérivant cette identité par rapport à θ, on en déduit que

$$\sum_y c(y,\phi) \cdot \frac{(y - b'(\theta))}{\phi} \cdot \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) = 0,$$

d'où

$$\sum_y (y - b'(\theta)) f(y) = 0,$$

d'où

$$\mathbb{E}(Y) = \sum_y y f(y) = \sum_y b'(\theta) f(y) = b'(\theta) \sum_y f(y) = b'(\theta).$$

En dérivant une deuxième fois, on aboutit à

$$\sum_y c(y,\phi) \cdot \left(\frac{(y - b'(\theta))^2}{\phi} - b''(\theta)\right) \cdot \exp\left(\frac{y\theta - b(\theta)}{\phi}\right) = 0,$$

d'où

$$\sum_y \left((y - b'(\theta))^2 - \phi b''(\theta)\right) \cdot f(y) = 0.$$

Comme $\mathbb{E}(Y) = b'(\theta)$, on a $\mathbb{V}(Y) = \sum_y (y - b'(\theta))^2 f(y)$, et l'on en déduit facilement que $\mathbb{V}(Y) = \phi b''(\theta)$.

Famille exponentielle : loi binomiale négative $\mathcal{NB}(k, p)$

On suppose que $Y \sim \mathcal{NB}(k, p)$, c'est-à-dire que, pour tout entier $n \geq 0$,

$$\mathbb{P}(Y = n) = \frac{\Gamma(n+k)}{\Gamma(k)n!} p^k (1-p)^n.$$

Alors la loi de Y se rattache à la famille exponentielle, avec les caractéristiques suivantes :

Nom	binomiale négative
Modèle	$Y \sim \mathcal{NB}(k, p)$
μ	$kp/(1-p)$
ϕ	1
$v(\mu)$	$\mu(1+\mu/k)$
θ	$\log(1-p)$
$b(\theta)$	$-k\log(1-e^{\theta})$

Ici, le paramètre k intervient dans l'expression de $v(\cdot)$, on le suppose fixé *a priori*.

Loi binomiale négative et mélange Poisson-Gamma

La loi binomiale négative apparaît naturellement comme un modèle de mélange, dans lequel le paramètre d'une loi de Poisson est modulé par une variable aléatoire représentant une hétérogénéité individuelle venant se superposer à l'effet des variables explicatives. Précisément, si :

- $\forall h \geq 0,\ \text{Loi}(Y|\ H = h) = \mathcal{P}(\lambda h)$,
- $H \sim \mathcal{G}amma(k, 1/k)$ avec donc $\mathbb{E}(H) = 1$,

la loi de Y (non-conditionnelle) est alors la loi binomiale négative d'espérance $\mu = \lambda$ et de variance $\mu(1 + \mu/k)$.

Modèle log-binomiale négative et expositions variables

Dans un modèle GLM reposant sur la loi binomiale négative et une fonction de lien logarithmique, on dispose d'au moins deux options **non-équivalentes** du point de vue de la modélisation pour prendre en compte des différences d'exposition :

(a) Utilisation du logarithme de l'exposition comme offset avec les réponses brutes

(b) Utilisation de réponses normalisées par l'exposition et de poids égaux à l'exposition

L'option (a) correspond, dans le cadre du modèle Poisson-Gamma, au fait que, conditionnellement à $H_i = h_i$, Y_i suit la loi de Poisson $\mathcal{P}(t_i\mu_i h_i)$, où t_i est l'exposition, et $\mu_i = \exp(\eta_i)$ est la moyenne prédite par le modèle pour une exposition égale à 1. On a alors $\mathbb{E}(Y_i) = t_i\mu_i$ et $\mathbb{V}(Y_i) = t_i\mu_i(1 + t_i\mu_i/k)$. Grouper des données possédant la même combinaison de valeurs des variables explicatives avec l'option (a) revient à supposer que celles-ci partagent la même valeur de H.

L'option (b) conduit quant à elle à $\mathbb{E}(Y_i) = t_i\mu_i$ et $\mathbb{V}(Y_i) = t_i\mu_i(1 + \mu_i/k)$, et correspondrait (!) dans le cadre du modèle Poisson-Gamma, au fait que, conditionnellement à $H_i = h_i$, Y_i suit la loi de Poisson $\mathcal{P}(t_i\mu_i h_i)$, tandis que H_i suit une loi Gamma d'espérance 1 et de paramètre kt_i.

