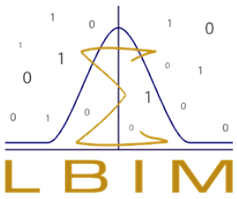
Choix d'une famille de distribution pour la modélisation de Score

Nelly Winzenrieth



Choix d'une famille de distribution pour la modélisation de Score

Laboratoire de Biostatistique et Informatique Médicale de la Faculté de Médecine de Strasbourg

WINZENRIETH Nelly
Master 2 Statistique
Année 2013-2014

Remerciements

Je tiens à remercier le Professeur Nicolas Meyer, directeur du laboratoire de Biostatistique et Informatique Médicale de la Faculté de Médecine de Strasbourg, de m'avoir accueillie au sein de son laboratoire.

Je souhaite remercier particulièrement le Docteur Erik-André Sauleau, mon maître de stage, pour ses conseils et sa disponibilité tout au long du stage.

Je tiens enfin à remercier Monsieur Schaeffer Mickaël pour son aide et son soutien durant le stage.

Table des matières

Chapitre 1

Introduction

1.1 Présentation du laboratoire

Le laboratoire de Biostatistique et Informatique Médicale de la Faculté de Médecine de Strasbourg se situe au sein de l'hôpital civil de Strasbourg. Il est dirigé par le professeur Nicolas Meyer. Le laboratoire travaille sur plusieurs axes de recherche tels que les méthodes PLS (Partial Least Squares), les modèles pour l'épidémiologie et les applications médicales des méthodes bayésiennes.

1.2 Définition d'un score

Dans le domaine médical, on est souvent amené à devoir modéliser des scores. Un score est un nombre exprimant le résultat d'un test, d'un questionnaire ou le ressenti de l'état de santé d'un patient. Par exemple, un score peut être représenté par une note ou une appréciation sur 20 ou sur 10 à l'aide d'entiers, ou encore par une échelle de douleur en millimètre sur une règle graduée en centimètres.

Le score d'Apgar est un exemple concret qui permet d'évaluer l'état de santé d'un nouveau-né. Le rythme cardiaque, la respiration, la tonicité musculaire, la couleur de la peau et la réactivité sont notés 0, 1 ou 2, puis additionnés. On obtient alors un résultat compris entre 0 et 10, appelé score d'Apgar. Un score compris entre 7 et 10 signifie que l'enfant est en bonne santé. En revanche, un score inférieur à 7 nécessite des soins immédiats adaptés.

Un autre exemple de score est l'indice de gravité simplifié (IGS 2). Ce score est calculé en fonction de différents paramètres (15 au total) relatifs au patient, qui sont recueillis dans les 24 heures suivant l'admission du patient en unité de soins intensifs. Par exemple, ce score prend en compte le type d'admission (programmée, médicale ou urgente), l'âge du patient (inférieur à 40 ans, entre 40 et 59 ans, ..., supérieur à 80 ans) ou encore la fréquence cardiaque (inférieure à 40, entre 40 et 69, ..., supérieure à 160). Pour chaque paramètre, un score est attribué selon la modalité correspondante à l'état du patient. On obtient l'IGS 2 en additionnant les différents scores obtenus pour les différents paramètres. C'est un score compris entre 6 et 163. A partir de ce score, des modèles statistiques prédisent le pourcentage de mortalité du patient. [7]

Un dernier exemple de score est l'échelle visuelle analogique (EVA). C'est une réglette de 10 centimètres qui permet au patient d'auto-évaluer sa douleur. Sur la face présentée au patient, se trouve un curseur le long d'une ligne droite dont l'une des extrémités correspond à "Absence de douleur" et l'autre à "Douleur insupportable". Le patient positionne le curseur

le long de cette ligne à l'endroit qui situe le mieux sa douleur. Sur l'autre face, le soignant lit l'intensité de la douleur ressentie par le patient à l'aide d'une graduation en millimètres sur la réglette.

1.3 Objectif de l'étude

On souhaite modéliser ces scores dans un cadre régressif. Il y a deux buts distincts dans la modélisation, un aspect descriptif d'une part, prédictif d'autre part. L'objectif d'un modèle descriptif est de donner des valeurs précises des estimations des paramètres du modèle alors que l'objectif d'un modèle prédictif est de prédire des valeurs le plus précisément possible. Par exemple pour les scores d'Apgar et de l'IGS 2, l'objectif est prédictif car on souhaite prédire au mieux l'état de santé du patient.

L'objectif du stage est, dans le cadre régressif, de trouver la meilleure famille de distribution pour les scores, au regard d'un certain nombre de critères.

1.4 Déroulement de l'étude

Les scores ont la particularité d'être bornés, sauf après une éventuelle transformation, et ils sont soit :
- continus : par exemple l'EVA
- discrets : par exemple le score d'Apgar
- ordonnés

Nous nous intéressons uniquement aux deux premiers cas.

Il n'existe pas, a priori, de modèles statistiques correspondant parfaitement à nos données. Afin de se fixer des modèles susceptibles d'être intéressants pour nos données, nous avons tout d'abord établi une liste de modèles :

- Le modèle linéaire Gaussien ne semble pas convenir à la modélisation des scores puisque la loi Normale est continue et non bornée.
- Le modèle linéaire généralisé de Poisson pourrait convenir pour certains scores. En effet, la loi de Poisson prend des valeurs discrètes dans $\mathbb{N}$. Les valeurs sont donc bornées à gauche mais non à droite.
- Le modèle linéaire généralisé Gamma sert à modéliser des données continues dans $\mathbb{R}_+^*$.
- Le modèle linéaire généralisé Binomial, quant à lui, permet de modéliser des données discrètes et bornées. Cependant la loi Binomiale est une répétition de la loi de Bernoulli. Ce modèle ne semble pas adapté à tous les types de scores.
- Le modèle de régression Beta permet de modéliser des données continues et bornées dans $]0,1[$.

Voulant explorer un maximum de possibilités, nous avons décidé d'étudier le modèle linéaire Gaussien, les modèles linéaires généralisés de Poisson, Gamma et Binomial ainsi que le modèle de régression Beta. Bien que certains modèles ne semblent pas convenir à nos données, les étudier nous permettra de confirmer notre idée ou au contraire de la contredire. L'idée est de trouver une famille idéale qui conviendrait aux scores continus et discrets. En effet la frontière entre les deux n'est pas toujours très nette, par exemple dans le cas de l'EVA, l'échelle

est censée être continue mais, étant une réglette de 10 centimètres graduée en millimètres, le soignant ne peut lire qu'un nombre fini de valeurs.

Dans un premier temps, nous avons travaillé sur un jeu de données réel, une étude qui mesure l'implication des médecins généralistes dans le dépistage précoce du diabète de l'enfant. Les différents modèles, cités précédemment, ont été étudiés sur ce jeu de données.

Par la suite nous avons simulé des données afin de pouvoir les contrôler et tester les différents modèles sur un nombre important de données, avec de l'inférence fréquentiste.

Enfin, nous avons travaillé sur des données simulées avec des modèles bayésiens.

Dans la suite de ce rapport, les différents modèles étudiés seront tout d'abord introduits. Puis les critères de comparaison des modèles seront détaillés. Par la suite l'étude du jeu de données réel sera présentée et le travail effectué sur des scores simulés sera développé. Enfin une introduction à l'inférence bayésienne ainsi que les résultats obtenus, avec cette méthode, sur des scores simulés seront expliqués.

Chapitre 2

Différents modèles de régression

2.1 Modèles linéaires généralisés

Cette partie est principalement basée sur les références [8], [5] et [6].

Il s'agit d'une famille de modèle de régression. Le but est de trouver une relation de nature stochastique entre la variable réponse que l'on notera Y et p variables explicatives que l'on notera $X^{(1)}, ..., X^{(p)}$. On notera également $\mathscr{L}$ la loi d'une variable aléatoire et $\mathbb{E}$ son espérance.

La spécification d'un modèle linéaire généralisé se fait ainsi :
- l'indépendance entre les individus : $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont mutuellement indépendants pour $i = 1, ..., n$ avec n la taille de l'échantillon
- $\mathscr{L}(Y|X^{(1)}, ..., X^{(p)})$ appartient à la famille exponentielle
- les covariables $X^{(1)}, ..., X^{(p)}$ influent sur le niveau moyen conditionnel de la variable réponse au travers d'un prédicteur linéaire

$$\eta = \beta_0 + \beta_1 \times X^{(1)} + ... + \beta_p \times X^{(p)} \tag{2.1}$$

- il existe une fonction de lien g qui relie le prédicteur linéaire au niveau moyen conditionnel de la variable réponse

$$g(\underbrace{\mathbb{E}(Y|X^{(1)}, ..., X^{(p)})}_{\mu}) = \eta \tag{2.2}$$

2.1.1 La famille exponentielle

La famille exponentielle est constituée des distributions dont la densité se met sous la forme :

$$f(y, \theta, \phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \tag{2.3}$$

où θ est le paramètre canonique
$b(\theta)$ est le cumulant
$c(y, \phi)$ est une constante de normalisation pour que la densité s'intègre à 1
ϕ est un paramètre de nuisance (peut être connu ou inconnu)

En général $a(\phi) = \frac{\phi}{\omega}$ avec ω le poids connu.

En notant $\mu = \mathbb{E}[Y]$ avec Y une variable aléatoire de densité se mettant sous la forme (2.3), la fonction de lien canonique s'obtient en posant $g(\mu) = \theta$.

Propriété : Expression des moments

Soit Y une variable aléatoire qui appartient à la famille exponentielle, de densité mise sous la forme canonique (2.3).
Alors $\mu = \mathbb{E}[Y] = b'(\theta)$ et $Var(Y) = a(\phi)b''(\theta)$.
La fonction $b''(\theta)$ définit de manière implicite la fonction de variance en posant $V(\mu) = b''(\theta)$.

La loi Normale

$Y \sim \mathcal{N}(\mu, \sigma^2)$ de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$

$$\begin{aligned} f(y, \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \times exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= exp\left(\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right)\right) \end{aligned} \tag{2.4}$$

Les paramètres de la loi Normale sont :

$$\begin{aligned} \theta &= \mu \\ b(\theta) &= \frac{\mu^2}{2} \\ a(\phi) &= \sigma^2 \quad \text{donc } \phi = \sigma^2 \text{ car } \omega = 1 \\ c(y, \phi) &= -\frac{1}{2}\left(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right) \end{aligned}$$

On en déduit que $b'(\theta) = \mu$ et $b''(\theta) = 1$.
On obtient la fonction de lien canonique en posant $g(\mu) = \theta$. Ce qui donne $g(\mu) = \mu$.
La fonction de lien canonique de la loi Normale est donc la fonction *identité*.
La fonction de variance s'obtient en posant $V(\mu) = b''(\theta)$ d'où $V(\mu) = 1$.

Dans le cas particulier du modèle linéaire Gaussien, on peut directement écrire :
$Y = \beta_0 + \beta_1 \times X^{(1)} + ... + \beta_p \times X^{(p)} + \epsilon$
avec ϵ indépendantes et identiquement distribuées suivant une loi $\mathcal{N}(0, \sigma^2 I)$

La loi Gamma

$Y \sim \Gamma(a, b)$, avec a et b positifs, de densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^+$

$$\begin{aligned} f(y, a, b) &= \frac{1}{\Gamma(a)} b^a e^{-by} y^{a-1} \\ &= exp\left(\frac{\frac{b}{a}y - log(b)}{\frac{-1}{a}} + (a-1)log(y) - log(\Gamma(a))\right) \end{aligned} \tag{2.5}$$

On pose $\theta = \frac{b}{a}$ et (a,b) $\rightarrow \left(\theta = \frac{b}{a}, a\right)$ est bijective.

$$f(y, \theta, a) = exp\left(\frac{y\theta - log(\theta)}{-\frac{1}{a}} - \frac{log(a)}{\frac{1}{a}} + (a-1)log(y) - log(\Gamma(a))\right)$$

Les paramètres de la loi Gamma sont :

$$\begin{aligned} \theta &= \frac{b}{a} \\ b(\theta) &= log(\theta) \\ a(\phi) &= -\frac{1}{a} \quad \text{donc } \phi = \frac{1}{a} \text{ car } \omega = 1 \\ c(y,\phi) &= -\frac{log(a)}{\frac{1}{a}} + (a-1)log(y) - log(\Gamma(a)) \end{aligned}$$

On en déduit que $b'(\theta) = \frac{1}{\theta}$ et $b''(\theta) = -\frac{1}{\theta^2}$. De plus $\mu = b'(\theta)$. Dans ce cas, $\mu = \frac{1}{\theta}$ d'où $\theta = \frac{1}{\mu}$.
On obtient la fonction de lien canonique en posant $g(\mu) = \theta$. Ce qui donne $g(\mu) = \frac{1}{\mu}$.
La fonction de lien canonique de la loi Gamma est donc la fonction *inverse*.
La fonction de variance s'obtient en posant $V(\mu) = b''(\theta)$ d'où $V(\mu) = -\frac{1}{\theta^2} = -\mu^2$.

La loi de Poisson

$Y \sim \mathcal{P}(\lambda)$ de densité par rapport à la mesure de comptage sur $\mathbb{N}$

$$\begin{aligned} f(y,\lambda) &= \frac{\lambda^y e^{-\lambda}}{y!} \\ &= exp\left(y \times log(\lambda) - \lambda - log(y!)\right) \end{aligned} \tag{2.6}$$

Les paramètres de la loi de Poisson sont :

$$\begin{aligned} \theta &= log(\lambda) \quad \Longrightarrow \quad e^\theta = \lambda \\ b(\theta) &= \lambda \quad \Longrightarrow \quad b(\theta) = e^\theta \\ a(\phi) &= 1 \quad \text{donc } \phi = 1 \text{ car } \omega = 1 \\ c(y,\phi) &= -log(y!) \end{aligned}$$

On en déduit que $b'(\theta) = b''(\theta) = e^\theta = \mu$. De plus $\mu = b'(\theta)$. Dans ce cas, $\mu = e^\theta$ d'où $\theta = log(\mu)$.
On obtient la fonction de lien canonique en posant $g(\mu) = \theta$. Ce qui donne $g(\mu) = log(\mu)$.
La fonction de lien canonique de la loi de Poisson est donc la fonction *logarithme*.
La fonction de variance s'obtient en posant $V(\mu) = b''(\theta)$ d'où $V(\mu) = \mu$.

La loi de Bernoulli

$Y \sim \mathcal{B}(p)$ de densité par rapport à la somme de Dirac $\nu = \delta_{(0)} + \delta_{(1)}$

$$\begin{aligned} f(y,p) &= p^y(1-p)^{(1-y)} \\ &= exp\left(y \times log\left(\frac{p}{1-p}\right) + log(1-p)\right) \end{aligned} \tag{2.7}$$

Les paramètres de la loi de Bernoulli sont :

$$\begin{aligned} \theta &= log\left(\frac{p}{1-p}\right) \quad \Longrightarrow \quad e^\theta = \frac{p}{1-p} \quad \Longrightarrow \quad p = \frac{e^\theta}{1+e^\theta} \\ b(\theta) &= -log(1-p) \quad \Longrightarrow \quad b(\theta) = log(1+e^\theta) \\ a(\phi) &= 1 \quad \text{donc } \phi = 1 \text{ car } \omega = 1 \\ c(y,\phi) &= 0 \end{aligned}$$

On en déduit que $b'(\theta) = \frac{e^\theta}{1+e^\theta}$ et $b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2}$. De plus $\mu = b'(\theta)$. Dans ce cas, $\mu = \frac{e^\theta}{1+e^\theta}$ $\Leftrightarrow \theta = log\left(\frac{\mu}{1-\mu}\right) = logit(\mu)$. Donc $b''(\theta) = \mu(1-\mu)$
On obtient la fonction de lien canonique en posant $g(\mu) = \theta$. Ce qui donne $g(\mu) = logit(\mu)$. La fonction de lien canonique de la loi de Bernoulli est donc la fonction *logit*.
La fonction de variance s'obtient en posant $V(\mu) = b''(\theta)$ d'où $V(\mu) = \mu(1-\mu)$.

La loi Binomiale

$Y \sim \mathcal{B}(N,p)$ avec N connu. La densité de $\frac{Y_i}{N_i} = \pi_i$ se met sous la forme :

$$\begin{aligned} f(\pi_i, N_i, p_i) &= \mathbb{P}\left(\frac{Y_i}{N_i} = \pi_i\right) \text{ avec } \mathbb{P} \text{ la probabilité} \\ &= \mathbb{P}(Y_i = \pi_i N_i) \\ &= \left(C_{N_i}^{\pi_i N_i}\right) p^{\pi_i N_i}(1-p)^{N_i(1-\pi_i)} \\ &= exp\left(N_i \pi_i log\left(\frac{p}{1-p}\right) + N_i log(1-p) + log\left(C_{N_i}^{\pi_i N_i}\right)\right) \\ &= exp\left(\frac{\pi_i log\left(\frac{p}{1-p}\right) + log(1-p)}{\frac{1}{N_i}} + log\left(C_{N_i}^{\pi_i N_i}\right)\right) \end{aligned} \tag{2.8}$$

Les paramètres de la loi Binomiale sont :

$$\begin{aligned} \theta &= log\left(\frac{p}{1-p}\right) \quad \Longrightarrow \quad e^\theta = \frac{p}{1-p} \quad \Longrightarrow \quad p = \frac{e^\theta}{1+e^\theta} \\ b(\theta) &= -log(1-p) \quad \Longrightarrow \quad b(\theta) = log(1+e^\theta) \\ a(\phi) &= \frac{1}{N_i} \quad \text{donc } \phi = 1 \text{ car } \omega = N_i \\ c(y,\phi) &= log\left(C_{N_i}^{\pi_i N_i}\right) \end{aligned}$$

Les paramètres θ et $b(\theta)$ de la loi Binomiale sont les mêmes que ceux de la loi de Bernoulli. D'après le cheminement effectué à partir de ces paramètres pour la loi de Bernoulli, on peut directement conclure que la fonction de lien canonique de la loi Binomiale est la fonction *logit* et que la fonction de variance est $V(\mu) = \mu(1-\mu)$.

2.1.2 Estimation des paramètres

On doit estimer le vecteur des paramètres $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ présent dans le prédicteur linéaire. On utilise la méthode du maximum de vraisemblance pour ses propriétés d'efficacité asymptotique.

