Sophie Baillargeon

Solutions, série d'exercices sur les GLM

1. Pour le lien logit $logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = exp(\boldsymbol{x}_i^t\boldsymbol{\beta}) \Longrightarrow \pi_i = \frac{exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}{1+exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}$. On a que :

$$\begin{aligned} \boldsymbol{x}_i^t\boldsymbol{\beta} &\in (-\infty, \infty) \\ exp(\boldsymbol{x}_i^t\boldsymbol{\beta}) &\in (0, \infty) \end{aligned}$$

Posons $a = exp(\boldsymbol{x}_i^t\boldsymbol{\beta})$. Étant donné que $\frac{a}{1+a}$ est une fonction croissante en a (dérivée $\frac{\partial}{\partial a}\frac{a}{1+a} = \frac{1}{(1+a)^2}$ toujours positive), on n'a qu'à calculer la valeur de $\frac{a}{1+a}$ au extrémités $a = 0$ et $a = \infty$ pour trouver l'intervalle des valeurs possibles de π_i.
En $a = 0$, $\frac{a}{1+a} = \frac{0}{1+0} = 0$.
De plus, $\lim\limits_{a\to\infty} \frac{a}{1+a} = \lim\limits_{a\to\infty} \frac{1}{1/a+1} = \frac{1}{1/\infty+1} = \frac{1}{0+1} = 1$.
Ainsi, on a bien que $\pi_i \in (0, 1)$.

Pour le lien probit, $\pi_i = \Phi(\boldsymbol{x}_i^t\boldsymbol{\beta})$ où Φ est la fonction de répartition d'une $N(0, 1)$. Par définition, une fonction de répartition retourne toujours des valeurs entre 0 et 1.

Pour le lien log-log, on a

$$\begin{aligned} \ln\left(-\ln\left(1-\pi_i\right)\right) &= \boldsymbol{x}_i^t\boldsymbol{\beta} \\ -\ln\left(1-\pi_i\right) &= \exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right) \\ \left(1-\pi_i\right) &= \exp\left(-\exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right)\right) \\ \pi_i &= 1-\exp\left(-\exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right)\right). \end{aligned}$$

L'intervalle des valeurs prédites est donc déduit ainsi :

$$\begin{aligned} \boldsymbol{x}_i^t\boldsymbol{\beta} &\in (-\infty, \infty) \\ \exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right) &\in (0, \infty) \\ -\exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right) &\in (-\infty, 0) \\ \exp\left(-\exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right)\right) &\in (0, 1) \\ 1-\exp\left(-\exp\left(\boldsymbol{x}_i^t\boldsymbol{\beta}\right)\right) &\in (0, 1) \\ \pi_i &\in (0, 1) \end{aligned}$$

2. On a trouvé en classe que le vecteur des dérivées partielles de la log vraisemblance dans une régression logistique simple était $S(\boldsymbol{\beta}) = \sum_{i=1}^n \left((y_i - \pi_i), x_i(y_i - \pi_i)\right)^t$. Le maximum de vraisemblance est atteint lorsque ce vecteur vaut $(0, 0)^t$. On doit donc résoudre le système à deux équations et 2 inconnues $S(\boldsymbol{\beta}) = (0, 0)^t$, où les inconnus sont $\pi(x = 0)$ et $\pi(x = 1)$. Ici, x ne peut prendre que deux valeurs. En conséquence, π ne peut prendre que deux valeurs aussi. On a que $\pi_i = \pi(x = 0)$ si $x_i = 0$ et $\pi_i = \pi(x = 1)$ si $x_i = 1$.

De la deuxième équation, on obtient

$$\sum_{i=1}^n x_i(y_i - \pi_i) = n_{0\bullet} \times 0 + \sum_{i \text{ avec } x_i=1} 1 \times (y_i - \pi_i) = n_{11} - n_{1\bullet}\pi(x = 1) = 0.$$

Ainsi $\hat{\pi}(x=1) = n_{11}/n_{1\bullet}$.

