Standardisation dans un modèle de Poisson : modélisation et conséquences sur les prédictions

Mickaël Schaeffer

Standardisation dans un modèle de régression de Poisson : modélisation et conséquences sur les prédictions.

SCHAEFFER Mickaël
mickael.schaeffer@etu.unistra.fr
Université de Strasbourg
Master 2 Statistiques

Année 2012

Tuteur professionnel :
Dr E.A. Sauleau
ea.sauleau@unistra.fr

Responsable pédagogique :
Pr. A. Guillou
armelle.guillou@math.unistra.fr

Référent universitaire :
Mr. N.Poulin
poulin@math.unistra.fr

Remerciements

En préambule de ce rapport, je souhaiterais remercier le Dr Erik-André Sauleau, mon superviseur professionnel, pour ses conseils, sa patience et sa disponibilité tout au long de ce stage. Je suis reconnaissant pour le temps qu'il m'a consacré tout au long de l'étude mais également pour son encadrement et sa participation au cheminement de ce rapport.

Je tiens également à remercier le Professeur Nicolas Meyer pour son accueil au sein du laboratoire de biostatistiques de l'hôpital civil de Strasbourg, et pour m'avoir ouvert les portes de la discipline en milieu hospitalier.

Personnellement, je tiens à leur adresser ma reconnaissance quant à la richesse de l'étude proposée, et à la confiance qui m'a été accordée.

Merci également à messieurs François Lefebvre et Nicolas Sananes que j'ai côtoyé, pour leur gentillesse et leur bonne humeur durant toute cette période.

Table des matières

Remerciements **3**

1 Introduction **1**
1.1 Présentation du laboratoire 1
1.2 Les indicateurs de risque en épidémiologie 1
1.3 Le S.I.R. 2
1.4 La population 2
1.5 Le calcul du risque 3
1.6 L'objectif de l'étude 3

2 Définition d'un offset **4**

Le modèle de Poisson avec Offset **6**

3 Modèle à deux variables **6**
3.1 Prédicteur à variables qualitatives 6
3.1.1 Le modèle 6
3.1.2 La simulation des observations 6
3.2 Les équations : les modèles 7
3.3 Ajustement et Résultats 8
3.3.1 Les coefficients des modèles 8
3.3.2 Test de surdispersion de Dean 9
3.3.3 Le critère d'Akaïke 10
3.3.4 L'écart des prédictions aux observations 11
3.4 Prédictions identiques 12
3.4.1 La simplification des facteurs ? 12
3.4.2 Les résultats de l'algorithme 13
3.5 Compensation des coefficients 13
3.6 Conclusions 18
3.7 Modification du prédicteur linéaire 19
3.8 Conclusion 20

4 Le modèle à trois variables **21**
4.1 L'ajout d'une variable temporelle 21
4.2 Méthodes de simulation 22
4.2.1 La simulation de la population 22
4.2.2 Simulation des observations 23
4.3 Ajustement et résultats 25

4.3.1 Les coefficients estimés . . . 26
4.3.2 Les critères AIC . . . 26
4.4 L'effet du temps . . . 26
4.5 Estimation des risques . . . 27
4.6 L'analyse des résidus . . . 29
4.6.1 Modèle sous population constante . . . 29
4.6.2 Modèle sous population linéairement croissante dans le temps . . . 30
4.7 Test de surpdispersion . . . 33
4.8 Les interactions . . . 33
4.9 Reflexion sur la notion de temps . . . 39

Discussion **41**

5 Annexes **43**
5.1 Le critère de la déviance . . . i
5.2 Standardisation Âge & Sex . . . ii
5.3 La méthode IRLS . . . iv
5.4 Vérifications, simulations . . . vi
5.4.1 Le modèle à deux variables . . . vi
5.4.2 L'effet "Balance" de l'information . . . vi
5.4.3 Modèle sans effet simple, avec offset . . . vi
5.4.4 Le modèle à trois variables . . . vii

1 Introduction

1.1 Présentation du laboratoire

Le présent rapport décrit le travail réalisé dans le cadre d'un stage de deuxième année de Master en Statistiques à l'université de Strasbourg et dans les locaux du laboratoire de biostatistiques et d'informatique médicale de la faculté de médecine de Strasbourg.

Le laboratoire est situé au coeur de l'Hôpital Civil de Strasbourg, principal et historique site des hôpitaux universitaires de Strasbourg, dirigé par le docteur Nicolas MEYER.(MCU-PH).

Sous la tutelle du Dr Erik-André Sauleau (MCU-PH), et durant une période de 6 mois, nous avons étudié un des modèles usuels visant à expliquer un indicateur de mortalité essentiel en épidémiologie. L'objectif de ce rapport est de résumer, de manière quasi-chronologique, le déroulement de ce stage, ainsi que les différents résultats, mathématiques, parfois surprenants et inattendus, rencontrés au cours de cet apprentissage.

Introduisons dès lors le sujet de notre étude.

1.2 Les indicateurs de risque en épidémiologie

L'épidémiologie est l'étude de la distribution des déterminants des problèmes de santé dans les populations humaines (Type maladie et autres états concernant la santé). Dans ce contexte peuvent être utilisés différents indicateurs (de l'état de santé, ou encore de l'exposition à une maladie) comme les risques, les taux, les ratio, les proportions etc.

Pour pouvoir comparer des taux, par exemple de type mortalité ou morbidité, il peut être nécessaire de neutraliser un facteur de confusion. Prenons par exemple le cas des taux d'incidence des cancers du poumon pour les deux catégories : hommes et femmes. Pour pouvoir comparer des proportions, il est impératif de prendre en compte l'âge des individus, c'est à dire prendre l'âge comme facteur de confusion.
La comparaison des taux bruts est impossible si la distribution des classes d'âge est différente dans les populations.

La standardisation est une méthode utilisée par les épidémiologistes[9] pour prendre en compte ce phénomène. Le principe est de corriger le déséquilibre entre les populations à comparer, en utilisant les taux spécifiques d'une population de référence[5].

Dans le prolongement d'une étude précédente sur la comparaison de différents modèles avec offsets (S.Columbu,M.Musio,E.A.Sauleau,2011)[3], nous nous intéresserons ici aux différentes

méthodes de standardisation, et plus particulièrement à un indicateur de risque fréquent en épidémiologie : le $S.I.R.$ Après en avoir donné la définition, nous étudierons certains modèles de régression statistique, dans le but d'analyser les facteurs explicatifs, les propriétés ainsi que les méthodes utilisées.

1.3 Le S.I.R.

Souvent utilisé pour l'étude des cancers, le $S.I.R.$ (Standardized Incidence Ratio) est un des indicateurs les plus employés pour quantifier l'incidence d'une maladie sur une population. Il est défini par la formule suivante :

$$S.I.R. = \frac{O}{E}$$

où O (O *pour "Observed"*) est le nombre de cas observés (déclarés) de la maladie et E (E *pour "Expected"*) le nombre de cas attendus.

O est une quantité observée, disponible dans les bases de données des centres de soin, ou des registres des cancers dans le cas des cancers. E est une quantité calculée, définie à partir d'un risque calculé comme le rapport du nombre de cas observés sur la taille de la population.
Le détail de la méthode de calcul de E est défini dans la section 1.4

1.4 La population

La population utilisée ici est un échantillon de la population totale, dont on connaît certaines données : par exemple la catégorie d'âge d'une personne ainsi que son sexe. On présente alors les classes de populations sous la forme suivante :

Sex	1	1	1	1	1	1	1	1	1
Age	1	2	3	4	5	6	7	8	9
N	30000	25000	24000	23589	17523	12542	9587	6500	6453

Sex	2	2	2	2	2	2	2	2	2
Age	1	2	3	4	5	6	7	8	9
N	32500	28579	25412	25613	18456	15236	11587	8459	7452

Notations
On note i l'indice représentant le sexe, $i \in [\![1 : I]\!]$, par exemple hommes et femmes, et j l'indice représentant la catégorie d'âge, $j \in [\![1 : J]\!]$, comme par exemple les classes d'âge 0-10 ans, 10-20 ans etc.. Dans notre exemple ici, $I = 2$ et $J = 9$.

1.5 Le calcul du risque

Le risque peut être calculé de trois façons :

Calcul d'un risque global

La première méthode de calcul du risque est dite "*globale*". Il s'agit de la méthode de calcul intuitive qui consiste à considérer l'ensemble de la population, pour connaitre le risque de contracter la maladie. On effectue le rapport sur toute la population, du nombre de cas observés sur le nombre de personnes de la population.
Une autre façon de calculer le risque à appliquer à une population est de dissocier le risque pour les hommes, du risque pour les femmes, de contracter une maladie. On obtient un risque par sexe. Enfin on peut dissocier les risques pour chaque catégorie d'âge de la population.
On obtient respectivement ;
$\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!]$:

$$p = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij}}{\sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}} \qquad p_i = \frac{\sum_{j=1}^{J} O_{ij}}{\sum_{j=1}^{J} N_{ij}} \qquad p_j = \frac{\sum_{i=1}^{I} O_{ij}}{\sum_{i=1}^{I} N_{ij}}$$

Risque Global *Risque Sexe* *Risque Âge*

Selon le risque appliqué à la population, nous obtenons trois méthodes de calcul du nombre de cas attendus E, que nous noterons respectivement $E.glob$, $E.sex$ et $E.age$.

En notant N_{ij} l'effectif de population de sexe i et de classe d'âge j, on a :

$$E.glob_{ij} = p \times N_{ij}$$

$$E.sex_{ij} = p_i \times N_{ij}$$

$$E.age_{ij} = p_j \times N_{ij}$$

1.6 L'objectif de l'étude

Définissons maintenant les objectifs de cette étude et quels sont les résultats attendus. Nous allons considérer une variable réponse (le ratio $S.I.R.$) en prenant en compte différentes variables dans notre modèle, comme l'âge, ou le sexe. Des variables explicatives supplémentaires pourraient être ajoutées au modèle, mais nous nous limiterons ici au cas simplifié des deux variables

citées. On étudiera alors les estimations, les prédictions, ainsi que les diagnostics classiques du modèle statistique.

Quels sont les impacts des différentes méthodes de calcul de E sur les coefficients de nos variables ? Peut-on, à l'aide des critères de sélection habituels, comparer les modèles ajustés ?

Nous simulerons des données représentant des cas observés d'une maladie donnée. La méthode de simulation sera détaillée dans le paragraphe 3.1.2. Le nombre de cas observés pourra soit être simulé avec un risque pour les hommes différent de celui des femmes, soit en utilisant un risque différent pour chaque classe d'âge. L'objectif est de pouvoir comparer les modèles contenant les offsets $E.glob$, $E.age$, et $E.sex$ et d'étudier plus précisémment l'impact de cette variable sur l'estimation des coefficients, en fonction de la méthode de simulation.

On utilisera un modèle de poisson dans le cadre de la régression linéaire généralisée appliquée à la variable $S.I.R.$. Le modèle de poisson a la particularité d'être, entre autre, utilisé dans le cas où la variable à expliquer est un ratio de données de comptage[4].
En effet les propriétés de ce modèle permettent de simplifier le problème du ratio en un problème d"offset" à introduire dans le prédicteur linéaire. Cette particularité est détaillée dans le paragraphe ci-dessous.

2 Définition d'un offset

Dans certains cas, comme celui de la régression log-linéaire ici étudiée, la fonction de lien *logarithme* appliquée à la variable réponse permet de simplifier l'ajustement sur le ratio en un problème d'ajustement sur le numérateur.

En considérant une variable réponse sous la forme $\frac{A}{B} > 0$, on a la relation suivante :

$$\log\left(\frac{A}{B}\right) = \beta_0 + \sum_{i=1}^{n} X_i \times \beta_i$$

$$\Leftrightarrow \log(A) - \log(B) = \beta_0 + \sum_{i=1}^{n} X_i \times \beta_i$$

$$\Leftrightarrow \log(A) = \beta_0 + \sum_{i=1}^{n} X_i \times \beta_i + \log(B)$$

$$\Leftrightarrow A = \exp\left(\beta_0 + \sum_{i=1}^{n} X_i \times \beta_i + \log(B)\right)$$

$$\Leftrightarrow A = B \times \exp\left(\beta_0 + \sum_{i=1}^{n} X_i \times \beta_i\right)$$

Ajuster un modèle de poisson sur une variable réponse sous forme de ratio, revient à ajuster un même modèle de poisson sur le numérateur, en incluant le dénominateur dans le prédicteur linéaire $\log(B)$. (Ou encore en incluant le dénominateur comme facteur multiplicatif de l'exponentielle, dans le prédicteur linéaire, comme le montre la dernière équation.)

Dans toute la suite du rapport, i se rapporte aux modalités de la variable X_1 ($Sexe$) et j aux modalités de la variable X_2 ($\hat{A}ge$). Ainsi, l'observation O_{ij} correspond aux nombres de cas observés pour la *i-ème* modalité de X_1 et la *j-ème* modalités de X_2 : Dans notre cas, les groupes sont Sex et $\hat{A}ge$: l'observation O_{23} correspond par exemple aux nombres de cas observés du deuxième groupe sexe (ici les femmes), et de la troisième classe d'âge.
On a alors $\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!]$,

$$\log\left(\frac{O_{ij}}{E_{ij}}\right) = \beta_0 + \beta_{1i} + \beta_{2j}$$

$$\Leftrightarrow \log\left(\frac{O_{ij}}{E_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}$$

$$\Leftrightarrow \log(O_{ij}) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \log(E_{ij})$$

$$\Leftrightarrow O_{ij} = \exp\left(\beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \log(E_{ij})\right)$$

$$\Leftrightarrow O_{ij} = E_{ij} \times \exp\left(\beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}\right)$$

On peut également noter, au travers de ces équations que :

$$SIR_{ij} = \exp\left(\beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}\right)$$

Etudions l'impact de cet offset sur les prédictions et quelles sont les conséquences du choix de la méthode de calcul sur les coefficients du modèle.

Le modèle de Poisson avec Offset

3 Modèle à deux variables

3.1 Prédicteur à variables qualitatives

3.1.1 Le modèle

Comme décrit précédemment, deux variables explicatives qualitatives seront utilisées dans un premier temps : *Age* et *Sexe* à respectivement 9 et 2 modalités. On veut expliquer un nombre de cas observés en prenant un compte un offset que nous noterons E.

