

# Modèles linéaires généralisés

## TD

M1 Actuariat, année 2022-2023

**Exercice 1** Le but de cet exercice est de vous faire manipuler les commandes permettant d'effectuer une régression linéaire simple et multiple sous R et SAS mais également d'analyser vos résultats et effectuer des représentations graphiques.

a) Nous allons utiliser le jeu de données *gavote* qui contient des informations sur les élections présidentielles aux Etats Unis, en particulier en Géorgie. Nous retrouvons, par ligne, les différentes comtés de la Géorgie et, par colonne, les variables suivantes :

- equip : type de matériel de vote utilisé
- econ : le niveau économique de la comté
- perAA : le pourcentage de afro-américains
- rural : comté rurale ou urbaine
- atlanta : la comté fait partie de l'aire métropolitaine de Atlanta ou pas
- gore : nombre d'électeurs pour Al Gore
- bush : nombre d'électeurs pour George Bush
- other : nombre d'électeurs pour les autres candidats
- votes : nombre de suffrages exprimés
- ballots : bulletins de vote émis

Après avoir analysé les données (par des graphiques et des résumés statistiques des variables par exemple), un modèle de régression doit être mis en place pour déterminer les facteurs qui affectent les bulletins nuls (variable *undercount*). Le modèle estimé doit être ensuite validé par des tests statistiques, une analyse des résidus et l'étude des différentes mesures d'influence.

b) Un deuxième jeu de données (*TermLife*) sera analysé par les étudiants afin de proposer un modèle de régression pour expliquer la variable *FACE* en prenant seulement les observations pour lesquelles  $FACE > 0$ . Y a-t-il besoin de transformer au préalable la variable *FACE*? Pourquoi? Quelle transformation appliquer? Un descriptif de ce jeu de données sera fourni en séance.

**Exercice 2** Nous considérons la transformée de Box-Cox. Écrire le modèle et calculer, à  $\lambda$  fixé, les estimateurs du maximum de vraisemblance de  $\beta$  et  $\sigma^2$ . Trouver ensuite la procédure d'estimation du paramètre  $\lambda$  de la transformée.

**Exercice 3** Soit  $Y$  une variable aléatoire dont la densité peut se mettre sous la forme exponentielle :

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

Montrer que :

a)

$$E(Y) = b'(\theta), \quad Var(Y) = b''(\theta)a(\phi)$$

où ' et '' désignent les dérivées premières et secondes par rapport à  $\theta$ .

b) la déviance peut s'exprimer comme :

$$D = 2 \sum_{i=1}^n \left\{ \frac{y_i(\tilde{\theta} - \hat{\theta}) - b(\tilde{\theta}) + b(\hat{\theta})}{a(\phi)} \right\}$$

avec  $\tilde{\theta}$  l'estimateur du maximum de vraisemblance sous le modèle saturé et  $\hat{\theta}$  l'estimateur du maximum de vraisemblance sous le modèle estimé.

**Exercice 4** Nous considérons les lois de Poisson, Bernoulli et Gamma.

- a) Montrer que ces lois appartiennent à la famille exponentielle en déterminant le paramètre de la moyenne  $\theta$ , le paramètre de dispersion, les fonctions  $b$  et  $c$ .
- b) Trouver la fonction lien canonique ainsi que la fonction variance  $V(\mu)$ .

**Exercice 5** Montrer qu'en cas de variable réponse Bernoulli, la déviance est donnée par :

$$D = -2 \sum_{i=1}^n \left\{ \hat{\pi}_i \ln \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} + \ln(1 - \hat{\pi}_i) \right\}$$

Dans ce cas, peut-on utiliser la déviance comme mesure de la qualité d'ajustement du modèle ?

Calculer également la déviance pour des variables réponses Normale et Poisson.

**Exercice 6** À l'aide de la procédure SAS PROC GENMOD, construire un modèle linéaire généralisé pour les variables :

- nombre d'enfants (jeu de données *enfants.txt*) ;
- nombre de décès dus au diabète (jeu de données *diabete.txt*) ;
- montant des sinistres (jeu de données *car.csv*).