De la première équation, on obtient

$$\sum_{i=1}^{n}(y_i - \pi_i) = \sum_{i \text{ avec } x_i=0}(y_i - \pi_i) + \sum_{i \text{ avec } x_i=1}(y_i - \pi_i) = n_{01} - n_{0\bullet}\pi(x=0) + n_{11} - n_{1\bullet}\pi(x=1) = 0.$$

Dans cette équation, on remplace $\pi(x=1)$ par son maximum $n_{11}/n_{1\bullet}$ trouvé ci-dessus.

$$n_{01} - n_{0\bullet}\pi(x=0) + n_{11} - n_{1\bullet}n_{11}/n_{1\bullet} = n_{01} - n_{0\bullet}\pi(x=0) + n_{11} - n_{11} = n_{01} - n_{0\bullet}\pi(x=0) = 0$$

Ainsi $\hat{\pi}(x=0) = n_{01}/n_{0\bullet}$.

3. Pour le lien logit, $\pi(x+1) = \frac{exp(\alpha+\beta x+\beta)}{1+exp(\alpha+\beta x+\beta)} = \frac{exp(\alpha+\beta x)exp(\beta)}{1+exp(\alpha+\beta x)exp(\beta)} = \frac{exp(\alpha+\beta x)}{1/exp(\beta)+exp(\alpha+\beta x)}$.
Si β est positif, $exp(\beta) > 1$ et $1/exp(\beta) < 1$ donc $\pi(x+1) = \frac{exp(\alpha+\beta x)}{1/exp(\beta)+exp(\alpha+\beta x)} > \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)} = \pi(x)$.

Pour le lien probit, $\pi(x+1) = \Phi(\alpha+\beta x+\beta) = P(N(0,1) < \alpha+\beta x+\beta) > P(N(0,1) < \alpha+\beta x) = \pi(x)$ si $\beta > 0$.

Pour le lien log-log, $\pi(x+1) = 1 - \exp(-\exp(\alpha+\beta x+\beta))$. Si $\beta > 0$, on a que

$$\begin{aligned}
\exp(\alpha+\beta x+\beta) &> \exp(\alpha+\beta x) \\
-\exp(\alpha+\beta x+\beta) &< -\exp(\alpha+\beta x) \\
\exp(-\exp(\alpha+\beta x+\beta)) &< \exp(-\exp(\alpha+\beta x)) \\
-\exp(-\exp(\alpha+\beta x+\beta)) &> -\exp(-\exp(\alpha+\beta x)) \\
1-\exp(-\exp(\alpha+\beta x+\beta)) &> 1-\exp(-\exp(\alpha+\beta x)) \\
\pi(x+1) &> \pi(x)
\end{aligned}$$

Donc pour les trois fonctions de lien, une augmentation de la valeur de x implique une augmentation de la valeur de π et on dit que l'association entre la variable réponse et la variable explicative est positive.

4.
```
data transp;
input rejet x @@;
datalines;
5 1 1 1 1 1 3 1 1 1 2 1 2 1 3 1 6 1 4 1
7 0 5 0 2 0 6 0 3 0 4 0 2 0 5 0 1 0 3 0
;
run;
proc genmod data=transp;
model rejet = x / dist=Poisson link=log;
run;
\end{SASinput}
\begin{SASoutput}[,fontsize=\scriptsize, frame=single,framesep=3mm]
The GENMOD Procedure

        Model Information

Data Set              WORK.TRANSP
Distribution              Poisson
```

```
Link Function                 Log
Dependent Variable          rejet


Number of Observations Read          20
Number of Observations Used          20


             Criteria For Assessing Goodness Of Fit

Criterion                     DF           Value        Value/DF

Deviance                      18         19.0471          1.0582
Scaled Deviance               18         19.0471          1.0582
Pearson Chi-Square            18         18.6992          1.0388
Scaled Pearson X2             18         18.6992          1.0388
Log Likelihood                           13.5594
Full Log Likelihood                     -38.7825
AIC (smaller is better)                  81.5650
AICC (smaller is better)                 82.2709
BIC (smaller is better)                  83.5564


Algorithm converged.


             Analysis Of Maximum Likelihood Parameter Estimates

                           Standard       Wald 95%             Wald
Parameter  DF  Estimate     Error    Confidence Limits  Chi-Square  Pr > ChiSq

Intercept   1    1.3350    0.1622    1.0171    1.6529       67.72      <.0001
x           1   -0.3054    0.2491   -0.7935    0.1828        1.50      0.2201
Scale       0    1.0000    0.0000    1.0000    1.0000

NOTE: The scale parameter was held fixed.
\end{SASoutput}
Note : c'est équivalent à :
\begin{SASinput}[,fontsize=\footnotesize]
data transp;
input rejet medic$ @@;
datalines;
5 N 1 N 1 N 3 N 1 N 2 N 2 N 3 N 6 N 4 N
7 A 5 A 2 A 6 A 3 A 4 A 2 A 5 A 1 A 3 A
;
run;
proc genmod data=transp;
class medic (descending);
model rejet = medic / dist=Poisson link=log;
run;
```