3.1.2 La simulation des observations

Dans le but de pouvoir maîtriser les effets, et pouvoir évaluer la qualité d'un ajustement, nous choisissons de simuler nos données selon le schéma suivant : le nombre de cas observés noté O_{ij} pour une combinaison $X_1 = i, X_2 = j$ est simulé selon une loi de poisson de moyenne un risque fixé appliqué à une population, c'est à dire
$\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!]$:

$$\mu_{ij} = Risque \times N_{ij}$$

$$O_{ij} \sim> \mathcal{P}(\mu_{ij})$$

$$\iff O_{ij} \sim> \mathcal{P}(Risque \times N_{ij})$$

Les tailles de population chez les hommes et les femmes ont été simulé selon un modèle réel de *pyramide des âges* ; c'est à dire une décroissance des effectifs de population en fonction de l'âge.
On se fixe alors des risques pour chaque classe d'âge. A titre d'exemple, nous utilisons ici également une répartition des risques proche de celle connue des épidémiologistes, plus faible pour des classes plus jeunes, et croissant en fonction de l'âge. C'est-à-dire :

Risque	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8	r_9
Valeur	0.025	0.02	0.03	0.057	0.08	0.1	0.12	0.0115	0.105

Les observations obtenues se représentent graphiquement sous la forme suivante :

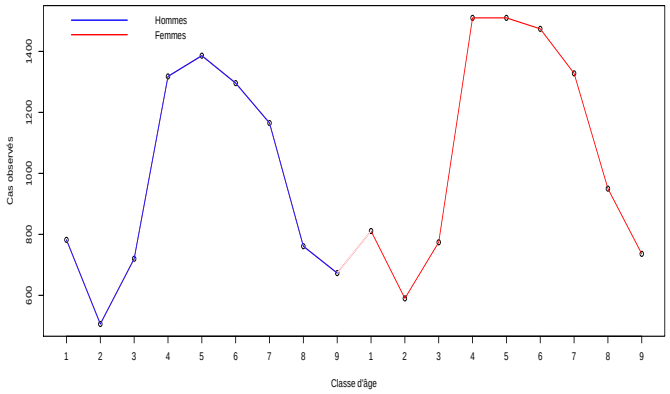


FIGURE 1: Observation la première année

3.2 Les équations : les modèles

En notant $\mathbb{X}_{ij}$ la matrice des variables explicatives, η_{ij} le prédicteur linéaire contentant une partie commune β_0, un effet $Sexe$ et un effet $\hat{A}ge$, i.e $\eta_{ij} = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}$ on a les hypothèses suivantes :

$$\begin{cases} \mathcal{L}(O_{ij}) = \mathcal{P}(\lambda_{ij}) \\ log(\lambda_{ij}) = \eta_{ij} + \log(E_{ij}) \\ O_{ij} \amalg X_{1i} \amalg X_{2j} \qquad \text{indépendance des variables} \end{cases}$$

L'offset (le nombre de cas attendus) peut etre calculé, comme nous l'avons montré au paragraphe 1.4, de plusieurs façons : en standardisant par rapport à l'âge, au sexe, ou à aucune des deux variables.
En utilisant respectivement chacun de ces trois offsets, pour définir trois modèles différents, on obtient les équations suivantes :

$$\log\left(\frac{O_{ij}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} \tag{1}$$

$$\log\left(\frac{O_{ij}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} \tag{2}$$

$$\log\left(\frac{O_{ij}}{E.global_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} \tag{3}$$

Etant donné la méthode de simulation choisie, on s'attend à observer une différence, au sens du critère d'Akaike (défini en 3.3.3), entre le modèle (1) et les modèles (2) et (3). En effet, l'écart entre chaque classe d'âge, en plus de la variable X_2 est pris en compte par l'offset $E.age$

Effectuons alors les diagnostics classiques sur ces trois modèles pour étudier l'impact de cette variable.

3.3 Ajustement et Résultats

3.3.1 Les coefficients des modèles

Le modèle *global*

En notant $x_{ab} = \beta_{ab}$ dans les résultats ci-dessous, les valeurs des coefficients estimés pour le modèle dit *Global* sont :

$$\log\left(\frac{O_{ij}}{E.global_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}$$

	coef.glob	2.5 %	97.5 %
(Intercept)	0.59487761	0.55626938	0.63301212
x12	0.03953723	-0.01607812	0.09508244
x13	0.05948663	0.00341213	0.11547416
x14	-0.01691108	-0.07443512	0.04047211
x15	0.01543243	-0.04697784	0.07751045
x16	0.03228866	-0.03677095	0.10073166
x17	0.09629130	0.02185716	0.16985869
x18	0.04770770	-0.04042854	0.13425910
x19	0.01613581	-0.07402035	0.10459103
x110	0.01850755	-0.07591747	0.11097915
x111	0.04610905	-0.06145331	0.15087118
x112	0.02484098	-0.15283424	0.19350719
x21	-2.08733978	-2.13611118	-2.03915929

FIGURE 2: Coefficients estimés pour le modèle *global* et intervalle de confiance

Le modèle *Âge*

Les valeurs des coefficients (estimées) pour le modèle dit *Age* ci-dessous sont les suivantes :

$$\log\left(\frac{O_{ij}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}$$

	coef.age	2.5 %	97.5 %
(Intercept)	0.6079082062	0.56929997	0.64604272
x12	0.0217927553	-0.03382260	0.07733796
x13	-0.0091475607	-0.06522206	0.04683997
x14	0.0009146608	-0.05660939	0.05829785
x15	-0.0112530418	-0.07366331	0.05082498
x16	0.0469216351	-0.02213798	0.11536463
x17	0.0447928671	-0.02964126	0.11836026
x18	0.0759823313	-0.01215390	0.16253374
x19	0.0259499454	-0.06420621	0.11440516
x110	-0.0478015485	-0.14222657	0.04467004
x111	-0.0321296657	-0.13969202	0.07263247
x112	0.4309502259	0.25327500	0.59961643
x21	-2.0873397795	-2.13611118	-2.03915929

FIGURE 3: Coefficients estimés pour le modèle $\hat{Age}$ et intervalle de confiance

Le modèle *Sex*

Les valeurs des coefficients (estimées) pour le modèle dit *Sex*, ci-dessous sont les suivantes :

$$\log\left(\frac{O_{ij}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}$$

	coef.sex	2.5 %	97.5 %
(Intercept)	-0.0266613983	-0.06526963	0.01147311
x12	0.0395372323	-0.01607812	0.09508244
x13	0.0594866292	0.00341213	0.11547416
x14	-0.0169110783	-0.07443512	0.04047211
x15	0.0154324321	-0.04697784	0.07751045
x16	0.0322886635	-0.03677095	0.10073166
x17	0.0962912953	0.02185716	0.16985869
x18	0.0477076952	-0.04042854	0.13425910
x19	0.0161358129	-0.07402035	0.10459103
x110	0.0185075542	-0.07591747	0.11097915
x111	0.0461090470	-0.06145331	0.15087118
x112	0.0248409796	-0.15283424	0.19350719
x21	-0.0006393995	-0.04941080	0.04754109

FIGURE 4: Coefficients estimés pour le modèle *Sex* et intervalle de confiance

3.3.2 Test de surdispersion de Dean

Une des propriétés fondamentales du modèle de poisson est l'égalité de la moyenne et de la variance. Il est nécéssaire de tester cette hypothèse avant d'interpréter les résultats de nos modèles. Le test de Dean[2] permet de tester cette égalité.

On obtient, pour les trois modèles, les p-valeurs suivantes au test de surdispersion :

	modèle Global	modèle Sex	modèle Âge
Statistique	-3.59	-3.59	-3.59
p-valeur	0.99	0.99	0.99

Il est important de noter que la valeur de la statistique de test est identique dans les trois modèles, ainsi que la p-valeur. Ces p-valeurs étant toutes supérieures à 0.05, il nous est impossible ici de mettre en avant une différence significative (au risque de 5%) entre la moyenne et la variance du modèle.

Il n'y a alors pas de problème de surdispersion dans ces trois modèles.

3.3.3 Le critère d'Akaïke

Pour pouvoir comparer ces modèles, il est nécéssaire d'utiliser un critère de compraison. Le critère d'Akaïke (*AIC* pour Akaïke Information Criterion[1]) est l'un des plus utilisé dans ce contexte. Voyons alors s'il est possible de départager ces modèles selon ce critère.

AIC modèle Global	AIC modèle Sex	AIC modèle Âge
219.277	219.277	219.277

FIGURE 5: *Critères d'Akaïke pour les trois modèles*

Il est ici impossible de départager ces trois modèles en utilisant ce critère. En effet, ces trois modèles ont la même valeur du critère d'Akaïke.

D'après la définition de la déviance d'un modèle statistique[8], en notant D la déviance du modèle, $\mathcal{L}(y, y)$ la vraisemblance du modèle saturé, $\mathcal{L}(y, \mu)$ la vraisemblance du modèle, on a :

$$\mathcal{L}(y, \mu) = \mathcal{L}(y, y) - \frac{D}{2}$$

D'après la définition de l'AIC, en notant k le nombre de paramètres à estimer dans le modèle, et par substitution de l'équation ci-dessus, on a :

$$AIC = D - 2\mathcal{L}(y, y) + 2k$$

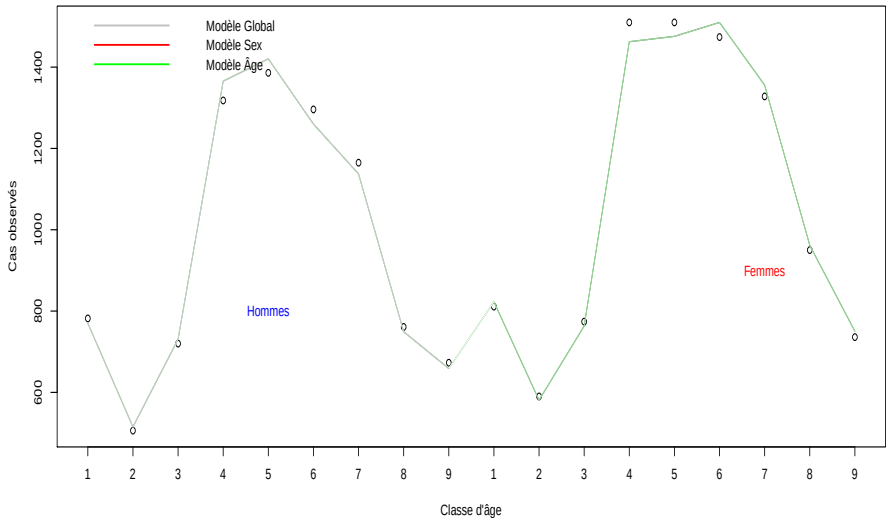
L'AIC et la déviance diffèrent donc d'une constante : $2\mathcal{L}(y, y) + 2k$. Ainsi, l'égalité des AIC dans les modèles implique donc une égalité des déviances, ici égales à 11.19. Il est ainsi impossible d'utiliser la déviance pour comparer ces modèles.

3.3.4 L'écart des prédictions aux observations

Regardons maintenant quelles sont les prédictions de nos trois modèles, et quelles sont celles qui sont les plus cohérentes avec les observations. A l'aide du logiciel `R`, nous obtenons les observations et les prédictions suivantes pour les trois modèles :

N°	1	2	3	4	5	6	7	8	9
O	782	506	720	1318	13860	1296	1165	761	673
Pred	770	515	731	1365	1420	1260	1137	749	659

N°	10	11	12	13	14	15	16	17	18
O	811	590	774	1510	1510	1474	1328	950	736
Pred	823	581	763	1463	1476	1510	1356	962	750



Les courbes sont indiscernables. Ceci est dû au fait que les prédictions sont identiques dans les 3 modèles. Par conséquent les critère d'Akaïke, donc la déviance, statistique de surdispersion, et résidus coïncident également.

3.4 Prédictions identiques

Notre conjecture de départ semble ici ne pas pouvoir être vérifiée.
En effet, aucun des 3 modèles ne se différencie des autres, ni par un critère de comparaison, ni par la qualité de ses prédictions.

Comment expliquer l'égalité des ces prédictions, alors que les coefficients estimés sont différents pour chaque modèle ?

Deux possibilités sont ici envisagées.

• Une première hypothèse est la simplifaction d'une variable ou d'un facteur, en raison des méthodes de calcul de E. En effet, étant donné le fait que les coefficients soient différents pour les trois modèles, il est raisonnable de penser que deux facteurs puissent s'annuler entre eux lors du calcul des prédictions. Celles-ci seraient alors rendues indépendantes de l'offset, ce qui expliquerait les égalités observées précédement.

• Une autre possibilité est la redondance de facteurs explicatifs. En effet, la présence de l'offset permet également de prendre en compte une différence entre des classes données, et d'expliquer partiellement un effet. Le rôle joué est proche de celui d'une variable. Nous étudierons alors les coefficients estimés par le modèle.

3.4.1 La simplification des facteurs ?

Pour répondre à cette question, on peut dans un premier temps s'interroger sur la nature de nos données.

Pour vérifier cette hypothèse *de simplification de facteurs*, nous allons nous intéresser à la méthode employée par le logiciel `R` au travers de sa fonction `glm`.

La méthode d'ajustement implémentée par défaut dans le logiciel `R` est la méthode des moindres carrés repondérés : la méthode IRLS *Iteratively Reweighted Least Squares*, ici appliquée au modèle de poisson avec offset. Nous rappelons brièvement les étapes de cet algorithme en annexe 5.3 : *La méthode IRLS*
Cet algorithme à été ré-implémenté manuellement, en prenant en compte la présence de l'offset (*cf Annexes*) dans le but d'évaluer la valeur d'une prédiction donnée, itération par itération, *à l'intérieur de la fonction* `glm`.

3.4.2 Les résultats de l'algorithme

Sous l'hypothèses de simplifcation de deux variables dans le prédicteur linéaire, les valeurs observées dans l'algorithme d'estimation devraient être égales à partir d'une itération donnée. C'est dans le but de vérifier cette hypothèse que nous définissions les étapes de l'algorithme du Fisher-Scoring en annexe 5.3.
Après avoir recodé la fonction `glm` en y introduisant un offset, on stocke les valeur des prédictions, à chaque itération.