Famille exponentielle : loi gaussienne inverse

On suppose que $Y \sim \mathcal{GI}(m, \lambda)$, c'est-à-dire que Y possède la densité sur $]0, +\infty[$ définie par

$$f_Y(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-m)^2}{2m^2 x}\right).$$

Alors la loi de Y se rattache à la famille exponentielle, avec les caractéristiques suivantes :

Nom	gaussienne inverse
Modèle	$Y \sim \mathcal{GI}(m, \lambda)$
μ	m
ϕ	$1/\lambda$
$v(\mu)$	μ^3
θ	$-1/(2\mu^2)$
$b(\theta)$	$-\sqrt{-2\theta}$

Famille exponentielle : loi(s) de Tweedie

Le terme de « lois de Tweedie » désigne les lois de la famille exponentielle pour lesquelles la fonction de variance est de la forme $v(\mu) = \mu^p$. On retrouve la loi normale ($p = 0$), de Poisson ($p = 1$), Gamma ($p = 2$) et gaussienne inverse ($p = 3$).

Le cas où $1 < p < 2$ correspond à une loi de Poisson composée avec une loi Gamma. Si $Y = \sum_{k=1}^{K} C_k$, où $K \sim \mathcal{P}(\lambda)$, et $(C_k)_{k\geq 1}$ est une suite de v.a. i.i.d. de loi $\mathcal{G}amma(\text{forme} = k, \text{échelle} = \alpha)$, indépendante de K, la loi de Y se rattache à la famille exponentielle, avec les caractéristiques suivantes (ici, Y est un cas mixte avec $\mathbb{P}(Y = 0) > 0$ et loi continue sur $]0, +\infty[$), en posant $p = \frac{k+2}{k+1}$:

Nom	Tweedie $(1 < p < 2)$
Modèle	$Y \sim \mathcal{PC}(\lambda, \mathcal{G}amma(\text{forme} = k, \text{éch.} = \alpha))$
μ	$\lambda k \alpha$
ϕ	$\mu^{2-p}/(\lambda(2-p))$
$v(\mu)$	μ^p
θ	$-\mu^{-(p-1)}/(p-1)$
$b(\theta)$	$(-(p-1)\theta)^{-\frac{2-p}{p-1}}/(2-p)$

Ici, le paramètre k est supposé fixé *a priori*, ainsi que le paramètre ϕ, ce qui entraîne que le rapport $\frac{(k\alpha)^{2-p}}{\lambda^{p-1}} = \frac{\mathbb{E}(C)^{2-p}}{\mathbb{E}(K)^{p-1}}$ est fixé *a priori*.

Equations de vraisemblance pour un modèle log-Gamma

L'écriture des équations de vraisemblance pour un modèle log-Gamma fournit les équations suivantes.

- pour le terme constant $\beta_{\mathbf{0}}$, l'annulation du gradient de la log-vraisemblance par rapport à $\beta_{\mathbf{0}}$ se réécrit :

$$\sum_{i=\mathbf{1}}^{N} w_i \left(\frac{y_i}{\mu_i} \right) = \sum_{i=\mathbf{1}}^{N} w_i .$$

- pour une variable catégorielle $x^{(\ell)}$ à K modalités $\{a_{\mathbf{0}}, \ldots, a_{K-\mathbf{1}}\}$, encodée via les indicatrices des $K-1$ modalités $a_{\mathbf{1}}, \ldots, a_{K-\mathbf{1}}$, l'annulation du gradient de la log-vraisemblance par rapport au coefficient $\beta_{(\ell,k)}$ associé à la modalité a_k se réécrit :

$$\sum_{i=\mathbf{1}}^{N} w_i \left(\frac{y_i}{\mu_i} \right) \mathbf{1}(x_i^{(\ell)} = a_k) = \sum_{i=\mathbf{1}}^{N} w_i \mathbf{1}(x_i^{(\ell)} = a_k).$$

- par différence avec le total (équation pour $\beta_{\mathbf{0}}$), on en déduit la même identité sur les cas où $x^{(\ell)} = a_{\mathbf{0}}$.

Autrement dit, ces équations imposent que la moyenne pondérée des ratios $\frac{\text{réponse observée}}{\text{moyenne prédite}}$ est égale à 1.

Equations de vraisemblance pour un modèle log-NB

L'écriture des équations de vraisemblance pour un modèle log-NB fournit les équations suivantes.