On suppose que l'on dispose des données $(y_i, x_i^{(1)}, ..., x_i^{(p)})$ pour la $i^{ème}$ observation avec $i = 1, ..., n$. On notera $\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ et la variable aléatoire associée $\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ ainsi que

$$\mathbb{X} = \begin{pmatrix} 1 & X_1^{(1)} & \cdots & X_1^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n^{(1)} & \cdots & X_n^{(p)} \end{pmatrix}.$$

Soit $L(\mathbf{y}, \beta, \phi)$ la vraisemblance du modèle en les observations $\mathbf{y}$ (conditionnellement aux prédicteurs). Il s'agit de trouver la valeur de β qui maximise $L(\mathbf{y}, \beta, \phi)$, ou de manière équivalente *log* $L(\mathbf{y}, \beta, \phi) = \mathcal{L}(\mathbf{y}, \beta, \phi)$ la log-vraisemblance, pour obtenir $\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{argmax}\ \mathcal{L}(\mathbf{y}, \beta, \phi)$.

$$L(\mathbf{y}, \beta, \phi) = \prod_{i=1}^{n} f(y_i, \beta, \phi) \text{ par indépendance des individus}$$

$$\begin{aligned} \mathcal{L}(\mathbf{y}, \beta, \phi) &= log\left(L(\mathbf{y}, \beta, \phi)\right) \\ &= log\left(\prod_{i=1}^{n} f(y_i, \beta, \phi)\right) \\ &= \sum_{i=1}^{n} log\left(f(y_i, \beta, \phi)\right) \\ &= \sum_{i=1}^{n} log\left(exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right)\right) \\ &= \sum_{i=1}^{n} \underbrace{\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right)}_{\mathcal{L}_i(\mathbf{y}, \beta, \phi)} \end{aligned} \tag{2.9}$$

Afin de maximiser $\mathcal{L}(\mathbf{y}, \beta, \phi)$, on commence par déterminer les équations du score :

$$\frac{\partial \mathcal{L}}{\partial \beta_j}(\mathbf{y}, \beta, \phi) = 0 \text{ pour } j = 0, 1, ..., p \tag{2.10}$$

$$\frac{\partial \mathcal{L}}{\partial \beta_j}(\mathbf{y}, \beta, \phi) = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i}{\partial \beta_j}(\mathbf{y}, \beta, \phi) \qquad \text{avec} \quad \frac{\partial \mathcal{L}_i}{\partial \beta_j} = \frac{\partial \mathcal{L}_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} \tag{2.11}$$

On calcule les quatre termes séparément :

$$\begin{aligned} \frac{\partial \mathcal{L}_i}{\partial \theta_i}(\mathbf{y}, \beta, \phi) &= \frac{1}{a_i(\phi)}(y_i - b'(\theta_i)) \\ &= \frac{1}{a_i(\phi)}(y_i - \mu_i) \quad \text{car } \mu_i = b'(\theta_i) \end{aligned}$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{\frac{\partial \eta_i}{\partial \mu_i}} = \frac{1}{g'(\mu_i)} \quad \text{car } g(\mu_i) = \eta_i \text{ d'après (2.2)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_i^{(j)} \quad \text{car } \eta_i = \sum_{j=0}^{p} \beta_j x_i^{(j)} \text{ avec } x_i^{(0)} = 1 \text{ d'après (2.1)}$$

En remplaçant dans (2.11) on trouve :

$$\frac{\partial \mathcal{L}}{\partial \beta_j}(\mathbf{y}, \beta, \phi) = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}(y_i - \mu_i)\frac{1}{V(\mu_i)}\frac{1}{g'(\mu_i)}x_i^{(j)} \tag{2.12}$$

D'après (2.10), on trouve les équations du score suivantes :

$$\sum_{i=1}^{n} \frac{1}{a_i(\phi)}\frac{(y_i - \mu_i)}{V(\mu_i)g'(\mu_i)}x_i^{(j)} = 0 \text{ pour } j = 0, 1, ..., p$$

Les équations du score sont non linéaires, elles se résolvent numériquement grâce à l'algorithme de Newton-Raphson (voir annexe A.1) ou sa variante l'algorithme du Fisher-scoring.

Propriété sur le comportement asymptotique de $\hat{\beta}$

$\hat{\beta}$, l'estimateur du maximum de vraisemblance du paramètre β, suit approximativement la loi $\mathcal{N}(\beta, I_n(\beta)^{-1})$ lorsqu'il est calculé à partir d'un nombre d'observations n assez grand.

La matrice d'information de Fisher du modèle à n observations est

$$I_n(\beta) = \frac{1}{\phi}\mathbb{X}'.A(\beta).\mathbb{X} \tag{2.13}$$

où ϕ est le paramètre de dispersion, $\mathbb{X}$ est la matrice définie page précédente et $A(\beta)$ est la matrice diagonale dont le $i^{ème}$ terme diagonal est

$$A_{ii}(\beta) = \frac{\omega_i}{V(\mu_i)g'(\mu_i)^2}$$

où ω_i=1 sauf dans le cas de données binomiales, V est la fonction de variance, g la fonction de lien et μ_i la fonction moyenne du modèle.

2.1.3 Diagnostics et tests

■ **Analyse des résidus**

∗ **Cas général**

Les résidus de base sont : $\widehat{\epsilon_i} = Y_i - \underbrace{\widehat{\mu_i}}_{\widehat{Y_i}}$

où $\widehat{Y_i}$ est la prédiction de Y_i par le modèle : $\widehat{\mathbb{E}}(Y_i|X_i^{(1)}, ..., X_i^{(p)}) = \widehat{\mu_i} = g^{-1}(\widehat{\beta_0} + \widehat{\beta_1} \times X_i^{(1)} + ... + \widehat{\beta_p} \times X_i^{(p)})$

Ces résidus sont hétéroscédastiques, on utilise alors :

- Les résidus de Pearson : $\widehat{\epsilon_i}^p = \sqrt{\omega_i}\frac{Y_i - \widehat{\mu_i}}{\sqrt{V(\widehat{\mu_i})}}$

- Les résidus de Pearson studentisés : $\widehat{\epsilon_i}^{ps} = \sqrt{\omega_i}\frac{Y_i - \widehat{\mu_i}}{\sqrt{V(\widehat{\mu_i})(1-h_i)}}$

avec h_i le $i^{ème}$ coefficient diagonal de H, la matrice des leviers, qui est définie par :
$H = A(\widehat{\beta})^{\frac{1}{2}}\mathbb{X}(\mathbb{X}'A(\widehat{\beta})\mathbb{X})^{-1}\mathbb{X}'A(\widehat{\beta})^{\frac{1}{2}}$ où $A(\widehat{\beta}) = diag\left(\frac{\omega_i}{V(\widehat{\mu_i})g'(\widehat{\mu_i})^2}\right)$

∗ **Cas particulier du modèle linéaire Gaussien**

On a : $\begin{pmatrix} \widehat{\epsilon}_1 \\ \vdots \\ \widehat{\epsilon}_n \end{pmatrix} \sim \mathcal{N}_n(0, \sigma^2(I_n - H))$ où $H = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$ et I_n la matrice identité de taille n

D'où $\widehat{\epsilon}_i \sim \mathcal{N}(0, \sigma^2(1-h_i))$ où h_i est le i^e coefficient diagonal de H

On travail avec :

- Les résidus standardisés $\widehat{\epsilon_i}^s = \frac{\widehat{\epsilon_i}}{\sqrt{\widehat{\sigma}^2(1-h_i)}}$ avec $\widehat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \widehat{\mu_i})^2$

- Les résidus studentisés $\widehat{\epsilon_i}^{st} = \frac{\widehat{\epsilon_i}}{\sqrt{\widehat{\sigma}^2_{(-i)}(1-h_i)}}$ avec $\widehat{\sigma}^2_{(-i)} = \frac{1}{n-(p+1)}\sum_{j=1, j\neq i}^{n}(y_i - \widehat{\mu}_{j,(-i)})^2$

car $\hat{\epsilon_i}^s$, $\hat{\epsilon_i}^{st} \sim \mathcal{N}(0,1)$

Les résidus standardisés d'un modèle linéaire Gaussien correspondent aux résidus de Pearson studentisés qui sont définis ci-dessus.

Sur les résidus standardisés ou studentisés, on effectue le test de Shapiro-Wilk, on teste :

$$\begin{gathered} \text{H}_0 : \text{les résidus suivent une loi normale} \\ \text{contre} \\ \text{H}_1 : \text{les résidus ne suivent pas une loi normale} \end{gathered}$$

La statistique de test est :

$$W = \frac{\left[\sum_{i=1}^{\left[\frac{n}{2}\right]} a_i(x_{(n-i+1)} - x_{(i)})\right]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.14}$$

avec : - $x_{(i)}$ correspond à la série des données triées
- $\left[\frac{n}{2}\right]$ est la partie entière du rapport $\frac{n}{2}$
- a_i sont des constantes générées à partir de la moyenne et de la matrice de variance co-variance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont fournies dans des tables spécifiques.

La région critique, rejet de la normalité, s'écrit : $\text{W} < \text{W}_{crit}$. Les valeurs seuils W_{crit} pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk. [10]

Sur les résidus standardisés ou studentisés, on effectue également le test de Breusch-Pagan qui permet de tester l'hypothèse d'homoscédasticité du terme d'erreur, on teste :

$$\begin{gathered} \text{H}_0 : \text{Var}(Y_i | X_i^{(1)}, ..., X_i^{(p)}) = \sigma^2 \ \forall \ i = 1,..,\text{n} \\ \text{contre} \\ \text{H}_1 : \exists \ i_0 \text{ tel que } \text{Var}(Y_{i_0} | X_{i_0}^{(1)}, ..., X_{i_0}^{(p)}) \neq \sigma^2 \\ \text{et } \exists \ i_1 \text{ tel que } \text{Var}(Y_{i_1} | X_{i_0}^{(1)}, ..., X_{i_0}^{(p)}) = \sigma^2 \end{gathered}$$

Se référer à [1] pour des explications concernant ce test.

▪ Tests d'adéquation globale du modèle aux données

∗ Test de la déviance

Considérons le test de : $\mathrm{H}_0 : \beta_1 = ... = \beta_p = 0$

contre

H_1 : au moins un des coefficients est différent de 0

La déviance est définie par : $\mathcal{D}(\beta) = \sum_{i=1}^{n} \mathcal{D}_i(\beta) = \sum_{i=1}^{n} 2\phi(\mathcal{L}(y_i, y_i, \phi) - \mathcal{L}(y_i, \mu_i, \phi))$

On a $\dfrac{\mathcal{D}(\hat{\beta})}{\phi} \xrightarrow{\mathscr{D}} \chi^2(n-(p+1))$

Soit $\hat{\beta}_r$ l'estimateur par maximum de vraisemblance de β calculé sous H_0, $\hat{\beta}_r = \begin{pmatrix} \hat{\beta}_{0r} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

On a également $\dfrac{\mathcal{D}(\hat{\beta}_r) - \mathcal{D}(\hat{\beta})}{\phi} \underset{\text{sous } \mathrm{H}_0}{\xrightarrow{\mathscr{D}}} \chi^2(p)$

De plus $(\mathcal{D}(\hat{\beta}))$ et $(\mathcal{D}(\hat{\beta}_r) - \mathcal{D}(\hat{\beta}))$ sont indépendants.

On obtient alors l'approximation suivante :

$$\begin{aligned} F &= \frac{(\mathcal{D}(\hat{\beta}_r) - \mathcal{D}(\hat{\beta}))/(\phi p)}{(\mathcal{D}(\hat{\beta}))/(\phi(n-(p+1)))} \overset{\text{sous } \mathrm{H}_0}{\sim} F(p, n-(p+1)) \\ &= \frac{(\mathcal{D}(\hat{\beta}_r) - \mathcal{D}(\hat{\beta}))/p}{(\mathcal{D}(\hat{\beta}))/(n-(p+1))} \overset{\text{sous } \mathrm{H}_0}{\sim} F(p, n-(p+1)) \end{aligned}$$

On rejette H_0 lorsque $F > F^{-1}_{F(p,n-(p+1))}(1-\alpha)$ au niveau de risque de première espèce α.

∗ Test de Wald

Considérons le test de : $\mathrm{H}_0 : \beta_1 = ... = \beta_p = 0$

contre

H_1 : au moins un des coefficients est différent de 0

Ce test peut s'écrire sous la forme :

$\mathrm{H}_0 : \psi(\beta) = 0$ contre $\mathrm{H}_1 : \psi(\beta) \neq 0$ avec $\psi : \left\{ \begin{array}{c} \mathbb{R}^{p+1} \to \mathbb{R}^p \\ \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \to \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \end{array} \right.$

La statistique de test de Wald s'écrit :

$$W = \psi(\hat{\beta})^t \left(\frac{\partial \psi}{\partial(\beta^t)}(\hat{\beta}) I_n(\hat{\beta})^{-1} \frac{\partial(\psi^t)}{\partial \beta}(\hat{\beta}) \right)^{-1} \psi(\hat{\beta}) \tag{2.15}$$

avec $I_n(\hat{\beta})$ définie par (2.13)

Cette statistique de test converge vers un $\chi^2(r)$ où r = rang$\left(\frac{\partial \psi}{\partial(\beta^t)}\right)$.
On rejette H_0 lorsque $W > F^{-1}_{\chi^2(r)}(1-\alpha)$ au niveau de risque de première espèce α.

∗ **Test d'adéquation de Pearson**

Considérons le test de : H_0 : adéquation du modèle aux données
contre
H_1 : non adéquation du modèle aux données

La statistique de test est définie de la manière suivante :

$$K^2 = \sum_{i=1}^{n} \omega_i \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

On a sous H_0 : $\frac{K^2}{\phi} \underset{\text{sous } H_0}{\xrightarrow{\mathscr{D}}} \chi^2(n-(p+1))$

On rejette H_0 lorsque $K^2 > F^{-1}_{\chi^2(n-(p+1))}(1-\alpha)$ au niveau de risque de première espèce α.

Ce test est utile seulement si ϕ est connu c'est à dire pour les modèles linéaires généralisés de Poisson, de Bernoulli et Binomial.

■ **Tests de sur-dispersion dans le cas de la régression de Poisson**

On s'intéresse à ces tests uniquement dans le cas de la régression de Poisson car en cas de sur-dispersion il existe une alternative qui n'existe pas dans le cas de la régression Binomial.

Lorsqu'un modèle de régression de Poisson est justifié on s'attend à ce que $\hat{\phi}$ approche la valeur attendue qui est 1.

On effectue un premier test, on teste $H_0 : \nu = 0$ contre $H_1 : \nu \neq 0$ dans le modèle binomial négatif dont la fonction de variance est $V(\mu) = \mu + \nu\mu^2$. Puis on effectue un deuxième test, on teste $H_0 : \nu = 1$ contre $H_1 : \nu \neq 1$ dans le modèle binomial négatif dont la fonction de variance est $V(\mu) = \nu\mu$.

On effectue ces tests grâce aux statistiques du score :

$$T_1^2 = \frac{\left(\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2 - (1-h_i)\hat{\mu}_i\right)^2}{2\sum_{i=1}^{n}\hat{\mu}_i{}^2} \text{ et } T_2^2 = \frac{1}{2n}\left(\sum_{i=1}^{n}\left(\frac{(Y_i - \hat{\mu}_i)^2 - (1-h_i)\hat{\mu}_i}{\hat{\mu}_i}\right)\right)^2$$

On a : T_1^2 et $T_2^2 \xrightarrow{\mathscr{D}} \chi^2(1)$

■ **Test de la significativité des prédicteurs**

Considérons le test de : $H_0 : \beta_1 = ... = \beta_s = 0$ avec $s < p$
contre
H_1 : au moins un des coefficients est différent de 0

La statistique de test est définie de la manière suivante :

$$F = \frac{(\mathcal{D}(\hat{\beta}_r) - \mathcal{D}(\hat{\beta}))/s}{(\mathcal{D}(\hat{\beta}))/(n-(p+1))} \overset{\text{sous } H_0}{\sim} F(s, n-(p+1))$$

On rejette H_0 lorsque $F > F^{-1}_{F(s,n-(p+1))}(1-\alpha)$ au niveau de risque de première espèce α.

On peut également tester ces hypothèses grâce au test de Wald défini ci-dessus.

2.2 Le modèle de régression Beta

Cette partie est principalement basée sur les références [3] et [11].

La loi beta a pour densité : $f(y,p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, p>0, q>0, y \in]0,1[$

Ferrari and Cribari-Neto ont proposé une paramétrisation différente : $\mu = \frac{p}{p+q}$ et $\phi = p+q$

$$f(y,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, 0<\mu<1, \phi>0, y \in]0,1[$$

de sorte que $\mathbb{E}[Y]=\mu$ et $Var(Y)= \frac{\mu(1-\mu)}{1+\phi} = \frac{V(\mu)}{1+\phi}$
La fonction de lien g est la fonction $logit$.

Cette densité ne se met pas sous la forme canonique (2.3), cette loi n'appartient pas à la famille exponentielle, ce n'est donc pas un modèle linéaire généralisé.

Cependant c'est un modèle régulier et on peut y faire de l'inférence par maximum de vraisemblance. On peut donc utiliser les mêmes outils que ceux pour les modèles linéaires généralisés excepté les outils basés sur la déviance qui utilise la loi exponentielle.

Le modèle de régression Beta est utilisé pour modéliser une variable réponse Y continue sur]0,1[. D'après Smithson and Verkuilen [11], si la variable Y prend des valeurs dans [a,b], on peut modéliser la variable $\frac{Y-a}{b-a}$ sur [0,1]. De plus, si Y prend des valeurs dans [0,1], on peut utiliser la transformation $\frac{Y\times(n-1)+0.5}{n}$ sur]0,1[, où n est la taille de l'échantillon.

Dans le cas de la régression Beta, les résidus de base sont également :

$$\widehat{\epsilon_i} = Y_i - \underbrace{\widehat{\mu_i}}_{\widehat{Y_i}} \tag{2.16}$$

Ces résidus sont hétéroscédastiques, on utilise alors :

- Les résidus de Pearson : $\widehat{\epsilon_i}^p = \frac{Y_i-\widehat{\mu_i}}{\sqrt{Var(\widehat{\mu_i})}}$
- Les résidus de Pearson studentisés : $\widehat{\epsilon_i}^{ps} = \frac{Y_i-\widehat{\mu_i}}{\sqrt{Var(\widehat{\mu_i})(1-h_i)}}$

avec h_i le $i^{ème}$ coefficient diagonal de H, la matrice des leviers, qui est définie par :
$H = W^{\frac{1}{2}}\mathbb{X}(\mathbb{X}'W\mathbb{X})^{-1}\mathbb{X}'W^{\frac{1}{2}}$ où $W^{\frac{1}{2}} \approx diag\left(V(\widehat{\mu_i})^{1/2}g'(\widehat{\mu_i})\right)^{-1}$

Tous les tests seront réalisés au seuil $\alpha = 5\%$.

Chapitre 3

Critères de comparaison des modèles

Les méthodes utilisées dans le but de comparer les modèles sur le jeu de données réel ainsi que sur une simulation d'un jeu de données seront expliqués dans une première partie. Dans une seconde partie, les méthodes appliquées à la comparaison des modèles sur les S simulations d'un jeu de données seront détaillées.

