Les Modèles Linéaires Généralisés

Carine TOURE

LMA36-MFA13

Université des Lagunes

Outline

- Déroulé du cours
- Introduction au modèle linéaire généralisé
- Théorie
  - Modélisation
  - Estimation des paramètres
  - Tests d'hypothèses
  - Evaluation et choix des modèles
  - Régression logistique
- Cas pratiques
  - Loi de Bernoulli
  - Loi binomiale
  - Loi de Poisson
  - Loi multinomiale

Déroulé du cours

- 9 h CM (3 séances de cours) / TD / TP
  - pas de rapporteurs pour ce cours
  - pas de contrôle continu – remplacé par une(des) note(s) de TD/TP
  - évaluation finale

- Objectifs du cours
  - dans la première partie du cours : comprendre les bases théoriques du MLG
  - dans la deuxième partie du cours : applications pratiques avec des exemples

Outline

- Déroulé du cours
- Introduction au modèle linéaire généralisé
- Théorie
  - Modélisation
  - Estimation des paramètres
  - Tests d'hypothèses
  - Evaluation et choix des modèles
  - Régression logistique
- Cas pratiques
  - Loi de Bernoulli
  - Loi binomiale
  - Loi de Poisson
  - Loi multinomiale

Introduction au modèle linéaire généralisé

- Le recours au modèle linéaire général est soumis aux hypothèses suivantes :
  - la relation entre l'espérance de la variable de réponse y et les variables explicatives est linéaire
    - on rencontre cependant de nombreux exemples où la relation n'est pas linéaire
  - les observations sont distribuées suivant une loi normale
    - hypothèse essentielle pour la réalisation des tests
  - la variance des variables aléatoires est constante
    - on rencontre cependant des cas où la variance varie en fonction de la moyenne
  - les variables aléatoires sont non-corrélées
    - hypothèse pas toujours vérifiée du fait notamment des conditions d'expérimentation
- Le modèle linéaire généralisé (GLM) est une extension du modèle linéaire permettant de s'affranchir des trois premières hypothèses et de traiter des observations dont la loi de probabilité appartient à une famille de lois élargie.

Introduction au modèle linéaire généralisé

- Soit :
  - $Y = (Y_1, \ldots, Y_n)$ le vecteur des observations,
  - X la matrice du plan d'expérience (aussi appelée matrice de design) regroupant les variables explicatives,
  - x_i est le vecteur ligne des variables explicatives associées à l'observation i,
  - θ est le vecteur des paramètres.

- Le modèle linéaire s'écrit :
  - $Y = X\theta + E$
  - avec Yi $\sim \mathcal{N}(x_i\theta, \sigma^2)$ pour l'observation i, ce qui conduit à $\mathbb{E}[Y_i] = x_i\theta$

- Le modèle linéaire généralisé s'écrit :
  - $g(\mathbb{E}[Y_i]) = x_i\theta$
  - établit une relation non-linéaire entre l'espérance de la variable à expliquer et les variables explicatives
  - permet d'envisager des observations de nature plus variées

Introduction au modèle linéaire généralisé

- Enoncé par Nelder & Wedderburn en 1972 => Etudier la liaison entre une variable de réponse Y_i et un ensemble de variables explicatives X_k

- Le MLG comprend 3 composantes :
  - une composante stochastique (Y) : associée à une loi de probabilité
  - une composante systématique (X) : utilisé commme prédicteur, défini sous forme d'une combinaison linéaire θX
  - une fonction de lien (g) : relation fonctionnelle entre la composante stochastique et celle systématique
    Exemple : régression linéaire simple => le lien est la fonction identité : $g(E(Y)) = E(Y)$

- Propositions :
  - Y est indépendamment distribué (observations indépendantes)
  - Y est associée à la famille dite de distributions exponentielles (Normal, Binomial, Poisson,...)
  - La relation linéaire n'est plus entre la variable de réponse et les variables explicatives mais plutôt entre la variable de réponse transformée par le lien et les variables explicatives
  - Les variables explicatives peuvent être des transformations non-linéaires d'autres variables
  - L'homogénéité de la variance n'est pas requise
  - Les résidus sont indépendants mais non-linéairement distribués
  - Les paramètres sont estimés à partir du maximum de vraisemblance plutôt que par la méthode des moindres carrés ordinaires

Outline

- Déroulé du cours
- Introduction au modèle linéaire généralisé
- Théorie
  - Modélisation
  - Estimation des paramètres
  - Tests d'hypothèses
  - Evaluation et choix des modèles
  - Régression logistique
- Cas pratiques
  - Loi de Bernoulli
  - Loi binomiale
  - Loi de Poisson
  - Loi multinomiale

La famille exponentielle naturelle

- Famille de lois de probabilité contenant des lois usuelles
  - loi normale, la loi de Bernoulli, la loi binomiale, la loi de Poisson, la loi Gamma ...
  - Ces lois ont une écriture sous forme exponentielle permettant d'unifier la présentation des résultats.
- **Définition** :
  - Soit f_Y (resp.P_Y) la densité (resp. loi) de probabilité de la variable Y . f_Y (resp. P_Y) appartient à la famille exponentielle naturelle si elle s'écrit sous la forme :

$$f_Y(y)(resp.\ P_Y(Y=y)) = \exp\left(\frac{1}{\gamma(\phi)}(y\omega - b(\omega)) + c(y,\phi)\right)$$

  - c est une fonction dérivable, b est trois fois dérivable et sa dérivée première b′ est inversible. Le paramètre ω est appelé paramètre naturel de la loi. ϕ est un paramètre appelé paramètre de nuisance ou de dispersion.

- **Propriété** :
  - Si la densité f_Y appartient à la famille exponentielle naturelle, alors

$$\mathbb{E}(Y) = \mu = b'(\omega)$$
$$\mathbb{V}(Y) = \gamma(\phi)b''(\omega)$$

Exercice

- Vérifier que les lois suivantes appartiennent à la famille exponentielle :
  - loi normale $\mathcal{N}(\mu, \sigma)$: $f_Y(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)$
  - loi de Bernoulli $\mathcal{B}(1, p)$: $P[Y = y; p] = p^y(1-p)^{1-y}$
  - loi de poisson Pois(λ): $P[Y = y; \lambda] = \frac{\lambda^y}{y!}e^{-\lambda}$
- Pour chacun des exemples, identifier les paramètres ω, b(ω) et $\gamma(\phi)$

Pour montrer qu'une loi de probabilité appartient à la famille exponentielle naturelle, il suffit de l'écrire sous sa forme exponentielle et d'identier les termes.

Rappel de l'écriture exponentielle :

$$f_Y(y)(resp.\ P_Y(Y=y)) = \exp\left(\frac{1}{\gamma(\phi)}(y\omega - b(\omega)) + c(y,\phi)\right)$$

Exemples classiques

- Pour montrer qu'une loi de probabilité appartient à la famille exponentielle naturelle, il suffit de l'écrire sous la forme d'une exponentielle et d'identier les termes. C'est le cas pour les lois de probabilité classiques suivantes :

— Loi gaussienne

$$f_Y(y;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) = \frac{e^{-y^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{1}{\sigma^2}(y\mu - \frac{\mu^2}{2})\right)$$

ce qui donne $\omega = \mu$, $b(\omega) = \frac{\omega^2}{2}$, $\gamma(\phi) = \phi = \sigma^2$.

— Loi de Poisson

$$P[Y = y;\lambda] = \frac{\lambda^y}{y!}e^{-\lambda} = \frac{1}{y!}\exp\left(y\log\lambda - \lambda\right)$$

ce qui donne $\omega = \log\lambda$, $b(\omega) = \lambda = \exp(\omega)$, $\gamma(\phi) = 1$.

