2018

La régression de poisson et ses applications

Njejimana, Gérard

UB, IPA

UNIVESITE DU BURUNDI

INSTITUT DE PEDAGOGIE APPLIQUEE

DEPARTEMENT DE MATHEMATIQUES

LA REGRESSION DE POISSON ET SES APPLICATIONS

Par :

NJEJIMANA Gérard

Sous la direction de :

M. MPUNIKIYE Léonce

Mémoire soutenu et défendu publiquement en vue de l'obtention du grade de Licencié en Pédagogie Appliquée, agrégé de l'enseignement secondaire en Mathématiques

Bujumbura, Avril 2018

DEDICACE

A Dieu tout Puissant ;

A ma chère Epouse NDAYISENGA Aline ;

A nos enfants IZERE Lesslie Labrune et IRIWACU Tonny-Andy;

A mes chers parents ;

A ma grand-mère paternelle ;

A mes frères et sœurs,

Je dédie ce Mémoire

Gérard NJEJIMANA

REMERCIEMENTS

Au terme de ce travail, fruit d'un grand effort, qu'il nous soit permis d'exprimer nos sentiments de profonde gratitude envers toutes les personnes sans le concours desquelles ce travail n'aurait pas eu un tel aboutissement.

Nos sentiments de reconnaissance s'adressent plus particulièrement à Monsieur MPUNIKIYE Léonce, Directeur et promoteur de ce travail qui, malgré ses multiples obligations, a accepté volontier d'orienter et de guider ce travail. Sa rigueur scientifique et ses directives bienveillantes nous ont été d'une grande utilité pendant la réalisation de ce travail.

Nos remerciements s'adressent à tous les professeurs de l'Institut de Pédagogie Appliquée, spécialement ceux des départements de Mathématiques et de Physique-Technologie pour la formation tant humaine qu'intellectuelle que nous avons bénéficié de leur part.

Nos remerciements vont à l'endroit de mes parents qui ont serré les ceintures dès l'école primaire jusqu'à la fin de mes études universitaires et également à mon épouse qui n'a pas cessé de m'encourager. Nous remercions également les familles NJEJIMANA Albert, HAVYARIMANA Frédéric pour leurs contributions tant morales que matérielles durant nos études.

Nous témoignons notre gratitude aux nombreuses personnes qui nous ont enseigné et nous ont aidé jusqu'à ce que nous voyons un tel aboutissement.

Enfin, que tous les frères et sœurs de la communauté des universitaires anglicans qui ont rendu agréable notre séjour dans cette communauté par leur soutien spirituel ainsi que ceux qui ont de près ou de loin, prêté main forte dans la réalisation de ce travail trouvent ici l'expression de notre profonde gratitude.

Gérard NJEJIMANA

RESUME DU MEMOIRE

SUJET : « **LA REGRESSION DE POISSON ET SES APPLICATIONS** »

Les lois de probabilité sont des éléments essentiels dans le calcul de la probabilité et de la statistique. Les éléments de base tels que les variables aléatoires qui sont subdivisées en variables aléatoires discrètes et en variables aléatoires continues ont aussi une part très considérable dans l'étude des différents événements de la nature.

Etant donné une loi de Poisson qui peut se présenter comme sous une forme générale $P(X=x)=e^{-\mu}\frac{\mu^{y}}{y!}$, cette loi est adaptée aux calculs des événements rares qui se présentent d'une façon accidentelle au cours de la vie. Aussi, l'espérance de ces événements et la variance sont en valeur égale, ce qui signifie que dans le calcul, l'espérance est égale à la variance.

Dans notre travail, nous nous sommes intéressés sur l'étude de la régression de Poisson et ses applications. Notre travail est subdivisé en trois chapitres :

Le premier chapitre énonce les différentes lois de probabilités.

Le deuxième présente de la régression de Poisson.

Le troisième qui met en application un exemple des événements de la régression de Poisson.

TABLE DE MATIERES

INTRODUCTION

La régression est un ensemble de Méthodes statistiques très utilisées pour analyser la relation d'une variable par rapport à une ou plusieurs autres. Pendant longtemps la régression d'une variable aléatoire Y sur le vecteur de variables aléatoires X désignait la moyenne conditionnelle de Y sachant X. Dans la pratique de la régression, on peut partir de la régression linéaire simple qu'on peut représenter sur une droite de régression passant par le centre du nuage des points représentés au cours de la résolution de la fonction $f(x) = ax + b$; il existe ensuite une autre forme de régression appelée régression log-linéaire ou tout simplement « la régression de poisson ». Cette régression de poisson sera étudiée en détail, on verra sa fonction de répartition, sa représentation, ses propriétés, et ses domaines d'applications. C'est la partie de la régression qui sera étudiée en profondeur car c'est l'objet même de mon travail.

La régression linéaire est utilisée en pratique afin de trouver une relation entre une variable réponse et une ou plusieurs variables explicatives. Une lacune de cette méthode est qu'elle est inappropriée à la variable réponse de dénombrement. Dans un tel cas, la régression de Poisson doit être utilisée.

Ce travail décrira d'une façon détaillée la régression de Poisson. Les propriétés de la loi de Poisson sont énoncées dans le but d'expliquer la régression de Poisson.

L'hypothèse fondamentale de la régression de Poisson est que la moyenne et la variance soient égales. On verra que cette condition est vérifiable à l'aide de la statistique de Pearson.

Ce travail est subdivisé en trois chapitres. Le premier parle de quelques lois de probabilité et des notions des variables aléatoires. Le second chapitre parle de la régression de Poisson tandis que le troisième parle d'application de la régression de Poisson.

CHAPITRE I. QUELQUES LOIS DE LA PROBABILITE

I.1. Notions de variables aléatoires [1], [4]

Dans le calcul de la probabilité et de la statistique, nous présentons quelques éléments de base. Nous parlerons de variables aléatoires.

Considérons une expérience dont le résultat est incertain et supposons que l'ensemble des résultats possibles lui soit connu. On appelle cet ensemble fondamental de l'expérience et le note par Ω. On peut s'intéresser à une fonction du résultat plutôt qu'au résultat lui-même. Les événements auxquels on s'intéresse sont liés à des fonctions réelles définies sur l'ensemble fondamental et qui sont appelés variables aléatoires.