Commenter les résultats et valider le modèle proposé.

**Exercice 7** Nous considérons un portefeuille d'assurance automobile belge comprenant 20354 polices, observées durant une période de 3 ans. Pour chaque police et pour chaque année sont renseignés le nombre de sinistres et certaines caractéristiques de l'assuré : le sexe du conducteur (homme - femme), l'âge du conducteur (trois classes d'âge : 18 – 22 ans, 23–30 ans et > 30 ans), la puissance du véhicule (trois classes de puissance : < 66kW, 66–110kW et > 110kW), la taille de la ville de résidence du conducteur (grande, moyenne ou petite, en fonction du nombre d'habitants) et la couleur du véhicule (rouge ou autre).

La procédure PROC GENMOD de SAS permet de réaliser la régression de Poisson du nombre de sinistres sur les 5 variables explicatives présentées ci-dessus. Les résultats sont présentés dans le Tableau suivant :

Paramètre	Estimation	Erreur Standard	Wald 95% limites de confiance	Khi 2	Pr>Khi 2
Intercept	-1.9242	0.0302	-1.9833 -1.8650	4063.54	<.0001
Sexe F	-0.0581	0.0265	-0.1100 -0.0063	4.82	0.0281
Sexe H	0	0	0 0	.	.
Age 17-22	0.6651	0.0583	0.5508 0.7793	130.23	<.0001
Age 23-30	0.2525	0.0261	0.2015 0.3036	93.87	<.0001
Age >30	0	0	0 0	.	.
Puissance >110kW	-0.0116	0.0750	-0.1586 0.1353	0.02	0.8769
Puissance 66-110kW	0.0563	0.0275	0.0024 0.1102	4.19	0.0406
Puissance <66kW	0	0	0 0	.	.
Ville grande	0.2549	0.0306	0.1949 0.3150	69.27	<.0001
Ville moyenne	0.0756	0.0311	0.0147 0.1364	5.92	0.0150
Ville petite	0	0	0 0	.	.
Couleur Rouge	-0.0236	0.0416	-0.1052 0.0580	0.32	0.5710
Couleur Autre	0	0	0 0	.	.

- 1) Indiquez clairement ce que chaque colonne représente. A quoi correspondent les lignes où apparaissent des 0 ?
- 2) Commentez à la fois le signe du coefficient obtenu et sa significativité.
- 3) Est-ce que au vu du tableau ci-dessus certaines variables explicatives pourraient être omises sans nuire à la qualité du modèle ?

La valeur de  $L(\hat{\beta})$  pour le modèle reprenant les 5 variables explicatives est  $-19282.6$ . L'analyse de type 3 fournit les résultats présentés au tableau suivant :

Source	DF	Khi 2	Pr>Khi 2
Sexe	1	4.85	0.0276
Age	2	173.56	<.0001
Puissance	2	4.38	0.1120
Ville	2	74.10	<.0001
Couleur	1	0.32	0.5698

- 4) Commentez soigneusement. Est-ce que l'on peut raisonnablement diminuer le nombre de variables explicatives ?
- 5) Et si on avait effectué une analyse de type 1 aurait-on obtenu les mêmes résultats ? Oui ? Non ? Pourquoi ?

Dans une deuxième étape nous regroupons les niveaux de puissance ‘‘66-110kW’’ et ‘‘>110kW’’ en une seule classe. Nous en arrivons au modèle décrit au tableau suivant :

Paramètre	Estimation	Erreur Standard	Wald 95% limites de confiance	Khi 2	Pr>Khi 2
Intercept	-1.9277	0.0299	-1.9862 -1.8692	4165.69	<.0001
Sexe F	-0.0575	0.0265	-0.1093 -0.0056	4.72	0.0299
Sexe H	0	0	0 0	.	.
Age 17-22	0.6668	0.0582	0.5526 0.7809	131.02	<.0001
Age 23-30	0.2547	0.0260	0.2038 0.3056	96.09	<.0001
Age >30	0	0	0 0	.	.
Puissance >66kW	0.0508	0.0269	-0.0019 0.1034	3.57	0.0587
Puissance <66kW	0	0	0 0	.	.
Ville grande	0.2545	0.0306	0.1944 0.3145	69.03	<.0001
Ville moyenne	0.0757	0.0311	0.0148 0.1365	5.93	0.0148
Ville petite	0	0	0 0	.	.