— Loi de Bernoulli

$$P[Y = y;p] = p^y(1-p)^{1-y} = \exp\left(y\log\frac{p}{1-p} + \log(1-p)\right)$$

ce qui donne $\omega = \log\frac{p}{1-p}$, $b(\omega) = -\log(1-p) = \log(1+\exp\omega)$, $\gamma(\phi) = 1$.

Modèle

- La démarche d'écriture d'un modèle linéaire généralisé est constituée de deux étapes :
  - Choix d'une loi de probabilité pour les variables aléatoires Y_i au sein de la famille exponentielle naturelle.
  - Modélisation du lien entre l'espérance des Y_i et les variables explicatives au travers d'une function de lien g inversible :

$$g(\mu_i) = x_i\theta$$

- En notant de manière générale pour un vecteur y, f(y) le vecteur $(f(y_1), \ldots, f(y_n))$ où f est une fonction, on peut écrire le modèle linéaire généralisé sous la forme matricielle suivante :

$$g(\mathbb{E}(Y)) = X\,\theta$$

- Le choix de la loi de la probabilité appartenant à la famille exponentielle est dicté par la nature des données
  - loi de Bernoulli : type binaire
  - loi binomial : nombre de succès
  - loi de poisson : comptage
  - loi exponentielle : données de survie

Choix de la fonction de lien

- Très souvent, on choisit la fonction de lien g comme étant la fonction qui transforme l'espérance μ en le paramètre naturel de la loi (c'est la fonction de lien naturel).

- D'après la propriété de l'espérance μ donnée sur les familles exponentielles, cela revient à choisir :

$$g(\mu) = (b')^{-1}(\mu).$$

- Ainsi,
  - pour les observations de loi normale du modèle linéaire, la fonction de lien naturel est la function identité
  - pour la loi de poisson, la fonction de lien naturel est la fonction log
  - pour la loi de Bernoulli, la fonction de lien naturel est la fonction logit = log (p/(1-p))
    - toute fonction bijective de $]0;1[$ dans $\mathbb{R}$ peut être candidate

- De manière générale le choix de la fonction de lien est une liberté supplémentaire dans la démarche de modélisation
- En pratique, si aucune raison de choisir une fonction de lien spécique ne s'impose, le choix par défaut consiste à choisir la fonction de lien naturel.

Vraisemblance

- L'estimation des paramètres consiste à estimer le vecteur θ et le paramètre de dispersion ϕ
  - ϕ n'est pas le paramètre d'intérêt car n'apparaissant pas dans la composante systématique
  - on utilise la méthode du maximum de vraisemblance

- Vraisemblance :
  - Considérons un échantillon de n variables aléatoires indépendantes $Y_1, \ldots, Y_n$ dont les densités de probabilité f_{Y_i} sont issues de la famille exponentielle et $y = (y_1, \ldots, y_n)$ une réalisation de cet échantillon. Y_i est la réponse au point $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$.
  - Si la fonction de lien utilisée est celle du lien naturel, on a $\omega_i = g(E(Y_i)) = x_i\theta$, la vraisemblance en y s'écrit :

$$\mathcal{L}(y;\theta,\phi) = \prod_{i=1}^{n} f(y_i;\omega_i,\phi) = \prod_{i=1}^{n} f(y_i;x_i\theta,\phi)$$

  - Et la log-vraisemblance

$$\ell(y;\theta,\phi) = \sum_{i=1}^{n}\left[\frac{1}{\gamma(\phi)}(y_i x_i\theta - b(x_i\theta)) + c(y_i,\phi)\right]$$

Vraisemblance

- Les valeurs de θ et de ϕ qui rendent maximale cette fonction de log-vraisemblance sont les solutions du système d'équations aux dérivées partielles suivant :

$$\begin{cases} \dfrac{\partial \ell(y;\theta,\phi)}{\partial \theta_j} = 0 \quad \text{pour } j = 1,\dots,p \\ \dfrac{\partial \ell(y;\theta,\phi)}{\partial \phi} = 0 \end{cases}$$

- Pour simplifier l'écriture, on pose $\gamma(\theta) = 1$ dans cette partie, ce qui ne change pas les résultats obtenus. On a alors :

$$\frac{\partial \ell(y;\theta,\phi)}{\partial \theta_j} = \sum_{i=1}^{n} x_i^j \left[y_i - b'(x_i\theta) \right] \qquad j = 1,\dots,p$$

  - Et donc :

$$\sum_{i=1}^{n} x_i \left[y_i - b'(x_i\theta) \right] = 0.$$

Vraisemblance

- Ce système n'est linéaire que si b'(x) = x, ie f est une densité gaussienne et le modèle est un modèle linéaire.
- Pour tous les autres modèles linéaires généralisés, ce système à p équations est un système non linéaire en θ et il n'y a pas d'expression explicite pour les estimateurs.
- On utilise des algorithmes d'optimisation itératifs tels que l'algorithme de Newton-Raphson, l'algorithme du Fisher-scoring ou encore des algorithmes de descente de gradient.

- Prédiction :
  - A partir d'une estimation $\hat{\theta}$ de θ, on obtient une estimation de $\hat{\omega}_i = x_i\hat{\theta}$. La prédiction par le modèle au point x_i est alors l'estimation de la moyenne :

$$\hat{\mu}_i = g^{-1}(\hat{\omega}_i)$$

  - Dans le cadre du modèle linéaire uniquement, cette prédiction est aussi la prédiction de y_i.

Propriétés de l'estimateur de vraisemblance

- Notons $\hat{\theta}_n$ l'estimateur du maximum de vraisemblance (EMV). Cet estimateur vérifie les propriétés suivantes :
- Théorème :
  - Sous certaines conditions de régularité de la densité de probabilité, l'EMV possède les propriétés suivantes :

— $\hat{\theta}_n$ *converge en probabilité vers* θ *(ce qui implique que* $\hat{\theta}_n$ *est asymptotiquement sans biais)*

— $\hat{\theta}_n$ *converge en loi vers une loi gaussienne*

$$I_n(\theta,\phi)^{1/2}\left(\hat{\theta}_n-\theta\right) \xrightarrow{loi} \mathcal{N}(0,Id)$$

où $I_n(\theta,\phi) = -\mathbb{E}[\frac{\partial^2 \ell(y;\theta,\phi)}{\partial^2\theta}]$ *est la matrice d'information de Fisher évaluée en* θ *et* ϕ *(vraie valeur des paramètres) sur un échantillon de taille* n.

  - Lorsque g est la fonction de lien naturel, l'information de Fisher vaut

$$I_n(\theta,\phi) = \frac{1}{\gamma(\phi)} X'\mathbb{V}(Y)X.$$

Propriétés de l'estimateur de vraisemblance

- La matrice d'information de Fisher dépend des vraies valeurs des paramètres θ et ϕ qui sont inconnues.
- Classiquement, on évalue l'information de Fisher en $\hat{\theta}$ et $\hat{\phi}$.

- Ce résultat permet d'établir des intervalles de confiance de niveau asymptotique $1-\alpha$ pour les paramètres θ_j

- De $I_n(\theta,\phi)^{1/2}\left(\hat{\theta}_n-\theta\right) \xrightarrow{loi} \mathcal{N}(0,Id)$, on déduit en prenant l'information de Fisher au point $(\hat{\theta}, \hat{\phi})$

$$IC_{1-\alpha}(\theta_j) = \left[\hat{\theta}_j - u_{1-\alpha/2} I(\hat{\theta},\hat{\phi})^{-1/2}_{jj} \; ; \; \hat{\theta}_j + u_{1-\alpha/2} I(\hat{\theta},\hat{\phi})^{-1/2}_{jj}\right]$$

  - où $u_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$.