- pour le terme constant $\beta_{\mathbf{0}}$, l'annulation du gradient de la log-vraisemblance par rapport à $\beta_{\mathbf{0}}$ se réécrit :

$$\sum_{i=\mathbf{1}}^{N} w_i y_i = \sum_{i=\mathbf{1}}^{N} w_i \mu_i \left(\frac{k + y_i}{k + \mu_i} \right).$$

- pour une variable catégorielle $x^{(\ell)}$ à K modalités $\{a_{\mathbf{0}}, \ldots, a_{K-\mathbf{1}}\}$, encodée via les indicatrices des $K-1$ modalités $a_{\mathbf{1}}, \ldots, a_{K-\mathbf{1}}$, l'annulation du gradient de la log-vraisemblance par rapport au coefficient $\beta_{(\ell,k)}$ associé à la modalité a_k se réécrit :

$$\sum_{i=\mathbf{1}}^{N} w_i y_i \mathbf{1}(x_i^{(\ell)} = a_k) = \sum_{i=\mathbf{1}}^{N} w_i \mu_i \left(\frac{k + y_i}{k + \mu_i} \right) \mathbf{1}(x_i^{(\ell)} = a_k).$$

- par différence avec le total (équation pour $\beta_{\mathbf{0}}$), on en déduit la même identité sur les cas où $x^{(\ell)} = a_{\mathbf{0}}$.

Ces équations possèdent une interprétation intéressante dans le cas où la loi binomiale négative est vue comme le résultat d'un mélange Poisson-Gamma : la somme (pondérée) des moyennes prédites conditionnelles aux réponses observées doit être égale à la réponse totale observée.

Estimation du paramètre de dispersion utilisée par `mgcv`

Le paquet `mgcv` utilise une modification de l'estimation par la statistique de Pearson basée sur [Fle12], visant à réduire l'instabilité de cette estimation, notamment en présence de valeurs très faibles de $\hat{\mu}_i$.
L'estimation classique via la statistique de Pearson est donnée par :

$$\hat{\phi} = \frac{\mathcal{X}^2}{N - (d+1)}, \ \mathcal{X}^2 = \sum_{i=1}^{N} w_i \cdot \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}.$$

En introduisant

$$\bar{s} = \max \left[-0.9, \frac{1}{n} \sum_{i=1}^{n} v'(\hat{\mu}_i)(y_i - \hat{\mu}_i) \frac{\sqrt{w_i}}{v(\hat{\mu}_i)} \right],$$

la modification proposée est de remplacer $\hat{\phi}$ par :

$$\frac{\hat{\phi}}{1 + \bar{s}}.$$

Courbe de Lorenz et courbe ROC

Pour un modèle de régression logistique avec réponse 0/1 et exposition de 1 pour chaque donnée (x_i, y_i), la courbe de Lorenz relie les points de la forme :

$$\left(i/N, \hat{N}_1(i)/N_1 \right)_{i=1,\ldots,N}$$

tandis que la courbe ROC relie les points de la forme :

$$\left(1 - \frac{i - \hat{N}_1(i)}{N - N_1}, \frac{N_1 - \hat{N}_1(i)}{N_1} \right)_{i=1,\ldots,N},$$

où N_1 désigne le nombre total de réponses 1 dans les données, et $\hat{N}_1(i)$ le nombre de réponses 1 dans les i données présentant les i plus petites valeurs de $\hat{\mu}_i$.
Les deux courbes servent à visualiser les performances prédictives du modèle, et sont clairement reliées, mais pas identiques.

Deux types d'asymptotique « grand nombre de données »

Pour simplifier, on considère le cas d'un modèle GLM comportant uniquement des variables explicatives catégorielles, en supposant que les observations $(y_i)_{1\leq i\leq N}$ constituent une réalisation de variables aléatoires indépendantes $(Y_i)_{1\leq i\leq N}$ dont les lois exactement de la forme postulée par le modèle GLM.
Il faut alors distinguer les deux situations asymptotiques suivantes.

(a) limite d'un « grand nombre de données » : $N \to +\infty$ **et** hypothèses de régularité [33] ; suffisant pour garantir les propriétés de consistance et de normalité asymptotique lors de l'estimation des paramètres du modèle (les β_j et ϕ) ;

(b) limite d'un « grand nombre de données groupées » : $N \to +\infty$ **et** hypothèses de régularité précédentes **et** les données sont groupées de telle sorte que chaque ligne du jeu de données groupées possède une exposition qui tend vers $+\infty$; garantissant les propriétés de normalité des résidus, d'approximation χ^2 pour la déviance, etc.