3.1 Critères de comparaison sur un jeu de données

Dans l'objectif de trouver le meilleur modèle pour nos données, nous avons utilisé plusieurs critères. Ils permettent de comparer les différents résultats obtenus sur les modèles étudiés.

3.1.1 Erreurs de prédictions

L'erreur absolue moyenne (EAM), qui est la moyenne arithmétique des valeurs absolues des écarts entre valeur prédite et valeur observée, est définie de la manière suivante :

$$\frac{1}{n}\sum_{i=1}^{n}|\widehat{Y}_i - Y_i| \tag{3.1}$$

L'erreur quadratique moyenne (EQM), qui est la moyenne arithmétique des carrés des écarts entre valeur prédite et valeur observée, est définie de la manière suivante :

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2 \tag{3.2}$$

Avec Y_i la valeur du score observée pour la i^e observation et $\widehat{Y}_i$ la valeur du score prédite pour la i^e observation.

Afin de quantifier les erreurs de prédiction, ces deux critères sont utilisés.

L'erreur absolue moyenne est à l'échelle de nos données, elle est donc simple à interpréter. L'erreur quadratique moyenne pénalise les écarts importants entre valeur observée et valeur prédite.

Plus ces erreurs sont faibles, plus les prédictions du modèle sont proches du score observé.

3.1.2 Prédictions

Les quantiles des prédictions du score, à 0%, 25%, 50%, 75%, 100% et également à 2.5% et 97.5%, sont comparés aux quantiles du score observé.
Les quantiles des prédictions à 2.5% et 97.5% permettent de ne pas tenir compte du minimum et du maximum de ces prédictions qui ne sont pas des valeurs très représentatives.

Dans le but de comparer les effectifs prédits et les effectifs observés pour les différentes valeurs du score, on arrondit à l'unité les prédictions obtenues, par un modèle, et on trace les diagrammes en bâtons de ces effectifs.

Un test du χ^2 d'adéquation entre les effectifs observés et les effectifs prédits est également effectué.

Test du χ^2 d'adéquation

La statistique du Khi-deux sert à mesurer l'écart qui existe entre la distribution des effectifs théoriques et la distribution des effectifs observés. Elle permet alors de tester si cet écart est suffisamment faible pour être imputable aux fluctuations d'échantillonnage.

On veut comparer les effectifs observés pour chaque valeur du score aux effectifs théoriques obtenus par un modèle. L'hypothèse testée est la suivante :

H_0 : la distribution observée est conforme à la distribution théorique
contre
H_1 : la distribution observée ne s'ajuste pas à la distribution théorique

La statistique de test du Khi-deux est la suivante : $\chi^2_{obs} = \sum_{k=1}^{K} \frac{(o_k - t_k)^2}{t_k}$

avec : - K le nombre de classes
- o_k l'effectif observé dans la k-ième classe
- t_k l'effectif théorique dans la k-ième classe

Pour que l'approximation par la loi du χ^2 soit correcte, il est nécessaire que les effectifs théoriques dans chacune des classes soit au moins égal à 5. Si ce n'est pas le cas, il faut au préalable regrouper les classes contigües afin d'avoir un effectif suffisant. La valeur du nombre de classe K intervenant dans le nombre de degrés de liberté de la loi du χ^2 est celle obtenue après les éventuels regroupements. Dans notre cas, une classe correspond à une valeur du score.

On a alors χ^2_{obs} qui suit une loi du χ^2 à K-1 degrés de liberté.
On rejette H_0 si $\chi^2_{obs} > F^{-1}_{\chi^2(K-1)}(1-\alpha)$ au niveau de risque de première espèce α.

3.1.3 Résidus

On trace les résidus de Pearson studentisés, détaillés dans la section 2.1.3. Pour le modèle linéaire Gaussien, 95% des résidus doivent se trouver dans une bande de confiance (-2,2) et être centrés et symétriques. De plus ils doivent suivre une loi normale et être homoscédastiques. Les tests vérifiant ces hypothèses, présentés dans la section 2.1.3, seront effectués sur ces résidus. Pour les autres modèles, les résidus de Pearson studentisés doivent être centrés et symétriques.

3.1.4 AIC et BIC

Le critère d'information d'Akaike (Akaike Information Criterion - AIC [2]) est défini par :

$$AIC = -2 \times log(L(\hat{\theta}, \mathbf{y})) + 2 \times k \tag{3.3}$$

avec L la vraisemblance du modèle et k le nombre de paramètres à estimer dans le modèle.

Quand le nombre de paramètres k est grand par rapport au nombre d'observations $\left(\frac{n}{k} < 40\right)$, il est recommandé d'utiliser l'AIC corrigé.

$$AIC_c = AIC + \frac{2 \times k \times (k+1)}{n - (k+1)}$$

avec k le nombre de paramètres à estimer dans le modèle et n la taille de l'échantillon.

Le critère d'information bayésien (Bayesian Information Criterion - BIC [2]) est défini par :

$$BIC = -2 \times log(L(\hat{\theta}, \mathbf{y})) + log(n) \times k \tag{3.4}$$

avec L la vraisemblance du modèle, n la taille de l'échantillon et k le nombre de paramètres à estimer dans le modèle.

On sélectionne le modèle ayant le plus petit AIC ou BIC. L'AIC et le BIC permettent de comparer des modèles proposant différentes lois pour Y. En revanche on ne peut pas comparer des modèles pour Y et pour une transformation de Y.

3.1.5 Puissance et risque d'erreur des tests

Les différentes situations que l'on peut rencontrer dans le cadre des tests d'hypothèse sont résumées dans le tableau suivant :

Réalité hypothétique / Décision	H_0 vraie	H_1 vraie
H_0 acceptée	Pas d'erreur	Acceptation de H_0 à tort Risque de 2^{nd} espèce β
H_1 acceptée	Rejet de H_0 à tort Risque de $1^{ère}$ espèce α	Pas d'erreur Puissance du test 1-β

On appelle risque d'erreur de première espèce, noté α, la probabilité de rejeter H_0 à tort c'est à dire la probabilité de rejeter H_0 et d'accepter H_1 alors que H_0 est vraie.

On appelle risque d'erreur de seconde espèce, noté β, la probabilité d'accepter H_0 à tort c'est à dire la probabilité d'accepter H_0 alors que H_1 est vraie.

On appelle puissance d'un test la probabilité de rejeter H_0 à raison c'est à dire la probabilité de rejeter H_0 et d'accepter H_1 alors que H_1 est vraie. Sa valeur est 1-β.

Si lors d'un test on espère que H_1 soit vraie, on espère rejeter H_0 et donc que la puissance du test soit égale à 1. Lorsque le test permet de rejeter H_0 et donc d'accepter H_1, la puissance du test sera égale à 1. En revanche, si le test ne permet pas de rejeter H_0, la puissance du test sera égale à 0.

A l'inverse, si lors d'un test on espère que H_0 soit vraie, on espère accepter H_0 et donc que le risque d'erreur de première espèce du test soit égal à 0. Lorsque le test ne permet pas de rejeter H_0, le risque d'erreur de première espèce du test sera égal à 0. En revanche, si le test permet de rejeter H_0, le risque d'erreur de première espèce du test sera égal à 1.

Nous nous intéressons ici à différents tests, détaillés dans la section 2.1.3.

Pour tous les modèles, le test de la déviance ou le test de Wald est utilisé pour tester la significativité globale de la régression. On s'attend à ce que l'hypothèse H_1 soit vraie et donc que la puissance du test soit égale à 1. Ces tests sont également utilisés pour tester la significativité des différents coefficients du modèle. Si on pense qu'une variable explicative a de l'influence sur la variable réponse, on s'attend à ce que l'hypothèse H_1 soit vraie et donc que la puissance du test soit égale à 1. Si au contraire, on pense qu'une variable explicative n'a pas d'influence sur la variable réponse, on s'attend à ce que l'hypothèse H_0 soit vraie et donc que le risque d'erreur de première espèce du test soit égal à 0.

De plus, pour le modèle linéaire Gaussien, on effectue le test de Shapiro-Wilk pour tester la normalité des résidus de Pearson studentisés et le test de Breusch-Pagan pour tester leurs homoscédasticité. Pour le modèle linéaire généralisé de Poisson, on effectue des tests de sur-dispersion. Enfin, pour les modèles linéaires généralisés de Poisson et Binomial, on effectue également le test d'adéquation de Pearson. Pour tous ces tests, on s'attend à ce que l'hypothèse H_0 soit vraie et donc que le risque d'erreur de première espèce de ces tests soit égal à 0.

3.2 Critères de comparaison sur les S simulations

Afin de résumer les comparaisons des modèles sur les S simulations du point de vue des critères détaillés dans la section précédente, on va expliciter les méthodes utilisées.

3.2.1 Erreurs de prédictions

On calcule la moyenne des erreurs absolues moyennes, sur les S simulations, de la façon suivante :

$$\frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{n}\sum_{i=1}^{n}|\widehat{Y_{si}} - Y_{si}|\right)$$

On calcule également la moyenne des erreurs quadratiques moyennes, sur les S simulations, de la façon suivante :

$$\frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{Y_{si}} - Y_{si})^2\right)$$

Avec Y_{si} la valeur du score observée pour la i^e observation de la s^e simulation et $\widehat{Y_{is}}$ la valeur du score prédite pour la i^e observation de la s^e simulation.

On récupère également sur chaque simulation le modèle qui a l'erreur absolue moyenne la plus faible ainsi que le modèle qui a l'erreur quadratique moyenne la plus faible. On obtient sur les S simulations le nombre de fois où chaque modèle a les différentes erreurs les plus faibles.

3.2.2 Prédictions

On récupère sur chaque simulation les quantiles des prédictions du score, aux différents niveaux. Pour chaque modèle, on a S quantiles par niveau. Cela permet d'obtenir la moyenne des quantiles des prédictions du score, aux différents niveaux, calculée sur les S simulations. On pourra les comparer aux quantiles du score simulé.

On calcule également sur chaque simulation les effectifs prédits pour les différentes valeurs du score. Pour chaque modèle, on a S effectifs par valeur du score. Cela permet, pour chaque modèle, de tracer les boîtes à moustaches des effectifs prédits pour les différentes valeurs du score obtenus sur les S simulations. On pourra les comparer aux effectifs des valeurs du score simulé.

Sur chaque simulation on récupère le nombre de fois où chaque modèle obtient l'adéquation du test du χ^2, c'est à dire le nombre de fois où l'hypothèse H_0 du test du χ^2 est vraie.

3.2.3 Résidus

On récupère sur chaque simulation les résidus de Pearson studentisés pour chaque modèle. Ces résidus doivent être centrés et symétriques.

Dans le but d'observer s'ils sont centrés, on calcule la moyenne des résidus pour chaque simulation. A partir des S moyennes de résidus, on obtient le minimum, le 1^{er} quartile, la médiane, la moyenne, le $3^{ème}$ quartile et le maximum de ces S moyennes des résidus.

Afin d'observer s'ils sont symétriques, on calcule les quantiles à 2.5% et 97.5% des résidus sur chaque simulation. On obtient le minimum, le 1^{er} quartile, la médiane, la moyenne, le $3^{ème}$ quartile et le maximum des quantiles à 2.5% et 97.5% des résidus sur les S simulations.

3.2.4 AIC et BIC

Sur chacune des simulations on récupère le modèle qui a l'AIC le plus faible ainsi que le modèle qui a le BIC le plus faible. On obtient alors, sur les S simulations, le nombre de fois où chaque modèle a ces deux critères les plus faibles.

3.2.5 Puissance et risque d'erreur empirique des tests

Sur les S simulations, on calcule la puissance empirique ou le risque d'erreur empirique de première espèce des différents tests. Pour cela on calcule la moyenne des puissances et des risques d'erreur de première espèce obtenus sur les S simulations. On espère obtenir une puissance empirique égale à 1 et un risque d'erreur empirique de première espèce égal à 0.

Pour tous les modèles, on s'intéresse à la puissance empirique des tests de la déviance ou de Wald pour tester la significativité globale de la régression. On s'intéresse également à la puissance empirique ou au risque d'erreur empirique de première espèce (suivant les cas) de ces tests pour tester la significativité des coefficients du modèle. On calcule également le risque d'erreur empirique de première espèce des tests de Shapiro-Wilk et de Breusch-Pagan pour le modèle linéaire Gaussien, des tests de sur-dispersion pour le modèle linéaire généralisé de Poisson ainsi que du test de Pearson pour les modèles linéaires généralisés de Poisson et Binomial.

Chapitre 4

Étude d'un jeu de données réel

Nous nous intéressons à une étude qui a mesuré (par autoquestionnaire) l'implication des médecins généralistes dans le dépistage précoce du diabète de l'enfant, sur une échelle discrète de 0 à 10. La valeur 0 signifie que le médecin ne se sent pas du tout impliqué dans le dépistage précoce du diabète de l'enfant et la valeur 10 signifie au contraire qu'il se sent entièrement impliqué. Des questions de connaissances sur le diabète ont été posées au médecin et 287 questionnaires ont été récoltés. Le but de cette étude est de relier l'implication des médecins généralistes dans le dépistage précoce du diabète de l'enfant aux réponses à des questions de connaissance sur le diabète. Notre objectif est de définir, parmi les différents modèles étudiés, le meilleur modèle pour ce jeu de données.

Afin de pouvoir étudier ce jeu de données avec le modèle linéaire généralisé Gamma, on effectue un décalage des données de [0,10] dans [1,11]. De même pour pouvoir étudier ce jeu de données avec le modèle de régression Beta, on effectue une transformation des données de [0,10] dans]0,1[comme détaillée dans la section 2.2. Par conséquent, comme expliqué dans la section 3.1.4, on ne pourra pas comparer ces deux modèles aux autres du point de vue des critères de l'AIC et du BIC.

Pour ces modèles, une fois les prédictions obtenues nous effectuons les transformations inverses sur celles-ci dans le but de pouvoir comparer ces prédictions à celles des autres modèles.

La variable réponse est un score, qui peut prendre des valeurs de 0 à 10, réparti de la façon suivante :

Valeurs du score	0	1	2	3	4	5	6	7	8	9	10
Effectifs	3	6	9	11	14	20	26	57	69	43	29

On notera Y_i le score du médecin n°i avec $i = 1, ..., 287$.

Les variables explicatives sont au nombre de 6 et sont explicitées ci-dessous :

$X_i^{(1)}$ = âge du i^e médecin

$$X_i^{(2)} = \begin{cases} 0 & \text{si le } i^e \text{ médecin suit en majorité des patients diabètiques depuis moins de 5 ans} \\ 1 & \text{si le } i^e \text{ médecin suit en majorité des patients diabètiques depuis plus de 5 ans} \end{cases}$$

$$X_i^{(3)} = \begin{cases} 0 & \text{si le } i^e \text{ médecin estime que la liste de symptômes qu'on lui présente nécessite la mise en place d'un suivi du patient} \\ 1 & \text{si le } i^e \text{ médecin estime que la liste de symptômes qu'on lui présente ne nécessite pas la mise en place d'un suivi du patient} \end{cases}$$

$$X_i^{(4)} = \begin{cases} 0 & \text{si le } i^e \text{ médecin effectue l'examen adapté dans un délai adéquat} \\ 1 & \text{si le } i^e \text{ médecin effectue l'examen adapté dans un délai non adéquat} \\ 2 & \text{si le } i^e \text{ médecin n'effectue pas l'examen adapté} \end{cases}$$

$$X_i^{(5)} = \begin{cases} 0 & \text{si le } i^e \text{ médecin confirme le diagnostic par une analyse d'urine et un autre examen} \\ 1 & \text{si le } i^e \text{ médecin confirme le diagnostic uniquement par une analyse d'urine} \end{cases}$$

$$X_i^{(6)} = \begin{cases} 0 & \text{si le } i^e \text{ médecin prescrit l'analyse d'urine immédiatement} \\ 1 & \text{si le } i^e \text{ médecin prescrit l'analyse d'urine pour le lendemain matin} \\ 2 & \text{si le } i^e \text{ médecin prescrit l'analyse d'urine dans la semaine} \end{cases}$$

La variable $X_i^{(1)}$ est quantitative alors que les variables $X_i^{(2)}$, $X_i^{(3)}$, $X_i^{(4)}$, $X_i^{(5)}$ et $X_i^{(6)}$ sont toutes qualitatives à respectivement 2, 2, 3, 2 et 3 modalités.

Nous définissons nos modèles de la façon suivante :

- indépendances entre les individus : $(Y_i, X_i^{(1)}, ..., X_i^{(6)})$ sont mutuellement indépendants pour $i = 1, ... , 287$
- $\mathscr{L}(Y_i|X_i^{(1)}, ..., X_i^{(6)})$ appartient à la famille exponentielle
- les covariables $X_i^{(1)}, ..., X_i^{(6)}$ influent sur le niveau moyen conditionnel de la variable réponse au travers d'un prédicteur linéaire

$$\begin{aligned} \eta_i &= \beta_0 + \beta_1 \times X_i^{(1)} + \beta_2 \times \mathbb{1}(X_i^{(2)} = 1) + \beta_3 \times \mathbb{1}(X_i^{(3)} = 1) \\ &\quad + \beta_{4,1} \times \mathbb{1}(X_i^{(4)} = 1) + \beta_{4,2} \times \mathbb{1}(X_i^{(4)} = 2) + \beta_5 \times \mathbb{1}(X_i^{(5)} = 1) \\ &\quad + \beta_{6,1} \times \mathbb{1}(X_i^{(6)} = 1) + \beta_{6,2} \times \mathbb{1}(X_i^{(6)} = 2) \end{aligned} \tag{4.1}$$

- il existe une fonction de lien g qui relie le prédicteur linéaire au niveau moyen conditionnel de la variable réponse

$$g(\underbrace{\mathbb{E}(Y_i|X_i^{(1)}, ..., X_i^{(6)})}_{\mu_i}) = \eta_i \tag{4.2}$$

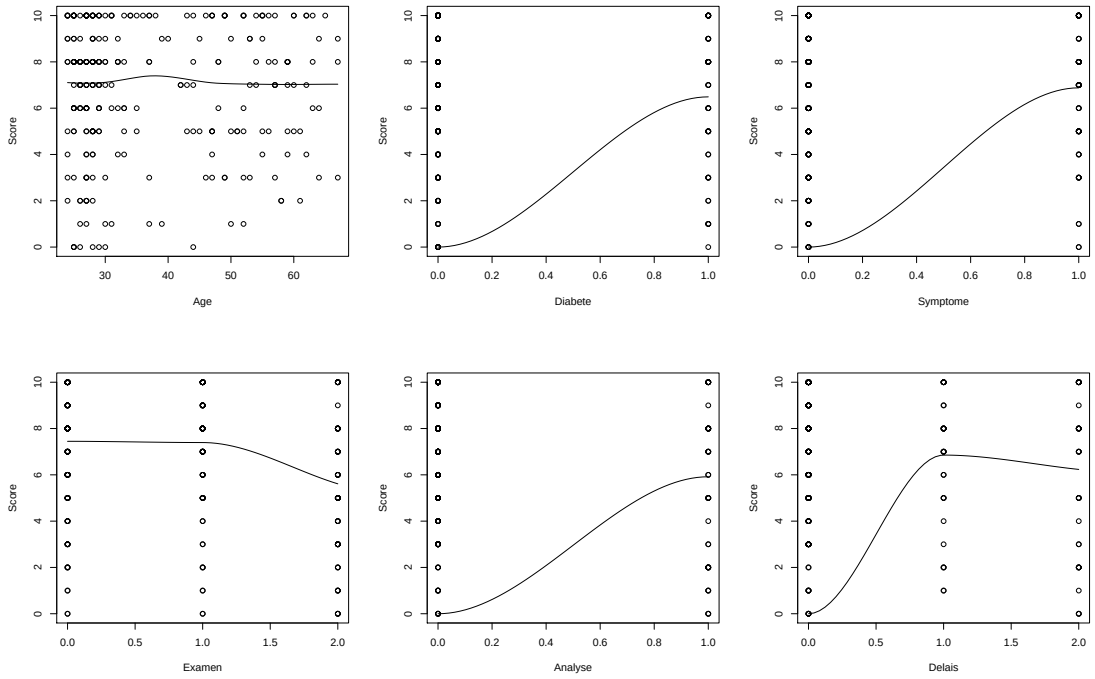
$(\star)$

Il y a 9 paramètres à estimer. Nous n'introduisons pas d'interactions pour ne pas avoir trop de paramètres à estimer par rapport à la taille de l'échantillon.

