Chapitre 4: Les modèles linéaires généralisés.

4.1) Introduction:

Le modèle de régression linéaire multiple étudié lors des derniers chapitres peut parfois avoir certaines limitations:

- Distribution normale
  - ↳ Inappropriée dans la plupart des contextes (surtout en actuariat)
- Variance constante
  - ↳ Hypothèse très contraignante
- Valeurs possibles de Y entre $-\infty$ et $+\infty$
  - ↳ Plusieurs contextes positifs seulement (ex: réclamations d'assurances)
  - ↳ Possible que Y soit une v.a discrète
    - ex: - Nombre de réclamations $(0, 1, 2, \ldots)$
    - - Renouvellement de police (0 ou 1)

Le modèle linéaire généralisé (ou encore GLM pour "generalized linear models") est une généralisation de la régression linéaire multiple dont l'objectif est de palier aux limitations précédentes.

4.2) Background: La famille exponentielle :

De manière générale, une v.a. Y obéit à une distribution faisant partie de la famille exponentielle si :

$$f_Y(y) = \exp\left\{ \frac{y \times \theta - b(\theta)}{a(\phi)} - c(y, \phi) \right\}$$

; où

- θ : Paramètre canonique
- ϕ : Paramètre de dispersion
- $a(\phi)$, $b(\theta)$, $c(y, \phi)$: 3 fonctions générales de y, θ et ϕ.

exemples:

(1) Loi Normale: En posant:

- $\theta = \mu$
- $\phi = \sigma^2$
- $b(\theta) = \theta^2/2$
- $a(\phi) = \phi$
- $c(y,\phi) = \frac{-1}{2}\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$

... alors

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

et

$$Y \sim N(\mu, \sigma^2)$$

Preuve:

$$f_Y(y) = \exp\left\{ \frac{y\mu - \mu^2/2}{\sigma^2} + \left(\frac{-1}{2}\right)\left(\frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2)\right) \right\}$$

$$= \exp\left\{ -\frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) \right\}$$

$$= \exp\left\{ -\frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2}\ln(2\pi\sigma^2) \right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\cdot\left(\frac{y-\mu}{\sigma}\right)^2}$$

(2) Loi Poisson: En posant:

- $\theta = \ln(\mu)$
- $\phi = 1$
- $b(\theta) = e^{\theta}$
- $a(\phi) = \phi$
- $c(y, \phi) = -\ln(y!)$

... alors

$$f_Y(y) = \frac{e^{-\mu}\mu^y}{y!}$$

et

$$Y \sim \text{Poisson}(\mu)$$

Preuve:

(73)

$$f_y(y) = \exp\left\{ \frac{y\ln(\mu) - e^{\ln(\mu)}}{1} + (-1)\ln(y!) \right\}$$

$$= \exp\left\{ y\ln(\mu) - \mu + \ln\left(\frac{1}{y!}\right) \right\}$$

$$= \frac{e^{-\mu}\mu^{y}}{y!}$$

(3) Loi Bernouilli: En posant:

- $\theta = \ln(\pi/(1-\pi))$
- $\phi = 1$
- $b(\theta) = \ln(1+e^{\theta})$
- $a(\phi) = \phi$
- $c(y,\phi) = 0$

... alors

$$f_y(y) = \pi^{y} \times (1-\pi)^{1-y}$$

et

$$Y \sim Bern(\pi)$$

Preuve:

$$f_Y(y) = \exp\left\{ \frac{y \times \ln\left(\frac{\pi}{1-\pi}\right) - \ln\left(1 + e^{\ln\left(\frac{\pi}{1-\pi}\right)}\right)}{1} + 0 \right\}$$

$$= \exp\left\{ y \times \ln(\pi) - y \times \ln(1-\pi) + \ln\left(1-\pi\right) \right\}$$

$$= \exp\left\{ y \times \ln(\pi) + (1-y) \times \ln(1-\pi) \right\}$$

$$= \pi^{y} \times (1-\pi)^{1-y}.$$

autres exemples :

- Loi beta
- Loi binomiale
- Loi gamma
- Loi inverse-Gaussienne
- Loi binomiale négative
- Loi pareto
- Loi weibull

... sont toutes des distributions qui appartiennent à la famille exponentielle.

4.3) Généralités sur les modèles de régression avec la famille exponentielle:

4.3.1) Contexte:

Le contexte est très similaire à celui de la régression multiple :

- $\underset{n\times 1}{Y}$: Vecteur des observations de Y_i $(i=1,\dots,n)$
- $X_{n\times(p+1)}$: Matrice shéma contenant n lignes d'observations, et (p+1) colonnes de variables explicatives
- $\beta_{(p+1)\times 1}$: Vecteur des (p+1) paramètres $\beta_0, \beta_1, \dots, \beta_p$ à estimer

4.3.2) Structure du modèle:

On suppose maintenant que

- $(Y_i \mid X_i) \sim$ Famille exponentielle

et que

- $$g\big(\mathbb{E}(Y_i \mid X_i)\big) = X_i \beta \;=\; \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}$$

où $g(\cdot)$ est une fonction continue appelée **fonction de lien**

⊛ Important (fonction de lien):

- **Domaine de $g(\cdot)$:** Domaine des valeurs possible de $\mu \equiv \Omega(\mu)$

  ex: • loi gamma $\Rightarrow \Omega(\mu) = [0, +\infty[$

  • loi bernoulli $\Rightarrow \Omega(\mu) = [0, 1]$

  ⋮

- **Image de $g(\cdot)$:** $\mathbb{R} =]-\infty, +\infty[$.

- **Conclusion:**

  Le but de la fonction de lien est d'obtenir des valeurs de μ qui "font du sens" dans le contexte du modèle, à partir du "prédicteur linéaire" $X\beta$ qui peut prendre valeurs dans $\mathbb{R}$.

  Ainsi, on obtient μ en inversant $g(\cdot)$:

$$g(\mathbb{E}(Y_i | X_i)) = X_i \beta$$

$$\Rightarrow \boxed{\mathbb{E}(Y_i | X_i) = g^{-1}(X_i \beta)}$$

*Il est donc nécessaire de choisir une fonction inversible pour la fonction de lien!

4.4) Approche générale:

4.4.1) Procédure avec les GLM:

I) Choisir une distribution pour Y dans la famille exponentielle

II) Choisir une fonction de lien $g(\cdot)$

III) Estimer les paramètres $\underline{\beta}$ et ϕ

IV) Valider le modèle

4.4.2) Estimation des paramètres:

Dans le cas des GLM, on estime les paramètres en utilisant la méthode du maximum de vraisemblance

On souhaite donc choisir $\underline{\beta}$ qui maximise la fonction de vraisemblance suivante:

$$\ell(\underline{\beta}) = \sum_{i=1}^{n} \ln\left(f_Y(y_i) \right)$$

... nouvelle "métrique" de "distance"

12-11-2012