I.1.1. Variables aléatoires discrètes

Définition : Une variable aléatoire discrète définie sur (Ω, A), une application $X:\Omega \to \mathbb{R}$ telle que $X(\Omega)$ est dénombrable (en général $X(\Omega) \subset \mathbb{N}$ ou $X(\Omega) \subset \mathbb{Z}$ et dans tous les cas $X(\Omega)$ est en correspondance bijective avec $\mathbb{N}$) et telle que pour tout x réel : $X^{-1}(x) = \{w \in \Omega / X(w) = x\} \in A$: ce qui exprime tout simplement que $X^{-1}(x)$ est un événement.

I.1.2. Variables aléatoires continues [1]

Si une variable aléatoire X peut prendre les valeurs réelles continues, dans un intervalle, elle est dite continue.

Une variable aléatoire X est définie par son intervalle de variation $[a,b]$ et par la fonction $f(x)$ appelée fonction de densité. L'intervalle peut être ouvert ou fermé, même seulement d'un côté, borné ou non. Il est commode de prolonger f à l'intervalle $(-\infty, +\infty)$ en posant $f(x) = 0$ en déhors de l'intervalle de définition originale. La fonction de densité $f(x)$ doit satisfaire deux conditions :

a) La fonction de densité $f(x)$ est non négative : $f(x) \geq 0$

b) L'intégral de la fonction de densité $f(x)$ est égal à 1

$\int_{\mathbb{R}} f(x)dx = 1$

Dans mon travail, je me suis référé aux différents types des lois que je juge aussi importantes dans la résolution des exercices.

I.2. Exemple de lois discrètes [1], [4]

I.2.1. La loi de probabilité discrète

Soit $x_1, x_2, ..., x_n$ qui sont les valeurs possibles de la variable aléatoire X et les $p_1, p_2, ..., p_n$ sont des probabilités associées aux valeurs correspondantes. La fonction $f(x)$ qui associe à chaque x la valeur $P_r(X = x)$ s'appelle aussi loi de probabilité ou densité. Par conséquent, la probabilité que la variable aléatoire X prenne la valeur x_i est égale à P_i et on écrit $P_r(X = x) = P_i$. Les probabilités $p_1, p_2, ..., p_n$ doivent satisfaire deux conditions :

- Tous les P_i sont positifs : $P_i > 0$
- La somme de tous les P_i est égale à 1

Il est toujours possible d'associer à une variable aléatoire une probabilité et définir ainsi une loi de probabilité. Lorsque le nombre d'épreuves augmente indéfiniment les fréquences observées vers le phénomène étudié tendent vers les probabilités et les distributions observées vers les distributions de probabilité ou loi de probabilité.

Identifier la loi de probabilité suivie par une variable aléatoire donnée est essentiel car cela conditionne le choix des méthodes employées pour répondre à une question donnée. Les variables aléatoires discrètes prennent des valeurs

entières sur un intervalle donné. Ce sont généralement les résultats de dénombrement.

I.2.1.1. Loi uniforme

Définition : Une distribution de probabilité suit une loi uniforme lorsque toutes les valeurs prises par la variable aléatoire sont équiprobables. Si n est le nombre de valeurs différentes prises par la variable aléatoire.

$\forall i, P(X = x_i) = \frac{1}{n}$ pour $i = 1,2,...,n$ dans ce cas, on dit que la variable aléatoire X est distribuée selon une loi discrète uniforme. Si X est une variable aléatoire discrète, alors l'espérance mathématique de X notée $E(X)$ est finie par : $E(X) = x_1P_1 + x_2P_2 + ... + x_nP_n = \sum_{i=1}^{n} x_iP_i$

Exemple : La distribution des chiffres obtenus au lancer de dé si ce dernier est non pipé suit une loi uniforme dont la loi de probabilité est la suivante :

x	1	2	3	4	5	6
$P(x=x_i)$	1/6	1/6	1/6	1/6	1/6	1/6

avec pour espérance : $E(X) = \frac{1}{6}\sum_{i=1}^{6} xi = 3,5$ et variance

$$V(X) = \frac{1}{6}\sum_{i=1}^{6}\left[E\left(xi^2\right)\right] - \left[E(X)\right]^2 = 2,92$$

où les valeurs x_i correspondent au rang i de la variable X dans la série.

Dans le cas particulier d'une loi discrète uniforme où les valeurs de la variable aléatoire X correspondent au rang

$$x_i = i\left(\forall i \in [1,n]\right).\ E(X) = \frac{n+1}{2} \text{ et } V(X) = \frac{n^2-1}{12}.$$

I.2.1.2. Loi de Bernoulli

Soit un univers constitué de deux éventualités : P pour Pile et F pour face sur lequel, on construit une variable aléatoire discrète. «nombre de succès » telle qu'au cours d'une épreuve, on ait :

$$\begin{cases} \text{Si P est réalisé } X=1 \\ \text{Si F est réalisé } X=0 \end{cases} \Leftrightarrow X=\begin{cases} 1 & \text{si P est réalisé} \\ 0 & \text{sinon} \end{cases}$$

La variable indicatrice X est telle que :

$$X:\Omega \to \mathbb{R}$$
$$X(\Omega)=\{0,1\}$$

La loi de probabilité associée est que cette variable a été définie X telle que

$$\begin{matrix} P(X=0)=q \\ P(X=1)=p \end{matrix} \text{ avec } p+q=1 \text{ est appelée loi de Bernoulli notée } \beta(1,p).$$

La loi de Bernoulli est utilisée lorsqu'une expérience aléatoire n'a que deux résultats possibles : Le succès avec une probabilité p et l'échec avec une probabilité $q=1-p$

L'espérance de la variable de Bernoulli est par définition :

$$E(X)=P \text{ car par définition } E(X)=\sum_{i=1}^{2} x_i p_i=(0\times q)+(1\times p)=p.$$

La variance de la variable de Bernoulli est $V(X)=p.q$ car par définition

$$V(X)=\sum_{i=1}^{2} x_i^2.p_i-\left[E(X)\right]^2=(0\times q)+(1\times p)-p^2$$

d'où $V(X)=p-p^2=p(1-p)=p.q.$

I.2.1.3. Loi Binomiale

Décrite pour la première fois par Isaac Neutron en 1676. Et démontrée pour la première fois par mathématicien suisse. Jacob Bernoulli en 1713, la Loi binomiale est l'une des distributions de probabilité les plus fréquemment rencontrées en statistiques appliquées.