La log-vraisemblance vaut  $-19283.2$  et l'analyse de Type 3 fournit les résultats suivants :

Source	DF	Khi 2	Pr>Khi 2
Sexe	1	4.74	0.0294
Age	2	176.07	<.0001
Puissance	1	3.56	0.0593
Ville	2	73.82	<.0001

5) Que fait-on au vu des nouveaux résultats ? Commentez.

Exercice 3:

Y une VA tq sa densité  $f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$

a) Compose  $\ell = \ln(f(y; \theta, \phi))$

$$\text{On a alors } \frac{\partial \ell(y; \theta, \phi)}{\partial \theta} = \frac{1}{f(y; \theta, \phi)} \frac{\partial f(y; \theta, \phi)}{\partial \theta}$$

$$\mathbb{E}\left[\frac{\partial \ell(Y)}{\partial \theta}\right] = \int \frac{1}{f(y; \theta, \phi)} \frac{\partial f(y; \theta, \phi)}{\partial \theta} f(y; \theta, \phi) dy = \frac{\partial}{\partial \theta} \underbrace{\int f(y; \theta, \phi) dy}_{=1} = 0$$

$$\begin{aligned} \text{Dans notre cas, } \frac{\partial \ell(Y)}{\partial \theta} &= \frac{Y - b'(\theta)}{a(\phi)} \Rightarrow \mathbb{E}\left[\frac{\partial \ell(Y)}{\partial \theta}\right] = \frac{\mathbb{E}[Y] - b'(\theta)}{a(\phi)} = 0 \\ &\Rightarrow \mathbb{E}[Y] = b'(\theta). \end{aligned}$$

Pour la variance, on a que  $\frac{\partial \ell}{\partial \theta} = \frac{1}{f} f'$  et on obtient  $\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{1}{f^2} (f')^2 + \frac{1}{f} f''$

$$\begin{aligned} \Rightarrow \mathbb{E}\left[\frac{\partial^2 \ell(Y)}{\partial \theta^2}\right] &= - \int \underbrace{\frac{(f'(y))^2}{(f(y))^2}}_{=\mathbb{E}[f'(y)]^2} f(y) dy + \int \underbrace{\frac{f''(y)}{f(y)}}_{=\frac{\partial^2}{\partial \theta^2} \int f(y) dy = 0} f(y) dy \\ &= \left(\frac{f'}{f}\right)^2 = \left(\frac{\partial \ell}{\partial \theta}\right)^2 \end{aligned}$$

$$\Rightarrow -\mathbb{E}\left[\frac{\partial^2 \ell(Y)}{\partial \theta^2}\right] = \mathbb{E}\left[\left(\frac{\partial \ell(Y)}{\partial \theta}\right)^2\right]$$

$$\Leftrightarrow \frac{b''(\theta)}{a(\phi)} = \mathbb{E}\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right]$$

$$\text{On a que } \frac{\partial \ell(Y)}{\partial \theta} = \frac{Y - b'(\theta)}{a(\phi)} \text{ et } \frac{\partial^2 \ell(Y)}{\partial \theta^2} = -\frac{b''(\theta)}{a(\phi)}$$

$$\text{Donc } \frac{b''(\theta)}{a(\phi)} = \frac{1}{(a(\phi))^2} \underbrace{\mathbb{E}[Y - b'(\theta)]^2}_{=\text{Var}(Y)}$$

$$\Rightarrow \text{Var}(Y) = b''(\theta) a(\phi)$$

b)  $D = 2 \ln(\lambda) = 2 \ln\left(\frac{\mathcal{L}_{\text{SAT}}}{\mathcal{L}}\right)$

$$\mathcal{L} = \prod_{i=1}^m f(y_i; \hat{\theta}, \phi) = \exp\left(\sum_{i=1}^m \left[\frac{y_i \hat{\theta} - b(\hat{\theta})}{a(\phi)} + c(y_i, \phi)\right]\right)$$