33. En simplifiant, on demande que l'exposition totale associée à l'occurrence de chaque paire de modalités de chaque paire de variables explicatives, tende vers $+\infty$; une formulation mathématique correcte des conditions est sensiblement plus compliquée, voir par exemple [FT01] et les références qui s'y trouvent).

Mesure des performances prédictives sur données de test

Si les données $(y_i, x_i, w_i)_{i=1,\dots,N}$ sont issues d'une suite i.i.d. de variables aléatoires $(Y_i, X_i, W_i)_{i=1,\dots,N}$, la loi des grands nombres fournit l'approximation, valable dans la limite d'un grand nombre de données de test :

$$\frac{\rho_{\text{Test}}}{|I_{\text{Test}}|} \approx \mathbb{E}\left(\rho(Y', \hat{\mu}^{[\text{Aj.}],\text{obs.}}, W')\right),$$

où (Y', X', W') suit la loi commune des variables (Y_i, X_i, W_i), et est indépendante de celles-ci, et où $\hat{\mu}^{[\text{Aj.}],\text{obs.}}$ est la valeur moyenne prédite par le modèle ajusté sur les données de test à partir de X' et W'.
On estime ainsi directement les performances prédictives du modèle ajusté.

Mesure des performances prédictives par validation croisée

Si les données $(y_i, x_i, w_i)_{i=\mathbf{1},\ldots,N}$ sont issues d'une suite i.i.d. de variables aléatoires $(Y_i, X_i, W_i)_{i=\mathbf{1},\ldots,N}$, ρ_{LOOCV} apparaît comme une somme de N variables aléatoires $\rho(Y_i, \hat{\mu}_i^{[-i]}, W_i)$, de même loi, mais non-indépendantes, ayant chacune pour espérance

$$\mathbb{E}\left(\rho(Y', \hat{\mu}^{[\mathsf{N\text{-}1}]}, W')\right),$$

où (Y', X', W') suit la loi commune des variables (Y_i, X_i, W_i), et est indépendante de celles-ci, et où $\hat{\mu}^{[\mathsf{N\text{-}1}]}$ est la valeur moyenne prédite par un modèle ajusté sur un jeu de données constitué d'une suite de $N-1$ variables aléatoires i.i.d. de même loi que les (Y_i, X_i, W_i). Attention : ici, l'espérance porte également sur le jeu de données utilisé pour ajuster le modèle, et pas seulement sur (Y', X', W').

Si la dépendance entre les variables aléatoires $\rho(Y_i, \hat{\mu}_i^{[-i]}, W_i)$ n'est pas trop forte, on peut s'attendre à avoir l'approximation :

$$\frac{\rho_{\mathsf{LOOCV}}}{N} \approx \mathbb{E}\left(\rho(Y', \hat{\mu}^{[\mathsf{N\text{-}1}]}, W')\right).$$

De plus, on peut s'attendre à ce que l'ajout d'une donnée supplémentaire dans le jeu de données utilisé pour l'ajustement ne modifie pas substantiellement les performances prédictives du modèle, si bien que l'on aurait

$$\mathbb{E}\left(\rho(Y', \hat{\mu}^{[\mathsf{N\text{-}1}]}, W')\right) \approx \mathbb{E}\left(\rho(Y', \hat{\mu}^{[\mathsf{N}]}, W')\right).$$

Si l'on suppose de plus que les performances prédictives d'un modèle ajusté sur un jeu de N données i.i.d. sont peu affectées par l'aléa présent dans les données lorsque N est suffisamment grand, on s'attend à ce que

$$\frac{\rho_{\mathsf{LOOCV}}}{N} \approx \mathbb{E}\left(\rho(Y', \hat{\mu}^{\mathsf{obs.}}, W')\right),$$

où l'espérance porte uniquement sur (Y', X', W'), et où $\hat{\mu}^{\mathsf{obs.}}$ désigne la valeur moyenne prédite par le modèle ajusté sur les données $(y_i, x_i, w_i)_{i=\mathbf{1},\ldots,N}$. Cette discussion s'étend au cas de la validation croisée $K-$fold.