Soit l'application $S_n : \Omega^n \to \mathbb{R}^n$

Avec $S_n = X_1 + X_2 + \cdots + X_i + \cdots + X_n$, X_i est une variable de Bernoulli.

La variable binomiale S_n représente le nombre de succès obtenus lors de la répétition de n épreuves identiques et indépendantes. Chaque épreuve ne pouvant donner que deux résultats possibles.

Ainsi la loi de probabilité suivie par la somme de n variables de Bernoulli où la probabilité associée au succès est p est la loi Binomiale de paramètres n et p.

$S_n : \Omega^n \to \mathbb{R}^n$

$$S^n = \sum_{i=1}^{n} X_i \to \beta(n, p)$$

La probabilité $S_n = k$ c'est-à-dire l'obtention de k succès au cours de n épreuves indépendantes est :

$P(S_n = k) = C_n^k p^k q^{n-k}$.

Il est facile de démontrer que l'on a bien une loi de probabilité car :

$$\sum_{k=0}^{n} P(S_n = k) = \sum_{k=0}^{n} C_n^k p^k q = (p+q)^n = 1.$$

L'Espérance d'une variable binomiale S_n est égale à $E(S_n) = np$.

En effet, $E(S_n) = E(X_1 + X_2 + \cdots + X_i + \cdots + X_n)$

or $E(X_1 + X_2 + \cdots + X_i + \cdots + X_n) = \sum_{i=1}^{n} E(X_i)$ (Propriété d'espérance) et $E(X_i) = P$ (variable de Bernoulli) d'où $E(S_n) = np.$

La variance d'une variable binomiale S_n est égal à $V(S_n) = np.q.$

En effet,

$V(S_n) = V(X_1 + X_2 + \cdots + X_i + \cdots + X_n).$

$V(X_1 + X_2 + \cdots + X_i + \cdots + X_n) = \sum_{i=1}^{n} V(X_i)$

et $V(S_n) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} p.q$, avec $V(X_i) = p.q$, car variable de Bernoulli , d'où $V(S_n) = np.q.$

I.2.1.4. Loi de Poisson[1], [4]

On dit qu'une variable aléatoire, comptant le nombre de réalisations d'un certain évènement par unité de temps ou par exemple par unité de surface, est de Poisson, de paramètre λ et on note $P(\lambda)$, si la probabilité de chaque valeur possible est définie par : $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$ avec $k = 1, 2, ..., n$

Si x suit une loi de Poisson de paramètre $\lambda > 0$, on note $X \sim P(\lambda)$ avec la loi de probabilité : $P(\lambda) = e^{-\lambda} \frac{\lambda^k}{k!}$

λ est un nombre positif quelconque. Nous savons que la somme de toute probabilité vaut l'unité et que $e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$

Vérifions que c'est bien une distribution de probabilité et calculons l'espérance et la variance. En effet, 1° $\sum p_k = 1$

$$\lambda_0 = \sum_{k=0}^{\infty} p_k = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} e^{\lambda} = 1$$

$$E(P(n)) = \lambda_1 = \sum_{k=0}^{\infty} k p_k \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

$$= \lambda_2 = \sum_{k=0}^{\infty} k^2 p_k \lambda e^{-\lambda} \sum_{k=1}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} + \sum_{k-1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

2° Si $X \sim P(\lambda)$, on a :

$$E(X) = \sum x_i p_i = \lambda$$

$$V(X) = \sum p_i (x_i - E(X))^2 = \lambda$$

Tableau des lois discrètes usuelles [7]

Loi	Loi de probabilité	Espérance	Variance
Uniforme discrète	$P(X=x) = \frac{1}{n} \quad x \in \{1,2,...,n\}$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$
Bernoulli	$P(X=x) = P^x (1-p)^{1-x} \quad x \in \{0,1\}$	p	$p(1-p)$
Binomiale	$P(X=x) = C_n^x p^x (1-p)^{n-x} \quad x \in \{0,1\}$ $= C_n^x p^x q^{n-x}$	np	$np(1-p)$ ou npq
Poisson	$P(X=x) = e^{-\mu} \frac{\mu^y}{y!} \quad x \in \{0,1,...,n\}$	μ	μ

I.3. Exemples de lois continues [1], [4]

Les lois de probabilités continues sont caractérisées par les densités de probabilités.

I.3.1. Loi uniforme continue

On dit qu'une variable X suit une loi uniforme sur $[a,b]$si les valeurs prises par Xsont dans $[a,b]$ et si la probabilité pour que les valeurs de X appartiennent à un intervalle $I=[\alpha,\beta]\subset[a,b]$.est proportionnelle à la longueur de l'intervalle I, c'est-à-dire $P(\{X\subset I\})=\dfrac{\beta-\alpha}{b-a}$. D'autre part, $\forall x\in[a,b]$, on a $P\left[\{a\le X\le x\}\right]=\dfrac{x-a}{b-a}$.

$E(X)=\dfrac{a+b}{2}$ et $V(X)=\left(\dfrac{b-a}{12}\right)^2$

I.3.1. La loi exponentielle :

Un système complexe peut tomber en panne : la détermination des probabilités de son fonctionnement est fondamental par la gestion de ses relations avec l'environnement. La variable aléatoire « instant ou un système complexe tombe en panne » suit souvent une loi exponentielle.

On dit qu'une variable X de support $\mathbb{R}^+$, suit une loi exponentielle de paramètre $\lambda>0$ si sa fonction de densité F est donnée par :

$$f(x)=\begin{cases}\lambda e^{-\lambda x} & si \quad 0\le x<\infty\\ 0 & \text{sinon}\end{cases}$$

Cela se fait dans l'ensemble qui respecte la fonction $f:\mathbb{R}\to\mathbb{R}^+$on note : $X\simeq Exp(\lambda)$, cette loi est souvent utilisée pour représenter la durée de vie d'un objet.

$E(X)=\frac{1}{\lambda}$ et $V(X)=\frac{1}{\lambda^2}$

I.3.2. La loi Gaussienne ou Normale

Une variable aléatoire X suit une loi normale de paramètre μ et σ, $(X \sim N(\mu,\sigma))$ lorsque sa densité de probabilité est sa fonction f définie par :

$f:\mathbb{R}\to\mathbb{R}$

$x \to \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$ où $f(x)$ est la densité de probabilité

X : Variable quantitative continue $(-\infty<x<+\infty)$

μ : Moyenne de la variable X, $\mu \in \mathbb{R}$.