(a) $Y_i|x_i \xrightarrow{\text{ind.}} Poisson(\mu_i)$ avec $\ln(\mu_i) = \alpha + \beta x_i$

(b) $\ln \mu_N - \ln \mu_A = \alpha + \beta \times 1 - (\alpha + \beta \times 0) = \beta$

(c) $\mu_N = \mu_A \Leftrightarrow \ln \mu_N - \ln \mu_A = 0$. Il s'agit donc de tester $H_0 : \beta = 0$ versus $H_0 : \beta < 0$.
Test unilatéral : nous devons utilisez le test de Wald
Statistique du test : $z = \hat{\beta}/ASE(\hat{\beta}) = -0.3054/0.2491 = -1.226$
À comparer à la valeur critique $-Z_\alpha = -1.645$. Ici la statistique du test est plus grande que la valeur critique (i.e. moins extrême) et par conséquent on accepte H_0. Il n'y aurait donc pas de différence significative entre les deux médicaments.

(d) L'int. de conf. pour β est $\hat{\beta} \pm 1.96 ASE(\hat{\beta})$. L'int. de conf. pour $\mu_B/\mu_A = e^\beta$ est donc $\exp\{\hat{\beta} \pm 1.96 ASE(\hat{\beta})\}$. Ici on obtient donc l'int. de conf. suivant pour μ_B/μ_A : $[\exp(-0.7935), \exp(0.1828)] = [0.4523, 1.2001]$.

5. (a) On a que $\mu_i = e^{\cdots + \ln v_i} = v_i e^{\cdots}$, c.-à-d. que le nombre d'éclairs est proportionnel au nombre d'orages, ce qui est tout-à-fait raisonnable. Il s'agit d'une variable offset.

(b) Si on multiplie v_i par 1.05, alors μ_i est multiplié par 1.05 et donc le nombre d'éclairs augmente également de 5%.

(c) Ce nombre sera multiplié par $e^{0.24} \approx 1.27$, donc une hausse de 27%.

(d) Test du rap. des vrais. : $\lambda = D_0 - D_1 = 123.4 - 120.1 = 3.2$, pour un seuil de $P[\chi^2_1 > 3.2] \approx 0.07$, donc une évidence faible pour l'hypothèse nulle qu'il n'y a pas d'interaction.

(e) La réponse en (b) ne change pas.
Par contre, la réponse en (c) devient intéressante en présence d'une interaction. Si x_{i1} augmente d'une unité, alors le nombre moyen d'éclairs est multiplié par $e^{\beta_1 + \beta_2 x_{i2}}$, il dépend donc maintenant de la valeur de x_{i2}. Ainsi pour $x_{i2} = 1$ on a une multiplication par $e^{0.24 - 0.01 \times 1} \approx 1.26$, donc une hausse de 26% alors que pour $x_{i2} = 10$, on multiplie par $e^{0.24 - 0.01 \times 10} \approx 1.15$, donc une hausse de seulement 15%.

6. (a) La proportion de "Oui" va en augmentant avec le revenu pour les revenus faibles à élevés, mais diminue ensuite en allant de revenu élevé à revenu très élevé. On n'a donc pas une relation monotone entre la proba. de "Oui" et le revenu, et donc si on veut entrer le revenu comme variable continue dans le modèle, il faudrait la faire entrer de façon quadratique, certainement pas de façon linéaire.

(b)
```
data model;
input srevenu noui n;
datalines;
1 10 35
3 15 35
5 20 35
6  5 35
;
run;
proc genmod data=model;
model noui / n = srevenu / dist=bin link=logit type3;
output out=num pred=pnum;
run;
proc genmod data=model;
class srevenu;
model noui / n = srevenu / dist=bin link=logit type3;
lsmeans srevenu / ilink;
output out=catego pred=pcat;
run;
```

Les valeurs prédites pour le modèle avec revenu numérique son présentées dans le tableau suivant :

Revenu	Probabilité de "Oui" estimée	Proportion de "Oui" observée
Faible ($x = 1$)	$e^{-0.5-0.021\times1}/(1+e^{-0.5-0.021\times1}) = 0.37$	10/35=0.29
Moyen ($x = 3$)	$e^{-0.5-0.021\times3}/(1+e^{-0.5-0.021\times3}) = 0.36$	15/35=0.43
Élevé ($x = 5$)	$e^{-0.5-0.021\times5}/(1+e^{-0.5-0.021\times5}) = 0.35$	20/35=0.57
Très élevé ($x = 6$)	$e^{-0.5-0.021\times6}/(1+e^{-0.5-0.021\times6}) = 0.35$	5/35=0.14

Les valeurs prédites pour le modèle avec revenu catégorique sont égales aux valeurs observées puisqu'il s'agit d'un modèle saturé (le même nombre de paramètres que d'observations).