Tests de modèles emboités

- Cette section présente différents tests d'hypothèses qui vont permettre d'examiner les qualités du modèle, de déterminer si les différentes variables explicatives présentes dans le modèle sont pertinentes ou non.
- Les tests de modèles emboîtés sont les plus généraux et permettent de répondre à la majorité des questions qui se posent.
- Test de modèles emboités :
  - M_1 et M_0 respectivement définis par $g(\mu) = X_1\theta_1$ et $g(\mu) = X_0\theta_0$ sont dits emboités si le modèle M_0 est un cas particulier du modèle plus général M_1, c'est a dire si le sous-espace engendré par les colonnes de X_0 est inclus dans le sous-espace linéaire engendré par les colonnes de X_1.
  - Le test des hypothèses $H_0 = \{M_0\}$ contre $H_1 = \{M_1\}$, est alors réalisé à l'aide du test du rapport de vraisemblance dont la statistique de test s'écrit :

$$T = -2 \log \frac{\mathcal{L}(y; \hat{\theta}_0)}{\mathcal{L}(y; \hat{\theta}_1)} = -2(\ell(y; \hat{\theta}_0) - \ell(y; \hat{\theta}_1)),$$

    - où $\hat{\theta}_0$ et $\hat{\theta}_1$ sont respectivement les estimateurs du maximum de vraisemblance de θ dans les modèles M_0 et M_1.

Tests de modèles emboités

- Test de modèles emboités :

$$T = -2 \log \frac{\mathcal{L}(y; \hat{\theta}_0)}{\mathcal{L}(y; \hat{\theta}_1)} = -2(\ell(y; \hat{\theta}_0) - \ell(y; \hat{\theta}_1)),$$

- On montre que sous certaines conditions, cette statistique de test converge en loi vers une loi du χ^2 à $p_1 - p_0$ degrés de liberté, où p_0 et p_1 sont respectivement les dimensions des espaces engendrés par les colonnes de X_0 et X_1
- Si on effectue le test au niveau α, on rejettera H_0 au profit de H_1 si $T \geq \chi^2_{1-\alpha,p1-p0}$ où $\chi^2_{1-\alpha,p1-p0}$ est le quantile d'ordre $1 - \alpha$ de la loi du χ^2 à $p_1 - p_0$ degrés de liberté.

- Ce test est parfois présenté sous une forme différente reposant sur la déviance (ou déviance résiduelle), qui est l'écart entre le logarithme de la vraisemblance du modèle d'intérêt M et celui du modèle le plus complet possible, appelé modèle saturé, et note M_S . Le modèle saturé est le modèle comportant n paramètres, ie autant que d'observations. La déviance du modèle M est alors définie par :

$$D(M) = -2 \left(\ell(y; \hat{\theta}) - \ell(y; \hat{\theta}_S) \right).$$

- La statistique de test T présentée précédemment peut être réécrite en terme de déviance sous la forme $T = D(M_0) - D(M_1)$
- Le test global du modèle consiste à tester $H_0 = \{g(\mu_i) = a\}$ contre $H_1 = \{g(\mu_i) = x_i\theta\}$

à l'aide du test du rapport de vraisemblance. Il permet de tester si toutes les variables sont inutiles pour expliquer la variable réponse Y .

Tests de $\theta_j = \theta_{0j}$

- Tests de $\theta_j = \theta_{0j}$:
  - Si la réponse au test global est positive, la suite logique consiste à tester quelles sont les variables ou facteurs qui ont une influence. La connaissance de la loi asymptotique de $\hat{\theta}$ nous permet de construire des tests sur les parametres θ, sur des combinaisons linéaires de θ ou encore de μ_i ainsi que des intervalles de confiance. Tous ces résultats sont asymptotiques.
  - On souhaite tester l'hypothèse

$$\mathrm{H}_0 = \{\theta_j = \theta_{0j}\} \quad \text{contre} \quad \mathrm{H}_1 = \{\theta_j \neq \theta_{0j}\}$$

  - où θ_{0j} est une valeur définie a priori. D'après le théoreme de l'EMV sous H_0,

$T_j = I(\theta, \phi)_{jj} \, (\hat{\theta}_j - \theta_{0j})^2$ converge en loi vers un $\chi^2(1)$ et $P(T_j > t_j)$ donne une p-valeur asymptotique du test. En général on utilise ce test avec $\theta_{0j} = 0$ afin de déterminer si le paramètre θ_j est significativement non nul. Ce test est appele test de Wald.

  - En pratique, l'information de Fisher est calculée non pas en les vrais paramètres qui sont inconnus mais en $\hat{\theta}$ et $\hat{\phi}$. La statistique de test est donc :

$$T_j = I_n(\hat{\theta}, \hat{\bar{\phi}})_{jj} \, (\hat{\theta}_j - \theta_{0j})^2.$$

Tests de Cθ = 0

- **Tests de Cθ = 0:**
  - On peut, comme avec le modèle linéaire, être amené à effectuer un test sur une combinaison linéaire des paramètres.
  - Les hypothèses à tester sont

$$\mathrm{H}_0 = \{C\theta = 0\} \quad \text{contre} \quad \mathrm{H}_1 = \{C\theta \neq 0\}$$

  - où C est un vecteur ligne de dimension p (dimension de θ). La construction d'un tel test nécessite de déterminer la loi de Cθ̂. Sachant que θ̂ suit asymptotiquement une loi normale, on l'obtient, en utilisant la methode Delta :

$$[C\ I_n(\theta,\phi)^{-1}C']^{-1/2}(C\hat{\theta} - C\theta) \xrightarrow{loi} \mathcal{N}(0, I_p)$$

  - Ici, l'information de Fisher est évaluée en θ̂ et $\hat{\phi}$
  - Le test de région de rejet $\{|\mathrm{T}| > \mathrm{u}_{1-\alpha/2}\}$ avec $T = [C\ I_n(\hat{\theta},\hat{\phi})^{-1}C']^{-1/2}C\hat{\theta},$

est un test de niveau asymptotique α pour les hypotheses H_0 et H_1 données plus haut

  - De même, l'intervalle $IC_{1-\alpha}(C\theta) = \left[C\hat{\theta} - u_{1-\alpha/2}/\sqrt{C\ I_n(\hat{\theta},\hat{\phi})^{-1}C'}\ ;\ C\hat{\theta} + u_{1-\alpha/2}/\sqrt{C\ I_n(\hat{\theta},\hat{\phi})^{-1}C'}\right]$
  - où $\mathrm{u}_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$ est un intervalle de confiance de Cθ de niveau asymptotique $1-\alpha$.

Exercice

- Soit la variable aléatoire X~ $\mathcal{N}$(m, σ) et (X_1, X_2, ..., X_n), un échantillon i.i.d de même loi que X.
  - Calculer la fonction de maximum de vraisemblance de cet échantillon.
  - Estimer les paramètres m et σ de cette loi.
- Rappel
  - la vraisemblance en y s'écrit $$\mathcal{L}(y;\theta,\phi) = \prod_{i=1}^{n} f(y_i;\omega_i,\phi) = \prod_{i=1}^{n} f(y_i;x_i\theta,\phi)$$
  - les paramètres optimaux sont ceux qui s'annulent aux dérivées partielles de la log-vraisemblance

$$\begin{cases} \dfrac{\partial \ell(y;\theta,\phi)}{\partial \theta_j} = 0 \quad \text{pour } j = 1,\ldots,p \\ \dfrac{\partial \ell(y;\theta,\phi)}{\partial \phi} = 0 \end{cases}$$

Solution

- Rappel
  - $$\mathcal{L}(y;\theta,\phi)=\prod_{i=1}^{n} f(y_i;\omega_i,\phi)=\prod_{i=1}^{n} f(y_i;x_i\theta,\phi)$$

$$\begin{cases} \dfrac{\partial \ell(y;\theta,\phi)}{\partial \theta_j}=0 \quad \text{pour } j=1,\ldots,p \\ \dfrac{\partial \ell(y;\theta,\phi)}{\partial \phi}=0 \end{cases}$$