σ : Ecart-type de la variable X, aussi σ est un réel strictement positif.

$E(X)=\mu$ et $V(X)=\sigma^2$

Cette loi modélise une variable dont la valeur est le résultat de nombreuses causes produisant un effet moyen et une dispersion autour de cet effet moyen. Ces conditions se résument aux conditions dites de Borel : si une variable aléatoire dépend de nombreux facteurs indépendants entre eux agissant chacun avec une faible intensité de façon additive et avec le même ordre de grandeurs, alors cette variable suit une loi normale.

- Elle varie de $-\infty$ à $+\infty$
- Elle est symétrique par rapport à la valeur moyenne μ
- Elle présente deux points d'inflexion en $\mu-\sigma$ et $\mu+\sigma$.
- La densité $f(x)$ maximale correspond en abscisse à la moyenne.
- La densité $f(x)$ tend vers 0 quand x tend vers $-\infty$ ou $+\infty$

I.3.3. Loi Normale standard ou Normale centrée réduite

soit $X \sim N(\mu,\sigma)$ et $Z = \frac{X-\mu}{\sigma}$ alors, on $Z \sim N(0,1)$ et on dit que Z suit la loi Normale centrée réduite, si sa fonction de densité f est donnée par $f(z) = \frac{1}{\sqrt{2\pi}e^{\frac{-z^2}{2}}}$ la fonction de répartition F de la variable Z est donnée par

$$F(z) = \int_{-\infty}^{z} f(z)dz$$

$$= P(Z \leq z)$$

Tableau des lois continues

Loi	Densité	Espérance	Variance
Uniforme continue sur [a,b]	$f(x) = \frac{1}{b-a} \quad x \in [a,b]$	$\frac{b-a}{2}$	$\frac{(b-a)^2}{12}$
Gauss $LG(\mu,\sigma)$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$ $x \in]-\infty,+\infty[$	μ	σ^2
Exp de paramètre λ	$f(x) = \lambda \exp(-\lambda x), \quad x \in [0,+\infty]$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma $\gamma(\lambda,r)$	$f(x) = \frac{\lambda}{\sigma(r)} e^{-\lambda x}(\lambda x)^{r-1}$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$
Normale standard ou Normale centrée réduite	$f(z) = \frac{1}{\sqrt{2\Pi}} e^{-\frac{z^2}{2}}$		$V(Z) = E(Z^2) - E^2(z)$ $= E\left(\frac{X-\lambda}{\sigma}\right)^2 = \frac{V(X)}{\sigma^2} = 1$

I.4. Autres lois usuelles [1], [4]

I.4.1. Loi de Pearson : χ^2

Cette loi a été découverte par Karl Pearson (1857-1936).

Soit $Z_1, Z_2, ..., Z_n$, n variables aléatoires indépendantes normales centrées réduites. La somme de leurs carrés : $Z_1^2 + Z_2^2 + ... + Z_n^2$ est une variable aléatoire qui suit une loi de χ^2 à n ddl. On note $\chi^2 \sim \chi^2(n)$. Sa fonction de densité de la variable χ^2 est donnée par $f(x) = \frac{x^{\frac{n}{2}-1} e^{\frac{-x}{2}}}{2^{\frac{n}{2}} \sigma^{n/2}}$ pour $x \geq 0$

$$E(\chi^2) = n$$
$$V(\chi^2) = 2n$$

Aussi, d'une manière la plus simple, étant donné n variables aléatoires Gaussiennes $N(0,1)$: X_i la loi de la variable aléatoire $Y = \sum_{i=1}^{n} (X_i - \bar{X})^2$ est appelée loi du χ^2 à $n-1$ ddl.

C'est la loi la plus utilisée en statistiques après la loi normale. En effet, son statut de somme des carrées lui confère un rôle analogue à celui de notre distance Euclidienne. Elle servira donc parfois de « distance » entre les probabilités. Mais dans la résolution des problèmes statistiques et probabilistes, il n'existe pas d'expression analytique pour la fonction de répartition $f(x) = \frac{x^{\frac{n}{2}-1} e^{\frac{-x}{2}}}{2^{\frac{n}{2}} \sigma^{\frac{n}{2}}}$, on a donc recours à des tables pour effectuer des calculs.

I.4.2. Loi de Student :st(v)

Découvert par William Sealy Gosset dit Student en 1876-1937, cette loi se statifie sur l'étude de degré de liberté (ddl).

On appelle degré de liberté (ddl) le nombre de variables indépendantes mise en jeu moins le nombre de relations entre ces variables.

La loi de student est utilisée entre autres pour comparer des moyennes. Si X est une variable aléatoire suivant une loi χ^2 à n degrés de liberté et Z une variable indépendante de la première et suivant une loi normale centrée réduite $N(0,1)$, alors la variable, $T=\frac{z}{\sqrt{X}}$ suit une loi de student à n degrés de liberté (ddl).

Formule de la loi $$f(t)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)}\times\left(1+\frac{t^2}{n}\right)^{-\frac{(n+1)}{2}}$$

$f(t)$:densité de probabilité

T : Variable continue $(-\infty<t<+\infty)$

n : Nombre de ddl $(-\infty<n<+\infty)$

Γ : Fonction gamma

La moyenne de la loi est $\mu=0$ si $v>1$ et la variance $\sigma^2=\frac{n}{n-2}$ si $n>2$

I.4.3. Loi de Fisher-Snédécor : $f(n_1,n_2)$

Cette loi a été découverte par George W. Snédocor (1876-1974) et Sir Ronald Aylmer Ficher (1890-1962). Elle se présente comme suit : soit X_1, une variable aléatoire qui suit une loi de Pearson χ^2 à n_1 ddl, et soit X_2 une variable aléatoire

qui suit une loi χ^2 à n_2 ddl. Alors la loi de variable $F = \frac{\frac{X_1}{n_1}}{\frac{X_2}{n_2}}$ est appelée loi de snédécor à (n_1, n_2) ddl.