Si le phénomène décrit précédemment devait se vérifier, dès les premières itérations, les prédictions devraient être identiques dans les différents modèles.
Le résultat obtenu, pour une prédiction donnée, est le suivant :

	pred.age	pred.sex	pred.glob
[1,]	438107.3988	100624.9437	436056.3890
[2,]	160869.1405	36833.7175	160125.2974
[3,]	58885.4024	13380.5271	58622.2617
[4,]	21385.5265	4787.8288	21298.7433
[5,]	7636.9246	1699.5950	7613.7801
[6,]	2677.2723	660.5950	2675.0891
[7,]	977.6499	344.1256	980.2476
[8,]	438.7451	263.3093	440.9548
[9,]	284.2464	253.1668	285.1127
10,]	254.5634	252.9709	254.6600
11,]	252.9757	252.9708	252.9764
12,]	252.9708	0.0000	252.9708
13,]	252.9708	0.0000	252.9708
14,]	0.0000	0.0000	0.0000
15,]	0.0000	0.0000	0.0000
16,]	0.0000	0.0000	0.0000
17,]	0.0000	0.0000	0.0000
18,]	0.0000	0.0000	0.0000
19,]	0.0000	0.0000	0.0000
20,]	0.0000	0.0000	0.0000

FIGURE 6: Résultats d'une prédiction, itération par itération, pour chaque modèle

L'hypothèse de simplification ne se vérifie pas, les prédictions sont bien différentes pour chaque itération, cependant les trois modèles convergent vers la même limite.

L'explication de ce phénomène ne réside alors pas dans la simplification du modèle.

3.5 Compensation des coefficients

Comment expliquer l'égalité entre ces différentes prédictions ?

Cette égalité peut s'expliquer par le fait que les variables utilisées dans le prédicteur sont des variables qualitatives, à plusieurs modalités, et dont l'une d'entre elle est utilisée pour la standardisation.

En effet, un des effets à expliquer l'est par deux variables. Prenons par exemple notre population, une standardisation sur le sexe, nous avons alors le modèle :$\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!]$:

$$\log\left(\frac{O_{ij}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} \tag{4}$$

$$\log(O_{ij}) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \log(E.sex_{ij}) \tag{5}$$

Dans le prédicteur sont alors présents :
- La constante, commune à toutes les prédictions β_0
- Un vecteur de coefficients relatifs à l'effet $Sexe$, $\beta_1 = (\beta_{11}, \beta_{12})$
- Un vecteur de coefficients relatifs à l'effet $\hat{A}ge$, $\beta_2 = (\beta_{21}, ..., \beta_{29})$
- L'offset relatif à l'effet $Sexe$, $\log(E.sex)$

L'effet Sexe est ainsi expliqué par deux variables : l'offset $\log(E.sex)$ et le facteur β_1.
Plus précisémment, comme

$$log(E.sex_{ij}) = log(R.sex_i \times N_{ij}) = log(R.sex_i) + log(N_{ij})$$

on a :

$$\log(O_{ij}) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \log(R.sex_i) + \log(N_{ij})$$

La partie explicative propre à la distinction des groupes Hommes et Femmes ($i = 1$ ou $i = 2$) est divisée en deux ; d'une part la partie expliquée par la partie de l'offset $log(R.sex_i)$, d'autre part le coefficient β_{1i}, estimé par la méthode des moindres carrés, visant à expliquer "l'information restante".
Les parties explicatives sont les suivantes :$\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!]$:

$$\log(O_{ij}) = Partie\ commune + Effet\ Sexe\ + Effet\ \hat{A}ge\ + \log(N_{ij})$$

$$\log(O_{ij}) = (\beta_0) + \left(\sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \log(R.sex_i)\right) + \left(\sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j}\right) + \log(N_{ij}) \tag{6}$$

Pour mieux comprendre le phénomène présenté précédemment, simplifions l'écriture du modèle, en considérant deux variables qualitatives A et B, à respectivement I et J modalités, ainsi qu'une variable réponse quantitative O.

Le modèle sans offset

La prédiction donnée par un ajustement sans offset est fournie par l'équation suivante :
$\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!], A_1 = 0, B_1 = 0$:

$$\widehat{O_{ij}} = U + A_i + B_j \tag{7}$$

U est la partie explicative commune à tous les individus, appelée *intercept du modèle*.
Les contraintes d'identifiabilité du modèle imposent la valeur nulle aux premiers coefficients de chaque variable, permettant alors l'interprétation des autres valeurs, comme la différence expliquée à cette valeur de référence.

Un offset de type constant

Introduisons dans un premier temps, un offset de type constant, noté C.
C est interprétée par le modèle comme une partie commune à tous les individus, au même titre que U. Le cumul de ces deux valeurs représente donc la valeur commune à tous les individus.
On a :

Partie commune **à expliquer**	=	Partie commune **imposée**	+	Partie commune **restante**

C'est-à-dire, avec les notations utilisées précédemment, et en notant μ la partie commune à expliquer restante :

$$U = C + \mu \tag{8}$$

D'où les prédictions :$\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!], A_1 = 0, B_1 = 0$:

$$\widehat{O_{ij}} = \mu + A_i + B_j + C \tag{9}$$

Les prédictions par le modèle sans offset et le modèle contenant un offset de type constant sont ainsi égales.

Un offset à I modalités

Introduisons un offset à I modalités notées $(C_1, C_2, ..., C_I)$.
L'observation O_{11} étant considéré comme la valeur de référence par rapport aux autres individus, on a, de la même manière que précédemment, et avec les mêmes notations :

$$U = C_1 + \mu \tag{10}$$

D'autre part,considérant l'offset comme une information *imposée pour chaque groupe i*, on a :

Différence groupes 1 et i **à expliquer**	=	Différence groupes 1 et i **imposée**	+	Différence groupes 1 et i **restante**

La différence entre les groupes 1 et i imposée par l'offset est donnée par $C_i - C_1$, et on a :

$$A_i = C_i - C_1 + \alpha_i \tag{11}$$

En utilisant les notations précédentes, les prédictions fournies par le modèle sont données par :
$\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!], A_1 = 0, B_1 = 0$:

$$\widehat{O_{ij}} = \mu + \alpha_i + B_j + C_i \tag{12}$$

L'égalité des prédictions entre modèles sans offset et avec offset à I modalités
Montrons alors que, dans ce contexte, les prédictions sont identiques pour tous les modèles. Les prédictions fournies par le modèle contenant un offset à i modalités $(C_1, C_2, ..., C_I)$ sont

$$\widehat{O_{ij}} = \mu + \alpha_i + B_j + C_i$$

Or d'après les relations précédentes :

$$\begin{aligned}
\widehat{O_{ij}} &= \mu + \alpha_i + B_j + C_i \\
&= \mu + (\alpha_i + C_i) + B_j \\
&= \mu + (A_i + C_1) + B_j && \text{d'après(12)} \\
&= (\mu + C_1) + A_i + B_j \\
&= U + A_i + B_j && \text{d'après(11)}
\end{aligned}$$

Ainsi, pour tous les individus, on retrouve bien les prédictions données par le modèle sans offset.

Exemple numérique : prédiction $i = 1, j = 3$ et prédiction $i = 2, j = 3$
Pour fixer les idées, prenons par exemple des valeurs numériques, pour une prédiction donnée : Comparons la prédiction sans offset (i.e la valeur estimée des coefficients) aux modèles avec offset *global* et offset *Sex*

Modèle sans offset

Les prédictions pour les nombres de cas observés pour les catégories $i = 1, j = 3$ $i = 2, j = 3$ est donnée par :

$$\log(\hat{O_{13}}) = \hat{\beta_0} + \hat{\beta_{23}} \qquad \log(\hat{O_{13}}) = 4.79 + 0.66$$
$$\log(\hat{O_{23}}) = \hat{\beta_0} + \hat{\beta_{12}} + \hat{\beta_{23}} \qquad \log(\hat{O_{23}}) = 4.79 + (-1.73) + 0.66$$

Offset Constant 4

Introduisons maintenant un offset constant fixé à 4, identique à tous les individus.
D'après le raisonnement effectué ci-dessus, la somme de l'offset constant et de l'intercept doivent être équivalents à l'estimation de l'intercept dans le modèle sans offset (7). La partie explicative commune à tous les individus se divise en deux parties :

$$\log(\hat{O_{13}}) = \hat{\beta_0} + \hat{\beta_{23}} + \mathbf{4} \qquad \log(\hat{O_{13}}) = \mathbf{0.79} + 0.66 + \mathbf{4}$$

$$\log(\hat{O_{23}}) = \hat{\beta_0} + \hat{\beta_{12}} + \hat{\beta_{23}} + \mathbf{4} \qquad \log(\hat{O_{23}}) = \mathbf{0.79} + (-1.73) + 0.66 + \mathbf{4}$$

$Hommes$: **4+0.79+0.66=4.79** $+ 0.66 = 5.45$
$Femmes$: **4+(-1.73)+0.79+0.66=4.79** $+ (-1.73) + 0.66 = 3.72$

Offset Sex (4,8)

On introduit maintenant un offset *Sex*. Nous introduisons un offset de 4 pour les hommes et 8 pour les femmes. Ceci s'interprète donc comme une valeur commune de 4 pour les hommes et les femmes, et une différence supplémentaire de 4 pour les femmes.
Vérifions alors que la somme de 4 et de l'intercept correspond à l'estimation de l'intercept dans le modèle sans offset (8) et que la somme de l'effet *Sex* et de 4 correspond à la somme de l'effet *Sex* dans le modèle (7).

$$\log(\hat{O_{13}}) = \hat{\beta_0} + \hat{\beta_{23}} + \mathbf{4} \qquad \log(\hat{O_{13}}) = \mathbf{0.79} + 0.66 + \mathbf{4}$$

$$\log(\hat{O_{23}}) = \hat{\beta_0} + \hat{\beta_{12}} + \hat{\beta_{23}} + \mathbf{8} \qquad \log(\hat{O_{23}}) = \mathbf{0.79} + (-5.73) + 0.66 + \mathbf{8}$$

$Hommes$: $4 + 0.79 + 0.66 = 4.79 + 0.66 = 5.45$
$Femmes$: $0.79 +$ **(-5.73)** $+ 0.66 +$ **8** $= 0.79 +$ **(-1.73)** $+ 0.66 +$ **4** $= 3.72$

On retrouve bien les mêmes prédictions, indépendemment de l'offset introduit.

3.6 Conclusions

L'introduction d'un offset dans le prédicteur linéaire, calculé comme le produit d'un risque standardisé sur une variable, et de la taille de la population, ne présente à priori pas d'apport informatif lors de l'ajustement du modèle de Poisson (avec un lien logarithme).

En effet, la comparaison d'un modèle ne contenant que les effets simples de type qualitatifs, et d'un modèle contenant un offset de type $E = R \times N$ aboutissent aux mêmes conclusions : des prédictions identiques, des déviances identiques, et des critères de comparaison (de type AIC) identiques.

Dans le but de prendre en compte la taille de la population, il est d'usage en épidémiologie de ramener le nombre de cas observés au nombre de cas attendus pour une classe de population donnée, comme nous l'avons présenté lors de la définition du $S.I.R$.

Nous identifions ici un phénomène que l'on pourrait qualifier de *balance* entre les parties à expliquer, lors de l'introduction d'un offset. Comme nous l'avons noté précédemment, cette valeur se décompose en deux parties (un risque R et une taille de population N). La partie *Risque* joue alors le rôle de complémentaire de l'effet simple explicatif, n'offrant alors pas d'apport informatif.

Il y a donc une équivalence, au sens propre du terme, entre la prise en compte dans le modèle de l'offset usuel de type $E = R \times N$, et la simple prise en compte dans le modèle de la taille de la population en offset. (Au logarithme près).

On a : $\forall i \in [\![1 : I]\!], \forall j \in [\![1 : J]\!]$:

$$\log\left(\frac{\widehat{O_{ij}}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma} \mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta} \mathbf{1}_{\delta=j}$$

$$\Longleftrightarrow$$

$$\log(\widehat{O_{ij}}) = \psi_0 + \sum_{\gamma=1}^{I} \psi_{1\gamma} \mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \psi_{2\delta} \mathbf{1}_{\delta=j} + \log(N_{ij})$$

3.7 Modification du prédicteur linéaire

Pour vérifier les résultats précédents, et s'assurer de la non complémentarité des deux facteurs explicatifs, on se propose d'ajuster, sur ce même jeu de données, les modèles dont la variable explicative utilisée pour la standardisation est absente du prédicteur linéaire, c'est à dire les modèles suivants :

$$log\left(\frac{O_{ij}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} \tag{13}$$

$$log\left(\frac{O_{ij}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} \tag{14}$$

$$log\left(\frac{O_{ij}}{E.global_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} \tag{15}$$

Les résultats obtenus après ajustement sont les suivants :

Critère d'Aikake

Comparons, après ajustement, les critères d'Aikake pour les 3 modèles :

AIC modèle Global	AIC modèle Âge	AIC modèle Sex
207.6308	191.5866	207.5006

FIGURE 7: *Critères d'Akaike pour les trois modèles, sans répétition de la covariable*

On remarque alors que les trois critères sont bien différents, et que le meilleur des modèles, celui dont le critère d'Akaike est minimal (avec une différence d'au moins un point) est le modèle avec un offset Âge, en l'occurence celui dont l'offset est standardisé sur la même variable que dans les simulations.

Coefficients

Les coefficients sont différents pour les trois modèles :

Les trois modèles sont bien distincts les uns des autres, les estimations de leurs coefficients respectifs également, ainsi que leurs prédictions. Le meilleur modèle est bien celui dont la méthode de standardisation est identique à la méthode de simulation.

.	**Modèle Global**	2.5%	95%	**Modèle Age**	2.5%	95%	**Modèle Sex**	2.5%	95%
Int	-1.57	-1.66	-1.49	0.07	0.05	0.09	-1.65	-1.73	-1.57
Sex1	-0.14	-0.17	-0.11	-0.14	-0.17	-0.11	X	X	X
Age2	0.67	0.57	0.77	X	X	X	0.67	0.57	0.77
Age3	1.09	0.99	1.18	X	X	X	1.09	0.99	1.18
Age4	1.68	1.60	1.77	X	X	X	1.68	1.60	1.77
Age5	2.10	2.01	2.18	X	X	X	2.10	2.01	2.18
Age6	2.12	2.03	2.21	X	X	X	2.12	2.03	2.21
Age7	2.40	2.32	2.49	X	X	X	2.40	2.31	2.49
Age8	2.46	2.37	2.55	X	X	X	2.46	2.36	2.55
Age9	2.65	2.56	2.74	X	X	X	2.65	2.56	2.74

N°	1	2	3	4	5	6	7	8	9
O	294	513	716	1314	1437	1168	1139	879	945
Glob	315	514	751	1340	1504	1101	1119	800	962
Age	315	514	753	1342	1508	1099	1116	796	961
Sex	311	508	742	1325	1487	1088	1105	790	951

N°	10	11	12	13	14	15	16	17	18
O	316	510	723	1286	1438	1092	1150	822	978
Glob	295	509	688	1260	1371	1159	1170	901	961
Age	296	509	690	1262	1375	1156	1168	897	961
Sex	299	515	967	1275	1388	1172	1184	911	972

3.8 Conclusion

Dans un cadre épidémiologique, où la variable d'intéret est un ratio du type S.I.R., dans le but de prendre en compte un offset, il est nécessaire de soit :

• Retirer du prédicteur la variable ayant servi à standardiser, pour éviter les risques de compensation, et l'égalité des prédictions (si l'intérêt de l'utilisateur est la sélection d'un modèle.)