Nous appellerons la variable Y_i *Score*, la variable $X_i^{(1)}$ *Age*, la variable $X_i^{(2)}$ *Diabete*, la variable $X_i^{(3)}$ *Symptome*, la variable $X_i^{(4)}$ *Examen*, la variable $X_i^{(5)}$ *Analyse* et la variable $X_i^{(6)}$ *Delais*.

4.1 Analyse graphique exploratoire des données

Tout d'abord, nous nous intéressons aux tracés exploratoires des différentes variables explicatives en fonction de la variable réponse *Score*.



Les variables explicatives semblent avoir de l'influence sur la variable *Score*. Les variables *Age* et *Examen* semblent avoir moins d'influence sur la variable *Score* que les variables *Diabete*, *Symptome*, *Analyse* et *Delais*. Excepté la variable *Age*, ce sont toutes des variables qualitatives, une transformation fonctionnelle de ces prédicteurs n'est donc pas envisageable.

4.2 Résultats obtenus pour les différents modèles

Tous les modèles s'obtiennent à partir de $(\star)$ en précisant que la loi de Y_i conditionnellement aux covariables suit

- une loi Normale pour le modèle linéaire Gaussien
- une loi de Poisson pour le modèle linéaire généralisé de Poisson
- une loi Gamma pour le modèle linéaire généralisé Gamma
- une loi Binomiale pour le modèle linéaire généralisé Binomial
- une loi Beta pour le modèle de régression Beta

Diagnostics graphiques des résidus de Pearson studentisés

Les graphiques des résidus de Pearson studentisés des différents modèles se trouvent en annexe B.1. Pour les modèles linéaires généralisés Gaussien, Poisson, Gamma et Binomial ainsi que pour le modèle de régression Beta les résidus sont plutôt bien placés dans des bandes de confiance à différents niveaux. Ils semblent centrés et symétriques.

En ce qui concerne les variables, certaines semblent influer sur le niveau moyen des résidus. Les variables explicatives qui ont le plus d'influence sur ce niveau moyen sont détaillées, en fonction de différents modèles, dans le tableau ci-dessous :

Modèle	Gaussien	Poisson	Gamma	Binomial	Beta
Variables influentes	*Diabete* *Examen* *Delais*	*Diabete* *Examen* *Analyse* *Delais*	*Diabete* *Symptome* *Delais*	*Symptome* *Delais*	*Diabete* *Examen* *Analyse* *Delais*

Dans le cas du modèle linéaire Gaussien, d'après le graphique en annexe B.1. l'hypothèse de normalité des résidus ne semble pas aberrante. Afin de tester la normalité des résidus, on utilise le test de Shapiro-Wilk, détaillé dans la partie 2.1.3. On obtient une p-valeur égale à $2{,}2^{-7}$ qui est inférieure à α. On rejette donc l'hypothèse nulle H_0 au risque de première espèce α et on accepte l'hypothèse alternative H_1. Les résidus ne suivent pas une loi normale.

On effectue également un test d'homoscédasticité des résidus grâce au test de Breusch-Pagan, détaillé dans la partie 2.1.3. On obtient une p-valeur égale à 0.27 qui est supérieure à α. On ne rejette pas l'hypothèse nulle H_0. Les résidus sont homoscédastiques.

Test d'adéquation globale du modèle aux données

Pour chaque modèle, on teste

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_{4,1} = \beta_{4,2} = \beta_5 = \beta_{6,1} = \beta_{6,2} = 0$$
contre
$$H_1 : \text{au moins un des coefficients est différent de 0}$$

Pour les modèles linéaires généralisés, ce test est réalisé au moyen du test de la déviance et pour le modèle de régression Beta, ce test est réalisé au moyen du test de Wald . Ces tests sont détaillés dans la partie 2.1.3.

Pour tous les modèles, les p-valeurs obtenues sont inférieures à α. On rejette l'hypothèse nulle H_0 et on accepte l'hypothèse alternative H_1 au risque de première espèce α. Cela nous conforte dans l'idée que les variables explicatives ont une certaine influence sur le *Score*.

Test d'adéquation de Pearson

Pour les modèles linéaires généralisés de Poisson et Binomial, on teste

$$H_0 : \text{adéquation du modèle aux données}$$
contre
$$H_1 : \text{non adéquation du modèle aux données}$$

Les statistiques de test K^2 sont égales à 203 pour le modèle linéaire généralisé de Poisson et à 290 pour le modèle linéaire généralisé Binomial. Ces statistiques sont inférieures à 317, le quantile du χ^2 à $n - (p + 1)$ soit 278 degrés de liberté. On ne rejette donc pas H_0 et on conclut à l'adéquation des modèles aux données.

Tests de sur-dispersion

Pour le modèle linéaire généralisé de Poisson on effectue les deux tests de sur-dispersion détaillés dans la partie 2.1.3 de ce rapport. On obtient pour le 1^{er} test une p-valeur égale à 0.27 et pour le 2^e test une p-valeur égale à 0.43. Ces deux p-valeurs sont supérieures à α. On ne rejette pas l'hypothèse nulle H_0.

Sélection de variables

Pour tester si les variables $X_i^{(1)}$, $X_i^{(2)}$, $X_i^{(3)}$ et $X_i^{(5)}$ (quantitatives ou qualitatives à deux modalités) ont un effet sur la variable réponse, on teste pour $j = 1$, 2, 3 et 5, les hypothèses :

$$H_0 : \beta_j = 0$$
contre
$$H_1 : \beta_j \neq 0$$

Pour tester si les variables $X_i^{(4)}$ et $X_i^{(6)}$ (qualitatives à trois modalités) ont un effet sur la variable réponse, on teste pour $j = 4$ et 6, les hypothèses :

$$H_0 : \beta_{j,1} = \beta_{j,2} = 0$$
contre
$$H_1 : \text{au moins un des deux coefficients est différent de 0}$$

Pour les modèles linéaires généralisés, ces tests sont réalisés au moyen du test de la déviance et au moyen du test de Wald pour le modèle de régression Beta.

Pour chacun des modèles, après sélection de variables seules les variables $X_i^{(3)}$ *Symptome*, $X_i^{(4)}$ *Examen* et $X_i^{(5)}$ *Analyse* restent dans le modèle.

Choix de la fonction de lien

Chaque modèle a été ajusté avec différentes fonction de lien. On retient celui qui a une déviance minimale. Excepté pour le modèle linéaire généralisé Gamma où la fonction de lien *logarithme* est retenue, pour les autres modèles les fonctions de lien retenues sont celles définies dans la section 2.1.1.

Les résultats suivants sont obtenus à partir des modèles après sélection de variables et choix de la fonction de lien.

Erreurs de prédictions

	Gaussien	Poisson	Gamma	Binomial	Beta
EAM	1.69	1.69	1.69	1.68	1.70
EQM	2.74	2.79	2.77	2.72	2.79

Les erreurs absolues moyennes sont assez conséquentes mais les différences entre les modèles sont très faibles. Les erreurs quadratiques moyennes sont également assez importantes, cependant on observe cette fois-ci des différences plus marquées entre les modèles.

Bien que les différences observées entre les modèles du point de vue de ces deux erreurs sont faibles, il est à noter que le modèle linéaire généralisé Binomial a tout de même l'erreur absolue moyenne et l'erreur quadratique moyenne les plus faibles.

Prédictions

Les quantiles des prédictions du score, à 0%, 25%, 50%, 75%, 100% et également à 2.5% et 97.5%, obtenus pour les différents modèles, sont les suivants :

	0%	25%	50%	75%	100%	2.5%	97.5%
Score observé	0.00	6.00	7.00	8.5	10.00	1.00	10.00
Gaussien	1.15	6.43	7.19	8.38	9.86	2.88	9.46
Poisson	2.18	6.46	7.21	7.40	10.19	2.67	8.57
Gamma	2.01	6.51	7.19	7.40	10.03	2.64	8.24
Binomial	1.82	6.42	7.03	7.28	9.80	2.75	8.05
Beta	1.82	6.42	7.03	7.28	9.80	2.75	8.05

Nous comparons les quantiles des scores prédits aux quantiles du score simulé.

On observe que les modèles ne prédisent quasiment aucune valeur faible du score. En effet pour le quantile des prédictions à 0%, les valeurs dépassent fortement 0.

Pour le quantile des prédictions à 100%, les valeurs semblent acceptables pour tous les modèles. Les quantiles à 25%, 50% et 75% semblent convenables pour tous les modèles.

Pour le quantile des prédictions à 2.5% et à 97.5%, les valeurs obtenues pour les modèles sont fortement éloignés des valeurs pour les données réelles.

AIC - BIC

	Gaussien	Poisson	Binomial
AIC	1251	1334	1299
BIC	1277	1355	1321

Le modèle linéaire Gaussien a l'AIC et le BIC les plus faibles.

4.3 Conclusions

Les variables qui influent le plus sur l'implication des médecins généralistes dans le dépistage précoce du diabète de l'enfant sont les variables *Symptome*, *Examen* et *Analyse*. Tous les modèles donnent les mêmes résultats.

L'étude de ce jeu de données ne permet pas de mettre réellement en avant un modèle. L'erreur absolue moyenne et l'erreur quadratique moyenne sont quasiment équivalentes quelque soit le modèle avec un avantage pour le modèle linéaire généralisé Binomial. Du point de vue de l'AIC et du BIC, le modèle linéaire Gaussien est le meilleur.

Les prédictions obtenues, pour chacun des modèles, sont assez éloignées des données réelles. Les variables ne semblent pas expliquer toute l'information sur le score.

Le travail réalisé sur ce jeu de données réel a permis de se faire une première idée concernant les modèles et d'étudier un cas concret de score. Afin de confirmer ou de contredire ces premiers résultats nous avons travaillé sur des simulations que nous allons présenter dans le prochain chapitre.

Chapitre 5

Comparaison des modèles sur des scores simulés

Nous décidons de simuler des scores entiers de 0 à 10 et de taille d'échantillon égale à 300, car c'est le type de score le plus fréquent. Nous choisissons 6 scénarios de simulation pour ces scores. Nous simulons des scores avec une moyenne faible, une moyenne intermédiaire et une moyenne élevée, avec les données centrées ou étalées autour de cette moyenne pour chacun des 3 cas. Chaque score est simulé en effectuant un tirage aléatoire des proportions des valeurs du score, de telle sorte que la somme des proportions soit égale à 1 et que la moyenne et la dispersion des données soient celles attendues.

Nous travaillons également sur un 7^e scénario, un score simulé avec une distribution des valeurs identique à celle du jeu de données réel étudié précédemment, c'est à dire que les proportions des effectifs pour chaque valeur du score ont été conservées.

Au total, cela fait 7 scénarios de simulation de score. Nous décidons d'étudier ces différents scénarios afin de voir si le comportement des modèles est le même quelque soit la distribution du score étudié.

En ce qui concerne les variables explicatives nous en simulons deux :

- une variable liée au score de la façon suivante $X_i^{(1)} \sim Y_i + \mathcal{N}(0,1)$, de manière à avoir une variable explicative égale au score masquée par un bruit
- une variable indépendante du score de la façon suivante $X_i^{(2)} \sim \mathcal{N}(0,1)$

D'autres bruits suivant une loi Normale avec d'autres paramètres ou suivant d'autres lois ont été testés. La loi suivie par la $1^{\text{ère}}$ variable a une légère influence sur les résultats obtenus pour les modèles mais les conclusions tirées sur la comparaison des modèles sont néanmoins les mêmes. La loi suivie par la $2^{ième}$ variable n'a pas d'influence sur les résultats.

Pour chaque scénario, une fois le score simulé nous le fixons et nous simulons S fois les variables explicatives. Donc pour chaque scénario nous étudions S fois chaque modèle. Nous fixons S = 1000.

Les modèles sont définis à partir de la spécification d'un modèle linéaire généralisé dans la section 2.1. avec $n = 300$ et $p = 2$.

Comme détaillé page 21 de ce rapport, les différents modèles s'obtiennent en précisant la loi de Y conditionnellement aux variables $X^{(1)}$ et $X^{(2)}$.

Les transformations des scores nécessaires à l'utilisation du modèle linéaire généralisé Gamma et du modèle de régression Beta sont réalisées comme expliqué dans le chapitre précédent.

Le travail effectué sur les différents modèles sera détaillé pour un scénario de simulation de score et les résultats obtenus sur les autres scénarios de simulation de score seront résumés.

5.1 Étude détaillée pour un scénario de score simulé

L'étude détaillée porte sur le 7^e scénario, un score similaire à celui étudié sur le jeu de données réel. Ce score a une moyenne élevée égale à 6,96 et des valeurs davantage centrées sur la moyenne. Il est distribué de la façon suivante :

Valeurs	0	1	2	3	4	5	6	7	8	9	10
Effectifs	3	6	9	12	15	21	27	60	72	45	30

Dans une première partie, les résultats obtenus sur une simulation seront explicités, puis les résultats obtenus sur les S simulations seront détaillés dans une deuxième partie.

5.1.1 Résultats obtenus sur une simulation

Erreurs de prédictions

Les erreurs absolues moyennes et les erreurs quadratiques moyennes obtenues, pour les différents modèles, sont les suivantes :

	Gaussien	Poisson	Gamma	Binomial	Beta
EAM	0.67	0.86	0.92	0.65	0.67
EQM	0.70	1.10	1.28	0.67	0.71

Ces deux erreurs sont les plus faibles pour le modèle linéaire généralisé Binomial.

Ces erreurs sont bien évidemment plus faibles que celles obtenues sur le jeu de données réel car une des deux variables explicatives contient toute l'information sur le score recouvert par un bruit. De plus, les différences que l'on observe entre les modèles sont plus importantes que celles observées sur le jeu de données réel.

Prédictions

Les quantiles des prédictions du score, à 0%, 25%, 50%, 75%, 100% et également à 2.5% et 97.5%, obtenus pour les différents modèles, sont les suivants :

	0%	25%	50%	75%	100%	2.5%	97.5%
Score simulé	0.00	6.00	7.00	8.25	10.00	1.00	10.00
Gaussien	0.05	5.77	7.28	8.49	10.52	1.85	10.25
Poisson	1.96	5.42	6.94	8.47	12.84	2.94	11.29
Gamma	1.47	5.26	6.99	8.75	13.65	2.42	11.93
Binomial	0.49	5.88	7.61	8.59	9.66	1.64	9.60
Beta	0.34	5.85	7.78	8.75	9.76	1.37	9.53

Nous comparons les quantiles des scores prédits aux quantiles du score simulé.

Pour les quantiles des prédictions à 0% et à 100%, les valeurs semblent convenables pour les modèles Gaussien, Binomial et Beta, en revanche pour les modèles Poisson et Gamma les valeurs sont trop éloignées des valeurs du score simulé (respectivement 0 et 10). Pour les quantiles des prédictions à 25%, 50% et 75%, les valeurs semblent acceptables pour tous les modèles.

Pour les quantiles des prédictions à 2.5% et à 97.5%, les valeurs semblent une nouvelle fois raisonnables pour les modèles Gaussien, Binomial et Beta. Au contraire pour les modèles Poisson et Gamma, les valeurs ne sont de nouveau pas assez proches des valeurs du score simulé (respectivement 1 et 10).

Les modèles linéaire Gaussien, linéaire généralisé Binomial et de régression Beta donnent donc des prédictions convenables pour le score. A l'inverse, les modèles linéaires généralisés de Poisson et Gamma ne semblent pas donner des prédictions raisonnables pour le score.

Afin d'obtenir les effectifs des prédictions pour les différentes valeurs du score, les prédictions des différents modèles sont arrondies à l'unité. Les effectifs ainsi obtenus sont les suivants :

Valeurs du Score	0	1	2	3	4	5	6	7	8	9	10
Score simulé	3	6	9	12	15	21	27	60	72	45	30
Gaussien	1	4	9	11	17	21	36	65	70	41	27
Poisson	0	0	5	10	26	38	39	68	46	30	38
Gamma	0	2	5	16	23	36	37	60	37	32	52
Binomial	1	5	7	13	16	17	31	63	72	58	17
Beta	1	7	8	12	15	17	27	55	70	73	15

Les effectifs du modèle linéaire Gaussien sont assez proches des effectifs du score simulé. En revanche les effectifs des modèles linéaires généralisés de Poisson et Gamma sont trop éloignés des effectifs du score simulé. Enfin les effectifs du modèle linéaire généralisé Binomial et du modèle de régression Beta semblent assez proches des effectifs du score simulé excepté pour les scores 9 et 10.

Ces deux lois sont bornées, il n'y a donc pas de prédictions du score supérieures à 10. En arrondissant ces prédictions à l'unité cela explique le fait qu'il y ait peu d'effectif pour le score 10 et par conséquent plus d'effectif pour le score 9.

Les effectifs prédits du modèle linéaire Gaussien sont assez proches des effectifs du score simulé malgré le fait que la loi Normale soit continue et non bornée.