Enoncé : Le rapport de deux variables indépendantes X_1 et X_2 distribuées selon une loi χ^2, chacune d'elles divisée par ses degrés de liberté, définit une variable aléatoire de Fischer dont la densité de probabilité est donnée par :

$$f(n_1,n_2)(x) = \frac{\Gamma\left[\left(\frac{n_1+n_2}{2}\right)\right]}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{x^{\frac{n_1-2}{2}}}{\left(1+\frac{n_1}{n_2}\right)\left(\frac{n_1 n_2}{2}\right)}$$

n_1 et n_2 étant le nombre de ddl de X_1 et X_2 respectivement

Lois de probabilité

Loi	Loi de probabilité	Espérance	Variance
Student $T(n)$	$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{N}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{\left(\frac{n+1}{2}\right)}$	$\mu = 0$ si n> 1	$\sigma^2 = \frac{n}{n-2}$ si $n > 2$
Pearson : $\chi^2(1)$	$f(x) = \frac{x^{\frac{n}{2}-1} e^{\frac{-x}{2}}}{2^{\frac{n}{2}}\sigma^{\frac{n}{2}}}$ $pour x \geq 0$	$E(x^2) = n$	$V(x^2) = 2n$
Fisher-Snédecor $F(n_1,n_2)$	$f(n_1,n_2/x) = \frac{\Gamma\left[(n_1+n_2)\right]}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} \frac{x^{\frac{n_1-2}{2}}}{\left(1+\frac{n_1}{n_2}x\right)\left(\frac{n_1+n_2}{2}\right)}$	$\frac{n_2}{n_2-2}$ $(n_2 > 2)$	$\frac{n_2^2}{n_1}\frac{n+n_2-2}{(n_2-2)^2(n_2-4)}$ $n_2 > 4$

CHAPITRE II. LA REGRESSION DE POISSON

Bien que la régression linéaire soit fréquemment utilisée dans les applications de la statistique, il arrive à certaines occasions que celle-ci ne soit pas appropriée. Par exemple, si la variable réponse est une variable de dénombrement. C'est plutôt la régression de poisson ou régression logarithmique qui sera utilisé. Une hypothèse fondamentale de la régression de poisson est que la moyenne et la variance soient égales. Nous pouvons vérifier cette condition à l'aide de la statistique de Pearson. Dans ce chapitre, nous allons nous référer sur certains points essentiels :

En premier lieu, la famille exponentielle sera décrite, ensuite la loi de poisson et ses différentes propriétés seront énoncées. Les modèles linéaires généralisés seront introduit pour ensuite faire place au sujet principal de ce chapitre : la régression de poisson.

II.1. La Famille exponentielle [11]

On dit qu'une loi fait partie de la famille exponentielle si sa fonction de densité (ou de probabilité) peut être réexprimée sous la forme

$$F_y(\theta / y;\phi) = \exp\left\{ \frac{y(\theta) - b(\theta)}{a(\phi)} - C(y,\phi) \right\}$$

Où b (θ) ne dépend pas de y et où $C(y,\theta)$ne dépend pas du paramètreθ. Les formules d'espérance et de variance d'une loi faisant partie de la famille exponentielle seront présentées à la section.

Soient $y_1, y_2, ..., y_n$ les réalisations des variables aléatoires indépendantes $[y_1, y_2, ..., y_n]$, où l'on suppose que $y_i (i = 1,2,...,n)$ à comme fonction de densité $F_y(\theta_i / y_i : \phi)$ de vraisemblance est définie comme étant

$\prod_{i=1}^{n} F_y(\theta_i / y_i : \phi)$ où n est le nombre d'observations ou individus.

Ainsi, pour une loi faisant la famille exponentielle.

$$L(\theta_i / y_i : \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i(\theta_i) - b(\theta_i)}{a(\phi)} - C(y_i, \phi) \right\}$$

$$= \exp\left\{ \sum_{i=1}^{n} \frac{y_i(\theta_i) - b(\theta_i)}{a(\phi)} - \sum_{i=1}^{n} C(y_i, \phi) \right\} \text{ où } y = \left[y_1, y_2, ..., y_n \right]'$$

Quant à la fonction log-vraisemblance, elle s'obtient en prenant le logarithme naturel de la fonction de vraisemblance.

$$\text{Donc, } l(\theta_i / y_i : \phi) = \ln\left\{ L(\theta_i / y_i : \phi) = \sum_{i=1}^{n} \frac{y_i(\theta_i) - b(\theta_i)}{a(\phi)} - \sum_{i=1}^{n} C(y_i, \phi) \right\}$$

II.2. La loi de Poisson [11]

On dit que Y suit une loi de poisson de paramètre μ si sa fonction de probabilité

$$\mathrm{P}[Y = y] = \begin{cases} e^{-\mu} \sum_{t=0}^{[y]} \frac{\mu^y}{y!} & \text{si, } y = 0,1,2,...,n \\ 0 & \text{sinon} \end{cases}$$

Où μ est un nombre réel positif. De plus, on a que Y a comme fonction de

$$\text{répartition } \mathrm{P}[Y = y] = \begin{cases} e^{-\mu} \sum_{t=0}^{y} \frac{\mu^y}{y!} & \text{si, } y = 0,1,2,...,n \\ 0 & \text{sinon} \end{cases}$$

Où μ est un nombre réel positif. De plus, on a Y qui a comme fonction de

$$\text{répartition } P[Y \le y] = \left\{ e^{-\mu} \sum_{t=0}^{y} \frac{\mu^t}{t!} \quad si \quad y \ge 0 \right.$$

Où $[y]$ correspond, à la partie entière de y. Afin de démontrer que la loi de Poisson de paramètre μ fait partie de la famille exponentielle. On doit exprimer sa fonction de probabilité sous la forme de l'équation :

$L(\theta_i / y_i : \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{y_i(\theta_i) - b(\theta_i)}{a(\phi)} - C(y_i, \phi) \right\}$ suite à une transformation de la fonction de probabilité, on obtient

$$P[Y \le y] = \frac{e^{-\mu + y\ln(\mu)}}{e^{\ln(y!)}} = \exp\left\{ \frac{y\ln(\mu) - \mu}{1} - \ln(y!) \right\}$$

Alors, les paramètres de la famille exponentielle sont :

$\theta = \ln(\mu)$

$b(\theta) = \exp(\theta) = \exp(\ln(\mu)) = \mu$

$a(\phi) = 1$

$c(y, \phi) = \ln(y')$

II.2.1. Le lien entre la loi de Poisson et la loi binomiale [11]