On se rend compte que les valeurs prédites par le modèle numérique s'éloignent beaucoup des valeurs observées. Ce modèle ne semble pas bien s'ajuster aux données.

En plus de comparer les valeurs prédites aux valeurs observées, on peut faire un test de rapport de vraisemblance comparant les deux modèles. H_0 : le modèle avec revenu numérique (M_R) est équivalent au modèle avec revenu catégorique (M_C)
Valeur observée de la statistique de test :
$w_{rv} = 2\{max\ln(L_C) - max\ln(L_R)\} = 2(-7.5759 - -15.6984) = 16.245$
Si H_0 est vraie, la statistique W_{rv} devrait suivre une χ^2_2 car M_C comporte 4 paramètres et M_R en comporte 5.
On a que $w_{rv} = 16.245 > \chi^2_{2,0.05} = 5.99$ donc on rejette H_0. Le modèle avec variable revenu numérique ne semble donc pas bien s'ajuster à ces données.

(c)
```
data etude;
input revenu$ consom$ freq @@;
if revenu="Faible" then srevenu=1;
if revenu="Moyen"  then srevenu=3;
if revenu="Eleve"  then srevenu=5;
if revenu="TEleve" then srevenu=6;
datalines;
Faible Oui 10 Faible Non 25
Moyen  Oui 15 Moyen  Non 20
Eleve  Oui 20 Eleve  Non 15
TEleve Oui  5 TEleve Non 30
;
run;
proc freq data=etude;
weight freq;
tables srevenu*consom / chisq;
run;
```

Le test sur le paramètre β dans le modèle avec revenu numérique est équivalent au test de Mantel et Haenszel pour variables ordinales dans un tableau de fréquences à deux variables :
Résultats du test sur β : Stat de Wald = 0.05 (suit une loi χ^2_1 sous l'hypothèse d'absence de lien entre les variables), seuil observé = 0.8184.
Résultats du test de Mantel et Haenszel : Stat $M^2 = 0.0524$ (suit une loi χ^2_1 sous l'hypothèse d'absence de lien entre les variables), seuil observé = 0.8190.

Le test de rapport de vraisemblances (différence de déviances) pour le lien entre le revenu catégorique et la variable réponse est équivalent au test d'indépendance se basant sur la statistique G^2 de rapport de vraisemblances dans un tableau de fréquences à deux variables :
Résultats du test avec le modèle : Stat = 16.30 (suit une loi $\chi^2_{(4-1)}$ sous l'hypothèse d'absence de lien entre les variables), seuil observé = 0.0010.
Résultats du test dans un tableau de fréquences : Stat $G^2 = 16.2977$ (suit une loi $\chi^2_{(4-1)}$ sous l'hypothèse d'absence de lien entre les variables), seuil observé = 0.0010.

7. `data kyphosis;`

```
input kyphosis age @@;
age2 = age**2;
datalines;
1 12 1 15 1 42 1 52 1 59 1 73 1 82 1 91 1 96 1 105
1 114 1 120 1 121 1 128 1 130 1 139 1 139 1 157
0 1 0 1 0 2 0 8 0 11 0 18 0 22 0 31 0 37 0 61 0 72 0 81 0 97 0 112
0 118 0 127 0 131 0 140 0 151 0 159 0 177 0 206
;
run;

proc genmod data=kyphosis descending;
model kyphosis = age / dist=bin link=logit;
output out=out1 pred=pred1;
run;
proc genmod data=kyphosis descending;
model kyphosis = age age2 / dist=bin link=logit;
output out=out2 pred=pred2;
run;

proc sort data=out1;
by kyphosis age;
proc sort data=out2;
by kyphosis age;
data graph;
merge out1 out2;
by kyphosis age;
run;
legend POSITION =(TOP RIGHT INSIDE) MODE = SHARE ACROSS=1
       label=none VALUE=("valeurs prédites modèle sans terme quadratique"
       "valeurs prédites modèle avec terme quadratique" "valeurs observées");
axis label=none;
proc gplot data=graph;
plot pred1*age="star" pred2*age="circle" kyphosis*age /
     overlay legend=legend vaxis=axis;
run; quit;
GOPTIONS reset=all;
```

(a) Stat de Wald = 0.54, seuil observé = $0.4627 > 0.05$ donc on conclut que l'age n'a pas d'effet.