- Solution
  - la fonction de MV est donnée par

$$\begin{aligned} L(x_1,x_2,...,x_n;m,\sigma) &= P(X_1=x_1,X_2=x_2,...,X_n=x_n) \\ &= f(x_1)\,f(x_2)\,...f(x_n) \text{ car les } X_1,X_2,...,X_n \text{ sont indépendantes.} \\ &= \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x_i-m)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\left(\prod_{i=1}^{n}\exp\left(-\frac{(x_i-m)^2}{2\sigma^2}\right)\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2\right). \end{aligned}$$

  - en prenant le logarithme $\mathcal{L}(m,\sigma)=\ln L(x_1,x_2,...,x_n;m,\sigma)$ du produit, on obtient

$$\mathcal{L}(m,\sigma)=-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m)^2-n\ln\sigma\sqrt{2\pi}.$$

Solution

- Rappel
  - $$\mathcal{L}(y;\theta,\phi)=\prod_{i=1}^{n} f(y_i;\omega_i,\phi)=\prod_{i=1}^{n} f(y_i;x_i\theta,\phi) \qquad \begin{cases} \dfrac{\partial \ell(y;\theta,\phi)}{\partial \theta_j}=0 \quad \text{pour } j=1,\ldots,p \\ \dfrac{\partial \ell(y;\theta,\phi)}{\partial \phi}=0 \end{cases}$$
- Solution

Les dérivées partielles par rapport à m et à σ sont respectivement

$$\frac{\partial \mathcal{L}(m,\sigma)}{\partial m}=\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-m),$$

et

$$\frac{\partial \mathcal{L}(m,\sigma)}{\partial m}=\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-m)^2-\frac{n}{\sigma},$$

Ces dérivées s'annulent lorsque

$$m=\frac{1}{n}\sum_{i=1}^{n}x_i,$$

et

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-m)^2,$$

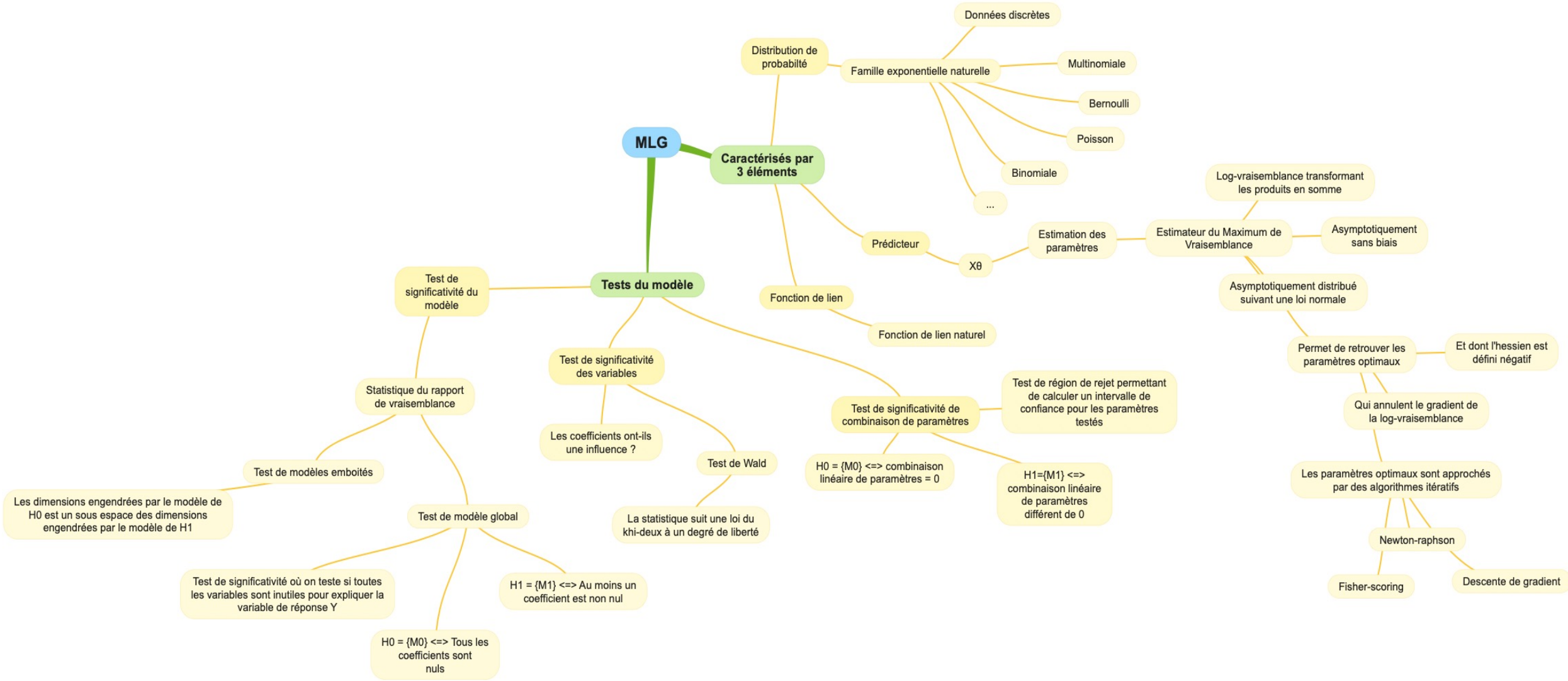
donc l'estimateur de MV du paramètre p est donné par

$$\widehat{m}=\frac{1}{n}\sum_{i=1}^{n}X_i,$$

et

$$\widehat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2.$$

Récapitulons



Outline

- Déroulé du cours
- Introduction au modèle linéaire généralisé
- Théorie
  - Modélisation
  - Estimation des paramètres
  - Tests d'hypothèses
  - Evaluation et choix des modèles
  - Régression logistique
- Cas pratiques
  - Loi de Bernoulli
  - Loi binomiale
  - Loi de Poisson
  - Loi multinomiale

Pseudo R^2

- Qualité d'ajustement du modèle (qualité de la prédiction)
  - Modèle linéaire => coefficient de determination R^2.
  - Modèle linéaire généralisé => mesures basées sur le même principe

- Pseudo R^2

$$pseudo\, R^2 = \frac{D(M_0) - D(M)}{D(M_0)}$$

  - On rappelle que $D(M) = -2\left(\ell(y;\hat{\theta}) - \ell(y;\hat{\theta}_S)\right)$.
  - M_0 est le modèle nul (le modèle à un seul paramètre)
  - Pseudo R^2 varie entre 0 et 1, plus il est proche de 1, meilleur est l'ajustement du modèle.

X^2 de Pearson généralisé

- Test de qualité d'ajustement de notre modèle
  - vérifie si les données sont susceptibles de provenir d'une distribution théorique spécifique

- Le $\mathcal{X}^2$ de Pearson généralisé est la statistique définie par :

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

  - où $\hat{\mu}_i = g^{-1}(x_i\widehat{\theta})$.

  - L'hypothèse nulle est que le modèle étudié est le bon modèle
  - Celle-ci est rejetée au niveau α si on obtient un X^2 supérieur au quantile $\chi^2_{n-p,1-\alpha}$

  - n est le nombre de données
  - p est le rang de la matrice de design X

Choix de modèle

- Lorsqu'on doit choisir entre plusieurs modèles candidats non-emboités :
  - la déviance est utilisée comme critère de sélection
  - un modèle sera qualifié de bon si sa déviance est proche du modèle saturé (pseudo R^2 proche de 1) et s'il est construit avec un faible nombre de paramètres
  - contraintes quelque peu antagonistes

- Des critères pénalisés permettent de prendre en compte ces deux contraintes antagonistes :
  - Le critère AIC (Akaike Information Criterion) dont une définition est :

$$AIC(M(\hat{\theta})) = -2\ell(y;\hat{\theta}) + 2p$$

    - où p est le rang de la matrice de design X. L'AIC est d'autant plus faible que la log-vraisemblance est élevée et que le nombre de paramètres est petit et permet donc d'établir un ordre sur les modèles en prenant en compte les deux contraintes

  - Le critère BIC (Bayesian Information Criterion) qui pénalise plus le sur-ajustement est déni par

$$BIC(M(\hat{\theta})) = -2\ell(y;\hat{\theta}) + np$$

Les résidus dans le MLG

- Les résidus se définissent comme l'écart entre l'observation y_i et sa prédiction par le modèle $\widehat{\mu}_i$

- Ainsi, les résidus bruts se définissent : $\varepsilon_i = y_i - \hat{\mu}_i$
- En les normalisant, on obtient les résidus de Pearson : $r_{p_i} = \dfrac{y_i - \hat{\mu}_i}{\sqrt{\mathbb{V}_{\hat{\theta}}(y_i)}}$,
  - où $\mathbb{V}_{\hat{\theta}}(y_i)$ est la variance théorique de y_i calculée en $\hat{\theta}$
  - si y_i suit une loi de Bernoulli alors, $\mathbb{V}_{\hat{\theta}}(y_i) = \hat{p}_i(1 - \hat{p}_i)$.