Voir [HTF09] (chapitre 7) pour une discussion plus approfondie de la validation croisée.

Terme(s) d'offset dans un GLM

Un terme d' « offset » (« décalage » ou « compensation ») est un terme que l'on ajoute tel quel dans la composante systématique du modèle. On peut voir un terme d'offset comme une variable explicative quantitative dont le coefficient est fixé à 1.
On rencontre ces termes principalement pour deux usages :

- gérer les différences d'exposition dans les modèles log-Poisson en utilisant le logarithme de l'exposition comme offset (c'est une alternative à l'utilisation de poids dans ce cas) ;
- gérer les effets d'une partie seulement des variables à l'aide du modèle GLM, les effets des autres variables (correspondant aux termes d'offset) ayant été estimés par ailleurs, et demeurant donc fixés lors de l'estimation des paramètres du GLM.

Matrice hessienne de $\log L$ et matrice d'information de Fisher.

En dérivant par rapport à β_k l'expression déjà obtenue pour la dérivée première $\frac{\partial \log L}{\partial \beta_j}$, on obtient l'expression explicite suivante :

$$\frac{\partial^2 \log L}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^{N} \frac{w_i}{\phi} \left(-\frac{1}{v(\mu_i) g'(\mu_i)^2} - \frac{(y_i - \mu_i) v'(\mu_i)}{v(\mu_i)^2 g'(\mu_i)^2} - \frac{(y_i - \mu_i) g''(\mu_i)}{g'(\mu_i)^3 v(\mu_i)} \right) \mathrm{x}_i^{(j)} \mathrm{x}_i^{(k)}.$$

En remplaçant les y_i par les Y_i et en prenant l'espérance, les termes en $Y_i - \mu_i$ ont une espérance nulle, et l'on obtient l'expression pour la matrice d'information de Fisher :

$$I_{j,k} = \sum_{i=1}^{N} \frac{w_i}{\phi} \frac{\mathrm{x}_i^{(j)} \mathrm{x}_i^{(k)}}{v(\mu_i) g'(\mu_i)^2}.$$

Dans les entrailles du Fisher scoring : IRLS

On décrit la méthode IRLS (Iteratively Reweighted Least Squares) pour résoudre les équations de vraisemblance d'un modèle GLM.

On rappelle les équations de vraisemblance vérifiées par $\beta^{\text{est.}}$:

$$\forall\ 0 \leq j \leq d,\ \sum_{i=1}^{N} \frac{w_i (y_i - \mu_i^{\text{est.}}) x_i^{(j)}}{v(\mu_i^{\text{est.}}) g'(\mu_i^{\text{est.}})} = 0.$$

Partant d'une solution approchée $\beta^{[r-1]} \approx \beta^{\text{est.}}$, on écrit que

$$\forall\ 0 \leq j \leq d,\ \sum_{i=1}^{N} \frac{w_i (y_i - \mu_i^{\text{est.}}) x_i^{(j)}}{v(\mu_i^{[r-1]}) g'(\mu_i^{[r-1]})} \approx 0.$$

En utilisant encore le fait que $\beta^{[r-1]} \approx \beta^{\text{est.}}$, on a l'approximation $\mu_i^{\text{est.}} \approx \mu_i^{[r-1]} + \frac{(\eta_i^{\text{est.}} - \eta_i^{[r-1]})}{g'(\mu_i^{[r-1]})}$, donc $\beta^{\text{est.}}$ vérifie également :

$$\forall\ 0 \leq j \leq d,\ \sum_{i=1}^{N} \frac{w_i \left(y_i - \mu_i^{[r-1]} - \frac{(\eta_i^{\text{est.}} - \eta_i^{[r-1]})}{g'(\mu_i^{[r-1]})} \right) x_i^{(j)}}{v(\mu_i^{[r-1]}) g'(\mu_i^{[r-1]})} \approx 0.$$

Dans les entrailles du Fisher scoring : IRLS

On choisit alors de définir $\beta^{[r]}$ comme solution de :

$$\forall\ 0 \leq j \leq d,\ \sum_{i=1}^{N} \frac{w_i \left(y_i - \mu_i^{[r-1]} - \frac{\eta_i^{[r]} - \eta_i^{[r-1]}}{g'(\mu_i^{[r-1]})} \right) \mathrm{x}_i^{(j)}}{v(\mu_i^{[r-1]}) g'(\mu_i^{[r-1]})} = 0.$$