(b) Stat de Wald = 4.40, seuil observé = $0.0360 < 0.05$ donc on conclut que l'age au carré a un effet. Ainsi, l'age a un effet quadratique et non linéaire sur le logit de la probabilité d'avoir une cyphose (traduction de kyphosis).

8.
```
data glm;
input ronfle attaque total;
cards;
0  24  1379
2  35   638
4  21   213
5  30   254
;
run;
proc genmod data=glm;
  model attaque/total = ronfle  / dist=bin  link=identity;
proc genmod data=glm;
```

```
  model attaque/total = ronfle  /  dist=bin  link=logit;
proc genmod data=glm;
  model attaque/total = ronfle  /  dist=bin  link=probit;
run;
```

	Test d'ajustement du modèle		
Fonction de lien	Déviance standardisée	ddl	seuil observé
identité	0.0692	2	0.9659917
logit	2.8089	2	0.245502
probit	1.8716	2	0.3922719

Ce test d'ajustement du modèle basé sur la déviance nous porte à croire que le modèle qui s'ajuste le mieux est celui avec le lien identité.

Étant donné que les données groupées ne comportent que 4 observations, comparons les valeurs prédites des 3 modèles :

Score de	Proportion	Probabilité prédite par le modèle		
ronflement	observée	linéaire	logistique	probit
0	0.017	0.017	0.021	0.020
2	0.055	0.057	0.044	0.046
4	0.099	0.096	0.093	0.095
5	0.118	0.116	0.132	0.131

On constate que les probabilités prédites avec la fonction de lien identité sont celles se rapprochant le plus des proportions observées. On conclut que le modèle avec lien identité présente le meilleur ajustement de l'association entre la risque d'attaque cardiaque et la fréquence de ronflement.

9. Une variable réponse ordinale Y = happy, à 3 modalités (1 = not, 2 = pretty, 3 = very)
   Une variable explicative numérique x = income
   (a) Si on dit que la modalité de référence est 3, qui signifie de se dire très heureux maritalement, le modèle logit généralisé est composé des deux équations suivantes :

$$\ln\left(\frac{P(Y=1|x)}{P(Y=3|x)}\right) = \beta_{01} + \beta_{11}x$$
$$\ln\left(\frac{P(Y=2|x)}{P(Y=3|x)}\right) = \beta_{02} + \beta_{12}x$$

   (b) Le modèle à rapports de cotes proportionnels est composé des deux équations suivantes :

$$logit(P(Y \leq 1|x)) = \beta_{01} + \beta_1 x$$
$$logit(P(Y \leq 2|x)) = \beta_{02} + \beta_1 x$$

   Remarquez que la pente β_1 est la même dans les deux modèles.

   (c)
```
data happy;
input happy income count;
datalines;
1 1 6
2 1 43
3 1 75
1 2 6
2 2 113
3 2 178
1 3 6
```

```
2 3 57
3 3 117
;
run;

* Modèle logit généralisé;
proc logistic data=happy;
freq count;
model happy = income / link=glogit;
run;

* Modèle à rapports de cotes proportionnels;
proc logistic data=happy;
freq count;
model happy = income / link=clogit;
run;
```

Notez que l'hypothèse de proportionnalité des rapports de cotes est acceptée.

(d) Le revenu n'a pas d'effet significatif sur Y. On arrive à cette conclusion avec les deux modèles.

(e) Modèle logit généralisé : $\hat{P}(Y = 1|x = 2) = \frac{exp(-2.555+-0.228\times 2)}{1+exp(-2.555+-0.228\times 2)+exp(-0.351+-0.096\times 2)} = 0.03024$.
Modèle à rapports de cotes proportionnels : $\hat{P}(Y = 1|x = 2) = \hat{P}(Y \leq 1|x = 2) - 0 = \frac{exp(-3.2467+-0.1117\times 2)}{1+exp(-3.2467+-0.1117\times 2)} = 0.03017$