- En normalisant les résidus bruts par l'effet levier, on obtient les résidus de Pearson normalisés :

$$r_{s_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{(1 - h_{ii})\mathbb{V}_{\hat{\theta}}(y_i)}},$$

  - où h_{ij} désigne le levier, c'est à dire le terme diagonal de la matrice $H = X\,(X'X)^{-1}X'$ dans le cas où la matrice de design X est de rang plein.

Les résidus dans le MLG (suite)

- Autre approche :
  - Les résidus de déviance :

$$r_{d_i} = signe(y_i - \hat{\mu}_i)\sqrt{2\ell(y_i;\hat{\theta}_S,\hat{\phi}) - 2\ell(y_i;\hat{\theta},\hat{\phi})}.$$

  - Les résidus de deviance standardisés :

$$r_{ds_i} = signe(y_i - \hat{\mu}_i)\sqrt{\frac{2\ell(y_i;\hat{\theta}_S,\hat{\phi}) - 2\ell(y_i;\hat{\theta},\hat{\phi})}{1 - h_{ii}}}.$$

    - on standardise les résidus de déviance pour pouvoir les comparer entre eux

- Intuitivement une observation ayant un résidu de déviance élevé est une observation ayant une grande influence sur l'estimation des paramètres du modèle et doit donc être examinée avec soin.
- Vérifier qu'il n'existe pas de structure inattendue dans les résidus et le cas échéant, reprendre le modèle proposé pour identifier la cause de cette structure.

Outline

- Déroulé du cours
- Introduction au modèle linéaire généralisé
- Théorie
  - Modélisation
  - Estimation des paramètres
  - Tests d'hypothèses
  - Evaluation et choix des modèles
  - Régression logistique
- Cas pratiques
  - Loi de Bernoulli
  - Loi binomiale
  - Loi de Poisson
  - Loi multinomiale

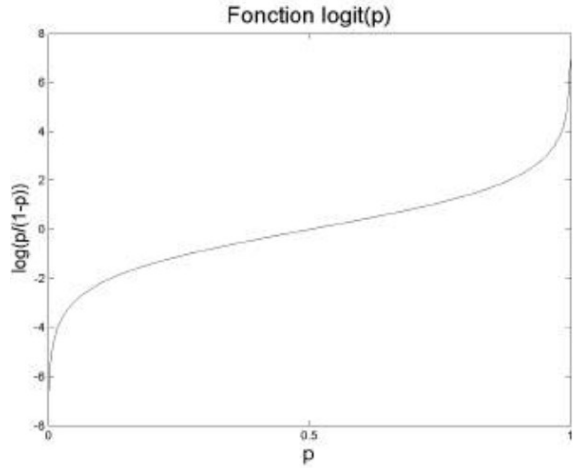
Cas de réponse binaire

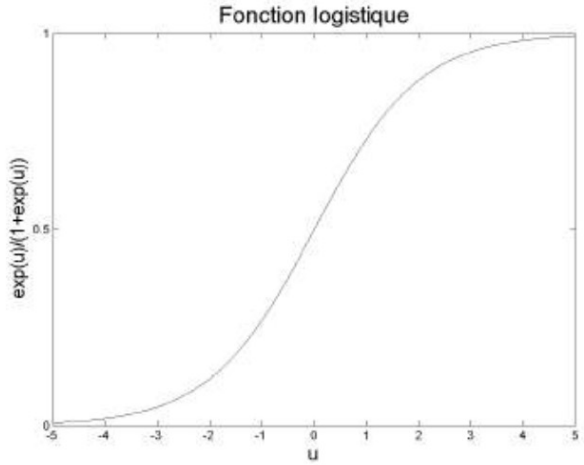
- Considérons le cas où Y est une variable binaire prenant 2 valeurs selon les variables explicatives :

$$Y_i \sim \mathcal{B}ernoulli(p_i)$$

$$g(\mathbb{E}(Y_i)) = g(p_i) = x_i\theta.$$

- Pour chaque individu i, Y suit une loi de Bernoulli de paramètre p_i qui est aussi son espérance.
- Lorsqu'on choisit comme fonction de lien la fonction logit, on appelle régression logistique le MLG associé.
- La fonction logit est une bijection de l'intervalle]0; 1[dans $\mathbb{R}$, et son graphe est symétrique par rapport au point (0.5, 0)





Cas de réponse binaire (suite)

- La probabilité P(X=1|x) = p(x) pour un individu pour lequel les variables explicatives forment un vecteur ligne x
- p(x) = h(θx) où h est la fonction logistique, la fonction inverse de la fonction logit : $h(u) = \frac{1}{1+e^{-u}}$
- Le modèle de régression logistique peut donc s'écrire : $p(x) = \frac{1}{1+e^{-x\theta}}$.
- Lorsque toutes les variables explicatives sont qualitatives,
  - les individus présentant les mêmes combinaisons de modalités sont regroupés pour former une seule observation, et la variable de réponse Y_i associée à ce groupe (nombre de succès pour cette combinaison de modalités) suit une loi binomiale de paramètres (n_i, p_i)
- Lorsque les variables explicatives sont continues, les valeurs de x_i sont différentes d'un individu à l'autre et aucun regroupement n'est possible.
- 2 types de modèles de régression logistique :
  - 1er cas : facteurs qualitatifs comme en analyse de variance
  - 2ème cas : modèle de type regression
- Si les modalités des variables qualitatives sont ordonnées et une relation linéaire est plausible entre les x et la fonction logit => on peut utiliser une régression linéaire
- Sinon, utiliser un modèle de type analyse de variance.

Cas de réponse binaire (1)

- Test d'ajustement de Hosmer-Lemeshow
  - lorsque les variables explicatives sont qualitatives, on utilise le test du khi-deux de Pearson ou le test du rapport de vraisemblance pour tester l'ajustement
  - lorsque les variables explicatives sont continues, on utilise le test de Homer-Lemeshow
    - ordonner les valeurs de $\widehat{p}_i$ par ordre croissant
    - on considère les classes d'individus définies par les déciles de la distribution des $\widehat{p}_i$
    - Pour chaque classe k on calcule N_k le nombre d'observations, A_k le nombre de cas $Y = 1$ et P_k la moyenne des $\widehat{p}_i$ dans la classe k
    - Le test est basé sur la statistique : $S = \sum_{k=1}^{10} \frac{(A_k - N_k P_k)^2}{N_k P_k (1 - P_k)}$
    - Sous l'hypothèse H_0 que le modèle est le vrai modèle, $S \sim \chi^2_8$.
    - Le choix de 10 classes est un compromis entre puissance et validité des conditions asymptotiques.
    - On peut choisir moins de classes si on a peu de données ou plus de classes si on a beaucoup de données.
    - Le nombre de degrés de liberté du χ^2 est égal au nombre de groupes moins 2.