La loi de Poisson peut être vue comme étant un résultat limite de la loi binomiale. En effet, soit $Y \sim$ binomiale (n, p) avec $n \to \infty$ et $p \to 0$ et posons $\mu = np$, on écrit :

$$P[Y = y] = \binom{n}{y} p^y (1-p)^{n-y} = \frac{n!}{y!(n-y)!} p^y (1-p)^{n-y}$$

$$= \frac{n!}{y!(n-y)!} \left(\frac{np}{n}\right)^y \left(1 - \frac{np}{n}\right)^{n-y}$$

$$= \frac{n(n-1)(...)(n-y+1)(np)^y \left(1 - \frac{np}{n}\right)^n}{n^y y! \left(1 - \frac{np}{n}\right)^y}$$

En faisant tendre $n \to \infty$ et $p \to 0$ de sorte que $np \to \mu$, on voit alors que

$$\left(1-\frac{np}{n}\right)^{y} \to \left(1-\frac{\mu}{n}\right)^{n} \simeq e^{-\mu}$$

$$\frac{n(n-1)...(n-y+1)}{n^{y}} \simeq 1, \qquad \left(1-\frac{np}{n}\right)^{y} \to \left(1-\frac{\mu}{n}\right)^{y} \simeq 1$$

Finalement, $P[Y=y]=\frac{\mu^{y}}{y!}e^{-\mu}$ qui est la fonction de probabilité d'une loi de poisson de paramètre μ.

II.2.2. La fonction génératrice des moments de la loi de Pearson [11]

La fonction génératrice des moments de la Loi de poisson notée $M_{y}(t)$ est utile afin de trouver les moments d'ordre K où $E[Y^{K}]=M_{y(t)}^{(K)}\Big|_{t=0}$.

On définit cette fonction génératrice des moments comme étant $E[e^{ty}]$. *On a* $M_{Y}(t)=E[e^{ty}]$

$$=\sum_{y=0}^{\infty} e^{ty}e^{-\mu}\frac{\left(\mu e^{t}\right)^{y}}{y!} = e^{-\mu+\mu e^{t}}$$

$$=e^{-\mu\left(e^{t}+1\right)}$$

A l'aide de cette fonction génératrice des moments, l'espérance et la variance peuvent être calculées : $E[Y]=\mu$ et $Var[Y]=E[Y^{2}]-E^{2}[Y]=\mu$

En sachant que la Loi de poisson fait partie de la famille exponentielle l'espérance et la variance avaient pu être trouvées à l'aide des paramètres de cette famille exponentielle. En effet, on a $E[Y]=b'(\theta)$, alors,

$$E[Y]=\frac{d}{d\theta}\exp(\theta)=\exp(\theta)=\mu$$

Comme on l'a obtenu précédemment avec la fonction génératrice des moments, de plus $Var[Y]=b''(\theta)a(\phi)$. Donc, on obtient

$$Var[Y]=\frac{d^2}{d\theta^2}\exp(\theta)=\exp(\theta)=\mu$$

Comme également a été obtenu avec la fonction génératrice des moments on obtient donc une propriété intéressante de la loi de poisson appelée propriété d'équidispersion, impliquant que $E[Y]=Var[Y]$, c'est-à dire qu'une loi est équidispersée dans le cas où son espérance et sa variance sont égales ; elle est sur dispersée (Sous dispersée) dans le cas où son espérance est inférieure (supérieure) à sa variance.

On mentionne que si $Y_j \approx \text{Poisson}\left(\mu_j\right)$; $(j=1,2,...)$ que les Y_j sont des variables aléatoires indépendantes et que $\sum_{j=1}^{\infty}\mu_j<\infty$ alors $Z_y=\sum_{j=1}^{\infty}Y_j \sim \text{Poisson}\left(\sum_{j=1}^{\infty}\mu_j\right)$

II.3. La régression de poisson [11]

La régression de poisson est utilisée dans le cas où la variable réponse Y_i et un vecteur de régresseurs x_i, on a $P[Y_i=y_i/x_i]=\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$ où $\ln(\mu_i)=x_i'\beta$ et où le lien logarithmique, $i=1,2,...,n$

C'est-à-dire $\ln(\mu_i)=x_i'\beta$ est le lien le plus commun en régression de poisson. De plus β est un vecteur de $p=p'+1$ composantes à estimer.

II.3.1. Les variables offset

Souvent en régression de poisson, des variables offset sont utilisées lorsque la variable de dénombrement Y est proportionnelle à une certaine autre variable exogène, que l'on veut inclure dans le modèle (dans le prédicateur linéaire). On

n'estime pas le coefficient β de cette nouvelle variable exogène : on le force plutôt à prendre la valeur unitaire.

Prenons l'exemple d'une régression où l'on voudrait prévoir le nombre de voitures dans un stationnement. La variable réponse est donc le nombre de voitures retrouvées dans un stationnement en particulier. On pourrait inclure, en variable offset, la superficie du stationnement Z_i.

Puisque plus le stationnement est grand, plus il est susceptible de loger un grand nombre de voitures. Ainsi, on pourrait utiliser $\mu_i = e^{x_i'+e^0+\ln(z_i)}$

Ici, X_i contient toutes les variables exogènes d'intérêt, à l'exception de la superficie du stationnent. Alors $\mu_i = z_i e^{x_i \beta z_i}$

Donc, si la superficie du stationnement est multipliée par une composante C quelconque, on obtient

$$\begin{aligned}\mu_i &= z_i e^{x_i\beta+\ln(z_i)} \\ &= e^{x_i'\beta+\ln c+\ln(z_i)} \\ &= ce^{x_i\beta}+\ln(z_i)\end{aligned}$$

La moyenne, μ_i est donc multipliée par C et elle est alors proportionnelle à la grandeur du stationnement, tel qu'il avait été supposé au préalable.

II.3.2. Le Modèle de la régression de Poisson

Supposons n observations indépendantes d'une variable réponse $Y_i(i=1,2,...,n)$ et p variables explicatives. De plus, supposons $Y_i / x_i \sim poisson(\mu_i)$ et que la fonction de lien et $g(\mu_i)=\ln(\mu_i)$

On tentera donc d'estimer μ_i l'espérance de la variable réponse l'estimation des différentes $\beta_k\,(k=1,2,...,p')$ est généralement faite par la méthode du maximum de vraisemblance. Cependant, en pratique l'emploi de la méthode numérique d'un logiciel sera nécessaire.