Nous traçons les diagrammes en bâtons de ces effectifs afin d'en avoir une représentation visuelle :

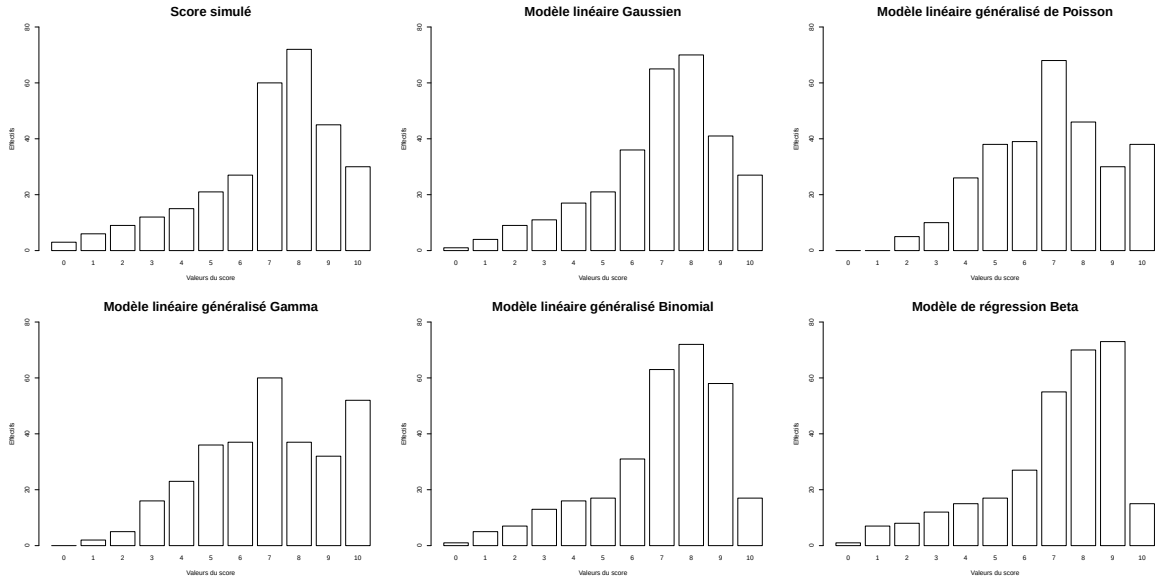


FIGURE 5.1 – Diagrammes en bâtons des effectifs des scores prédits

- Le diagramme des effectifs des prédictions du modèle linéaire Gaussien a la même forme que celui des effectifs du score simulé.

- Les diagrammes des modèles linéaires généralisés de Poisson et Gamma ont des formes similaires qui ne se rapprochent pas de celle du diagramme des effectifs du score simulé.

- Les diagrammes du modèle linéaire généralisé Binomial et du modèle de régression Beta ont une forme qui se rapproche de celui des effectifs du score simulé, excepté pour les effectifs des scores 9 et 10. En effet, pour le modèle linéaire généralisé Binomial, l'effectif prédit du score 10 est trop faible et l'effectif prédit du score 9 est trop élevé. Pour le modèle de régression Beta, l'effectif prédit du score 9 est supérieur à celui du score 8 et l'effectif prédit du score 10 est trop faible. Cependant nous pouvons dire que la forme du diagramme du modèle linéaire généralisé Binomial se rapproche plus de celle du score simulé que la forme du diagramme du modèle de régression Beta.

Le test du χ^2 d'adéquation des effectifs simulés et prédits conclut à l'adéquation pour les modèles Gaussien et Binomial et à la non-adéquation pour les modèles Poisson, Gamma et Beta.

Du point de vue des prédictions, le modèle linéaire Gaussien et le modèle linéaire généralisé Binomial se détachent clairement des autres. Bien que que la loi Normale soit continue et non bornée, le score prédit par le modèle linéaire Gaussien est assez proche du score simulé.

Un avantage se fait pour le modèle linéaire Binomial en ce qui concerne les prédictions telles quelles et un avantage pour le modèle linéaire Gaussien pour les effectifs des prédictions.

Résidus

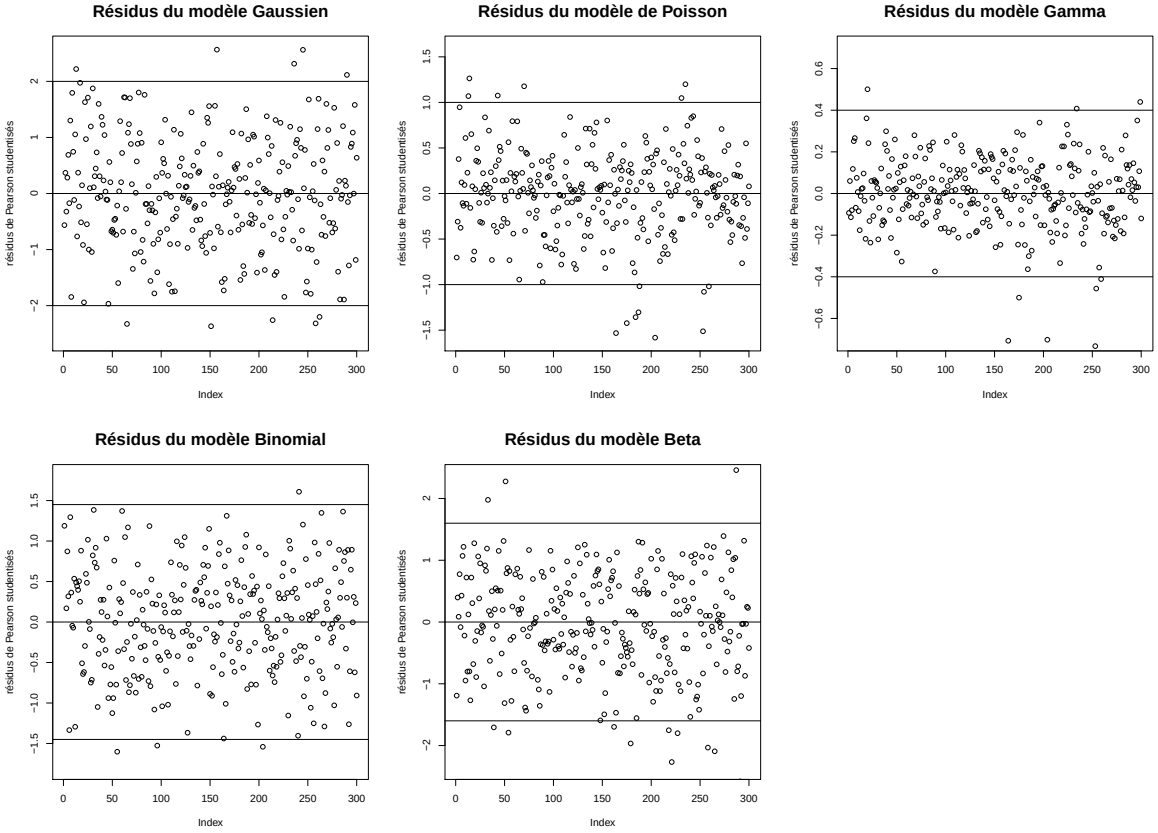


FIGURE 5.2 – Graphiques des résidus de Pearson studentisés

Les conclusions sur les résidus de Pearson studentisés, pour les différents modèles, sont les suivantes :

- Gaussien : ils sont bien placés dans une bande de confiance (-2,2), centrés et symétriques
- Poisson : ils sont centrés mais pas parfaitement symétriques, ils semblent se disperser du côté négatif
- Gamma : ils sont centrés mais pas parfaitement symétriques, trois résidus s'éloignent du côté négatif
- Binomial : ils sont centrés et symétriques
- Beta : ils sont centrés mais pas parfaitement symétriques, quelques résidus s'éloignent du côté négatif

AIC - BIC

	Gaussien	Poisson	Binomial
AIC	789	1192	828
BIC	804	1203	839

L'AIC et le BIC sont les plus faibles pour le modèle linéaire Gaussien.

Puissance et risque d'erreur des différents tests

La puissance des tests de la déviance ou de Wald de la significativité globale de la régression est égale à 1, pour tous les modèles, ce qui est très satisfaisant. La puissance des ces tests pour la significativité du premier paramètre est égale à 1 et le risque d'erreur de première espèce de ces tests pour la significativité du deuxième paramètre est égal à 0, pour tous les modèles, ce qui est très satisfaisant également.

Les risques d'erreur de première espèce du test de Shapiro-Wilk et celui du test de Breusch-Pagan, pour le modèle linéaire Gaussien sont égaux à 0, ainsi que ceux des deux tests de sur-dispersion, pour le modèle linéaire généralisé de Poisson. Il en va de même pour le test d'adéquation de Pearson, ils sont nuls pour les modèles linéaires généralisés de Poisson et Binomial. Tous ces résultats sont très satisfaisant.

Du point de vue des tests, aucun modèle ne se détache des autres.

Conclusions sur une simulation

Sur une simulation, le modèle linéaire Gaussien est meilleur pour les critères de l'AIC et du BIC et le modèle linéaire généralisé Binomial est meilleur en ce qui concerne les erreurs absolues moyennes et les erreurs quadratiques moyennes. Du point de vue des prédictions, du test du χ^2 d'adéquation et des résidus les deux modèles semblent équivalents.

5.1.2 Résultats obtenus sur S simulations

Erreurs de prédictions

Sur les S simulations, la moyenne des erreurs absolues moyennes et la moyenne des erreurs quadratiques moyennes obtenues, pour les différents modèles, sont les suivantes :

	Gaussien	Poisson	Gamma	Binomial	Beta
EAM	0.73	0.89	0.94	0.70	0.72
EQM	0.83	1.26	1.45	0.77	0.81

En moyenne sur les S simulations, l'erreur absolue moyenne la plus faible et l'erreur quadratique moyenne la plus faible s'obtiennent avec le modèle linéaire généralisé Binomial.

Sur les S simulations, on obtient le pourcentage de fois où chacun des modèles a l'erreur absolue moyenne ainsi que l'erreur quadratique moyenne la plus faible :

	Gaussien	Poisson	Gamma	Binomial	Beta
EAM	0.6 %	0 %	0 %	99.4 %	0.1 %
EQM	3.8 %	0 %	0 %	96.1 %	0 %

Le modèle Binomial a l'erreur absolue moyenne la plus faible dans 99.4% des cas et l'erreur quadratique moyenne la plus faible dans 96.1% des cas.

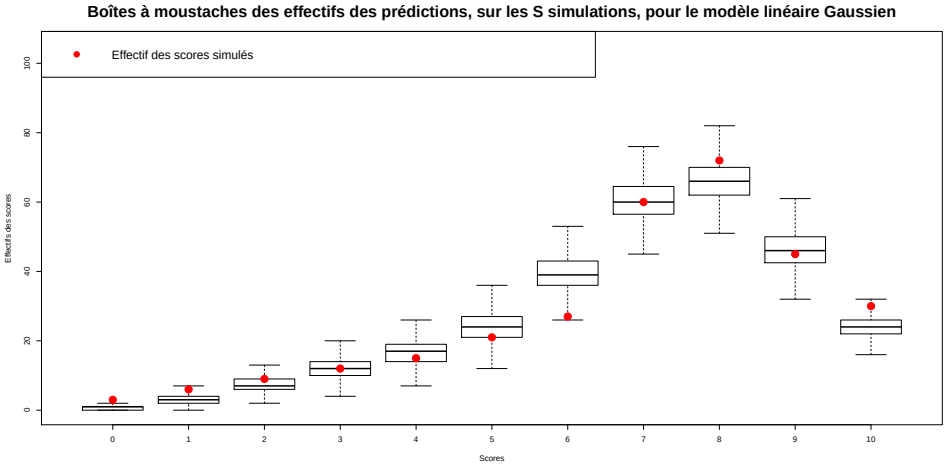
Prédictions

Sur l'ensemble des S simulations, en moyenne les quantiles des prédictions à 0%, 25%, 50%, 75%, 100% et également à 2.5% et 97.5% obtenues, pour les différents modèles, sont les suivants :

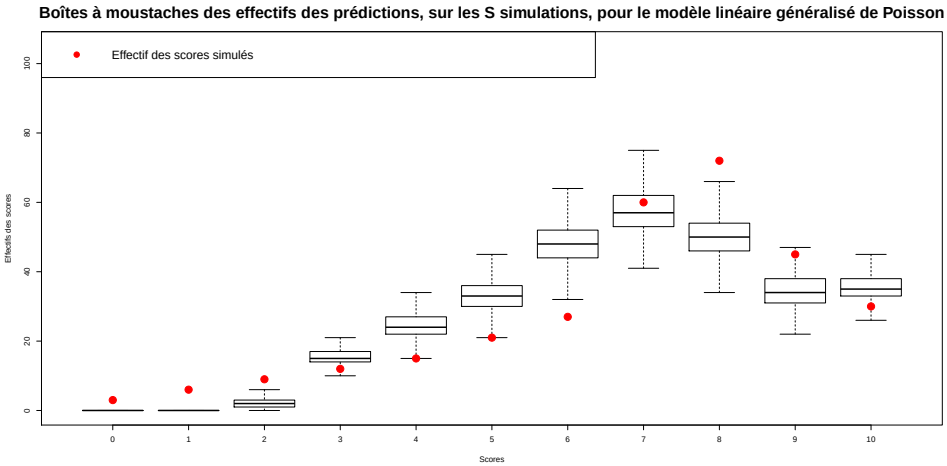
	0%	25%	50%	75%	100%	2.5%	97.5%
Données simulés	0.00	6.00	7.00	8.25	10.00	1.00	10.00
Gaussien	0.25	5.85	7.29	8.42	11.25	2.05	10.19
Poisson	2.22	5.51	6.98	8.37	13.33	2.97	11.17
Gamma	1.58	5.36	7.03	8.63	14.29	2.44	11.82
Binomial	0.72	5.98	7.62	8.54	9.73	1.66	9.57
Beta	0.54	5.98	7.76	8.70	9.83	1.39	9.61

Les quantiles des prédictions sont similaires à ceux obtenus pour une simulation. Les conclusions sont donc les mêmes.

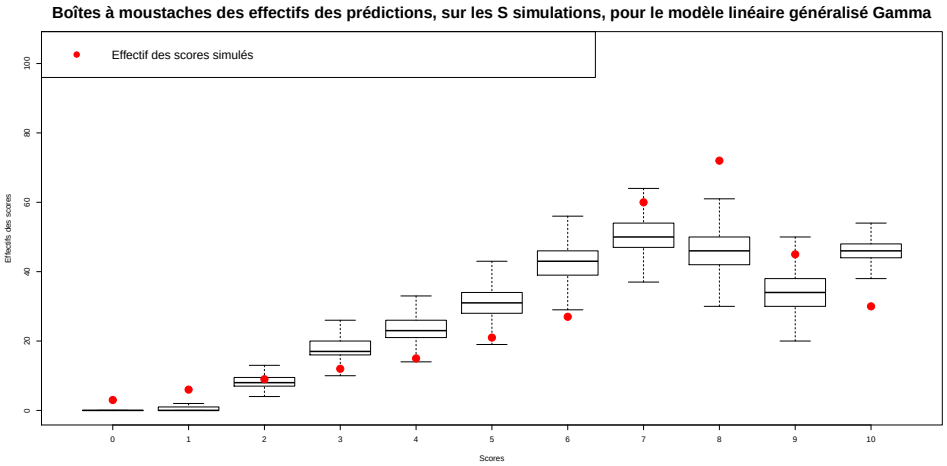
On trace les boîtes à moustaches des effectifs prédits des S simulations pour chaque valeur du score :



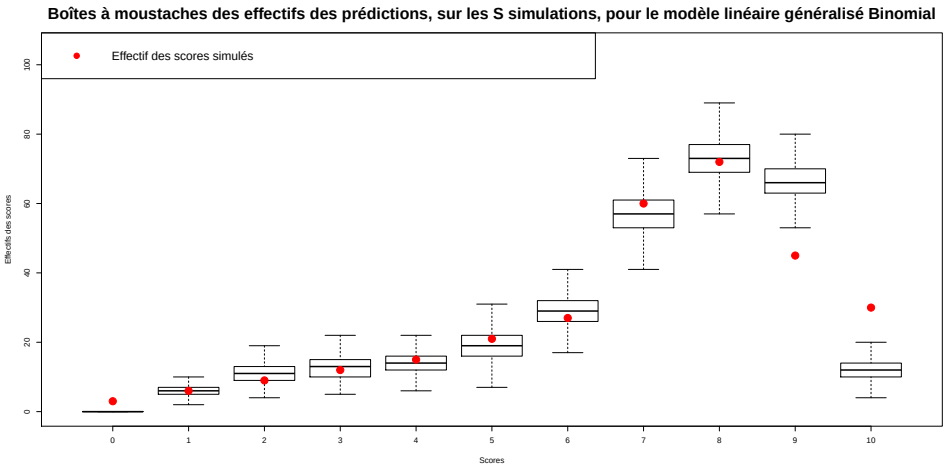
Le modèle linéaire Gaussien semble être assez satisfaisant du point de vue des prédictions des effectifs pour les différentes valeurs du score. Il sur-estime l'effectif pour la valeur 6 et sous-estime l'effectif pour la valeur 10.



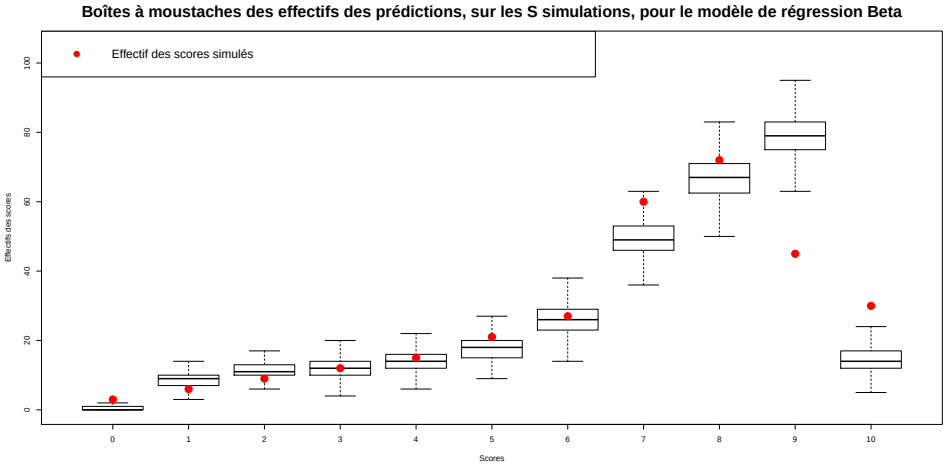
Le modèle linéaire généralisé de Poisson ne prédit aucun effectif pour les valeurs 0 et 1, par conséquent il sous-estime les effectifs pour ces valeurs du score. Il sur-estime les effectifs pour les valeurs 4, 5 et 6 et il sous-estime les effectifs pour les valeurs 8 et 9.



Le modèle linéaire généralisé Gamma sous-estime les effectifs pour les valeurs 0 et 1. Il sur-estime les effectifs pour les valeurs 4, 5 et 6 et il sous-estime les effectifs pour les valeurs 8 et 9. Enfin il sur-estime l'effectif pour la valeur 10.