Cas de réponse binaire (2)

- Cote, rapport de cote, risque relatif
  - La quantité $p(x)/1-p(x)$ est appelée "odds" en anglais et "cote" en français
  - Lorsque la cote est proche de 1, les deux évènements $(Y = 0)$ et $(Y = 1)$ ont des chances équivalentes de se produire
  - lorsqu'elle est proche de 0, c'est l'évènement $(Y = 0)$ qui est le plus probable,
  - lorsque l'odds est grand, l'évènement $(Y = 1)$ est plus probable
  - Dans le cadre du modèle logistique, on a $\text{odds}(x) = e^{x\theta}$
  - L'odd-ratio ou rapport de cotes est le rapport de deux odds associés à deux valeurs différentes des variables
$$OR(x,t) = \frac{p(x)(1-p(t))}{p(t)(1-p(x))}$$
  - Dans le cas où la variable x_j est quantitative, on obtient en posant $t_j = x_j + 1$ et en laissant inchangées les autres variables,
$$OR(x,t) = e^{\theta_j}$$
  - Le Risque Relatif ou Rapport de Risques entre 2 conditions x et t est : $RR(x,t) = \frac{p(x)}{p(t)}$
    - utilisé souvent en épidémiologie pour comparer les risques de maladies entre 2 conditions

Cas de réponse binaire (3)

- Cote, rapport de cote, risque relative (suite)
  - OR et RR ne sont pas identiques.
  - La relation entre odd-ratio et risque relatif est la suivante : $OR(x,t) = RR(x,t) \times \frac{1-p(t)}{1-p(x)}$
- Classement, courbe ROC
  - On pose Y = 0 si $\widehat{p}_i <$ s et Y = 1 sinon
  - Ce qui fait de la regression logistique un classifieur
  - On rappelle que p(x)=P(Y =1|x).
  - Si on prédit Y = 1 alors que Y = 0, on dit que l'on a un "faux positif".
  - Si on prédit Y = 0 alors que Y = 1, on dit que l'on a un "faux négatif".
  - On juge de la qualité des prédictions par la table de confusion

	$\#(\widehat{Y}_i = 0)$	$\#(\widehat{Y}_i = 1)$
$\#(Y_i = 0)$	a	b
$\#(Y_i = 1)$	c	d

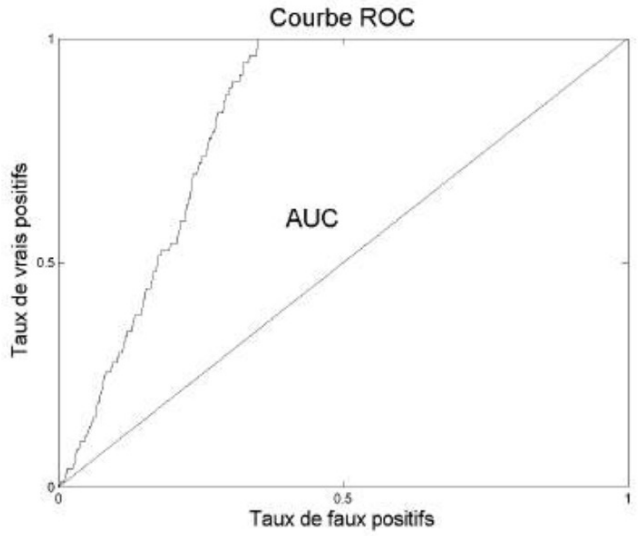
- a est le nombre de "vrais négatifs", d est le nombre de "vrais positifs", c'est le nombre de "faux négatifs" et b est le nombre de "faux positifs". On a de bonnes prédictions si les valeurs b et c sont petites devant a et d.

Cas de réponse binaire (4)

- Classement, courbe ROC
  - Il existe un grand nombre d'indicateurs synthétiques mesurant la qualité des prédictions à partir de la table de confusion : (b+c)/a+d, (b+c)/a+b+c+d ...

  - Courbe ROC
    - Supposons que les prédictions $\hat{Y}_i$ soient obtenues en utilisant un seuil s variant de 0 à 1
    - Pour chaque valeur de s, on a une table de confusion qui donne un taux de faux positifs $\tau_-(s) = \frac{b(s)}{a(s)+b(s)}$ et un taux de vrais négatifs $\tau_+(s) = \frac{d(s)}{c(s)+d(s)}$

    - La courbe ROC (receiver operating characteristic) est la courbe qui relie tous les points ($\tau_-(s)$, $\tau_+(s)$)



- Dans le meilleur des cas il existe un seuil, s_0 qui sépare parfaitement les positifs et les négatifs et $\tau_-(s_0) = 0$ et $\tau_+(s_0) = 1$. Dans ce cas la courbe ROC passe par le point (0,1)
- Dans le pire des cas les positifs et les négatifs sont complètement mélangés quel que soit le seuil, et les deux taux sont égaux. Dans ce cas la courbe ROC est proche de la bissectrice du carré.
- La quantité AUC (Aera Under the Curve) qui est égale à l'aire comprise sous la courbe ROC est un bon indicateur de la qualité des prédictions : si elle est proche de 1, les prédictions sont bonnes, si elle est proche de 0.5 les prédictions sont mauvaises.

Remarques

- Comme pour toute évaluation de la qualité d'une prédiction, il faut séparer le processus d'estimation des paramètres du modèle du processus d'estimation de la qualité de la prédiction.
- Pour ce faire on peut séparer les données en un échantillon d'apprentissage et un échantillon test, puis utiliser la validation croisée ou le re-échantillonnage.

Exercice

- Montrer la propriété suivante : $OR(x, \tilde{x}) > 1 \iff P(Y = 1|X = x) > P(Y = 1|X = \tilde{x})$.

- Quel est le meilleur modèle parmi les trois modèles suivants :
  - Modèle 1 : AIC = 51.09.
  - Modèle 2 : AIC = 42.87.
  - Modèle 3 : AIC = 43.10.

  - Rappel : L'odds ratio OR (rapport des chances) entre deux individus x et $\tilde{x}$ est

$$odds(x) = \frac{P(Y = 1|X = x)}{1 - P(Y = 1|X = x)}. \qquad OR(x, \tilde{x}) = \frac{odds(x)}{odds(\tilde{x})}.$$

$$AIC(M(\hat{\theta})) = -2\ell(y; \hat{\theta}) + 2p$$

  - où p est le rang de la matrice de design X. L'AIC est d'autant plus faible que la log-vraisemblance est élevée et que le nombre de paramètres est petit

Solution

- Montrer la propriété suivante : $OR(x,\widetilde{x}) > 1 \Longleftrightarrow P(Y=1|X=x) > P(Y=1|X=\widetilde{x})$.

  • Si $P(Y=1|X=x) > P(Y=1|X=\widetilde{x})$, alors

$$
\begin{aligned}
P(Y=1|X=x) \quad &> \quad P(Y=1|X=\widetilde{x}) \Longleftrightarrow 1-P(Y=1|X=x) < 1-P(Y=1|X=\widetilde{x}) \\
&\Longleftrightarrow \quad \frac{1}{1-P(Y=1|X=x)} > \frac{1}{1-P(Y=1|X=\widetilde{x})} \\
&\Longleftrightarrow \quad \frac{P(Y=1|X=x)}{1-P(Y=1|X=x)} > \frac{P(Y=1|X=\widetilde{x})}{1-P(Y=1|X=\widetilde{x})} \\
&\Longleftrightarrow \quad odds(x) > odds(\widetilde{x}) \\
&\Longleftrightarrow \quad OR(x,\widetilde{x}) = \frac{odds(x)}{odds(\widetilde{x})} > 1.
\end{aligned}
$$

- Quel est le meilleur modèle parmi les trois modèles suivants :

  · Modèle 1 : AIC $= 51.09$.

  · Modèle 2 : AIC $= 42.87 \longrightarrow Meilleur$.

  · Modèle 3 : AIC $= 43.10$.

le modèle choisi est celui qui aura la plus faible valeur d'AIC : le modèle choisi est "Modèle 2 : AIC $= 42.87$".