• N'introduire dans le préditeur linéaire que la taille de la population en tant qu'offset, comme nous l'avons décrit au paragraphe 3.6. (si l'intérêt de l'utilisateur est de prendre en compte la taille de la population sans souci de choix d'offsets.)

4 Le modèle à trois variables

4.1 L'ajout d'une variable temporelle

L'étude des taux d'incidence dans un cadre épidémiologique, nécéssite la connaissance d'un certain nombre de variables sur les individus.
Dans un contexte optimal, et théorique, la base de données se présente comme complète, contenant ainsi un maximum d'informations comme l'âge, le sexe, la date du relevé ou du diagnostic, l'activité professionnelle, le lieu de vie (donnée spatiale), le fait d'être fumeur (en particulier en cancérologie), etc..
Cependant, ces données sont très rarement entièrement connues des registres hôspitaliers, ou des registres des cancers.
Les variables les plus fréquemment utilisées pour modéliser les taux, ou encore les ratios, par les épidémiologistes sont généralement le nombre de cas observés, la taille de la population en fonction de l'âge, du sexe, et ce sur plusieurs années (ou par trimestre, semestre).

Nous avons vu précédémment le modèle comportant les variables qualitatives Âge et Sexe. Pour mieux modéliser une situation concrète d'étude de ratio, en fonction des données disponibles par les registres, ajoutons une variable temporelle explicative à notre modèle.

Généralement, l'étendue de la variable temporelle est de l'ordre de la dizaine d'années. Il est donc, selon la nature de cette covariable, possible de la considérer comme qualitative avec autant de modalités que de périodes relevées, ou quantitative continue à compter d'une date initiale.
En prenant en compte l'effet précédent, nous obtenons les trois modèles avec une variable temps qualitative et les trois modèles avec une variable temps quantitative suivants :
Notations : Dans toute la suite du rapport, on prendra 9 catégories d'âge ($J = 9$), 2 catégories pour le Sexe ($I = 2$), et si le temps est considéré comme variable qualitative $t \in [\![1 : T]\!]$ où $T = 5$.

Le temps comme variable qualitative :

$$log\left(\frac{O_{ij}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\tau=1}^{T} \beta_{3\tau}\mathbf{1}_{\tau=j} \tag{16}$$

$$log\left(\frac{O_{ij}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \sum_{\tau=1}^{T} \beta_{3\tau}\mathbf{1}_{\tau=j} \tag{17}$$

$$log\left(\frac{O_{ij}}{E.global_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \sum_{\tau=1}^{T} \beta_{3\tau}\mathbf{1}_{\tau=j} \tag{18}$$

Le temps comme variable quantitative :

$$log\left(\frac{O_{ij}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \beta_{Temps} \times Temps \tag{19}$$

$$log\left(\frac{O_{ij}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps \tag{20}$$

$$log\left(\frac{O_{ij}}{E.global_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps \tag{21}$$

L'étude sur le modèle à deux variables qualitatives ayant permis de mieux cerner l'impact de l'offset sur ce type de prédicteur, on s'intéressera plutôt aux modèles dont la variable Temps est considérée comme continue.
Détaillons alors la méthode de simulation des données.

4.2 Méthodes de simulation

4.2.1 La simulation de la population

Supposons que l'on ait, pour la première année, 9 catégories d'âge différents. On a alors 18 classes de populations, 9 catégories d'âge *Hommes* et 9 catégories d'âge *Femmes*.
On suppose maintenant que ces observations sont les observations de la première année. On cherche à simuler une taille de population pour les années suivantes.

Nous utilisons l'hypothèse que la taille d'une population est croissante au cours du temps, et ce, de manière linéaire. L'étendue de la variable Temps est de 5 ans.

Rappelons que la taille de la population la première année est connue. En notant α la pente traduisant l'évolution d'année en année ;
pour $\forall i \in [\![1, I]\!], \forall j \in [\![1, J]\!], \forall t \in [\![1, T]\!]$ on a :

$$N_{ijt} = N_{ij1} + \alpha \times (t - 1)$$

La taille du fichier de la population est donc désormais de 90 ($I \times J \times T$) dans notre exemple. On obtient ainsi une population pour les T années, par sexe et par catégorie d'âge :

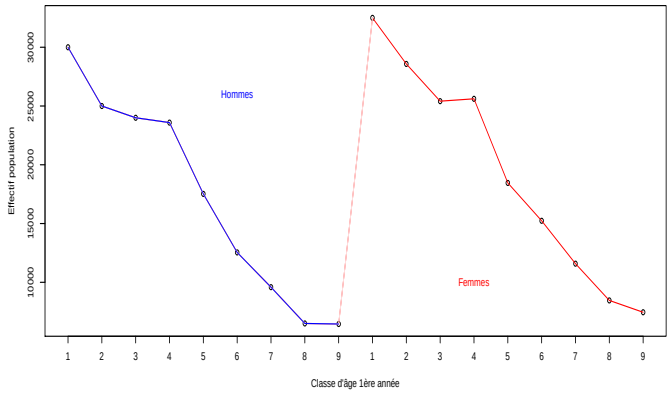


FIGURE 8: Population de la première année

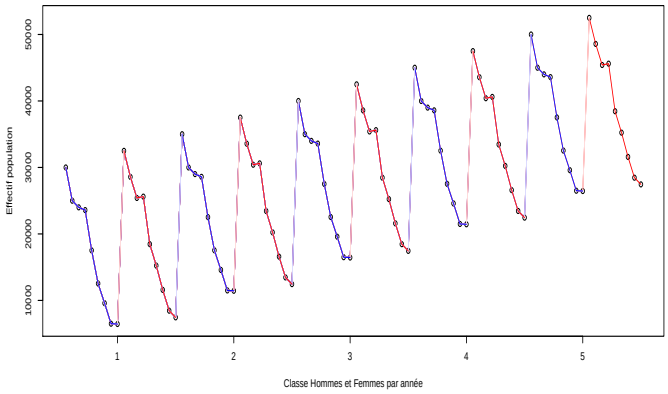


FIGURE 9: Population sur 5 années

4.2.2 Simulation des observations

Le nombre de cas observés la première année à été simulé selon une loi de Poisson autour d'une valeur moyenne pour chaque classe d'âge.

C'est-à-dire que pour un risque fixé r_j par catégorie d'âge, on avait :

$$O_{ij1} \sim> \mathcal{P}(r_j \times N_{ij1})$$

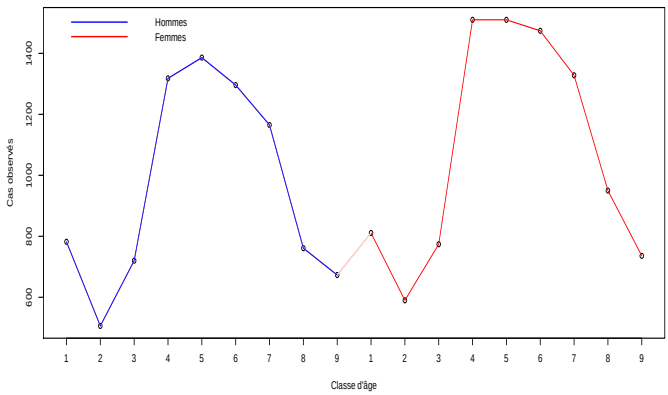


FIGURE 10: Observation la première année

Notons m_{ij1} la valeur moyenne pour chaque classe. On a $\forall i \in [\![1, I]\!], \forall j \in [\![1, J]\!]$:

$$m_{ij1} = (r_j \times N_{ij1})$$

Supposons maintenant, de la même manière que précédemment, que le nombre de cas observés augmente au fil du temps, et ce de manière linéaire.
On va alors simuler un nombre de cas observés autour d'une valeur moyenne chaque année, qui augmente linéairement en fonction du temps.

En notant p le coefficient directeur traduisant une augmentation du nombre de cas observés en fonction du temps on a :
$\forall i \in [\![1, I]\!], \forall j \in [\![1, J]\!], \forall t \in [\![1, T]\!]$;

$$m_{ijt} = (r_j \times N_{ijt}) + (t-1) \times p$$

Remarque :

On souhaite simuler les données de telle sorte que le nombre de cas observés augmente en fonction du temps, de manière linéaire, avec une pente supérieure à celle traduisant l'augmentation de la population.
C'est pourquoi on ne peut pas écrire $m_{ijt} = r_j.N_{ijt}$, au risque de ne pas avoir de proportionnalité entre la pente du nombre de cas observés et celle de N.

Avec ce modèle, deux pentes sont ici cumulées pour la simulation de O, on a en effet :

$$\begin{cases} m_{ijt} = (r_j \times N_{ijt}) + (t-1).p \\ N_{ijt} = N_{ij1} + \alpha \times (t-1) \end{cases}$$

$$\Rightarrow m_{ijt} = r_j(N_{ij1} + (t-1) \times \alpha) + (t-1)p$$
$$\Leftrightarrow m_{ijt} = r_j N_{ij1} + (t-1)(r_j \times \alpha + p)$$

Et finalement on a :

$$O_{ijt} \sim> \mathcal{P}(m_{ijt})$$
$$\Leftrightarrow O_{ijt} \sim> \mathcal{P}(r_j \times N_{ij1} + (t-1)(r_j \times \alpha + p))$$

En simulant les observations selon cette dernière équation on obtient :

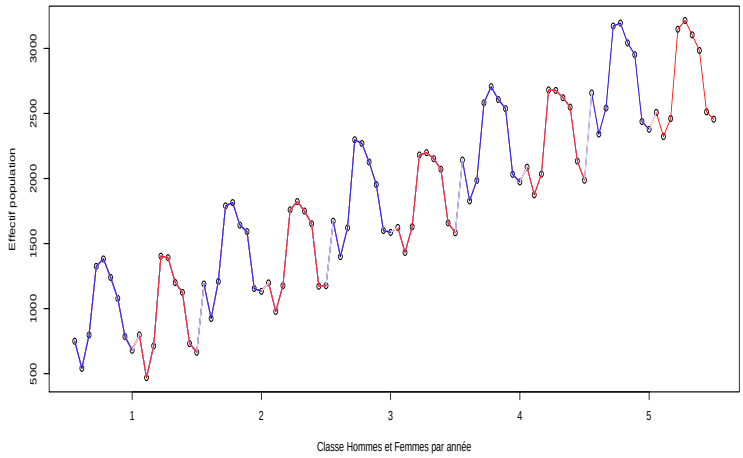


FIGURE 11: Observations sur 5 années

On a désormais un jeu de données avec une covariable $Temps$ supplémentaire. Ajustons alors les trois modèles de Poisson, calculés avec les différents offsets.

4.3 Ajustement et résultats

Rappelons les équations des 3 modèles que l'on souhaite ajuster :

$$log\left(\frac{O_{ij}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \beta_{Temps} \times Temps \tag{22}$$

$$log\left(\frac{O_{ij}}{E.sex_{ij}}\right) = \beta_0 + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps \tag{23}$$

$$log\left(\frac{O_{ij}}{E.global_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps \tag{24}$$

4.3.1 Les coefficients estimés

Après ajustement des trois modèles ci-dessus, les résultats pour les estimations des coefficients sont les suivants :

.	**Modèle Global**	2.5%	95%	**Modèle Age**	2.5%	95%	**Modèle Sex**	2.5%	95%
Int	-0.63	-0.65	-0.61	-0.14	-0.16	-0.13	-0.67	-0.69	-0.65
Sex1	-0.07	-0.08	-0.08	-0.07	-0.08	-0.06	X	X	X
Age2	-0.05	-0.08	-0.03	X	X	X	-0.05	-0.08	-0.03
Age3	0.15	0.13	0.17	X	X	X	0.15	0.13	0.17
Age4	0.47	0.45	0.49	X	X	X	0.47	0.45	0.49
Age5	0.71	0.69	0.73	X	X	X	0.71	0.69	0.73
Age6	0.82	0.79	0.84	X	X	X	0.81	0.79	0.84
Age7	0.91	0.89	0.93	X	X	X	0.91	0.89	0.93
Age8	0.83	0.81	0.86	X	X	X	0.83	0.81	0.86
Age9	0.82	0.79	0.84	X	X	X	0.82	0.79	0.84
Temps	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

4.3.2 Les critères AIC

Comparons, après ajustement, les critères d'Aikake pour les 3 modèles :

AIC modèle Global	AIC modèle Sex	AIC modèle Âge
900.452	903.674	884.514

FIGURE 12: *Critères d'Aikake pour les trois modèles*

On remarque alors qu'au sens de l'AIC, le meilleur modèle est le modèle dont l'offset a été calculé selon une standardisation sur la variable Age, ce qui est bien la méthode que nous avons choisi lors de la simulation de nos données.

4.4 L'effet du temps

On souhaite désormais étudier l'évolution de la population ainsi que l'évolution du nombre de cas observés au cours du temps. En supposant inconnue la méthode de simulation, à fortiori le coefficient, on estime la pente par le coefficient directeur de la droite des moindres carrés liant la population à la variable temporelle.

C'est à dire :

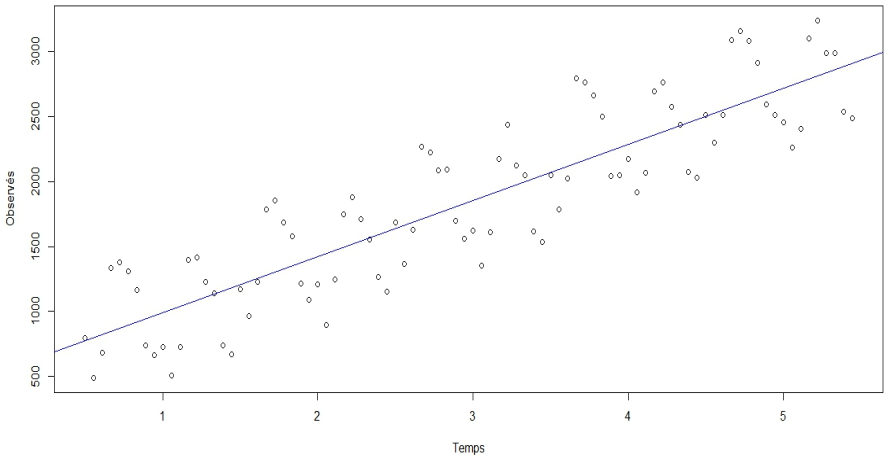


FIGURE 13: Droite des moindres carrés, estimation de la pente

IC 2.5%	Estimation	IC 97.5%
22.22	24.96	27.70

FIGURE 14: *Estimation de la pente par la méthode des moindres carrés*

L'estimation de la pente est alors satisfaisante (La valeur du coefficient à la simulation était de 25). Ainsi, on obtient une estimation de la pente de de l'évolution de la population que l'on notera respectivement $\hat{\alpha}$.