En posant $z_i = g'(\mu_i^{[r-1]})(y_i - \mu_i^{[r-1]}) + \eta_i^{[r-1]}$ et $\omega_i = \frac{w_i}{g'(\mu_i^{[r-1]})^2 v(\mu_i^{[r-1]})}$, et en se rappelant que $\eta_i^{[r]} = \sum_{j=0}^{d} \beta_j^{[r]} \mathrm{x}_i^{(j)}$, nos équations se réécrivent :

$$\forall\ 0 \leq j \leq d,\ \sum_{i=1}^{N} \omega_i \cdot \left(z_i - \left(\sum_{j=0}^{d} \beta_j^{[r]} \mathrm{x}_i^{(j)} \right) \right) \mathrm{x}_i^{(j)} = 0,$$

soit exactement les équations d'estimation d'un modèle **linéaire** dont les réponses sont les z_i, les variables explicatives les $\mathrm{x}_i^{(j)}$, les coefficients les $\beta_j^{[r]}$, et les poids les ω_i. Ainsi, le calcul de $\beta^{[r]}$ à partir de $\beta^{[r-1]}$ peut s'effectuer par estimation d'un modèle linéaire pondéré.

Dans les entrailles du Fisher scoring : IRLS

Il reste à voir que IRLS est exactement équivalent à la méthode de Fisher scoring. On note z le vecteur (colonne) constitué par les z_i, b le vecteur (colonne) constitué par les $g'(\mu_i^{[r-1]})(y_i - \mu_i^{[r-1]})$, $\boldsymbol{\eta}^{[r-1]}$ le vecteur (colonne) constitué par les $\eta_i^{[r-1]}$, Ω la matrice diagonale dont les coefficients sont les ω_i, $\mathfrak{X} = (\mathrm{x}_i^{(j)})_{\substack{\mathbf{1}\leq i\leq N\\ \mathbf{0}\leq j\leq d}}$ vu comme une matrice $N \times (d+1)$. L'expression matricielle des coefficients estimés d'un modèle linéaire pondéré s'écrit alors :

$$\beta^{[r]} = ({}^t\mathfrak{X} \times \Omega \times \mathfrak{X})^{-1} \times {}^t\mathfrak{X} \times \Omega \times \mathrm{z}.$$

On note ensuite que ${}^t\mathfrak{X} \times \Omega \times \mathfrak{X} = \phi \cdot I(\beta^{[r-1]})$. D'autre part, $\mathrm{z} = \mathrm{b} + \boldsymbol{\eta}^{[r-1]}$, et l'on constate, en écrivant $\boldsymbol{\eta}^{[r-1]} = \mathfrak{X} \times \beta^{[r-1]}$, que $({}^t\mathfrak{X} \times \Omega \times \mathfrak{X})^{-1} \times {}^t\mathfrak{X} \times \Omega \times \boldsymbol{\eta}^{[r-1]} = \beta^{[r-1]}$, et l'on vérifie par ailleurs que ${}^t\mathfrak{X} \times \Omega \times \mathrm{b} = \phi \nabla_{\beta^{[r-1]}}(\log L)$. On en déduit finalement que

$$\beta^{[r]} = \beta^{[r-1]} + I(\beta^{[r-1]}) \times \nabla_{\beta^{[r-1]}}(\log L),$$

ce qui est exactement l'itération du Fisher scoring.

Splines cubiques naturelles et optimisation d'écarts pénalisés

Interpolation

Etant donné des nombres réels $t_1 < \ldots < t_m$ et $c_1, \ldots, c_m$, il existe une, et une seule, fonction spline cubique naturelle construite sur les noeuds $t_1, \ldots, t_m$ vérifiant :

$$\forall 1 \leq i \leq m, \ f(t_i) = c_i \tag{3}$$

De plus, parmi les fonctions de classe C^2 vérifiant (3), cette fonction spline cubique naturelle est celle qui minimise la quantité :

$$\int_{-\infty}^{+\infty} \left[f''(t)\right]^2 dt.$$

Cette propriété entraîne la suivante, intéressante dans un contexte de régression.