Régression multilogistique

- Considérons maintenant une variable aléatoire Y prenant K > 2 modalités (catégories).
- Notons (Y = k) la variable qui vaut 1 si l'évènement (Y = k) est vrai et 0 sinon.
- La loi conjointe des K variables ((Y =1),...,(Y =K)) est une loi multinomiale de paramètre p=(p_1,...,p_K) où p_k = P[Y = k]

$$((Y = 1), \ldots, (Y = K)) \sim \mathcal{M}(1, p_1, p_2, \ldots, p_K),$$

  - tel que $\sum_{k=1}^{K} p_k = 1$. Par abus de notation, on notera Y, le vecteur ((Y = 1),...,(Y = K)).
  - La loi de probabilité s'écrit alors :

$$P[Y = y; p] = \prod_k {p_k}^{y_k} = \exp\left(\sum_k y_k \log p_k\right)$$

Régression multilogistique (1)

- L'approche usuelle consiste à choisir une modalité de référence, puis à modéliser les log-odds pour les autres. On choisie une modalité de référence, par exemple la première.
- On a :

$$\begin{aligned} P[Y = y; p] &= \exp\left(y_1 \log p_1 + y_2 \log p_2 + \dots y_K \log p_K\right) \\ &= \exp\left(\log p_1 + y_2 \log p_2/p_1 + \dots y_K \log p_K/p_1\right), \end{aligned}$$

- et on suppose que les "odds" sont linéaires en les variables explicatives, en notant $\theta^{(k/1)}$ la reparamétrisation on a :

$$\begin{aligned} \log p_2/p_1 &= x\theta^{(2/1)} \\ \log p_3/p_1 &= x\theta^{(3/1)} \\ \vdots &= \vdots \\ \log p_K/p_1 &= x\theta^{(K/1)} \end{aligned}$$

- p_1 est alors déduit de la contrainte $\sum_{k=1}^{K} p_k = 1$ $(p_1 = 1 - \sum_{k>1} p_k)$

Régression multilogistique (2)

- Ce modèle est appelé le modèle logistique multinomial. Il faut noter que le choix de la référence est important, puisqu'elle conditionne l'interprétation des résultats.
- On obtient l'expression des différentes probabilités p_k en fonction de x :

$$p_1 = \frac{1}{1+\sum_{l>1} e^{x\theta^{(l/1)}}}$$

et

$$p_k = \frac{1}{1+e^{-x\theta^{(k/1)}}+\sum_{l>1,l\neq k} e^{x(\theta^{(l/1)}-\theta^{(k/1)})}}, k = 2, K.$$

- Surdispersion
  - Il y a surdispersion lorsqu'on observe une plus grande variabilité que celle attendue par la loi utilisée. Cela peut provenir de différentes causes.
  - Exple : données de comptage
  - 2 possibilités pour modéliser cette surdispersion

Régression multilogistique (3)

- 2 possibilités pour modéliser la surdispersion :
  - on utilise un modèle standard (Poisson, Binomial) avec un paramètre de surdispersion, note φ. Les lois de Poisson et Binomiale avec surdispersion n'existent pas en tant que lois de probabilité, mais on montre qu'on peut utiliser une méthode dite de quasi-vraisemblance pour estimer les parametres θ et φ. De plus l'algorithme de Newton-Raphson suffit pour maximiser la quasi-vraisemblance. C'est une solution simple et économique.
  - on utilise un modèle plus complexe qui prend en compte le phénomène de sur-dispersion. Par exemple on remplace la loi Binomiale par une loi BetaBinomiale et on remplace la loi de Poisson par la loi Binomiale Négative. Cette solution d'une bonne modélisation se heurte au fait que ces nouvelles lois ne font souvent pas partie de la famille exponentielle et qu'on perd alors une partie des bonnes propriétés statistiques du modele linéaire généralisé standard.

- Détection de surdispersion
  - Pour détecter la surdispersion il suffit de calculer le rapport entre le χ^2 d'ajustement du modèle aux données, et son nombre de degrés de liberté.
  - S'il est nettement supérieur à 1, il y a surdispersion.

Exercice : Maximum de vraisemblance

- **Exercice 1.** Soit la variable aléatoire $X \rightsquigarrow \mathcal{B}(n,p)$ et $(X_1, X_2, ..., X_n)$ un échatillon $i.i.d$ de même loi que X.

a) Calculer la fonction de **MV** de cette échatillon.

b) Estimer le paramètre p de cette loi.

∘ *Le principe de la méthode MV :* Si un phénomène X a été l'objet de n observations indépendantes $x_1, x_2, ..., x_n$ les unes des autres, sa loi de probabilité $P(X = x)$(dans le cas discret : loi binomiale, loi de Poisson) ou sa densité (en cas de loi continue, comme la loi normale) est une fonction $f(x; \theta_1, ..., \theta_n)$ où $\theta_1, ..., \theta_n$ sont les paramètres de la loi.

· *Dans le cas discret :* on définit la fonction de **MV** est la probabilité de l'échatillon observé en fonction des paramètres $\theta = (\theta_1, ..., \theta_n)$:

$$\begin{aligned} L(x_1, x_2, ..., x_n; \theta) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \theta) \\ &= P_X\ (x_1, x_2, ..., x_n; \theta), \end{aligned}$$

avec $X = (X_1, X_2, ..., X_n)$.

· *Dans le cas continu :* on définit la fonction de **MV** est définie par

$$L(x_1, x_2, ..., x_n; \theta) = P_X\ (x_1, x_2, ..., x_n; \theta).$$

· On maximise la fonction $L(x_1, x_2, ..., x_n; \theta)$ sur l'ensemble des paramètres θ pour trouver $\widehat{\theta}$ l'estimateur de MV,

$$\widehat{\theta} = \arg\max_{\theta} L(x_1, x_2, ..., x_n; \theta).$$

Exercice : Maximum de vraisemblance (1)

- **Exercice 2.** Soit la variable aléatoire $X \rightsquigarrow \mathcal{N}(m, \sigma)$ et $(X_1, X_2, ..., X_n)$ un échatillon $i.i.d$ de même loi que X.

a) Calculer la fonction de **MV** de cette échatillon.

b) Estimer les paramètres de cette loi.

∘ *Le principe de la méthode MV :* Si un phénomène X a été l'objet de n observations indépendantes $x_1, x_2, ..., x_n$ les unes des autres, sa loi de probabilité $P(X = x)$(dans le cas discret : loi binomiale, loi de Poisson) ou sa densité (en cas de loi continue, comme la loi normale) est une fonction $f(x; \theta_1, ..., \theta_n)$ où $\theta_1, ..., \theta_n$ sont les paramètres de la loi.

· *Dans le cas discret :* on définit la fonction de **MV** est la probabilité de l'échatillon observé en fonction des paramètres $\theta = (\theta_1, ..., \theta_n)$:

$$\begin{aligned} L(x_1, x_2, ..., x_n; \theta) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n; \theta) \\ &= P_X\ (x_1, x_2, ..., x_n; \theta), \end{aligned}$$

avec $X = (X_1, X_2, ..., X_n)$.