II.3.3. L'interprétation des paramètres $\hat{\beta}_k$

Les paramètres $\hat{\beta}_k$ ont une interprétation particulière sans lien logarithmique. Ainsi, $\hat{\beta}_0$ représente le logarithme naturel de l'espérance de la variable réponse lorsque les p variables exogènes prennent simultanément la faveur 0 :

$$\hat{\mu} = e^{\hat{\beta}_0+\left(\beta_1\times 0+(\beta_2\times 0)+...+\left(\beta_{p'}\times 0\right)\right)} = e^{\hat{\beta}_0} \Rightarrow \hat{\beta}_0 = \ln(\mu_i)$$

Quant aux paramètres $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_{p'}$ si on augmente $x_i l\,(l<\mathrm{p}')$ d'une unité et que l'on maintient constante la valeur des autres variables exogènes, alors la valeur moyenne de Y_i est multipliée par $e^{\beta l}$: $\mu_i = e^{\hat{\beta}_0+\hat{\beta}_1 x_{i1}+\hat{\beta}_2 x_{i2}...+\hat{\beta}_{p'} x_{ip'}}$

Remarque : une loi de poisson est donnée par sa loi de probabilité

$(1)\,\forall k, p(x=k)>0$

$$(2)\sum_{k=0} p(X=k) = \sum_{k=0} e^{-\lambda}\frac{e^k}{k!} = e^{-\lambda}\sum_{k\geq 0}\frac{z^k}{k!}$$

$$= \sum_{k=0}\frac{z^k}{k!} = e^{\lambda}$$

$$= \sum_{k=0} p(X=k) = e^{-\lambda}e^{\lambda} = 1$$

Exemple : Une suspension bactérienne contient 5000 bactéries/litre. On en semence à partir de cette suspension, 50 boites de pétri à raison d'1cm^3 par boîte.

Si X, représente le nombre de colonies par boîte, alors la loi de probabilité de X est : $X \sim P(\lambda = 5)$ la probabilité pour qu'il y ait aucune colonie sur la boîte de pétri est : $P(X=0) = \frac{5^0 e^{-5}}{0!} = 0,0067$ soit approximativement 0,67% de chance.

La probabilité qu'n y ait au moins une colonie sur la boîte de pétri est $P(X>0) = 1\ p(x>0) = 1 - p(x=0) = 1 - 0,0067 = 0,9933 \quad soit 99,3\%$ de chance d'avoir au moins une colonie bactérienne qui se développe dans la boîte de pétri. Comme pour la loi binomiale il est possible d'utiliser une formule de récurrence pour calculer les valeurs de probabilités successives.

$$P(X=k) = \frac{\lambda}{k} p(X=k-1).$$

→L'Espérance d'une variable aléatoire de Poison est E(X)= λ par définition

$E(X) = \sum_{k\geq 0} kP_k = \sum_{k\geq 0} \frac{\lambda^k e^{-\lambda}}{k!}$, avec $k \in \mathbb{N}$, valeurs prises par la variable aléatoire X, $avec \sum_{k\geq 0} \frac{\lambda^k}{k!} = \left[\lambda + \frac{\lambda^2}{1!} + \frac{\lambda^3}{2!} + ... + \frac{\lambda^{k+1}}{k!}\right] = \left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + ... + \frac{\lambda^n}{k!}\right]$

D'où $E(X) = \lambda e^{-\lambda} \sum_{k>0} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$

→La variance d'une variable de Poisson est : $V(X) = \lambda$ par définition $V(X) = \sum_{k\geq 0} k^2 P_k - E(X)^2 = \sum_{k\geq 0} k^2 \lambda^k \frac{e^{-\lambda}}{k!} - \lambda^2$. En posant $k^2 = k + k(k-1)$, alors :

$$\sum_{k\geq 0} k^2 \lambda^{-k} \frac{e^{-\lambda}}{k!} = \sum_{k\geq 0} \frac{k}{k!} \lambda^k e^{-\lambda} + \sum_{k\geq 0} \frac{k(k-1)}{k!} \lambda^k e^{-\lambda}$$

D'où $\sum_{k\geq 0} k^2 \lambda^k e^{-\lambda} = \left[\lambda + \frac{\lambda^2}{1!} + \frac{\lambda^3}{2!} + ... + \frac{\lambda^{k+1}}{k!}\right] + e^{-\lambda}\left[\lambda^2 + \frac{\lambda^3}{1!} + \frac{\lambda^4}{2!} + ... + \frac{\lambda^{k+2}}{k!}\right]$

$$d'où \quad \sum_{k\geq 0} k^2 \frac{\lambda^k e^{-\lambda}}{k!} = \lambda^2 e^{-\lambda} \sum_{k\geq 0} \frac{\lambda^k}{k!} + \lambda e^{-\lambda} \sum_{k\geq 0} \frac{\lambda^k}{k!}$$

$$V(X) = \lambda^2 e^{-\lambda} e^{\lambda} + \lambda e^{-\lambda} e^{\lambda} - \lambda^2 = \lambda \Rightarrow V(X) = \lambda$$

II.3.4. Problème d'absence d'équidispersion [11]

Nous avons déjà dit que la loi de poisson repose sur la vérité qui insiste sur l'égalité entre son espérance et sa variance. Nous venons de donner un exemple illustrilatif ci-haut. Alors, cette condition appelée propriété d'équidispersion. Cependant, il arrive fréquemment que la variance de la variable réponse soit supérieure à son espérance. On parle alors d'un problème de surdispersion. Afin de tester si les données sont surdispersées ou équidispersées, l'une des deux statistiques :

$$\phi_p ou \quad \phi_\Delta où : \quad \phi_p = \frac{\chi^2 depearson}{nbre de \quad ddl}$$

$$\phi_\Delta = \frac{statistique de \quad déviance}{nbre de \quad ddl}$$

doit être calculée. Les statistiques du Khi-deux de Pearson et de déviance est déjà définies dans le premier chapitre. Dans le cas où $\phi \geq 1$ il y a sur dispersion et si $\phi \simeq 1$ il y a équidispersion. Lorsqu'il y a sur dispersion dans les données (aussi appelée variabilité extra-Poissonnienne) deux solutions s'offrent : continuer les analyses avec la loi binomiale négative ou tenir compte de la sur dispersion en modifiant les résultats obtenus en disant les statistiques du Khi-deux par ϕ et en multipliant les variances et les covariances par ϕ.