Optimisation d'écarts pénalisés

Etant donné des nombres réels $t_1 < \ldots < t_m$ et $c_1, \ldots, c_m$, des poids $w_1, \ldots, w_m$, une fonction d'écart d, et un nombre $\lambda > 0$, le minimum (lorsqu'il existe) sur les fonctions C^2 de la quantité :

$$\sum_{i=1}^{m} w_i \cdot d(f(t_i), c_i) + \lambda \int_{-\infty}^{+\infty} \left[f''(t)\right]^2 dt$$

est atteint en une fonction spline cubique naturelle construite sur les noeuds $t_1, \ldots, t_m$.

La base des B-splines

Partant de m noeuds $t_1 < \cdots < t_m$, la base des B-splines constitue une base de l'espace vectoriel des fonctions splines cubiques associée à cette suite de noeuds (fonctions définies sur $[t_1, t_m]$, n'incluant pas la contrainte supplémentaire des splines cubiques « naturelles » qui se traduit par des dérivées secondes nulles en t_1 et t_N).
Pour la définir, on étend la définition des t_i en posant $t_i = t_1$ pour $i < 1$, et $t_i = t_m$ pour $i > m$. On pose ensuite, pour tout i,

$$B_{i,1}(x) = \mathbf{1}(t_i \leq x \leq t_{i+1}),$$

puis, par récurrence, pour $p \geq 2$:

$$B_{i,p}(x) = \alpha_{i,p}(x) B_{i,p-1}(x) + (1 - \alpha_{i+1,p}(x)) B_{i+1,p-1}(x),$$

où

$$\alpha_{j,p}(x) = \begin{cases} \frac{x - t_j}{t_{j+p-\mathbf{1}} - t_j} \text{si } t_{j+p-1} > t_j \\ 0 \text{ sinon} \end{cases}.$$

La famille des fonctions $B_{0,4}, \ldots, B_{m+1,4}$ constitue la base en question.
Cette base est « locale », au sens où chaque fonction est nulle en-dehors d'un intervalle délimité par 5 noeuds consécutifs. On a pour tout x la relation $\sum_{i=0}^{m+1} B_{i,4}(x) = 1$, si bien que, pour l'utilisation dans une composante systématique comportant déjà par ailleurs un terme constant, il convient de supprimer l'une des fonctions (par exemple $B_{0,4}$).

Voir la fonction `bs` du paquet `splines`.

La base des B-splines pour les splines cubiques naturelles

Pour obtenir une base construite à partir de la base des B-splines incluant la contrainte des dérivées secondes nulles en t_1 et t_N (pour obtenir des fonctions qui se prolongent sur $\mathbb{R}$ en spline cubique naturelle), on peut par exemple partir de la matrice de taille $(m+1) \times 2$ définie par

$$A = \begin{pmatrix} B''_{0,4}(t_1) & \dots & B''_{m+1,4}(t_1) \\ B''_{0,4}(t_m) & \dots & B''_{m+1,4}(t_m) \end{pmatrix},$$

effectuer une décomposition QR de la forme

$${}^tA = Q \times R,$$

et prendre comme base les fonctions $(g_1, \dots, g_m)$, où

$$(g_0(x), \dots, g_{m+1}(x)) = (B_{0,4}(x), \dots, B_{m+1,4}(x)) \times Q.$$

(Hormis les coordonnées voisines des bords du vecteur, la base est identique à la base des B-splines.) Voir la fonction `ns` du paquet `splines`.

La base de splines naturelles paramétrée par les valeurs $f(t_i)$

Partant de m noeuds $t_1 < \cdots < t_m$, on peut définir explicitement une base $b_1, \ldots, b_m$ de l'espace vectoriel des fonctions splines cubiques associée à cette suite de noeuds de telle sorte que, pour toute fonction spline cubique naturelle f construite sur les noeuds $t_1, \ldots, t_m$, on ait :

$$f(x) = \sum_{i=1}^{m} f(t_i) \cdot b_i(x).$$

Cette base est celle utilisée par l'option `bs="cr"` du paquet `mgcv`. Voir [Woo17] p. 201–202 pour la définition explicite des fonctions b_i.