4.5 Estimation des risques

Supposons la méthode de simulation inconnue, on souhaite, après ajustement, estimer les risques associés à chaque strate de la population.
Les hypothèses établies sont les suivantes :

• L'évolution de la population est linéaire au cours du temps. On suppose connue la pente d'évolution de la population.
• L'évolution du nombre de cas observés est linéaire au cours du temps. On suppose connue l'estimation de la pente d'évolution du nombre de cas observés.

On avait : $\forall i \in [\![1..I]\!], \forall j \in [1..J], \forall t \in [\![1..T]\!]$

$$m_{ijt} = r_j \times N_{ij1} + (t-1)(r_j \times \alpha + p)$$

De la même manière que précédemment, il est possible d'estimer la pente d'une droite des moindres carrés au travers du nuage de points du nombre de cas observés.
Dans le but de quantifier l'évolution du nombre de cas observés en fonction du temps, ajustons une droite des moindres carrés pour chaque classe d'âge j et chaque groupe i : On a :

$$y_{ijt} = \widehat{b_{ij}} + \widehat{a_{ij}} \times (t-1)$$

On obtient ainsi une estimation de la pente et de l'ordonnée à l'origine pour chaque strate, i.e :

$$\widehat{a_{ij}} = r_j \widehat{\times \alpha + p} \qquad\qquad \widehat{b_{ij}} = \hat{r_j} \times N_{ij1}$$

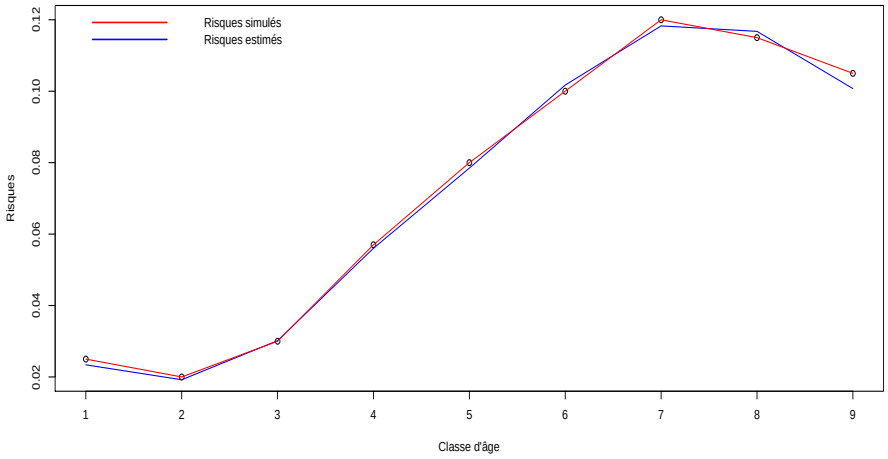
La pente estimée ne dépend pas du groupe i, en effet d'après les simulations ici réalisées, l'évolution est la même pour les hommes et les femmes. On s'attend alors à avoir, la même estimation de la pente, pour une catégorie d'âge fixée, dans les deux groupes sexe.
Quitte à utiliser la moyenne des deux estimations, on a alors $\hat{a_{ij}} = \hat{a_j},\ \forall j \in [\![1..J]\!]$
r_j et p sont ici inconnues. La connaissance de la pente $\hat{a_{ij}}$ ne permet alors pas d'évaluer les risques r_j.
Cependant, $\forall j \in [\![1..J]\!]$ on a une estimation de l'ordonnée à l'origine : $r_j \times N_{ij1}$, donnée par $\hat{b_{ij}}$.
Il est donc possible d'obtenir une estimation de r_j en utilisant la moyenne des estimations $\hat{b}$ par sexe. En effet :

$$\frac{m_{1jt} + m_{2jt}}{2} = r_j \frac{N_{1j1} + N_{2j1}}{2} + (t-1)(r_j \times \alpha + p)$$

Une estimation de r_j est alors obtenue par :

$$\hat{r_j} = \frac{\frac{1}{I}\sum_{i=1}^{I} \hat{b_{ij}}}{\frac{1}{I}\sum_{i=1}^{I} N_{ij1}} = \frac{\sum_{i=1}^{I} \hat{b_{ij}}}{\sum_{i=1}^{I} N_{ij1}}$$

Graphiquement, on a les estimations suivantes :



En supposant uniquement une évolution de la population et du nombre de cas observés, nous sommes ainsi en mesure d'estimer les risques pour chaque strate de la population.
Il est dès lors possible, en utilisant les valeurs de $\hat{a_{ij}}$ et les estimations $\hat{r_j}$, d'obtenir une estimation de p.
Quitte à prendre la moyenne, on s'attend à obtenir une même valeur estimée de la pente p pour tout i et pour tout j. On a alors :

$$\hat{p} = \frac{1}{J} \sum_{j=1}^{J} \hat{r_j} \hat{a_j} - \hat{\alpha}$$

Nous obtenons dans notre exemple une estimation $\hat{p} = 10.4$ pour une valeur fixée aux simulations de $p = 10$.

4.6 L'analyse des résidus

Intéressons nous, avant toute interprétation, à la validité de notre modèle ainsi qu'à sa qualité d'ajustement aux données.

4.6.1 Modèle sous population constante

Dans un premier temps, pour simplifier le problème, nous simulons les données comme mentionné précédemment, mais en considérant que la taille de la population reste constante dans le temps. C'est-à-dire :

$$\alpha = 0, \quad p \neq 0$$

Pour notre modèle contenant un offset standardisé par rapport à l'âge, on a les prédictions et les résidus suivants :

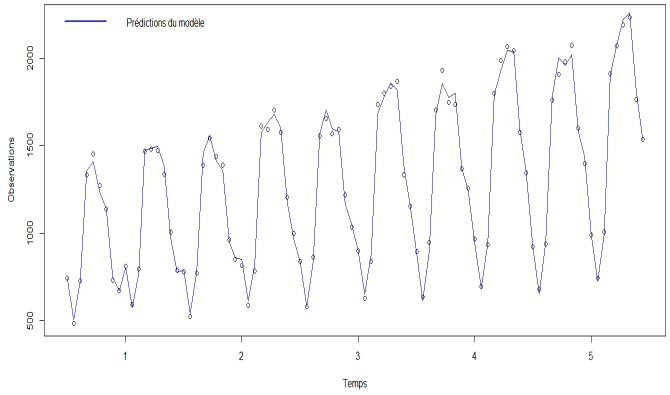


FIGURE 15: Prédictions du modèle, sous population constante

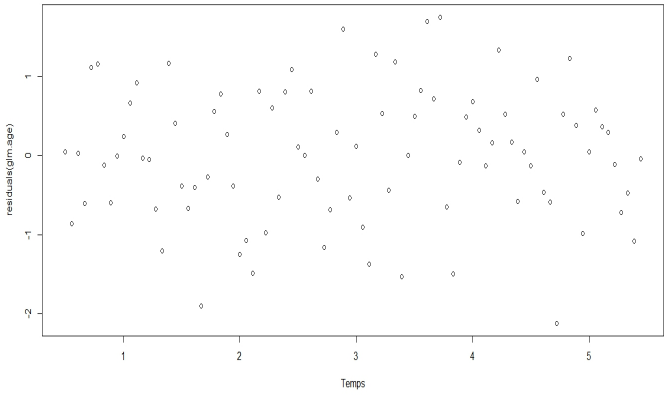


FIGURE 16: Résidus du modèle, sous population constante

L'analyse des résidus (ici les *Working residuals*) semble approuver la qualité des prédictions ci-dessus.
Les résidus sont ici faibles, homoscédastiques (égalité des variances), centrés autour de la valeur 0, et ne traduisant aucune tendance particulière.

Conclusion

Sous des hypothèses de linéarité d'évolution du nombre de cas observés, et d'une stabilité de la taille de la population, le modèle qui ajuste le mieux les données est le modèle attendu, comportant l'offset standardisé sur la variable de simulation.

4.6.2 Modèle sous population linéairement croissante dans le temps

Supposons maintenant, pour mieux cibler une situation concrète, une évolution de la population au cours du temps.
On a donc :

$$\alpha \neq 0, \quad p \neq 0$$

Comme nous l'avons vu précédemment, les deux pentes α et p se cumulent lors des simulations et $\forall i \in [\![1, I]\!], \forall j \in [\![1, J]\!], \forall t \in [\![1, T]\!]$:

$$m_{ijt} = r_j N_{ij1} + (t-1)(r_j.\alpha + p)$$

L'ajustement du modèle de Poisson nous amène alors aux prédictions suivantes :

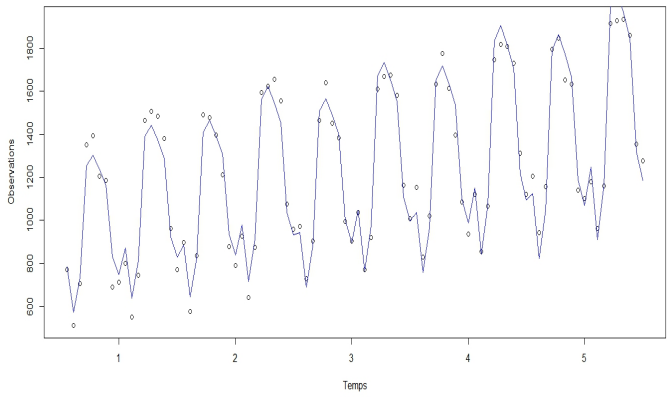


FIGURE 17: Prédictions du modèle, sous population croissante

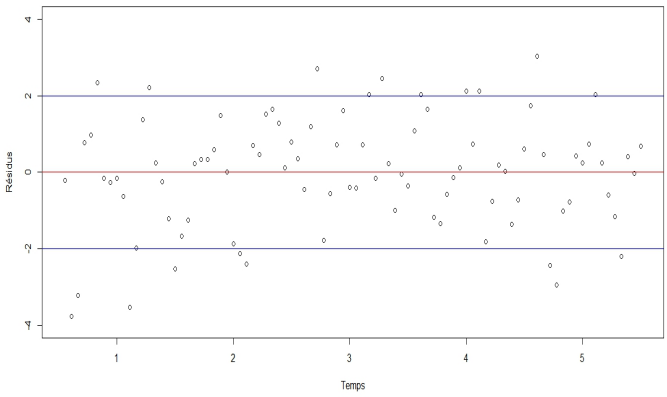


FIGURE 18: Résidus du modèle, sous population croissante

On voit sur le graphique ci-dessus l'évolution du nombre d'observés au cours du temps.
Le modèle ici semble s'ajuster de manière satisfaisante aux données. Les prédictions semblent être de bonne qualité, mais on note tout de même une dégradation par rapport à la qualité des prédictions du modèle sous population constante.
Les résidus sont raisonnablement centrés en 0, mais relativement importants, compris entre -4 et 4. Ils semblent également plus dispersés dans les premières périodes que dans les dernières, il sera ici nécéssaire de tester la surdispersion.
Avant d'effectuer des diagnostics supplémentaires, intéressons nous à un deuxième type de diagnostic graphique des résidus, à savoir la fonction de diagnostic résiduel `gamcheck`. [1]

1. Les fonctions `gamcheck` et `qqgam` sont des fonctions disponibles dans le package `mgcv` [10]

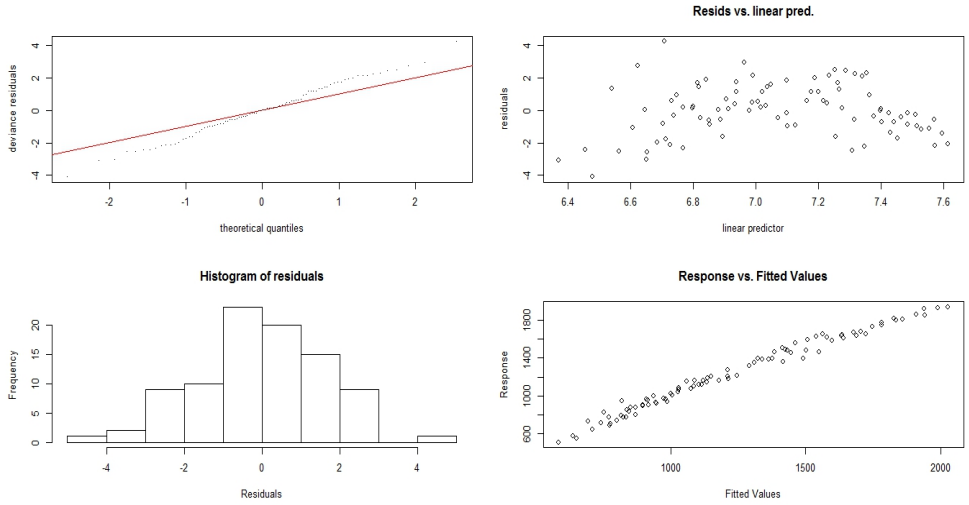


FIGURE 19: Résidus du modèle, sous population croissante

Commentaires

Graphe n°1 :
La représentation graphique des résidus de la déviance en fonction des quantiles théoriques issus d'une loi de Poisson.
Si le modèle est adéquat, les résidus de la déviance sont raisonnablement centrés, homoscédastiques, de loi s'approchant d'une loi normale. (Bien que non Gaussiens). Remarquons que les résidus de la déviance ne sont pas parfaitement alignés sur la première bissectrice.

Graphe n°2 :
Il s'agit de la représentation graphique des résidus du modèle. On reconnait le graphique précédemment commenté.

Graphe n°3 :
L'histogramme de la dispersion des résidus. Il est à noter que les résidus sont bien centrés en 0, comme nous le laissait supposer le 2ème graphique. Ils semblent également symétriquement répartis autour de leur valeur centrale 0. La dispersion en revanche, bien que symétrique, s'étend de -4 à 4. Ce qui reste relativement élevé.

Graphe n°4 :
La représentation graphique des observations en fonction des prédictions.
Si les prédictions sont bonnes, elles sont proches des observations, et par conséquent les points

du graphe alignés sur la première bissectrice.
On voit ici que ces points semblent plus ou moins alignés autour de la première bissectrice.