· *Dans le cas continu :* on définit la fonction de **MV** est définie par

$$L(x_1, x_2, ..., x_n; \theta) = P_X\ (x_1, x_2, ..., x_n; \theta).$$

· On maximise la fonction $L(x_1, x_2, ..., x_n; \theta)$ sur l'ensemble des paramètres θ pour trouver $\widehat{\theta}$ l'estimateur de MV,

$$\widehat{\theta} = \arg\max_{\theta} L(x_1, x_2, ..., x_n; \theta).$$

Solution exercice 1 : Maximum de vraisemblance

a) La fonction de **MV** de cette échatillon est donnée par

$$\begin{aligned} L(x_1, x_2, ..., x_n; p) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\ &= P(X_1 = x_1)(X_2 = x_2)...(X_n = x_n) \text{ car } X_i (i = 1.., n) \text{ sont indépendantes} \\ &= \prod_{i=1}^{n} C_n^{x_i} p^{x_i} (1-p)^{1-x_i} \\ &= \left(\prod_{i=1}^{n} C_n^{x_i} \right) p^{\sum_{i=1}^{n} x_i} (1-p)^{n - \sum_{i=1}^{n} x_i}. \end{aligned}$$

b) On a

$$\widehat{p} = \arg\max_{\pi} L(x_1, x_2, ..., x_n; p),$$

on maximise la fonction $L(x_1, x_2, ..., x_n; p)$ sur l'ensembles des paramètres, alors

$$\begin{aligned} \frac{\partial L}{\partial p} &= \left(\prod_{i=1}^{n} C_n^{x_i} \right) \left[\left(\sum_{i=1}^{n} x_i \right) p^{\left(\sum_{i=1}^{n} x_i \right) - 1} (1-p)^{n - \sum_{i=1}^{n} x_i} - \left(n - \sum_{i=1}^{n} x_i \right) p^{\sum_{i=1}^{n} x_i} (1-p)^{\left(n - \sum_{i=1}^{n} x_i \right) - 1} \right] \\ &= \left(\prod_{i=1}^{n} C_n^{x_i} \right) (1-p)^{n - \sum_{i=1}^{n} x_i} p^{\sum_{i=1}^{n} x_i} \left[\left(\sum_{i=1}^{n} x_i \right) p^{-1} - \left(n - \sum_{i=1}^{n} x_i \right) (1-p)^{-1} \right], \end{aligned}$$

$\frac{\partial L}{\partial \pi} = 0$, implique

$$\frac{\sum_{i=1}^{n} x_i}{p} = \frac{n - \sum_{i=1}^{n} x_i}{1-p},$$

alors

$$p \left(n - \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i,$$

$$p = \frac{1}{n} \sum_{i=1}^{n} x_i$$

donc l'estimateur de MV du paramètre p est donné par

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X}.$$

Solution exercice 2 : Maximum de vraisemblance

a) La fonction de **MV** de cette échatillon est donnée par

$$\begin{aligned} L(x_1, x_2, ..., x_n; m, \sigma) &= P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) \\ &= f(x_1) f(x_2) ... f(x_n) \text{ car les } X_1, X_2, ..., X_n \text{ sont indépendantes.} \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \left(\prod_{i=1}^{n} \exp\left(-\frac{(x_i - m)^2}{2\sigma^2}\right)\right) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2\right). \end{aligned}$$

b) Prenons le logarithme népérien $\mathcal{L}(m, \sigma) = \ln L(x_1, x_2, ..., x_n; m, \sigma)$ du produit, on obtient

$$\mathcal{L}(m, \sigma) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m)^2 - n \ln \sigma\sqrt{2\pi}.$$

Les dérivées partielles par rapport à m et à σ sont respectivement

$$\frac{\partial \mathcal{L}(m, \sigma)}{\partial m} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - m),$$

et

$$\frac{\partial \mathcal{L}(m, \sigma)}{\partial m} = \frac{1}{\sigma^3} \sum_{i=1}^{n} (x_i - m)^2 - \frac{n}{\sigma},$$

Ces dérivées s'annulent lorsque

$$m = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

et

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - m)^2,$$

donc l'estimateur de MV du paramètre p est donné par

$$\widehat{m} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

et

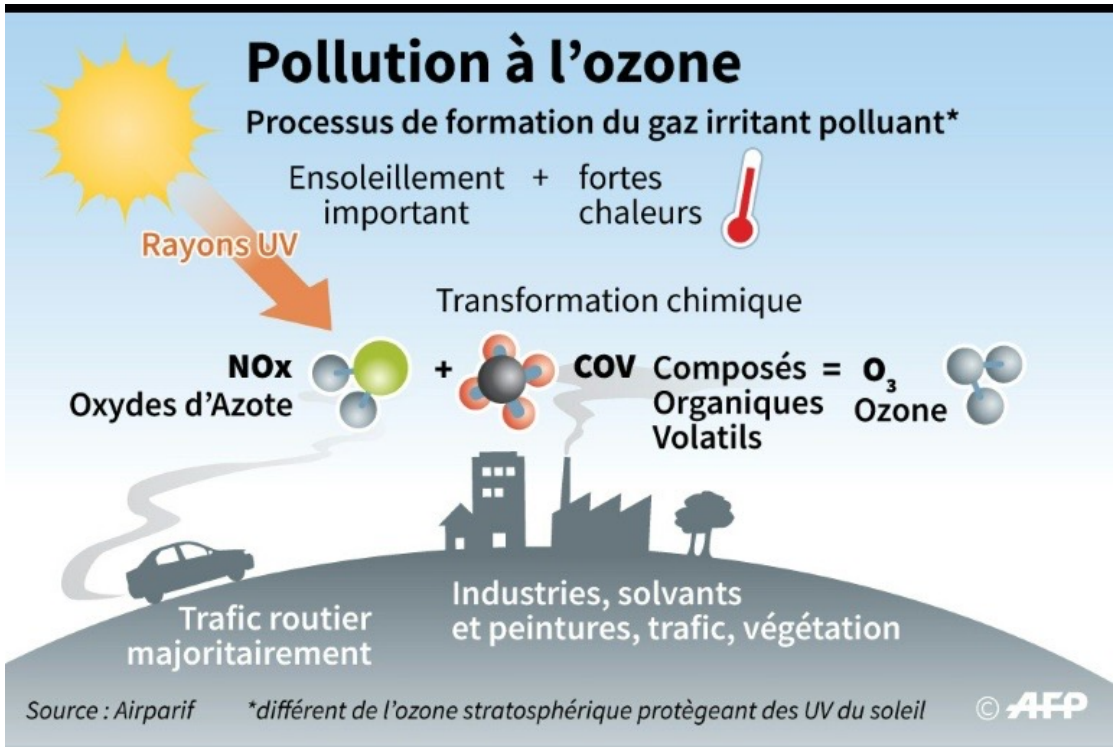
$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2.$$

Outline

- Déroulé du cours
- Introduction au modèle linéaire généralisé
- Théorie
  - Modélisation
  - Estimation des paramètres
  - Tests d'hypothèses
  - Evaluation et choix des modèles
  - Régression logistique
- Cas pratiques
  - Loi de Bernoulli
  - Loi binomiale
  - Loi de Poisson
  - Loi multinomiale

Contexte

- Pollution par l'ozone



- L'ozone : polluant secondaire formé à partir de monoxyde (NO) et dioxyde (NO2) d'azote en présence d'un fort rayonnement.
  - peut provoquer des insuffisances respiratoires, des céphalées ...
- Les agences de qualité de l'air :
  - surveille la concentration des polluants
  - prévienne la population à risque
- La procédure d'alerte :
  - si deux stations dépassent simultanément un seuil s1, l'alerte 1 est donnée,
  - si deux stations dépassent simultanément un seuil s2, l'alerte 2 est donnée.
- 2 variables climatiques considérées :
  - température maximale du jour
  - vitesse moyenne du vent

Contexte (1)

- On dispose d'une base de données contenant les concentrations maximales mesurées chaque jour en plusieurs stations ainsi que la température maximale et la vitesse moyenne du vent pour les mêmes dates sur une période de plusieurs années.
- Après avoir éliminé des dates pour lesquelles on a trop de données manquantes
  - on constitue une variable binaire Y qui est égale à 1 si au moins deux stations dépassent le seuil 180µ/m3 qui représente le seuil d'alerte de niveau 2, et 0 sinon.
  - Le graphique montre le lien entre les deux variables météorologiques et la variable indicatrice de l'alerte.
  - Les jours où l'alerte doit être déclenchée la température est élevée et la vitesse du vent est faible mais ce n'est pas toujours le cas