Le modèle linéaire généralisé Binomial semble être assez satisfaisant du point de vue des prédictions des effectifs pour les différentes valeurs du score. Il sous-estime l'effectif pour la valeur 0. Puis il sur-estime l'effectif pour la valeur 9 et sous-estime l'effectif pour la valeur 10.



Le modèle de régression Beta sous-estime légèrement les effectifs pour les valeurs 0, 7 et 8. Puis il sur-estime fortement les effectifs pour la valeur 9 et sous-estime l'effectif pour la valeur 10.

Sur les S simulations, le pourcentage de fois où le test du χ^2 conclut à l'adéquation des effectifs simulés et prédits des valeurs du score, pour les différents modèles, sont les suivants :

Gaussien	Poisson	Gamma	Binomial	Beta
87 %	0 %	0 %	73 %	2 %

Le test du χ^2 conclut à l'adéquation des effectifs simulés et prédits pour les valeurs du score, dans 87 % des cas pour le modèle linéaire Gaussien et dans 73 % des cas pour le modèle linéaire généralisé Binomial.

Malgré le fait que le modèle linéaire généralisé Binomial sur-estime l'effectif pour le score 9 et qu'il sous-estime l'effectif pour le score 10, les résultats du test du χ^2 sont plutôt satisfaisants. Cela s'explique par le fait que si l'effectif est trop faible pour une valeur donnée il est regroupé avec l'effectif d'une valeur contiguë.

Résidus

Un résumé des S moyennes des résidus de Pearson studentisés est le suivant :

	Gaussien	Poisson	Gamma	Binomial	Beta
Minimum	-1.5×10^{-3}	-7.0×10^{-3}	-3.3×10^{-4}	-7.7×10^{-3}	-1.1×10^{-1}
1^{er} Quartile	-8.4×10^{-4}	-6.1×10^{-3}	-2.6×10^{-4}	1.3×10^{-3}	-7.4×10^{-2}
Médiane	-6.4×10^{-4}	-5.9×10^{-3}	-2.4×10^{-4}	4.1×10^{-3}	-6.4×10^{-2}
Moyenne	-6.4×10^{-4}	-5.9×10^{-3}	-2.4×10^{-4}	4.2×10^{-3}	-6.3×10^{-2}
3^{e} Quartile	-4.4×10^{-4}	-5.6×10^{-3}	-2.2×10^{-4}	7.1×10^{-3}	-5.2×10^{-2}
Maximum	2.5×10^{-4}	-4.7×10^{-3}	-1.4×10^{-4}	1.7×10^{-2}	4.7×10^{-3}

Les résidus de Pearson studentisés semblent centrés pour tous les modèles.

Un résumé des S quantiles, à 2.5% et 97.5%, des résidus de Pearson studentisés est le suivant :

	Gaussien		Poisson		Gamma		Binomial		Beta	
	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
Min.	-2.30	1.58	-1.17	0.67	-0.49	0.26	-1.59	1.08	-2.01	1.21
1^{er} Qu.	-1.98	1.84	-0.99	0.78	-0.39	0.31	-1.34	1.21	-1.66	1.36
Méd.	-1.90	1.91	-0.95	0.81	-0.37	0.32	-1.28	1.25	-1.60	1.41
Moy.	-1.91	1.91	-0.95	0.82	-0.37	0.32	-1.28	1.26	-1.60	1.41
3^{e} Qu.	-1.84	1.99	-0.91	0.85	-0.35	0.34	-1.22	1.30	-1.53	1.46
Max.	-1.62	2.31	-0.76	0.99	-0.28	0.42	-1.09	1.52	-1.32	1.80

Pour observer la symétrie des résidus de Pearson studentisés de manière plus visuelle, on trace pour chaque modèle la boîte à moustaches de l'inverse des quantiles à 2.5% et la boîte à moustache des quantiles à 97.5%. Si ses quantiles sont symétriques, les deux boîtes à moustaches doivent se situer au même niveau. Ces graphiques se trouvent en annexe B.2.

Les résidus du modèle linéaire Gaussien sont parfaitement symétriques. En ce qui concerne les modèles linéaires généralisés de Poisson et Gamma ainsi que le modèle de régression Beta, les résidus ne semblent pas symétriques. Les résidus du modèle linéaire généralisé Binomial sont à peu près symétriques.

AIC - BIC

Sur les S simulations, on obtient le pourcentage de fois où chacun des modèles a l'AIC et le BIC le plus faible :

	Gaussien	Poisson	Binomial
AIC	98.6 %	0 %	1.4 %
BIC	97.4 %	0 %	2.6 %

Le modèle linéaire Gaussien a l'AIC et le BIC les plus faibles dans respectivement 98.6 % et 97.4 % des cas.

Puissance empirique et risque d'erreur empirique des différents tests

La puissance empirique des tests de la déviance ou de Wald de la significativité globale de la régression est égale à 1 pour tous les modèles. Cette puissance est très satisfaisante. En ce qui concerne ces tests pour la significativité des coefficients, la puissance empirique est égale à 1 pour le premier paramètre et le risque d'erreur empirique de première espèce est égal à 0 pour le deuxième paramètre, pour tous les modèles. Ces résultats sont très satisfaisants également.

Pour le modèle linéaire Gaussien, le risque d'erreur empirique de première espèce du test de Shapiro-Wilk sur les résidus de Pearson studentisés est de 0.071 et le risque d'erreur empirique de première espèce du test de Breusch-Pagan d'homoscédasticité des résidus de Pearson studentisés est de 0.082. Ces risques sont proches de 0, ce qui est satisfaisant.
Pour le modèle linéaire généralisé de Poisson, les risques d'erreur empiriques de première espèce des deux tests de sur-dispersion sont égaux à 0 ce qui est très satisfaisant.
Pour les modèles linéaires généralisés de Poisson et Binomial, les risques d'erreur empiriques du test d'adéquation de Pearson sont égaux à 0 ce qui est très satisfaisant également.

Une nouvelle fois les tests ne permettent pas de différencier les modèles.

Conclusions sur les S simulations

Les premières conclusions réalisées sur une simulation se confirment. Sur les S simulations, le modèle linéaire Gaussien est meilleur que les autres modèles du point de vue des résidus, de l'AIC et du BIC. En revanche, le modèle linéaire généralisé Binomial est meilleur que les autres modèles en ce qui concerne les erreurs absolues et quadratiques moyennes. Pour le test du χ^2 d'adéquation, le modèle linéaire Gaussien est légèrement meilleur que le modèle linéaire généralisé Binomial. Du point de vue des quantiles de prédictions, le modèle linéaire généralisé Binomial et le modèle linéaire Gaussien sont équivalents.

Les résultats obtenus sur les S simulations sont résumés de la manière suivante :

Critère	EAM	EQM	χ^2	Résidus	AIC - BIC
Meilleur	Binomial	Binomial	Gauss-Bino	Gauss-Bino	Gaussien

5.2 Conclusions sur les différents scénarios de score simulé

Le travail, détaillé précédemment pour un scénario de simulation de score, a été fait sur les 7 scénarios de simulation de score.

- Le 1^{er} score a une moyenne faible et des données centrées autour de cette moyenne.
- Le 2^{e} score a une moyenne faible et des données étalées autour de cette moyenne.
- Le 3^{e} score a une moyenne intermédiaire et des données centrées autour de cette moyenne.
- Le 4^{e} score a une moyenne intermédiaire et des données étalées autour de cette moyenne.
- Le 5^{e} score a une moyenne élevée et des données centrées autour de cette moyenne.
- Le 6^{e} score a une moyenne élevée et des données étalées autour de cette moyenne.
- Le 7^{e} score est semblable à celui du jeu de données réel et a été étudié en détail ci-dessus.

Les résultats obtenus sur les S simulations, pour chaque scénario, sont résumés ci-dessous :

	EAM	EQM	χ^2	Résidus	AIC - BIC
1^{er} score	Binomial	Binomial	Gauss-Bino	Gaussien	Gaussien
2^{e} score	Binomial	Binomial	Binomial	Gaussien	Gaussien
3^{e} score	Gaussien	Gaussien	Gauss-Bino	Gauss-Bino	Gaussien
4^{e} score	Bino-Gauss	Bino-Gauss	Gaussien	Gauss-Bino	Gauss-Bino
5^{e} score	Binomial	Binomial	Gauss-Bino	Gauss-Bino	Gaussien
6^{e} score	Binomial	Binomial	Gauss-Bino	Gaussien	Gaussien
7^{e} score	Binomial	Binomial	Gauss-Bino	Gauss-Bino	Gaussien

Quand un seul modèle est indiqué pour un critère cela signifie que c'est le meilleur dans plus de 85% des cas. Quand deux modèles sont indiqués pour un critère, cela signifie que le premier modèle inscrit est le meilleur mais plutôt dans un ordre de pourcentage de 70-60% pour le premier modèle et de 30-40% pour le second.

Les résultats diffèrent légèrement selon les scores simulés. Cependant, seuls le modèle linéaire Gaussien et le modèle linéaire généralisé Binomial ressortent du point de vue des critères étudiés. Le modèle linéaire généralisé Binomial est meilleur que les autres modèles en ce qui concerne les erreurs absolues et quadratiques moyennes pour les S simulations de 6 scénarios. Le 3^{e} score est un score de moyenne 5 avec les données réparties de façon symétrique par rapport à cette moyenne. Cela explique que le modèle linéaire Gaussien est le meilleur pour ce score, étant donné que l'on peut associer l'histogramme de ce score simulé à celui de la loi normale. Pour le test du χ^2 d'adéquation, le modèle linéaire Gaussien est légèrement meilleur que le modèle linéaire généralisé Binomial. Les résidus de Pearson studentisés sont centrés et symétriques pour le modèle linéaire Gaussien pour les 7 scénarios. Ces résidus sont centrés et à peu près symétriques pour le modèle linéaire généralisé Binomial. Du point de vue de l'AIC et du BIC, le modèle linéaire Gaussien est meilleur que les autres pour les S simulations des 7 scénarios.

Si le but de l'étude à réaliser est prédictif, on choisira le modèle linéaire généralisé Binomial. En revanche si le but est descriptif on choisira le modèle linéaire Gaussien.

De plus, du fait que la loi Binomiale est discrète et bornée, une préférence se fait pour le modèle linéaire généralisé Binomial.

Chapitre 6

Statistique Bayésienne

6.1 Principe de l'inférence bayésienne

Cette partie est principalement basée sur les références [4] et [9].

6.1.1 Définition

Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique, f($\underline{x}|\theta$), et d'une distribution a priori pour les paramètres, $\pi(\theta)$. La loi a priori $\pi(\theta)$ doit caractériser l'état des connaissances que l'on a a priori sur θ.

Le principe de l'analyse bayésienne est d'actualiser cet a priori à partir de l'information provenant des observations $\underline{x}$. L'idée est ensuite d'estimer la distribution a posteriori de θ conditionnellement au vecteur des observations $\underline{x}$. L'estimation de cette distribution a posteriori est basée sur le Théorème de Bayes, dont elle tire son nom.

Théorème de Bayes
Si A et B sont deux événements tels que P(B) $\neq 0$, P(A|B) et P(B|A) sont reliés par :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{6.1}$$

Si on suppose qu'en plus d'une distribution d'échantillonnage, f($\underline{x}|\theta$), on dispose d'une distribution a priori sur θ, $\pi(\theta)$, on dispose d'un modèle bayésien. On peut alors construire la distribution a posteriori de θ :

$$\pi(\theta|\underline{x}) = \frac{f(\underline{x}|\theta)\pi(\theta)}{\int f(\underline{x}|\theta')\pi(\theta')d\theta'} \tag{6.2}$$

Le dénominateur ne dépendant pas du paramètre θ, nous avons donc :

$$\pi(\theta|\underline{x}) \propto f(\underline{x}|\theta)\pi(\theta) \tag{6.3}$$

La loi a posteriori combine les informations a priori que l'on a sur θ et l'information apportée par $\underline{x}$. Toute l'inférence sur θ peut être menée à partir de $\pi(\theta|\underline{x})$. Par exemple, pour estimer θ, on peut prendre l'espérance de le loi $\pi(\theta|\underline{x})$, si elle existe. Cela constitue l'estimateur bayésien de θ.

Cependant, la loi a posteriori peut être difficile à étudier, notamment lorsque θ est multivarié. Dans ce cas, une solution consiste à utiliser les méthodes de Monte Carlo par chaîne de Markov (méthodes MCMC : Markov Chain Monte Carlo).

6.1.2 Méthode MCMC

Chaîne de Markov

Une chaîne de Markov est une suite de variables aléatoires (X_t, $t \in \mathbb{N}$) qui permet de modéliser l'évolution dynamique d'un système aléatoire : X_t représente l'état du système à l'instant t. La propriété fondamentale des chaînes de Markov est que son évolution future ne dépend du passé qu'au travers de sa valeur actuelle, c'est à dire :

$$\mathbb{P}(X_{t+1} \in A | X_0 = x_0, X_1 = x_1, ..., X_t = x_t) = \mathbb{P}(X_{t+1} \in A | X_t = x_t)$$

Le principe des méthodes MCMC est de générer une chaîne de Markov $(\theta_t)_{t=1,...,T}$ de loi stationnaire correspondant à la loi a posteriori $\pi(\theta|\underline{x})$. Un fois les variables θ_1, ..., θ_T de loi $\pi(\theta|\underline{x})$ simulées, on peut approximer de façon empirique $\mathbb{E}(\theta|\underline{x})$ par $\frac{1}{T}\sum_{i=1}^{T}\theta_i$.

L'utilisation des MCMC suppose de vérifier la convergence des chaînes. D'autres diagnostics existent, ils ne sont pas mentionnés ici.

Nous avons principalement utilisés deux outils :
- le tracé de l'historique des valeurs qui doit montrer une bonne exploration de tout le domaine
- des autocorrélations qui deviennent nulles après l'ordre 1

En annexe B.3. il y a un exemple de ces deux outils. Ces tracés étaient ceux attendus pour tous les modèles.

6.1.3 Méthode d'échantillonnage de Gibbs

L'algorithme de Gibbs est un algorithme classique et fréquemment utilisée pour simuler des chaînes de Markov (en particulier, cet algorithme est utilisé dans le logiciel OpenBUGS que nous avons utilisé).

Cette méthode est adaptée au cas où θ est multivarié, $\theta = (\theta^1, ..., \theta^p)$, et où les lois a posteriori conditionnelles $\pi(\theta^1|\theta^2, ..., \theta^p, \underline{x})$, ..., $\pi(\theta^p|\theta^1, ..., \theta^{p-1}, \underline{x})$ sont simulables.

L'algorithme de Gibbs génère une chaîne de Markov pour la transition : étant donné l'état $\theta_t = (\theta_t^1, ..., \theta_t^p)$ de la chaîne à l'instant t, on génère l'état $\theta_{t+1} = (\theta_{t+1}^1, ..., \theta_{t+1}^p)$ par :

$$\begin{array}{c} \theta_{t+1}^1 \sim \pi(\theta^1|\theta_t^2, \theta_t^3, ..., \theta_t^p, \underline{x}) \\ \theta_{t+1}^2 \sim \pi(\theta^2|\theta_{t+1}^1, \theta_t^3, ..., \theta_t^p, \underline{x}) \\ \vdots \\ \theta_{t+1}^p \sim \pi(\theta^p|\theta_{t+1}^1, \theta_{t+1}^2, ..., \theta_{t+1}^{p-1}, \underline{x}) \end{array}$$

DIC

Le critère de la déviance bayésienne (Deviance Information Criterion) est défini par :

$$\begin{aligned} DIC &= \mathbb{E}(D(\theta)|x) + p_D \\ &= \mathbb{E}(D(\theta)|x) + \mathbb{E}(D(\theta)|x) - D(\mathbb{E}(\theta)|x) \end{aligned} \tag{6.4}$$

avec $D(\theta)$ = -2 × log (f($\underline{x}|\theta$))

L'évaluation des modèles selon ce critère suit le principe que plus le critère DIC est faible, meilleur est le modèle.

6.2 Résultats

Nous avons utilisé une chaîne, 5000 itérations de burn-in (le burn-in désigne les itérations initiales, en général très instables, qui ne sont pas utilisées pour l'estimation des paramètres), pour garder 50000 itérations pour les estimations. Les paramètres que l'on estime sur chaque itération sont les paramètres présents dans le modèle. A chaque itération on calcule les prédictions ainsi que les erreurs absolues et quadratiques moyennes. Pour obtenir les estimateurs bayésiens de ces paramètres, on prend la moyenne des paramètres sur les 50000 itérations.

Nous ne présentons que les résultats obtenus sur les S simulations du 7^e scénario de score, présenté dans le chapitre précédent, car les résultats obtenus sont très proches en fréquentiste et en bayésien.

6.2.1 Résumé des résultats sur les S simulations

Erreurs de prédictions

Sur les S simulations, la moyenne des erreurs absolues moyennes et la moyenne des erreurs quadratiques moyennes obtenues, pour les différents modèles, sont les suivantes :

	Gaussien	Poisson	Gamma	Binomial	Beta
EAM	0.73	0.90	0.95	0.70	0.72
EQM	0.83	1.27	1.47	0.77	0.80

Ces résultats sont quasiment identiques à ceux obtenus dans la section 5.1.2, les conclusions concernant les modèles sont donc les mêmes.

Sur les S simulations, on obtient le pourcentage de fois où chacun des modèles a l'erreur absolue moyenne ainsi que l'erreur quadratique moyenne la plus faible :

	Gaussien	Poisson	Gamma	Binomial	Beta
EAM	3.1 %	0 %	0 %	96.8 %	0.1 %
EQM	1.2 %	0 %	0 %	98.8 %	0 %

A nouveau les résultats sont quasiment identiques à ceux obtenus dans la section 5.1.2, les conclusions concernant les modèles sont donc également les mêmes.

Prédictions

Sur l'ensemble des S simulations, en moyenne les quantiles des prédictions à 0%, 25%, 50%, 75%, 100% et également à 2.5% et 97.5% obtenues, pour les différents modèles, sont les suivants :

	0%	25%	50%	75%	100%	2.5%	97.5%
Score simulé	0.00	6.00	7.00	8.25	10.00	1.00	10.00
Gaussien	0.28	5.85	7.30	8.42	11.24	2.07	10.18
Poisson	2.23	5.51	6.97	8.38	13.29	2.98	11.16
Gamma	1.44	5.16	6.86	8.51	14.80	2.27	11.88
Binomial	0.73	5.98	7.62	8.54	9.73	1.67	9.57
Beta	0.59	5.96	7.72	8.66	9.80	1.45	9.67

Les quantiles des prédictions sont similaires à ceux obtenus dans la section 5.1.2, les conclusions sont donc les mêmes.

Les boîtes à moustaches des effectifs prédits, arrondis à l'unité, des S simulations pour chaque valeur du score se trouvent en annexe B.4. Elles sont quasiment identiques à celles de la section 5.1.2. Les conclusion sont les mêmes.