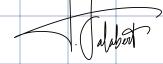
$$\begin{aligned} \Rightarrow \frac{\mathcal{L}_{\text{SAT}}}{\mathcal{L}} &= \exp\left(\sum_{i=1}^m \left[\frac{y_i \tilde{\theta} - b(\tilde{\theta})}{a(\phi)} + c(y_i, \phi)\right] - \sum_{i=1}^m \left[\frac{y_i \hat{\theta} - b(\hat{\theta})}{a(\phi)} + c(y_i, \phi)\right]\right) \\ &= \exp\left(\sum_{i=1}^m \frac{1}{a(\phi)} [y_i (\tilde{\theta} - \hat{\theta}) - b(\tilde{\theta}) + b(\hat{\theta})]\right) \end{aligned}$$

$$\Rightarrow D = 2 \ln\left(\frac{\mathcal{L}_{\text{SAT}}}{\mathcal{L}}\right) = 2 \sum_{i=1}^m \left[\frac{y_i (\tilde{\theta} - \hat{\theta}) - b(\tilde{\theta}) + b(\hat{\theta})}{a(\phi)}\right]$$

## Exercice 4:

Exercice 4 Nous considérons les lois de Poisson, Bernoulli et Gamma.

- Montrer que ces lois appartiennent à la famille exponentielle en déterminant le paramètre de la moyenne  $\theta$ , le paramètre de dispersion, les fonctions  $b$  et  $c$ .
- Trouver la fonction lien canonique ainsi que la fonction variance  $V(\mu)$ .

© Théo Jalabert 

Cas loi de Poisson :  $Y \sim P(\lambda) \quad \lambda > 0$

a) Donc pour  $y \in \mathbb{N}$ ,  $P(Y=y) = \frac{e^{-\lambda} \lambda^y}{y!}$

$$\begin{aligned} \Rightarrow \ln(P(Y=y)) &= -\lambda + y \ln(\lambda) - \ln(y!) \\ &= -\ln(y!) + y \ln(\lambda) - \lambda \end{aligned}$$

Donc par identification:  $c(y, \phi) = -\ln(y!)$  et  $\theta = \ln(\lambda) \Rightarrow \lambda = \exp(\theta)$

$$b(\theta) = \lambda = \exp(\theta) \text{ puis } a(\phi) = 1$$

D'où  $f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} = \exp\left\{y \frac{\ln(\lambda) - \lambda - \ln(y!)}{1}\right\}$

b) Ici  $\mathbb{E}[Y] = \lambda = \exp(\theta) \Rightarrow \theta = \ln(\lambda)$

Donc la fonction lien canonique est la fonction  $\ln$

$$\text{Var}(Y) = a(\phi) \overline{V(\mu)} \quad \text{car } Y \text{ appartient à la famille exponentielle.}$$

Fonct<sup>o</sup>  
variance

Sachant que  $\text{Var}(Y) = \lambda$  et  $a(\phi) = 1$   
 $\Rightarrow \overline{V(\mu)} = \lambda$

Cas de la loi de Bernoulli:  $X \sim B(p) \quad p \in [0, 1]$

a)  $X \in \{0, 1\}$  et pour  $i \in \{0, 1\}$ ,  $P(X=i) = p^i (1-p)^{1-i}$

$$\Rightarrow i \in \{0, 1\}, \ln(P(X=i)) = i \ln(p) + (1-i) \ln(1-p) = i \ln(\frac{p}{1-p}) + \ln(1-p)$$

Donc par identification:  $c(i, \phi) = 0$  et  $\theta = \ln(\frac{p}{1-p}) \Rightarrow p = \frac{e^\theta}{1+e^\theta}$

puis  $b(\theta) = -\ln(1-p) = \ln(1+e^\theta)$   
et  $a(\phi) = 1$

D'où  $f_X(i; \theta, \phi) = \exp\left[\frac{i\theta - b(\theta)}{a(\phi)} + c(i, \phi)\right] = p^i (1-p)^{1-i}$