Conclusions
Bien que d'assez bonne qualité de manière générale, la qualité de ces résidus reste discutable, leur dispersion est importante, leur répartition en revanche est assez symétrique. Il sera nécéssaire d'effectuer des diagnostics supplémentaires avant de pouvoir interpréter les coefficients de ce modèle

4.7 Test de surpdispersion

Testons désormais la surdispersion de notre modèle à l'aide du test de surdispersion de Dean. Les résultats obtenus sont les suivants :

Statistique de test	p-valeur	coefficient de surdispersion
4.339	7.155e-06	2.511

FIGURE 20: *Résultats du test de surdispersion de Dean*

D'après les hypothèses du test de Dean, la p-valeur étant inférieure strictement à 0.05, il nous est possible ici, (au risque de 5% de se tromper) d'affirmer qu'il y a surdispersion des données dans le modèle, c'est à dire la non-égalité entre le moyenne et la variance dans notre modèle.

Plusieurs facteurs peuvent être responsable de la surdispersion des données :

- La non inclusion d'un facteur explicatif, comme par exemple des interactions significatives.
- Le mauvais ajustement du modèle (mauvaise fonction de lien par exemple..)

4.8 Les interactions

Intéressons-nous, dans le but d'améliorer la qualité des prédictions, à l'inclusion des interactions, éventuellement significatives.
Elles peuvent en effet être la source de la surdispersion observée précédemment et responsable de la faiblesse de la qualité des résidus. Notons qu'ici, à titre d'exemple, nous utiliserons un offset de type âge (c'est-à-dire standardisé sur la variable $\hat{A}ge$). Remarquons cependant, comme nous l'avons vu dans la partie 3.5 *Compensation des coefficients*,qu'il n'est pas nécéssaire d'inclure l'offset comme le produit du risque et de la population, mais il suffit d'inclure la taille de la population en tant qu'offset.
Dans le but de ne pas cumuler les hypothèses, nous conserverons ici les notations initiales pour mieux schématiser la problématique.

Effectuons manuellement une procédure de sélection de modèle (de type *montante*), en comparant les modèles suivants :

Modèle sans interaction

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps \tag{25}$$

Modèles à 1 interaction

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps + \sum_{\gamma=1}^{I} \beta_{13\gamma}.Temps.\mathbf{1}_{\gamma=i} \tag{26}$$

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps + \sum_{\delta=1}^{J} \beta_{23\delta}.Temps.\mathbf{1}_{\delta=j} \tag{27}$$

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps + \sum_{\gamma=1}^{I}\sum_{\delta=1}^{J} \beta_{12\gamma\delta}.Temps.\mathbf{1}_{\gamma=i}\mathbf{1}_{\delta=j} \tag{28}$$

Modèles à 2 interactions

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps + \sum_{\delta=1}^{J} \beta_{23\delta}.Temps.\mathbf{1}_{\delta=j} + \sum_{\gamma=1}^{I}\sum_{\delta=1}^{J} \beta_{12\gamma\delta}.Temps.\mathbf{1}_{\gamma=i}\mathbf{1}_{\delta=j} \tag{29}$$

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I}\beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J}\beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps}\times Temps + \sum_{\gamma=1}^{I}\beta_{13\gamma}.Temps.\mathbf{1}_{\gamma=i} + \sum_{\gamma=1}^{I}\sum_{\delta=1}^{J}\beta_{12\gamma\delta}.Temps.\mathbf{1}_{\gamma=i}\mathbf{1}_{\delta=j} \tag{30}$$

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I}\beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J}\beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps}\times Temps + \sum_{\gamma=1}^{I}\beta_{13\gamma}.Temps.\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J}\beta_{23\delta}.Temps.\mathbf{1}_{\delta=j} \tag{31}$$

Modèle à 3 interactions

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I}\beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J}\beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps}\times Temps + \sum_{\gamma=1}^{I}\beta_{13\gamma}.Temps.\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J}\beta_{23\delta}.Temps.\mathbf{1}_{\delta=j} + \sum_{\gamma=1}^{I}\sum_{\delta=1}^{J}\beta_{12\gamma\delta}.Temps.\mathbf{1}_{\gamma=i}\mathbf{1}_{\delta=j} \tag{32}$$

Comparons les AIC respectifs de ces modèles

Effets principaux
988.8

FIGURE 21: *AIC modèle sans interaction*

+Age×$Temps$	+Sex×$Temps$	+Age×Sex
896.8	1012.3	988.5

FIGURE 22: *AIC modèles avec une interaction*

+Age×$Temps + Age \times Sex$	+Sex×$Temps + Age \times Sex$	+Sex×$Temps + Age \times Temps$
904.4	1012.1	896.7

FIGURE 23: *AIC modèles avec deux interactions*

+Age×$Temps + Age \times Sex + Sex \times Temps$
904.3

FIGURE 24: *AIC modèle avec 3 interactions*

Résultat attendu

D'après la méthode de simulation décrite précédemment, on avait :

$$m_{ijt} = r_j.N_{ij1} + (t-1)(r_j.\alpha + p)$$

On distingue un effet de l'âge (r_j) dans le coefficient directeur de l'équation ci-dessus.
Ainsi on s'attend à une interaction significative entre le temps et l'âge.

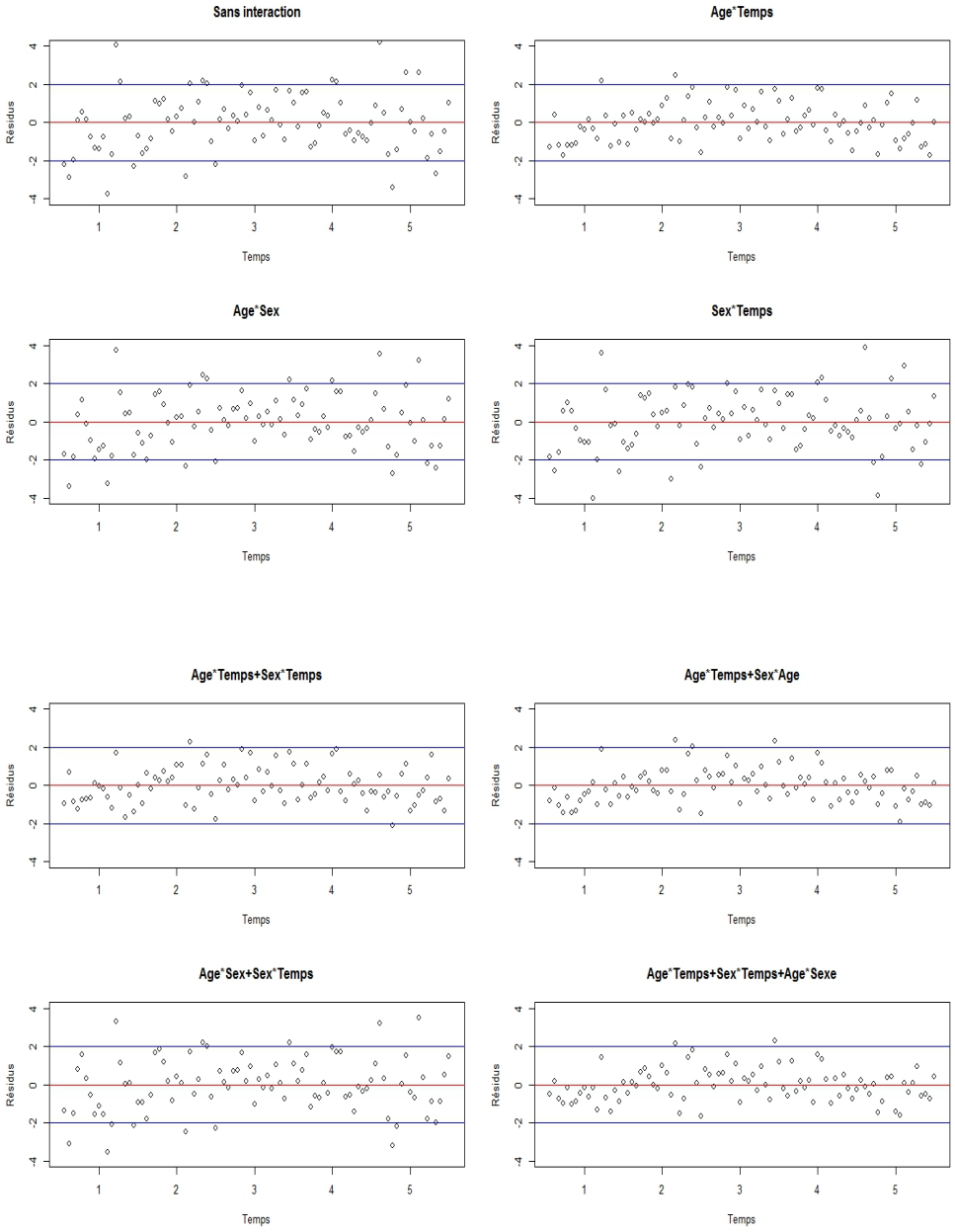
Commentaires

Dans une démarche de sélection de modèles selon une procédure montante, on distingue les étapes suivantes :
1. Modèle sans interaction
2. Modèle + Age × Temps

1. En effet, l'inclusion de l'interaction entre l'âge et le temps améliore de manière considérable la valeur de l'AIC. (Ce qui n'est pas le cas des deux autres interactions, du moins dans de moindres mesures. L'inclusion de l'interaction Age×Sex ne modifie pas la valeur de l'AIC de manière suffisament significative pour être exploitée.)

2. Il est possible d'ajouter une interaction supplémentaire à notre modèle contenant déjà l'interaction Age×Temps. L'ajout de Sex×Temps ne diminue pas la valeur de l'AIC, tandis que l'ajout de Age× Sex dégrade la qualité de notre critère.

Notons au passage, qu'il est possible d'observer les résidus pour les autres modèles décrits précédemment :



On voit alors que la qualité des modèles dont l'interaction Age×Temps est absente est moindre, comme nous le laissait entendre le critère d'Akaike de ces modèles.
L'apport en revanche de cette interaction a pour effet de diminuer de presque moitié l'étendue des résidus.
Ceci n'est pas vérifié avec les autres interactions, leur ajout dans le prédicteur linéaire n'a pas pour effet d'améliorer la qualité résiduelle.

Le modèle selectionné par valeur montante, basé sur la critère AIC ainsi que sur l'analyse résiduelle est donc le modèle contenant l'interaction Age×Temps, c'est-à-dire :

$$\log\left(\frac{\widehat{O_{ij}}}{E.age_{ij}}\right) = \beta_0 + \sum_{\gamma=1}^{I} \beta_{1\gamma}\mathbf{1}_{\gamma=i} + \sum_{\delta=1}^{J} \beta_{2\delta}\mathbf{1}_{\delta=j} + \beta_{Temps} \times Temps + \sum_{\delta=1}^{J} \beta_{23\delta}.Temps.\mathbf{1}_{\delta=j} \tag{33}$$

Test de surdispersion

Testons désormais la surdispersion de notre modèle comprenant l'interaction précédente, à l'aide du test de surdispersion de Dean.
Les résultats obtenus sont les suivants :

Statistique de test	p-valeur	coefficient de surdispersion
-3.9	0.99	0.64

FIGURE 25: *Résultats du test de surdispersion de Dean*

Il semblerait que la cause de la surpdispersion observée précédemment soit dûe à l'absence d'un terme explicatif significatif ; l'interaction entre l'âge et le temps.
On remarque en effet, qu'après inclusion de cette dernière interaction, et après analyse des résultats du test de surdispersion de Dean, il n'est plus possible de rejeter l'hypothèse nulle d'égalité entre la moyenne et la variance de notre modèle, au seuil de 5%.

L'analyse des résidus

Observons maintenant l'allure des résidus de ce modèle, contenant l'interaction entre $\hat{Age}$ et $Temps$:

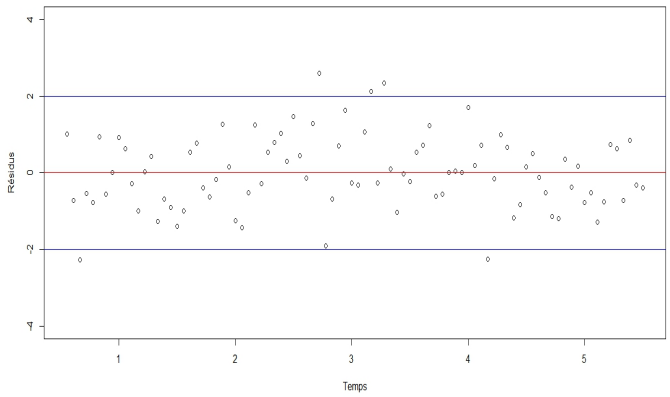


FIGURE 26: Résidus du modèle avec interaction

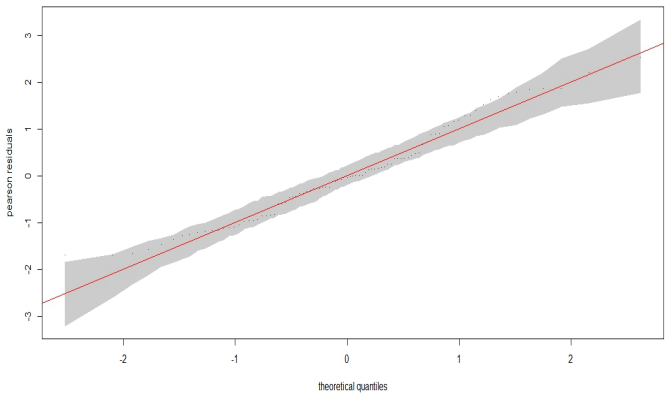


FIGURE 27: Diagramme Quantile-Quantile du modèle avec interaction

On observe alors une amélioration nette de la qualité des résidus. En effet ils sont toujours centrés symétriquement autour de la valeur moyenne 0, homoscédastiques, ne traduisant ainsi aucun effet résiduel, et d'amplitude réduite de moitié par rapport au précédent, puisqu'en majorité compris dans l'intervalle $[-2; 2]$.
La qualité du diagramme quantile-quantile[10] est satisfaisante, les points sont compris alignés sur la bissectrice, compris dans l'intervalle de confiance grisé entourant la droite.

4.9 Reflexion sur la notion de temps

La question de la nature du temps est un point essentiel de notre étude.
En effet, remarquons que le temps a été précédemment défini comme une variable continue lors de son introduction. Elle a également été considérée comme tel lors de l'ajustement des modèles.

Pourtant le temps semble ici de nature qualitative à 5 modalités (qui plus est ordonnées), étant donné que nous avons des données uniques pour chaque année.

Les épidémiologistes ont pour habitude de considérer le temps comme une variable quantitative. (Ils possèdent en effet une base de données de plus de 5 années).