Sur les S simulations, le pourcentage de fois où le test du χ^2 conclut à l'adéquation des effectifs simulés et prédits des valeurs du score, pour les différents modèles, sont les suivants :

Gaussien	Poisson	Gamma	Binomial	Beta
85 %	0 %	0 %	78 %	8 %

A nouveau les résultats sont quasiment identiques à ceux de la section 5.1.2.

DIC

Sur les S simulations, on obtient le pourcentage de fois où chacun des modèles a le DIC le plus faible :

	Gaussien	Poisson	Binomial
DIC	98.8 %	0 %	1.2 %

Le modèle linéaire Gaussien a le DIC le plus faible dans 98.8% des cas. Ce résultat est similaire à ceux obtenus avec les critères AIC et BIC dans le section 5.1.2.

6.2.2 Conclusions sur les S simulations

Sur les S simulations des 50000 itérations bayésiennes, le modèle linéaire Gaussien est meilleur que les autres modèles du point de vue du DIC. En revanche, le modèle linéaire généralisé Binomial est meilleur que les autres modèles en ce qui concerne les erreurs absolues moyennes et les erreurs quadratiques moyennes. Pour le test du χ^2 d'adéquation, le modèle linéaire Gaussien est légèrement meilleur que le modèle linéaire généralisé Binomial.

Sur les S simulations des autres scénarios de score, les conclusions sont identiques à celles de la section 5.2.

Les résultats obtenus en bayésien sont quasiment identiques à ceux obtenus en fréquentiste. Les conclusions sur les différents modèles sont les mêmes. Le travail effectué en statistique bayésienne a permis de confirmer les conclusions établies sur les modèles en fréquentiste.

Chapitre 7

Conclusions

L'étude des modèles sur le jeu de données réel a permis de se faire une première idée concernant ces modèles et d'étudier un cas concret de score. Nous avons pu tester les différents modèles sur un score réel.

Dans le but de confirmer ou de contredire les premiers résultats obtenus sur ce jeu de données nous avons continué le travail sur des données simulées. Travailler sur ce type des données permet de contrôler les données et de tester les modèles sur un nombre important de jeux de données. Cela permet par la suite d'émettre des conclusions que l'on peut juger fiables puisque les modèles sont répétés sur un nombre conséquent de jeux de données.

Les conclusions faites sur les modèles sont les mêmes en fréquentiste et en bayésien. L'approche bayésienne a permis de confirmer les résultats obtenus en fréquentiste.

Sur les S simulations, les résultats sont quasiment les mêmes quelque soit le scénario de simulation de score. Dans tous les cas, seuls le modèle linéaire Gaussien et le modèle linéaire généralisé Binomial ressortent de cette étude en fonction des critères choisis.

Globalement, le modèle linéaire Gaussien est meilleur que les autres modèles du point de vue des résidus, de l'AIC et du BIC. En revanche, le modèle linéaire généralisé Binomial est meilleur que les autres modèles en ce qui concerne les erreurs absolues moyennes et les erreurs quadratiques moyennes. Pour le test du χ^2 d'adéquation, le modèle linéaire Gaussien est légèrement meilleur que le modèle linéaire généralisé Binomial.

Cette étude a permis de mettre en avant le fait que, pour la modélisation de score borné et discret de 0 à 10, le modèle linéaire généralisé Binomial est meilleur que les autres modèles étudiés pour les prédictions et le modèle linéaire Gaussien est meilleur pour la sélection de variables.

Du fait que l'étude a été réalisée sur un score borné et discret, on choisira plutôt le modèle linéaire généralisé Binomial.

Annexes

Annexe A

Compléments mathématiques

A.1 Algorithme de Newton-Raphson

On s'intéresse aux équations du score suivantes :

$$\frac{\partial \mathcal{L}}{\partial \beta_j}(\mathbf{y}, \beta, \phi) = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \frac{(y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_i^{(j)} = 0 \text{ pour } j = 0, 1, ..., p \tag{A.1}$$

Pour résoudre ce système d'équations en β on approxime les fonctions $\frac{\partial \mathcal{L}}{\partial \beta_j}$ au moyen d'un développement de Taylor à l'ordre 1.

On veut résoudre :

$$\frac{\partial \mathcal{L}}{\partial \beta} = 0$$

La série de Taylor à l'ordre 1 est :

$$0_{p+1} = \frac{\partial \mathcal{L}}{\partial \beta} = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial \beta_0} \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial \beta_p} \end{pmatrix} = \frac{\partial \mathcal{L}}{\partial \beta}\left(\mathbf{y}, \beta^{(0)}, \phi\right) + D\left(\frac{\partial \mathcal{L}}{\partial \beta}(\mathbf{y}, \beta^{(0)}, \phi)\right) \times \left(\beta - \beta^{(0)}\right)$$

On obtient alors :

$$\hat{\beta} = \hat{\beta}^{(0)} - D\left(\frac{\partial \mathcal{L}}{\partial \beta}(\mathbf{y}, \hat{\beta}^{(0)}, \phi)\right)^{-1} \times \frac{\partial \mathcal{L}}{\partial \beta}\left(\mathbf{y}, \hat{\beta}^{(0)}, \phi\right)$$

On réitère le procédé, ce qui donne pour R=1,2,...

$$\hat{\beta}^{(R)} = \hat{\beta}^{(R-1)} - D\left(\frac{\partial \mathcal{L}}{\partial \beta}(\mathbf{y}, \hat{\beta}^{(R-1)}, \phi)\right)^{-1} \times \frac{\partial \mathcal{L}}{\partial \beta}\left(\mathbf{y}, \hat{\beta}^{(R-1)}, \phi\right)$$

avec $D\left(\frac{\partial \mathcal{L}}{\partial \beta}(\mathbf{y}, \hat{\beta}^{(R-1)}, \phi)\right)^{-1} = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 \mathcal{L}}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_0 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathcal{L}}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 \mathcal{L}}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 \mathcal{L}}{\partial \beta_p \partial \beta_p} \end{pmatrix}$

C'est la matrice hésienne de $\mathcal{L}(\mathbf{y}, \beta, \phi)$. On calcule le terme en position (j, k), à partir de (A.1) :

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k}(\mathbf{y}, \beta, \phi) &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_k} \left(\frac{1}{a_i(\phi)} \frac{(y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_i^{(j)} \right) \\
&= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \frac{\partial}{\partial \beta_k} \left(\frac{(y_i - \mu_i)}{V(\mu_i)} \times \frac{1}{g'(\mu_i)} \right) x_i^{(j)} \\
&= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left[\frac{1}{g'(\mu_i)} \times \frac{\partial}{\partial \beta_k} \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) + \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) \times \frac{\partial}{\partial \beta_k} \left(\frac{1}{g'(\mu_i)} \right) \right] x_i^{(j)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_k} \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) &= \frac{\partial}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) \\
&= \frac{\partial}{\partial \mu_i} \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \\
&= \frac{-V(\mu_i) - (y_i - \mu_i) V'(\mu_i)}{V(\mu_i)^2} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \\
&= \frac{-V(\mu_i) - (y_i - \mu_i) V'(\mu_i)}{V(\mu_i)^2} \frac{1}{g'(\mu_i)} x_i^{(k)} \text{ d'après (2.15) et (2.16)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \beta_k} \left(\frac{1}{g'(\mu_i)} \right) &= \frac{\partial}{\partial \mu_i} \left(\frac{1}{g'(\mu_i)} \right) \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \\
&= \frac{-g''(\mu_i)}{g'(\mu_i)^2} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_k} \\
&= \frac{-g''(\mu_i)}{g'(\mu_i)^2} \frac{1}{g'(\mu_i)} x_i^{(k)} \text{ d'après (2.15) et (2.16)} \\
&= \frac{-g''(\mu_i)}{g'(\mu_i)^3} x_i^{(k)}
\end{aligned}$$

En remplaçant, on trouve :

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}}{\partial \beta_j \partial \beta_k}(\mathbf{y}, \beta, \phi) &= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left[\frac{1}{g'(\mu_i)} \times \frac{\partial}{\partial \beta_k} \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) + \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) \times \frac{\partial}{\partial \beta_k} \left(\frac{1}{g'(\mu_i)} \right) \right] x_i^{(j)} \\
&= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left[\frac{1}{g'(\mu_i)} \times \frac{-V(\mu_i) - (y_i - \mu_i) V'(\mu_i)}{V(\mu_i)^2} \frac{1}{g'(\mu_i)} + \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) \times \frac{-g''(\mu_i)}{g'(\mu_i)^3} \right] x_i^{(j)} x_i^{(k)} \\
&= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left[\frac{1}{g'(\mu_i)^2} \times \frac{-V(\mu_i) - (y_i - \mu_i) V'(\mu_i)}{V(\mu_i)^2} - \frac{g''(\mu_i)}{g'(\mu_i)^3} \times \frac{(y_i - \mu_i)}{V(\mu_i)} \right] x_i^{(j)} x_i^{(k)} \\
&= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left[\frac{-1}{g'(\mu_i)^2 V(\mu_i)} - (y_i - \mu_i) \times \left(\frac{V'(\mu_i)}{V(\mu_i)^2 g'(\mu_i)^2} + \frac{g''(\mu_i)}{g'(\mu_i)^3 V(\mu_i)} \right) \right] x_i^{(j)} x_i^{(k)}
\end{aligned}$$

Annexe B

Résultats graphiques

B.1 Résidus de Pearson studentisés du jeu de données réel

Pour chaque modèle, les résidus de Pearson studentisés seuls et également en fonction de chaque variable explicative du jeu de données sont représentés.

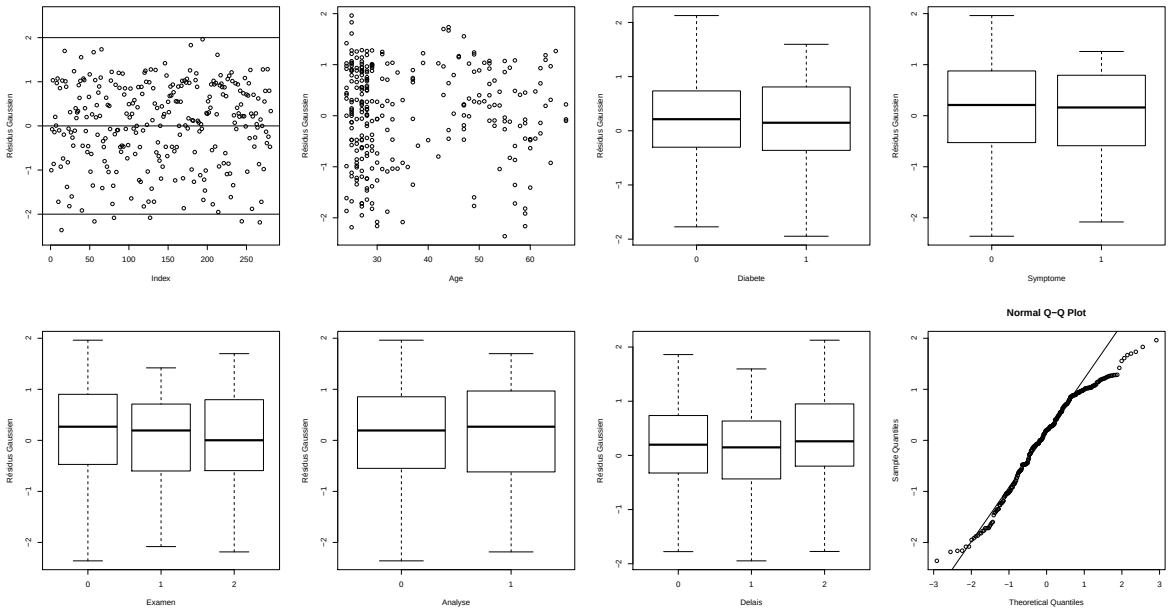


FIGURE B.1 – Résidus de Pearson studentisés du modèle linéaire Gaussien

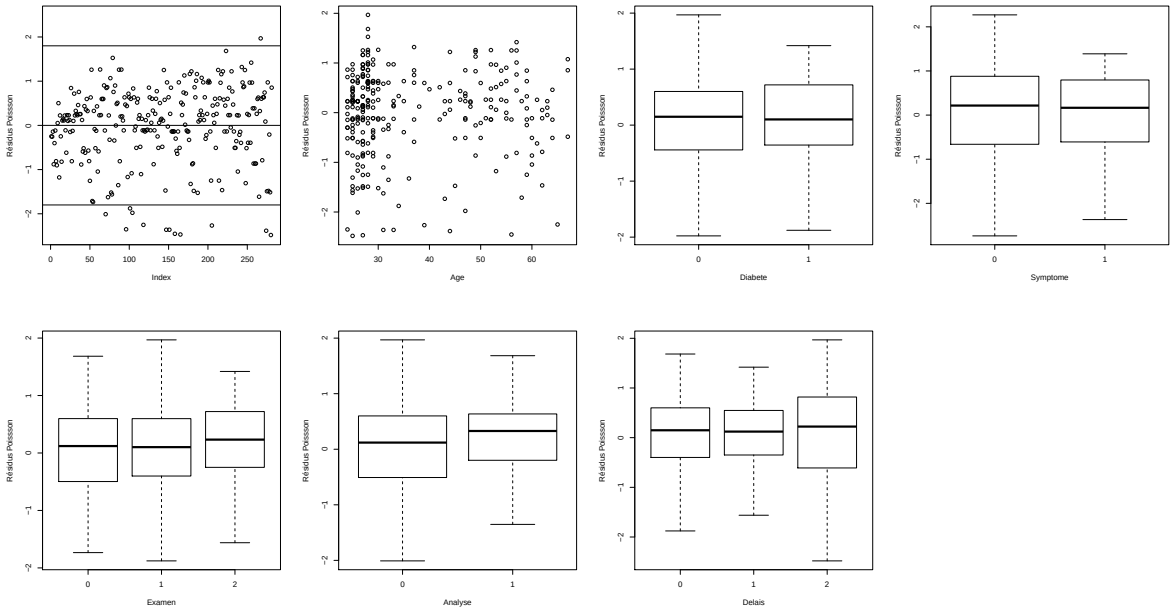


FIGURE B.2 – Résidus de Pearson studentisés du modèle linéaire généralisé de Poisson

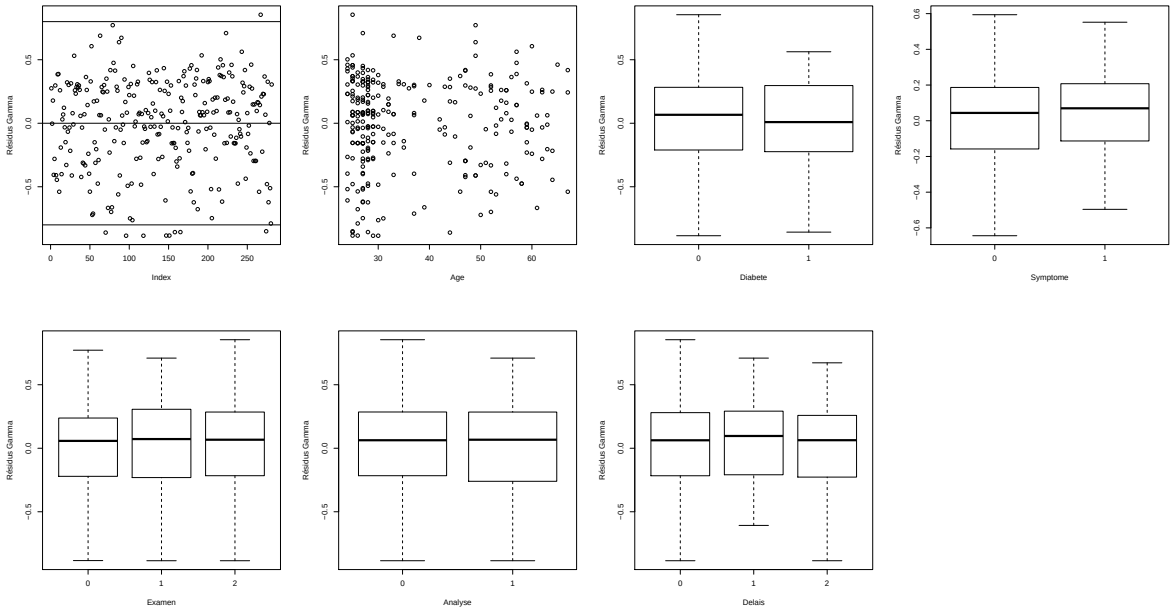


FIGURE B.3 – Résidus de Pearson studentisés du modèle linéaire généralisé Gamma

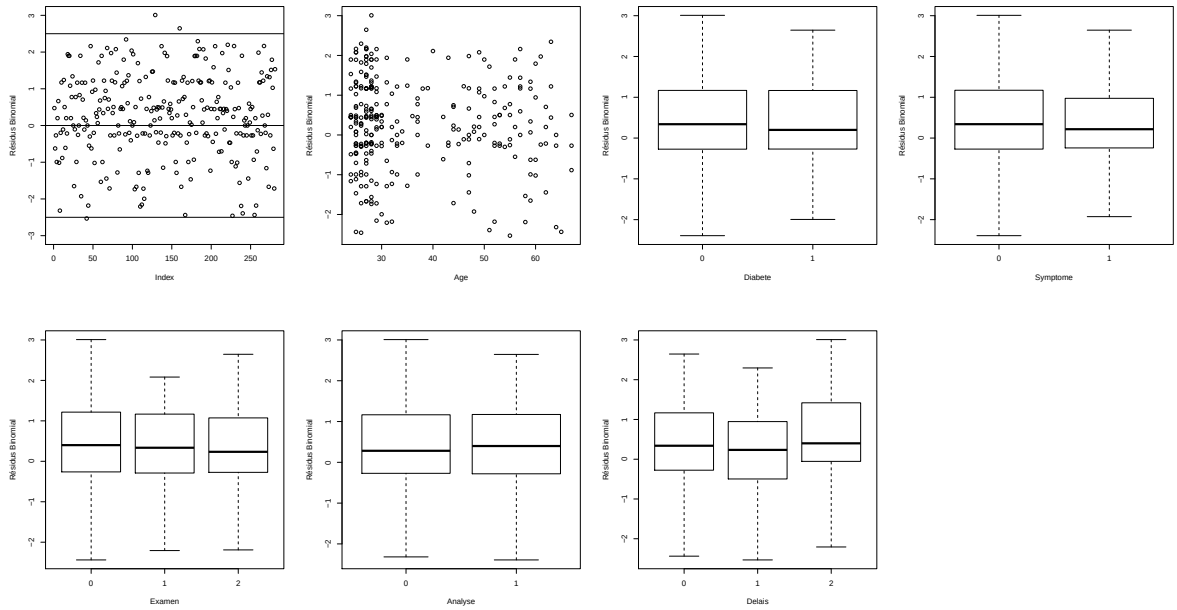


FIGURE B.4 – Résidus de Pearson studentisés du modèle linéaire généralisé Binomial