Splines de régression de type « plaque mince » en 1d

Partant de m noeuds $t_1 < \cdots < t_m$, on peut écrire toute fonction spline cubique f construite sur les noeuds $t_1, \ldots, t_m$ comme une combinaison linéaire :

$$f(x) = \sum_{i=1}^{m} \delta_i |x - t_i|^3 + \alpha_1 + \alpha_2 x,$$

la contrainte d'avoir une spline cubique « naturelle » se traduisant par les deux équations supplémentaires :

$$\sum_{i=1}^{m} \delta_i = 0 \text{ et } \sum_{i=1}^{m} \delta_i t_i = 0.$$

Plutôt que de limiter *a priori* l'ensemble des noeuds, une manière alternative de réduire la dimension du problème est de partir d'un « grand » jeu de noeuds, et de restreindre les valeurs possibles de $\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_m \end{pmatrix}$ au sous-espace engendré par les vecteurs propres associés à un nombre limité des plus grandes valeurs propres de la matrice $(|x_j - x_i|^3)_{1 \leq i,j \leq m}$.
Cette base est utilisée par l'option `bs="ts"` du paquet `mgcv`.

Les splines de régression de type « plaque mince » en 2d

Etant donnés m couples $(s_i, t_i)_{1 \leq i \leq m}$, et m nombres $c_1, \ldots, c_m$, et un nombre $\lambda > 0$, la fonction C^2 qui minimise la quantité

$$\sum_{i=1}^{m} |c_i - f(s_i, t_i)|^2 + \lambda \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[\left(\frac{\partial^2 f}{\partial s^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial s \partial t} \right)^2 + \left(\frac{\partial^2 f}{\partial t^2} \right)^2 \right] dsdt$$

se met sous la forme d'une combinaison linéaire :

$$f(s,t) = \sum_{i=1}^{m} \delta_i g(||(s,t) - (s_i, t_i)||) + \alpha_1 + \alpha_2 s + \alpha_3 t,$$

où $g(r) = r^2 \log(r)$, et où les δ_i satisfont les contraintes supplémentaires :

$$\sum_{i=1}^{m} \delta_i = 0, \ \sum_{i=1}^{m} \delta_i s_i = 0, \sum_{i=1}^{m} \delta_i t_i = 0.$$

En limitant les valeurs de $\begin{pmatrix} \delta_1 \\ \vdots \\ \delta_m \end{pmatrix}$ au sous-espace engendré par les vecteurs propres associés à un nombre limité des plus grandes valeurs propres de la matrice $(g(||(s_j, t_j) - (s_i, t_i)||))_{1 \leq i,j \leq m}$, on obtient une base de fonctions de deux variables batpisée « thin plate regression splines » dans `mgcv`.

Divergence de Kullback-Leibler

La divergence de Kullback-Leibler $d_{\mathsf{KL}}(\mu||\nu)$ fournit une manière intéressante de quantifier l'écart entre deux lois de probabilité μ et ν.

De manière très générale, on considère deux mesures de probabilité μ et ν sur un espace mesurable $(S, \mathscr{S})$, se mettant sous la forme :

$$\mu(A) = \int_A f(s) d\lambda(s), \ \ \nu(A) = \int_A g(s) d\lambda(s),$$

où λ est une mesure positive sur $(S, \mathscr{S})$, et f et g des fonctions mesurables de $(S, \mathscr{S})$ dans $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$, à valeurs positives. La divergence de Kullback-Leibler de μ par rapport à ν (également appelée entropie relative) est définie par :

$$d_{\mathsf{KL}}(\mu||\nu) = \int_S f(s) \log\left(\frac{f(s)}{g(s)}\right) d\lambda(s),$$

avec les conventions $f(s)/g(s) = +\infty$ si $f(s) > 0$ et $g(s) = 0$, $f(s)/g(s) = 1$ si $f(s) = 0$ et $g(s) = 0$, $\log(+\infty) = +\infty$, $\log(0) = -\infty$, et $0 \times \pm\infty = 0$.
On vérifie que :

- $d_{\mathsf{KL}}(\mu||\nu) \in [0, +\infty]$
- $d_{\mathsf{KL}}(\mu||\nu) = 0 \Leftrightarrow \mu = \nu$

Dans le cas où S est un ensemble fini ou dénombrable, on a :

$$d_{\mathsf{KL}}(\mu||\nu) = \sum_{s \in S} f(s) \log\left(\frac{f(s)}{g(s)}\right).$$

Dans le cas où $S = \mathbb{R}$ et où λ est la mesure de Lebesgue, f et g sont simplement les densités (au sens habituel) de μ et ν, et l'on a :

$$d_{\mathsf{KL}}(\mu||\nu) = \int_{-\infty}^{+\infty} f(s) \log\left(\frac{f(s)}{g(s)}\right) ds.$$