Nous avons ici choisi de considérer le temps comme une variable continue également, pour étudier l'impact d'une variable quantitative au milieu de variables qualitatives dans l'ajustement avec offset. Au vu de notre base de données, déclarer le temps comme une variable qualitative aurait amené à des prédictions plus précises encore, étant donné les 5 modalités en jeu.

La variable a donc été ici considérée volontairement continue, pour étudier le comportement des deux types de variables.

Remarque

Il est probable, comme par exemple dans la suite du rapport, dans le calcul d'un offset double, que l'on décide de stratifier sur la variable temporelle. Cette stratification se voit impossible avec nos définitions pour une variable continue. Nous la considérerons donc temporairement qualitatives, de manière à simplifier les calculs, en notant bien que les formules de standardisation basées sur des calculs de somme se généralisent au cas continu grâce au calcul intégral.

Notons la proximité des prédictions et des résidus, entre une variable temps qualitative et une variable temps quantitative :*(Rouge=Quali,Bleu=Quanti)*

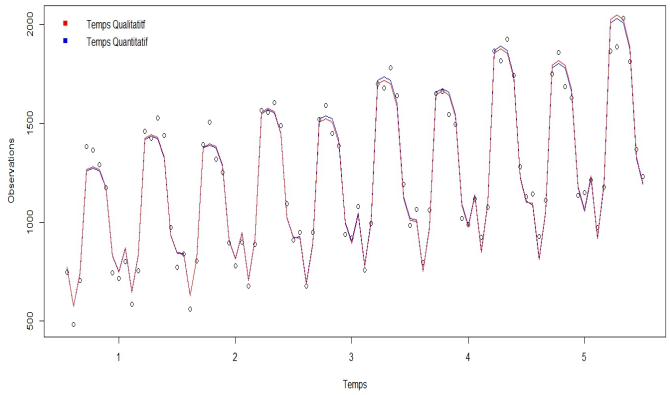


FIGURE 28: Comparaison des prédictions selon la nature du temps

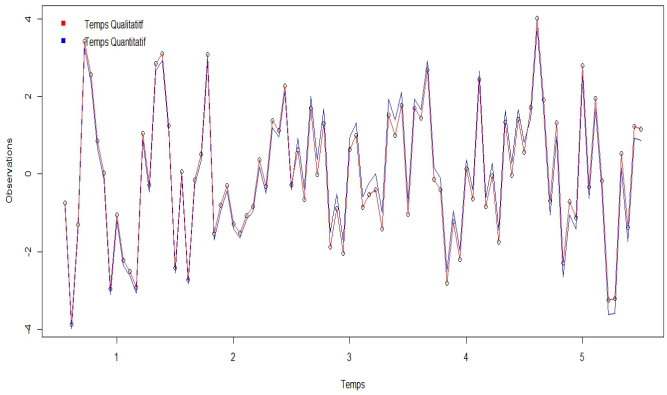


FIGURE 29: Comparaison des résidus selon la nature du temps

Il est donc raisonnable de parler de proximité entre les deux considérations, concernant la nature de la variable *Temps*.

Discussion : L'interprétation des résultats

L'étude de l'incidence d'une maladie peut amener l'utilisateur à employer les méthodes de standardisation de variables pour neutraliser un facteur dit de confusion (une différence inter-groupes), comme par exemple différentes tailles de populations. L'emploi de ces méthodes dans un modèle statistique de comptage tel que le modèle de poisson peut, sous certaines conditions, amener à des résultats inattendus. En effet, l'inclusion d'un effet sous la forme d'un offset impacte la qualité prédictive du modèle. Nous tentons ici, de résumer brièvement les conséquences de l'inclusion d'une telle variable, en fonction de la standardisation choisie.

La méthode de standardisation ?
La méthode de standardisation et les modèle habituellement ajustés ne permettent pas de distinguer les modèles contenant différents offsets, lorsque ceux-ci sont calculé comme une standardisation d'une variable. Les prédictions sont en effet équivalentes, et ce, indépendamment de la méthode de standardisation. Plusieurs démarches sont ici proposées pour pallier à ce problème.

• Si dans un cadre épidémiologue, l'objectif de l'utilisateur est de dissocier les modèles avec les différents offsets, dans le but de distinguer une différence importante entre plusieurs classes d'une même variable (par exemple une forte différence Hommes/Femmes), la variable utilisée pour la standardisation et le calcul de l'offset doit être retirée du prédicteur linéaire. Nous avons ainsi montré que le meilleur modèle parmi ceux ici proposés est celui dont le calcul d'offset est basé sur la même standardisation que lors des simulations.
Concrètement : Si le risque d'apparition d'une maladie est plus fort chez les hommes que chez les femmes, que le but de l'utilisateur est de mettre en avant cette différence, il est nécessaire d'ôter du prédicteur la variable qualitative *Sex*, lors de la comparaison du modèle contenant l'offset *Sex* avec d'autres offsets. Le modèle qui minimise l'AIC est alors le modèle contenant le même qu'offset que le choix des simulations.

• En revanche, si le but de l'utilisateur est de réaliser des estimations, le modèle contenant l'ensemble de ses variables est adéquat. Cependant, il peut s'affranchir du calcul de l'offset, par des méthodes de standardisation habituelles, et n'ajuster, de manière équivalente, un modèle n'ayant pour offset que la taille de la population en question.
En effet, indépendamment de la méthode de standardisation, le seul apport informatif fourni par l'offset est la taille de la population N_{ijt}.

Inclusion des interactions ?
Quel que soit le modèle choisi par l'utilisateur, avec ou sans variable, il se retrouve confronté à un problème de surdispersion et de qualité d'ajustement. Cette discussion est dûe à l'absence d'un facteur explicatif indispensable à la modélisation d'une situation concrète : l'intéraction

entre l'âge et le temps. En effet, la population actuelle est en constante évolution. (La discussion reste ouverte quant à la linéarité de cette évolution). De même, selon la maladie étudiée, pour des raisons diverses et variées (alimentation, mode de vie, stress, etc..) le nombre de cas déclarés d'une maladie M est également en constante évolution. (Il ne s'agit pas d'une généralité mais d'une globalité, prenons par exemple le nombre de cancers du sein qui a évolué de près de 197% entre 1980 et 2000.) Ici aussi, la discussion est ouverte quant à la linéarité de cette évolution.

La structure même de ces évolutions donne lieu à une interaction Âge & Temps non négligeable, comme nous avons pu le constater dans les chapitres précédents, nécéssaire à la correction de la surdispersion de nos modèles, ainsi qu'à l'amélioration de leur qualité résiduelle. Selon les besoins de l'utilisateur, et s'il s'autorise les interactions du premier ordre, il sera conseillé, dans le but d'interpréter, et d'améliorer la qualité d'ajustement, d'inclure l'interaction Âge & Temps à notre modèle.

Les limites de ces résultats

L'effet de complémentarité entre un offset et une variable est propre au modèle de Poisson avec un lien logarithme. En effet, il n'est pas affirmé ici qu'un autre modèle ou une autre fonction de lien offre les même propriétés à la variable modélisant le nombre de cas attendus.
D'autre part, les propriétés mises en avant ici sont propres aux modèles contenant des variables qualitatives (Sexe, ou Âge s'il est considéré comme tel), nous n'avons ici pas d'éléments propres au cas d'une variable continue et de la transposition de l'effet ici décrit.
Les méthodes de simulation ont été ici basées sur l'hypothèse de linéarité de l'évolution de la population et, de manière indépendante, l'évolution du nombre de cas observés. Le modèle linéaire ici supposé n'est pas forcément adéquat à la situation concrète de l'évolution d'une maladie. De plus on peut s'interroger sur le lien entre les deux pentes, l'évolution de la population impliquant de toute évidence une évolution du nombre de cas observés également, donc un lien entre les deux coefficients.

Conclusion

Ces résultats possèdent leurs atouts ainsi que leurs limites. Malgré la restriction de leur domaine d'applications très ciblés (un type de modèle, un type de fonction, un type de variable), ces résultats peuvent s'avérer utile dans la compréhension d'une structure de population. De plus, la méthode de standardisation ici présentée, est l'une des méthodes les plus employées en épidémiologie. Le produit d'un risque par une population en tant qu'offset explicatif est fréquent, c'est pourquoi il peut être intéressant d'évaluer l'impact de cette présence dans le prédicteur linéaire. Ces résultats peuvent également permettre la simplification d'un modèle, dans un cadre d'interprétation et de prédictions, pour l'épidémiologiste ou l'utilisateur s'intéressant à une maladie en constante évolution, parmi une population en pleine croissance..

Annexes

5.1 Le critère de la déviance

La déviance est un outil de comparaison entre plusieurs modèles. Elle compare la vraisemblance obtenue à celle que l'on obtiendrai dans un modèle parfait : le modèle saturé. C'est une mesure de l'ajustement qui prend en compte la complexité du modèle.

Dans le modèle saturé, la prévision est parfaite, il n'existe donc aucune incertitude.
La définition de ce critère est donc basée sur la comparaison à ce modèle complet :

$$D = 2\left[\sum_{i=1}^{n} \mathcal{L}_{\text{saturé}} - \mathcal{L}(\beta)\right] \geq 0$$

La déviance est égale à deux fois une différence de vraisemblance.
Elle constitue un écart en terme de log-vraisemblance entre modèle saturé et modèle étudié.

La déviance est donc nulle pour le modèle saturé sans incertitudes, on en déduit que le modèle considéré est d'autant meilleur, en terme d'ajustement, que sa déviance est faible.

La déviance peut être interprétée comme l'analogue de la somme des carrés résiduelle dans les modèles linéaires Gaussiens.
Il est alors possible de comparer si la différence entre deux déviances pour deux modèles emboités (notés *simple* et *complet*) est significative. Sous les hypothèses adéquates :

$$\Delta D = D_{simple} - D_{complet} \sim \chi^2_{p2-p1}$$

où $p1$ est le nombre de paramètre du modèle simple
et $p2$ le nombre de paramètres du modèle complet.

L'hypothèses nulle est alors la validité du modèle simple (une différence de déviance non significative) contre l'hypothèse alternative de déviance significative traduisant l'adéquation du modèle complet.

5.2 Standardisation Âge & Sex

Tous les modèles décrits dans le rapport contiennent un offset, calculé comme une stratification sur une variable.
Nous sommes cependant en possession de trois variables (ou plus).
Nous pourrions donc envisager un modèle dont l'offset serait calculé selon 2 variables.

"L'apriori" (que nous appelons ici offset) serait donc plus informatif, car il prendrait en compte deux facteurs au lieu d'un seul lors de la standardisation. Un offset stratifié sur deux variables permettrai ainsi d'obtenir un risque pour les catgéories d'âge *Hommes*, différents de celui pour les catégories d'âge *Femmes*. Le nombre de risque différents pour chaque classe serait ainsi plus important.

Le risque se calcule comme une somme sur les périodes (et par conséquent est ici considéré comme une variable qualitative) i.e ;

$$p_{ij} = \frac{\sum_{t=1}^{T} O_{ijt}}{\sum_{t=1}^{T} N_{ijt}}$$

Ce risque est appliqué aux strates de population pour obtenir une valeur du nombre de cas attendus "affinée" par rapport à la méthode précédente.
On s'attend par conséquent à un modèle qui ajuste mieux encore les données (au sens de l'AIC) que le modèle dont l'offset n'est calculé que par rapport à une variable unique.

Résultats
En notant E.agesex l'offset standardisé sur les deux variables Âge et Sexe, les prédictions du modèle sont les suivantes :

$$log\left(\frac{O_{ij}}{E.agesex_{ij}}\right) = \beta_0 + \beta_{Temps} \times Temps \tag{34}$$

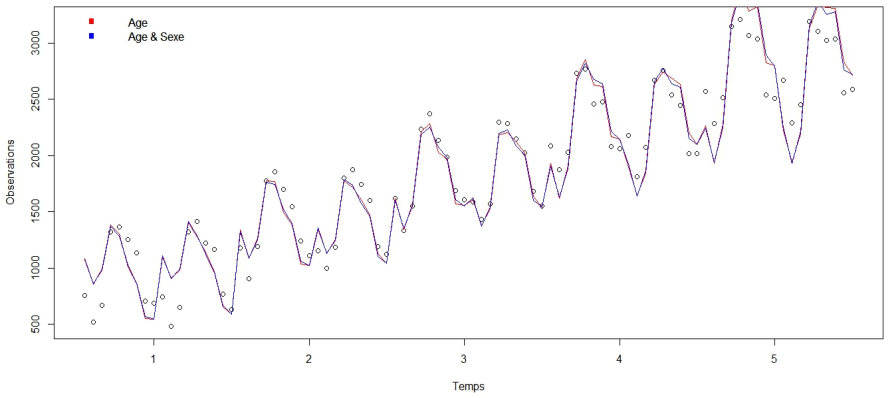


FIGURE 30: Risques estimés

Il est difficile de disntinguer une amélioration de la qualité d'ajustement au travers de ce graphique. Comparons alors les critères AIC des différents modèles :

AIC Stand.Glob	AIC Stand.Age	AIC Stand.Sex	AIC Stand.Age.Sex
3048	3036	3045	3017

FIGURE 31: *Critère d'Aikake des différents modèles*

A nouveau, le modèle Age a un critère inférieur aux modèles dont la variable a été standardisé sur le Sexe ou encore le modèle global. Mais on remarque également que le modèle utilisant la stratification affinée sur deux variables est encore de meilleure qualité au sens du critère d'Aikake que les autres modèles.

Le modèle standardisé sur plusieurs variables, bien qu'ayant un AIC légèrement inférieur à celui d'un modèle à standardisation simple, ne se distingue pas du point de vue de sa qualité d'ajustement. Les prédictions sont quasiment identiques à celle d'un modèle à standardisation simple. L'emploi d'un tel modèle est discutable : le choix sera laissé à l'utilisateur selon les besoins de l'étude.

5.3 La méthode IRLS

La méthode IRLS est une méthode itérative de maximisation de la vraisemblance dans le cas où les équations du score seraient non-linéaires. Cet algorithme (aussi appelé Algorithme du Fisher-Scoring), est une variante de la méthode de Newton-Raphson, où la matrice d'information de Fisher est utilisée au détriment de la matrice Hessienne. Celle-ci est en effet généralement plus simple à calculer car elle s'exprime comme un produit de matrices.[6]
Rappelons alors les étapes de l'algorithme de Newton-Raphson (Dans le cas simplifié d'une limite finie de fonction), et sous de bonnes conditions de régularité de la fonction.
On suppose qu'on cherche à résoudre l'équation $f(x) = 0$ pour une fonction f donnée.