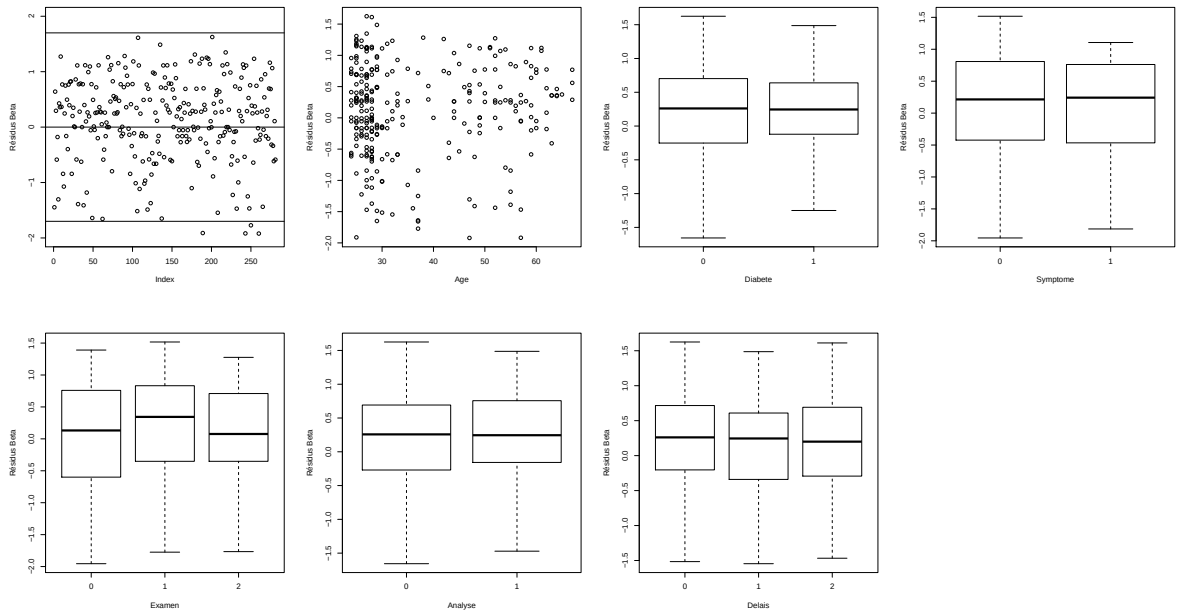
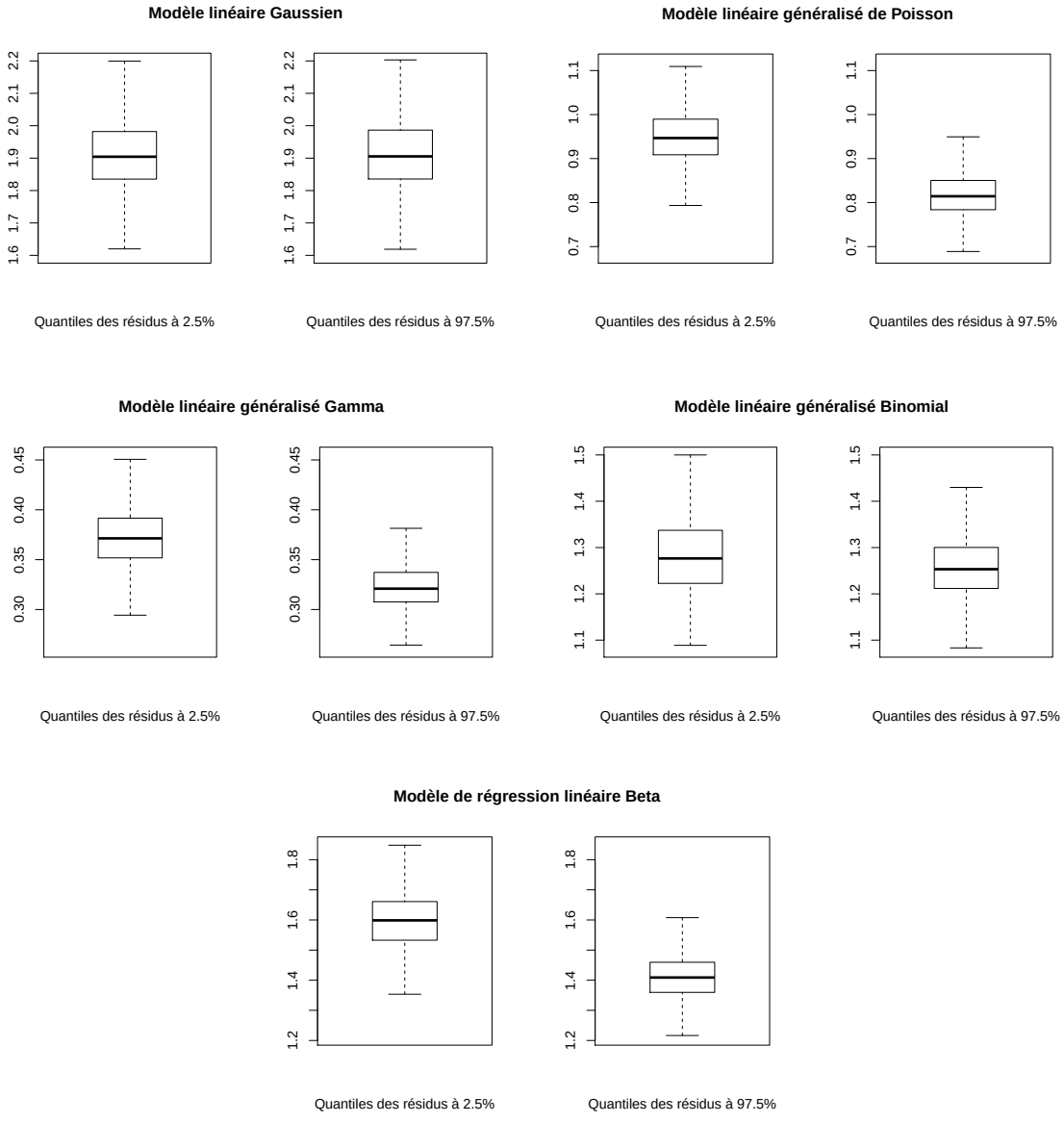


FIGURE B.5 – Résidus de Pearson studentisés du modèle de régression Beta

B.2 Boîtes à moustaches des résidus sur les S simulations

Dans le but d'observer si les résidus de Pearson studentisés, des différents modèles, sont symétriques sur les S simulations on trace pour chaque modèle la boîte à moustaches de l'inverse des quantiles à 2.5% et la boîte à moustache des quantiles à 97.5%. Si ses quantiles sont symétriques, les deux boîtes à moustaches doivent se situer au même niveau.



B.3 Diagnostics bayésien

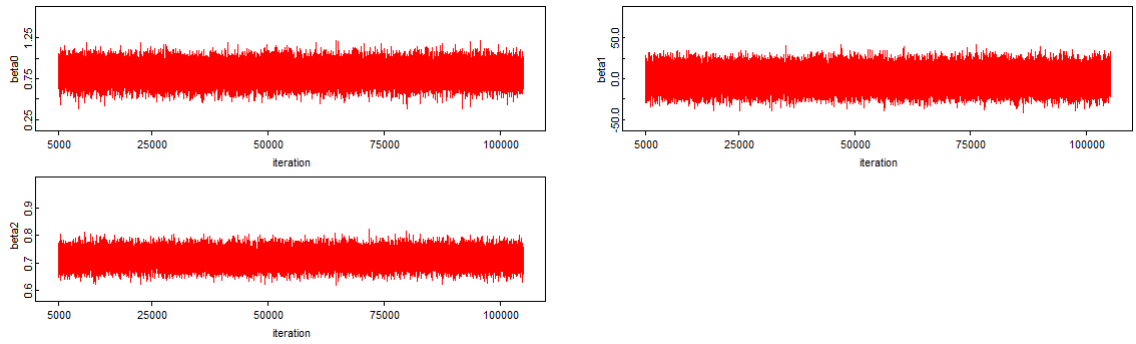


FIGURE B.6 – Historique des valeurs

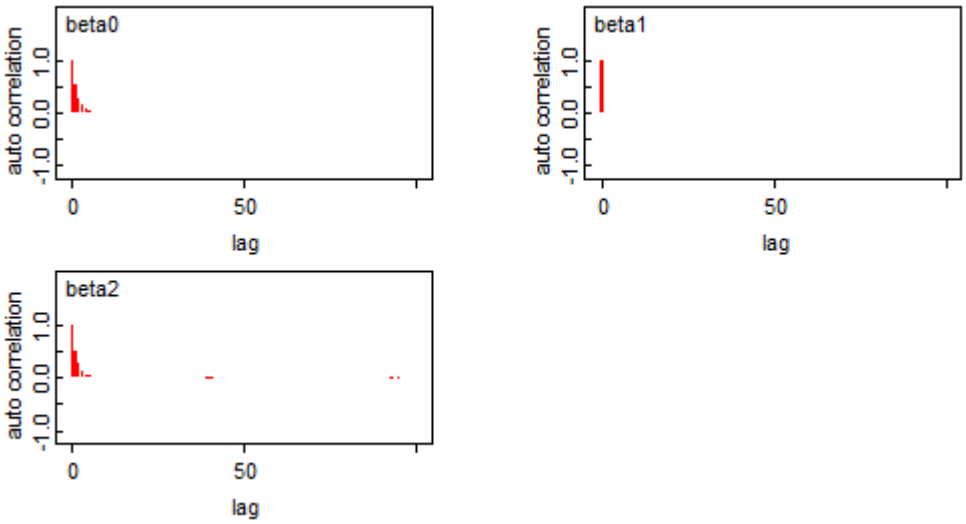
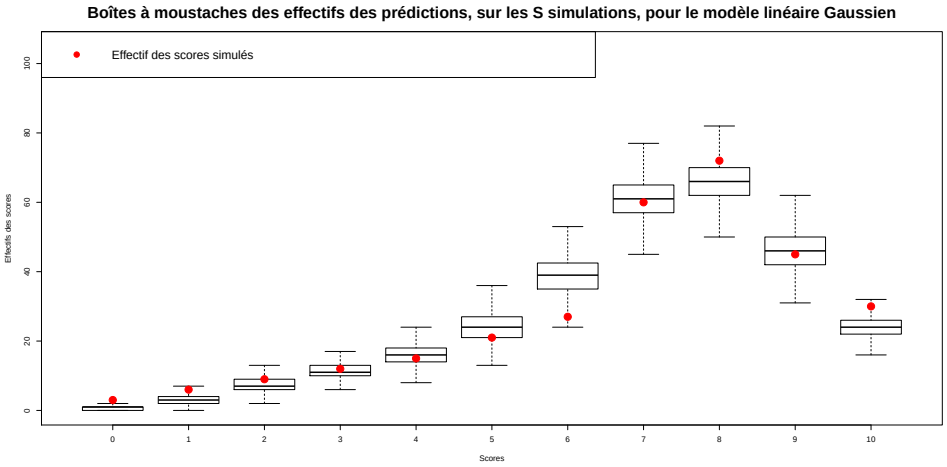
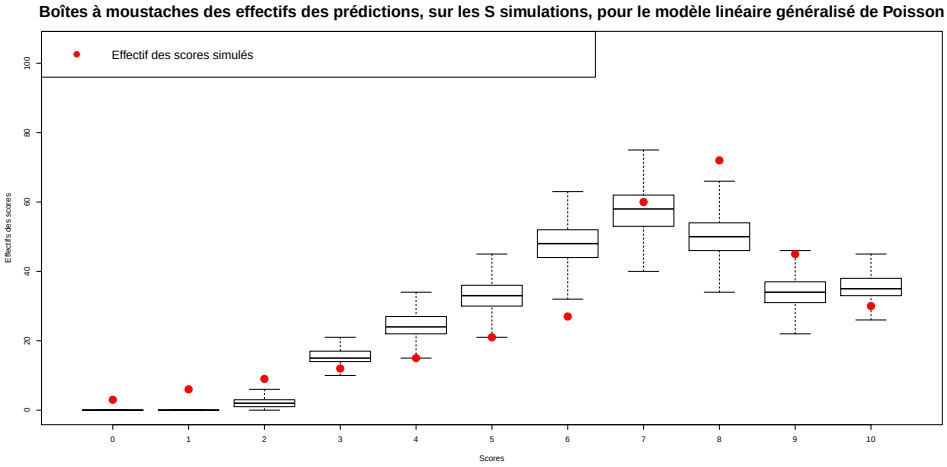


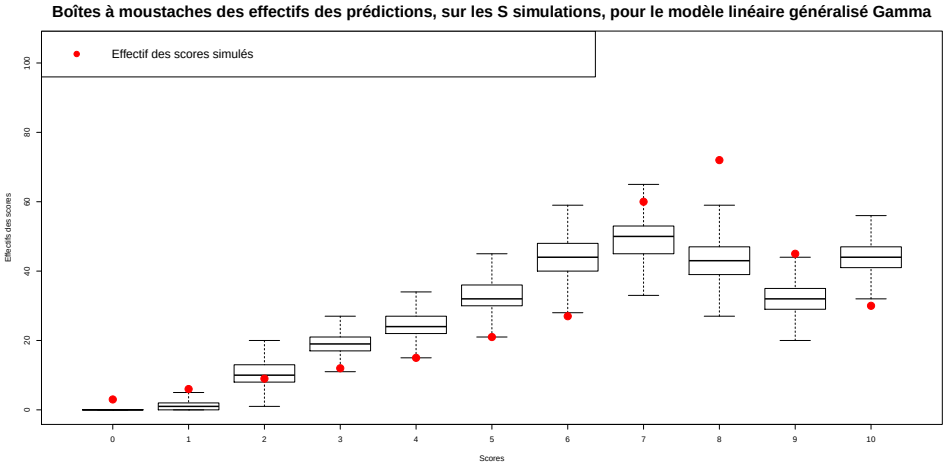
FIGURE B.7 – Autocorrélations

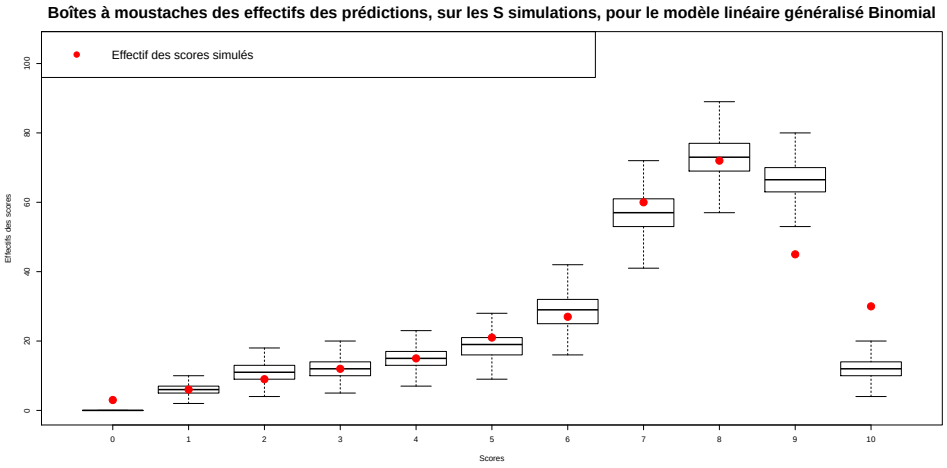
B.4 Boîtes à moustaches des effectifs des prédictions en bayésien

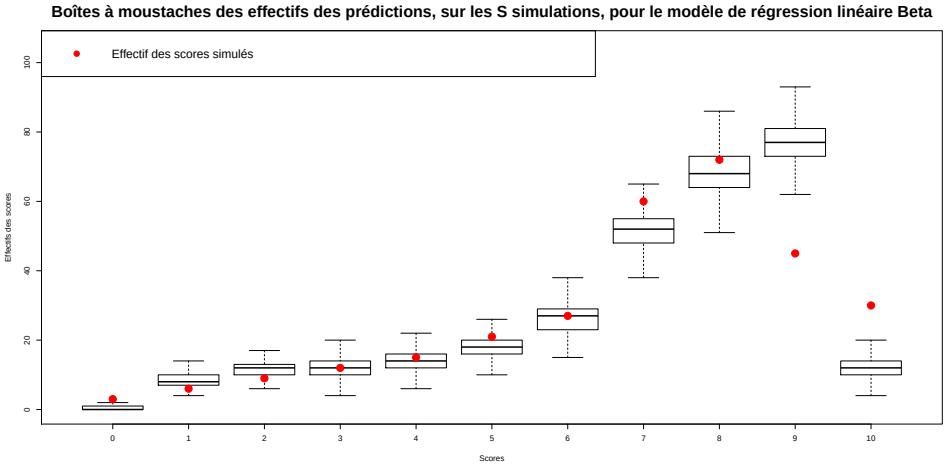
Pour chaque modèle, les boîtes à moustaches des effectifs prédits bayésiens, arrondis à l'unité, des S simulations pour chaque valeur du score sont les suivantes :








Boîtes à moustaches des effectifs des prédictions, sur les S simulations, pour le modèle linéaire généralisé Binomial
Effectif des scores simulés
Effectifs des scores
Scores
0
1
2
3
4
5
6
7
8
9
10
0
20
40
60
80
100


Boîtes à moustaches des effectifs des prédictions, sur les S simulations, pour le modèle de régression linéaire Beta
Effectif des scores simulés
Effectifs des scores
Scores
0
1
2
3
4
5
6
7
8
9
10
0
20
40
60
80
100

Bibliographie

[1] T. Breusch and A. Pagan. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5) :1287–1294, 1979.

[2] G. Claeskens and N. Lid Hjort. *Model Selection and Model Averaginig.* Cambridge University Press, 2009.

[3] F. Cribari-Neto and S. Ferrari. Beta regression. *Journal of Applied Statisctics*, 31(7) :799–815, 2004.

[4] J.L. Dortet. *Cours d'Analyse Bayésienne*, 2012.

[5] S. Geffray. *Cours de modéles linéaires généralisés*, 2013.

[6] J. W. Hardin and J. M. Hilbe. *Generalized Linear Models and Extensions.* STATA Press, 2007.

[7] J. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (saps 2) based on a european/north american multicenter study. *JAMA*, 270(24) :2957–2963, 1993.

[8] P. McCullagh and J.A. Nelder. *Generalized Linear Models.* London :Chapman and Hall., 1989.

[9] W. Robert. *Le choix bayésien : Principes et patrique.* Springer, 2006.

[10] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality. *Biometrika*, 52(3/4) :591–611, 1965.

[11] M. Smithson and J. Verkuilen. A better lemon squeezer ? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 11(1) :54–71, 2006.