Donc la loi de Bernoulli appartient à la famille exponentielle.

$$\text{b) Ici } \mathbb{E}[X] = p = \frac{e^\theta}{1+e^\theta} \Rightarrow \theta = \ln\left(\frac{p}{1-p}\right)$$

Fonction liens canonique

Puis par appartenance à la famille expo on a :

$$\text{Var}(X) = a(\phi)V(\mu)$$

$$\Rightarrow p(1-p) = V(\mu)$$

Donc la fonction Variance  $V(\mu)$  de la loi de Bernoulli satisfait :

$$V(\mu) = p(1-p).$$

Cas de la loi Gamma :  $Z \sim \Gamma(\alpha, \beta)$   $(\alpha, \beta) \in (\mathbb{R}_*)^2$

a) Il est évident que la loi Gamma appartient à la famille exponentielle puisque  $\forall m, \exists \theta$  si  $Y \sim \mathcal{E}(\theta)$  alors  $mY \sim \Gamma(m, \theta)$ .

$$f_Z(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z} \quad \text{avec } z \geq 0$$

Ici nous allons poser  $\mu = \frac{\alpha}{\beta}$  et  $\nu = \alpha$

On considère maintenant la v.a.  $X \sim \Gamma(\mu, \nu)$

↗ Il s'agit de la paramétrisation de la loi Gamma qui est utilisée en MLG.

$$\text{On a donc } f_X(x) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu x}{\mu}\right)^\nu \exp(-\frac{x\nu}{\mu}) \frac{1}{x} \quad x > 0$$

$$\Rightarrow \ln(f_X(x)) = -\ln(\Gamma(\nu)) + \nu \ln\left(\frac{\nu x}{\mu}\right) - \frac{x\nu}{\mu} - \ln(x)$$

$$\text{Donc par identification, on a: } c(x, \phi) = -\ln(\Gamma(\nu)) - \ln(x) \quad \text{et} \quad \theta = -\frac{\nu}{\mu} \\ = \ln\left(\frac{1}{x\Gamma(\nu)}\right)$$

$$\text{puis } b(\theta) = -\nu \ln\left(\frac{\nu x}{\mu}\right)$$

$$\text{et } a(\phi) = 1$$

$$\text{b) } \mathbb{E}[X] = \frac{\mu}{\nu} = -\frac{1}{\theta} \Rightarrow \theta = -\frac{\nu}{\mu} = -\frac{1}{\mathbb{E}[X]} \rightarrow \text{Fonction lien canonique}$$

$$\text{et par a) on a } \text{Var}(X) = a(\phi)V(X)$$

Fonction Variance

$$\text{Comme } \text{Var}(X) = \frac{\mu^2}{\nu} \text{ et } a(\phi) = 1$$

$$\Rightarrow V(X) = \frac{\mu^2}{\nu}$$

Calculons maintenant les déviances pour chacune de ces 3 lois :

\*  $Y \sim \mathcal{P}(\lambda)$   $\lambda > 0$ ,  $D = 2 \ln \left( \frac{\mathcal{L}_{SAT}}{\mathcal{L}} \right)$

$$\begin{aligned}\mathcal{L}_{SAT} &= \prod_{i=1}^m \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \quad \text{avec } \lambda_i = y_i \text{ car modèle saturé} \\ &= \prod_{i=1}^m \frac{y_i^{y_i} e^{-y_i}}{y_i!}\end{aligned}$$

$$\begin{aligned}\text{et } \mathcal{L} &= \prod_{i=1}^m \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \Rightarrow D &= 2 \left( \sum_{i=1}^m \left[ y_i \ln(y_i) - y_i - \ln(y_i!) \right] - \sum_{i=1}^m \left[ y_i \ln(\lambda_i) - \lambda_i - \ln(\lambda_i!) \right] \right) \\ &\quad \lambda \text{ car modèle estimé} \\ &= 2 \left( \sum_{i=1}^m \left[ y_i \ln\left(\frac{y_i}{\lambda_i}\right) - (y_i - \lambda_i) \right] \right) \\ &= 2 \sum_{i=1}^m \left[ y_i \ln\left(\frac{y_i}{\lambda_i}\right) - (y_i - \lambda_i) \right]\end{aligned}$$