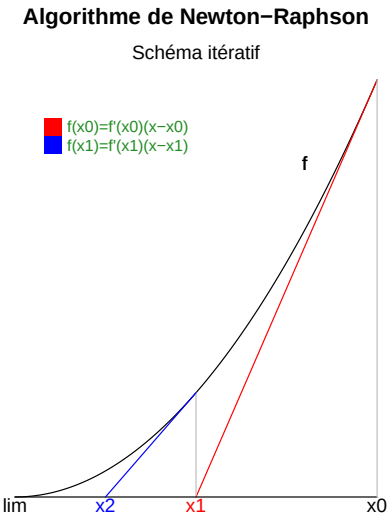


FIGURE 32: Représentation graphique des étapes d'itération de la méthode de Newton-Raphson

Le principe itératif est le suivant :
• On part d'une valeur initiale x_0
• On trace la tangente à la courbe f au point x_0.
Elle a pour équation :

$$y = f(x_0) + f'(x_0) \times (x - x_0)$$

Elle coupe l'axe des abscisses en un point x_1 tel que :

$$x_1 = x_0 \times \left(-\frac{f(x_0)}{f'(x_0)} \right)$$

• On réitère ce procédé en traçant la tangente en x_1 qui coupe l'axe $y = 0$ en x_2 : d'où

$$x_2 = x_1 \times \left(-\frac{f(x_1)}{f'(x_1)} \right)$$

.

On continue ré$_i$*tèreainsilasuite*(x_n) jusqu'à une tolérance fixée par l'utilisateur, pour obtenir la limite de convergence ; solution de notre équation.
Cette méthode peut s'étendre à la résolution des équations du score dans une recherche de maximisation de la vraisemblance.[7]
Le détail des étapes intermédiaires[7] ne sera pas développé ici.
La formule récursive utilisée par le logiciel `R` dans sa fonction `glm` est la suivante :

$$\boxed{(\mathbb{X}^t \mathbb{W} \mathbb{X})\beta^{(r)} = \mathbb{X}\mathbb{W}^{(r-1)} z^{(r-1)}}$$

où :

• $\mathbb{W}^{(r-1)} = diag\left(\frac{1}{V(\mu)a(\phi)} \left(\frac{\partial \mu}{\partial \eta} \right)^2 \right)_{(n \times n)}$
• $z^{(r-1)} = \left\{ (y - \mu) \left(\frac{\partial \eta}{\partial \mu} \right) + (\eta^{(r-1)} - \text{offset}) \right\}_{(n \times 1)}$
• $\eta_i^{(r-1)} = \sum_{k=1}^{N} x_{ki} \beta_k^{r-1} + \text{offset}$
• $\mu_i = g^{-1}(\sum_{k=1}^{N} x_{ki} \beta_k^{r-1}) = g^{-1}(\eta_i - \text{offset})$
• $\mathbb{X}$ est la design matrix du modèle

Dans le cas particulier d'un modèle de poisson, on a $V = id$, $\phi = 1$, $g = \log$:

Les seuils de convergence sont fixés à 10^{-10}.
La convergence de cet algorithme est généralement très rapide (ici 13 itérations suffisent) et il est généralement préféré à la méthode de convergence de Newton-Raphson car la matrice d'information de Fisher définie par :

$$I_n(\beta) = -\mathbb{E}\left[\frac{\partial}{\partial \beta^j} \frac{\partial}{\partial \beta^i} \mathcal{L}(y, \beta, \phi) \right]$$

est généralement plus simple à calculer que la matrice Hessienne.

5.4 Vérifications, simulations

5.4.1 Le modèle à deux variables

On souhaite également vérifier les résultats énoncés dans ce rapport sur plusieurs simulations. En particulier on souhaite vérifier si le modèle comprenant un offset $\hat{Age}$ ajuste mieux les données que les modèles *Sexe* et $\hat{Age}$ de façon générale. Nous avons également comparés les critères d'Akaike, les pentes, la conservation des effets, le choix de la méthode pour un, ou des, modèles donnés, sur un seul jeu d'observations. Vérifions nos hypothèses en simulant K jeu de données, et vérifions que les conclusions énoncées restent vérifiées. Nous utiliserons ici 1000 simulations différentes, c'est à dire : $K = 1000$.

5.4.2 L'effet "Balance" de l'information

Les conclusions énoncées précédemment concernant l'information contenue dans l'offset et l'ajustement ultérieur d'un modèle contenant toutes les variables est vérifié sur les K simulations. En incluant l'ensemble de nos variables, et en ajustant par rapport à l'une d'entre elles, il est systématiquement impossible de mettre en avant l'un des modèles. L'effet de complémentarité entre logaithme du risque et variable explicative est systématiquement vérifié dans les données simulées, et seul la taille de la population présente un apport informatif au modèle. Les AIC, les prédictions, ainsi que tous les critères de validation de modèle sont identiques,à chacune des 1000 simulations. Ces modèles sont systématiquement équivalents.

5.4.3 Modèle sans effet simple, avec offset

Comme nous l'avons vu dans le cas du modèle complet, il est impossible de mettre en avant un modèle sans en enlever la variable choisie pour le calcul de l'offset. Vérifions alors que dans K situations distinctes, en ôtant la variable de standardisation du prédicteur, le meilleur modèle est bien celui attendu : Comparons les valeurs des critères d'Aikake dans les différents modèles :

	Modèle Global	Modèle Âge	Modèle Sexe
Min.	177.60	161.60	177.70
1st Qu.	181.50	165.50	181.70
Median	184.00	168.00	184.10
Mean	184.60	168.60	184.80
3rd Qu.	187.00	171.00	187.20
Max.	201.40	185.40	201.40
Eacrt-Type	3.95	3.95	3.95

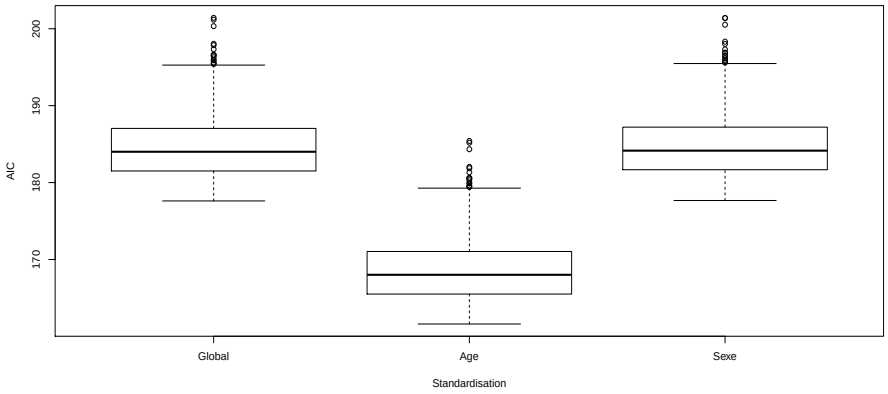


FIGURE 33: Diagrammes-boite des AIC pour les 3 modèles

La véritable valeur de l'AIC, (i.e celle dont le modèle ajustée sur les moyennes de simulation) est de 168. Les diagrammes-boite ci-dessus montrent que le meilleur modèle, au sens de l'AIC, est bien le modèle $\hat{Age}$: la méthode de simulation est bien reconnue pour le modèle à deux variables, sous condition que la répétition de l'information ait été évitée par le retrait d'une variable du prédicteur.

Notons que les modèles permettent également de retrouver les estimations des risques appliqués aux populations lors des simulations, comme nous l'avons vu dans le cas d'un seul modèle, ces valeurs se retrouvent également dans le cas des K modèles.

5.4.4 Le modèle à trois variables

On souhaite désormais vérifier les résultats énoncés précédemment pour le modèle à 3 variables : 2 qualitatives et le temps comme variable quantitative.
Les K modèles ont donc été simulés de la même manière que précédemment, dans le cas simple, en considérant un accroissement de la population au fil du temps, et un accroissement (indépendemment de l'accroissement de la population) du nombre de cas observés.

Critère d'Aikake
Après ajustement des 3 modèles, on s'intéresse dans un premier temps au critère d'Aikake pour les 3 modèles, selon les 3 standardisations. On a :

	Modèle Global	Modèle Âge	Modèle Sexe
Min.	854.00	838.00	862.20
1st Qu.	880.80	864.80	889.00
Median	889.20	873.20	897.60
Mean	889.50	873.50	897.90
3rd Qu.	897.60	881.60	906.00
Max.	949.50	933.50	957.70
Ecart-type	12.85	12.84	12.85

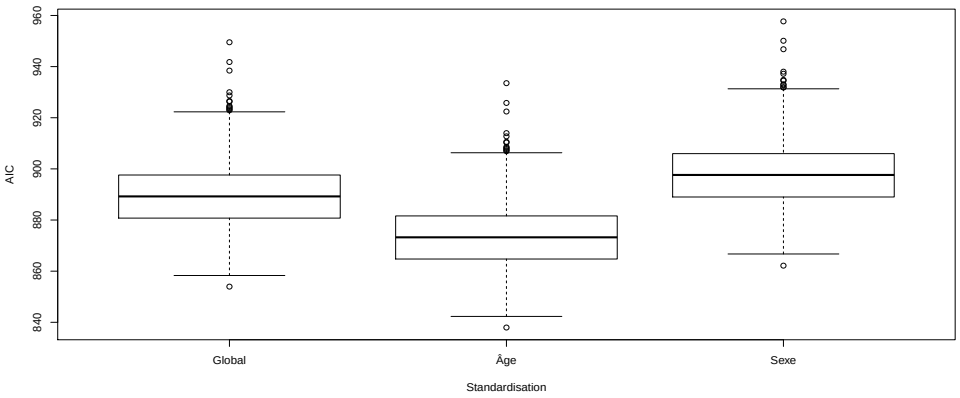


FIGURE 34: Diagrammes-boite des AIC pour les 3 modèles

La véritable valeur de l'AIC (celle dont le modèle est ajusté sur les moyennes de simulation) est de 820. On remarque, de la même manière que dans le cas isolé d'un modèle, que malgré l'ajout de la variable continue temporelle, le meilleur modèle reste celui dont le calcul de l'offset est propre à la méthode de simulation. Il nous est donc possible, dans ces conditions de sélectionner le meilleur modèle. Nos hypothèses sont alors vérifiées dans le cas de K simulations.

Estimation des risques

Intéressons-nous également à l'estimation des risques, après ajustement des 1000 modèles. La méthode utilisée pour estimer les risques est celle décrite au paragraphe 4.5 dans le cas d'un unique modèle.
(On utilise l'ordonnée à l'origine de la droite des moindres carrés, dans le nuage des observations).
Les résultats obtenus sont les suivants :

Âge	Min	$Quant_{25\%}$	$Quant_{50\%}$	$Moyenne$	$Quant_{75\%}$	Max	Sd
1	0.0225	0.0242	0.0246	0.0246	0.0250	0.0265	0.0006
2	0.0171	0.0191	0.0196	0.0196	0.0200	0.0216	0.0006
3	0.0266	0.0289	0.0294	0.0295	0.0300	0.0320	0.0008
4	0.0530	0.0556	0.0563	0.0563	0.0571	0.0599	0.0011
5	0.0738	0.0779	0.0790	0.0790	0.0801	0.0836	0.0016
6	0.0921	0.0971	0.0985	0.0985	0.0998	0.1058	0.0020
7	0.1107	0.1162	0.1179	0.1179	0.1195	0.1243	0.0024
8	0.1039	0.1102	0.1121	0.1122	0.1142	0.1220	0.0029
9	0.0927	0.1001	0.1020	0.1020	0.1040	0.1115	0.0029

TABLE 1: Estimations des risques pour 1000 Simulations

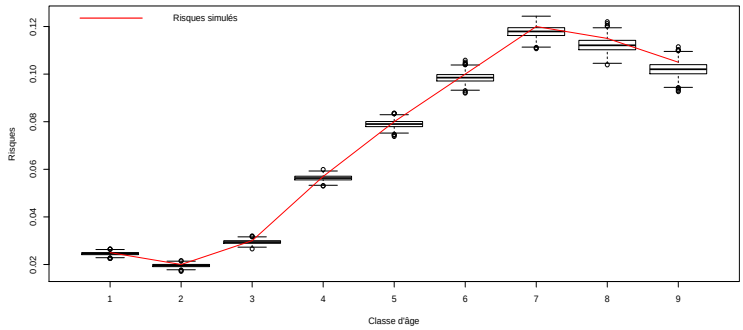


FIGURE 35: Estimations des risques pour 1000 Simulations

De manière globale, les résultats énoncés dans le cas isolé d'un seul modèle, de deux variables, ou trois variables, restent vérifiés au travers des K simulations. L'égalité des prédictions pour le modèle à prédicteur incluant l'effet simple reste vérifié, l'estimation de la pente dans un modèle à trois variables est de très bonne qualité, l'estimation des risques appliqués à la simulation est également de qualité satisfaisante.
Ainsi, l'ensemble des résultats énoncés, et des estimations réalisées se vérifient dans le cas de K (ici $K = 1000$) jeux d'observations différents.

Références

[1] Akaike.H, *A new look at statistical model identification*, IEEE Transactions on Automatic Control **AU-19** (1974), 716–722.

[2] C.B.Dean, *Testing for overdispersion in poisson and binomial regression models*, Journal of the American Statiscal Association **87 N° 418** (1992), 451–457.

[3] Silvia Columbu and Dr Erik-André Sauleau Prof ssa Monica Musio, *Régression de poisson - risque global et risque standardisé*, Universia' degli Studi di Cagliari, 2010.

[4] L.Eliott D.Loomis, D.B.Richardson, *Poisson regression analysis of ungrouped data*, Occup Environ Med **62** (2005), 325–329.

[5] Rachel A. Pedersen-Terry.M.Thernau Elizabeth.J.Atkinson, Cynthia S.Crowson, *Poisson models for person-years ans expected rates*, Tech. report, Mayo Foundation, 2008.

[6] Segolen Geffray, *Cours de modèles linéaires généralisés (chap.3)*, 2012.

[7] Joseph M.Hilbe James W.Hardin, *Generalized linear models and extensions*, 2007.

[8] Jospeh M.Hilbe James W.Hardin, *Generalized linear models and extensions*, ch. Derivation of the Poisson Algorithm, pp. 184–185, 2007.

[9] Organisation mondiale de la santé., *La standardisation : Une méthode épidémiologique classique pour la compraison des taux*, Bulletin Epidémiologique : Organisation panaméricaine de la santé **23 N°3** (2002), 1–7.

[10] Simon N.Wood Nicole H. Augustin, Dr Erik-André Sauleau, *On quantile quantile plots for generalized linear models.*, Computational Statistics & Data Analysis **56** (2012), 2404–2409.