Les résidus de déviance dans un modèle de Poisson s'écrivent:

$$r_i^D = \text{signe}(y_i - \lambda_i) \sqrt{d_i}$$

\*  $X \sim \mathcal{B}(p)$   $p \in [0, 1]$ .  $D = 2 \ln \left( \frac{\mathcal{L}_{SAT}}{\mathcal{L}} \right)$

$$\hat{p} = \frac{1}{m} \sum_{i=1}^m y_i$$

$$\ln(\mathcal{L}_{SAT}) = \sum_{i=1}^m \left[ y_i \ln(y_i) + (1-y_i) \ln(1-y_i) \right]$$

$$\mathcal{L} = \prod_{i=1}^m \hat{p}^{y_i} (1-\hat{p})^{1-y_i} \Rightarrow \ln(\mathcal{L}) = \sum_{i=1}^m y_i \ln(\hat{p}) + (1-y_i) \ln(1-\hat{p})$$

$$\begin{aligned}\Rightarrow D &= 2 \sum_{i=1}^m \left[ y_i \ln(y_i) + (1-y_i) \ln(1-y_i) - y_i \ln(\hat{p}) - (1-y_i) \ln(1-\hat{p}) \right] \\ &= 2 \sum_{i=1}^m \left[ y_i \left( \ln\left(\frac{y_i}{\hat{p}}\right) - \ln\left(\frac{1-y_i}{1-\hat{p}}\right) \right) + \ln\left(\frac{1-y_i}{1-\hat{p}}\right) \right]\end{aligned}$$

\*  $X \sim \Gamma(\mu, \nu)$

$$\mathcal{L} = \prod_{i=1}^m \left[ \frac{1}{\Gamma(\nu)} \left( \frac{\nu x_i}{\mu} \right)^{\nu} \exp\left(-\frac{\nu x_i}{\mu}\right) \frac{1}{x_i} \right] \text{ et } \mathcal{L}_{SAT} = \prod_{i=1}^m \left[ \frac{1}{\Gamma(\nu)} \left( \frac{\nu x_i}{\mu} \right)^{\nu} \exp\left(-\frac{\nu x_i}{\mu}\right) \frac{1}{x_i} \right]$$

modèle estimé les  $\mu_i$  sont remplacés par les  $\hat{\mu}_i$   
et dans le modèle saturé les  $\mu_i$  sont remplacés par les  $y_i$ .

$$\hat{P} = \pi$$

**Exercice 5** Montrer qu'en cas de variable réponse Bernoulli, la déviance est donnée par :

$$D = -2 \sum_{i=1}^n \left\{ \hat{\pi}_i \ln \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} + \ln(1 - \hat{\pi}_i) \right\}$$

Dans ce cas, peut-on utiliser la déviance comme mesure de la qualité d'ajustement du modèle ?

Calculer également la déviance pour des variables réponses Normale et Poisson.

$$\hat{\pi}_i = \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}}$$

$$\text{Où } D = 2 \sum_{i=1}^m \left[ y_i \left( h\left(\frac{y_i}{1-y_i}\right) - h\left(\frac{1}{1-\hat{\pi}_i}\right) \right) + h\left(\frac{1-y_i}{1-\hat{\pi}_i}\right) \right]$$

=

Le modèle de régression logistique s'écrit

$$g(\pi) = \ln \left( \frac{\pi}{1 - \pi} \right) = x' \beta$$

avec  $Y \sim \mathcal{B}(\pi)$  et  $\pi = P(Y = 1)$ . La fonction lien est la fonction lien logit ( $g(\mu) = \frac{\mu}{1-\mu}$ ) qui est la fonction lien canonique pour Bernoulli.

Le rapport  $\frac{\pi}{1-\pi}$  est appelé rapport de côtes (odds ratio) ou simplement côte.

Nous trouvons facilement

$$\pi = \frac{e^{x' \beta}}{1 + e^{x' \beta}}$$