II.4. Les différents types de résidus d'une régression de Poisson [11]

En régression linéaire classique, les résidus sont définis comme étant la différence entre la valeur observée de la variable endogène et sa valeur prédite

par le modèle obtenu. Dans ce cas, les résidus sont indépendants et identiquement distribués de moyenne nulle et de variance constante σ^2.

Cependant, lorsque la variable réponse en est une de dénombrement les résidus $Y_i - \hat{\mu}_i$ ne sont pas de même variance qui provient d'une distribution asymétrique. Ainsi aucun résidus ne provient d'une distribution symétrique et n'est de moyenne nulle et de variantes composantes. Trois types de résidus ont été définis pour remédier à ce problème :

- Résidus d'Ascombe ;
- Résidus de Pearson ;
- Résidus de Déviance.

Ces trois types de résidus seront respectivement définis tout de suite et leur utilisation est plutôt décrite ultérieurement.

I.4.1. Les résidus d'Ascombe

Les résidus de Ascombe sont définis comme étant la transformation de Y s'approchant le plus d'une loi normale centrée réduite. Dans le cas où y suit une loi de Person $Y^{\frac{2}{3}}$ est la transformation se rapprochant le plus de la normale centrée reduite (μ_c cullagh α, Nelder, 1989). Ainsi, on définit les résidus de Ascombe (r_{A_i}) par $r_{A_i} = 1{,}5\dfrac{\left(y_i^{\frac{2}{3}} - \hat{\mu}_i^{\frac{2}{3}}\right)}{\hat{\mu}_i^{\frac{1}{6}}}$

I.4.2. Les résidus de Pearson

Les résidus de pearson r_{P_i} sont quant à eux utilisés lorsque la propriété qui nous intéresse est l'homoscédasticité. Ceux-ci sont définis comme étant $r_{P_i} = \dfrac{Y_i - \mu_i}{\sqrt{\hat{w}_i}}$

On note que $\hat{w}_i$ est un estimé de la variance de y_i ; dans le cas de la loi de poisson

$\hat{w}_i = \hat{\mu}_i$ et alors $r_{P_i} = \dfrac{Y_i - \hat{\mu}_i}{\sqrt{\hat{w}_i}}$

Pour de grandes tailles d'échantillons, les résidus r_{Pi} sont de moyenne nulle et de variance unitaire. Cependant leur distribution est asymétrique. A l'opposé pour des petites tailles d'échantillons les résidus de Pearson studentisés sont utilisés.

Ceux-ci définis comme étant $r_{P_i} = \dfrac{r_{P_i}}{\sqrt{\hat{w}_i}} = \dfrac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i(1-h_{ii})}}$

I.4.3. Les résidus de déviance

Les résidus de déviance, r_{Di} sont souvent utilisés lorsque la variable réponse, Y provient d'une loi faisant partie de la famille exponentielle. Ils sont obtenus à l'aide de la fonction log- vraisemblance et sont définis comme étant :

$$r_{D_i} = signe(y_i - \mu_i)\sqrt{\left[2\alpha(\phi)\{(y/y:\phi) - l(\hat{\mu}_i/y_i)\}\right]}$$

Pour une loi de poisson

$$r_{D_i} = signe(y_i - \mu_i)\sqrt{\left[2\{-y_i + y_i\ln(y_i) - \ln(y_i!) + \hat{\mu}_i - y_i\ln(\hat{\mu}_i) + \ln(y_i!)\right]}$$

$$= signe(y_i - \mu_i)\sqrt{2\left\{y_i\ln\left(\frac{y_i}{\mu_i}\right) + (\hat{\mu}_i - y_i)\right\}}$$

Tout, comme pour les résidus de Pearson, on peut utiliser les résidus de déviance studentisés lorsque la taille d'échantillon est petite. Ceux-ci obtenus à l'aide de l'expression $r_{D_i} = \dfrac{r_{D_i}}{\sqrt{1-h_{ii}}} = \dfrac{signe(y_i - \hat{\mu}_i)\sqrt{2\left\{y_i\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (\hat{\mu}_i - y_i)\right\}}}{\sqrt{1-h_{ii}}}$

Les trois types de résidus peuvent s'exprimer en fonction de C comme suit :

$$r_{A_i} = \sqrt{\hat{\mu}_i \times 1{,}5\left(c^{2/3} - 1\right) \times 1{,}5\left(c^{2/3} - 1\right)}$$

$$= r_{P_i} = \sqrt{\hat{\mu}_i \times c - 1 \times c - 1}$$

$$= r_{D_i} = \sqrt{\hat{\mu} \times \left\{2\left(\ln(c)\right) - c + 1\right\}^{1/2} \times \left(\ln(c)\right) - c + 1\}^{1/2}}$$

On remarque que chacun de 3 types de résidus vaut 0 lorsque $C-1$, c'est-à-dire lorsque $Y = \hat{\mu}$ et ils augmentent lorsque C croît de cette figure, on voit que les résidus d'Ascombe et de Déviance prennent des valeurs semblables pour des valeurs données de C. Cependant les résidus de Pearson prennent des valeurs beaucoup plus grandes pour les mêmes valeurs données de C.

CHAPITRE III. EXEMPLE D'APPLICATION DE LA REGRESSION DE POISSON

Dans ce chapitre, nous allons essayer de donner quelques applications dans la vie courante de la régression de Poisson.

Il est très important de savoir dans quel domaine et dans quelles conditions la régression de poisson est applicable et comment l'appliquer.

III.1. Le modèle de comptage [12]

Dans les modèles de comptage, la variable endogène prend un petit nombre de valeurs positives. Cette variable discrète peut être selon les cas :

- Le nombre de brevets déposés par une Entreprise
- Le nombre d'accidents de travail dans une entreprise
- Le nombre de faillite dans le secteur bancaire
- Etc.

III.1.1. Le modèle de régression de Poisson

Le modèle de base de la littérature économique pour la représentation et l'analyse des données de comptage est le modèle de poisson.

La variable endogène par exemple, le nombre de fois qu'un pays africain H ait été sous ajustement durant la période (1980-1989) notée Y_i est supposé suivre une loi de Poisson.

La probabilité pour qu'un pays H soit sous ajustement est donc :

$$P(Y_i = y) = \frac{e^{-\lambda_i} \lambda_i^{\,y}}{y!};\ y \in \mathbb{N}, \lambda_i > 0, i = 1, \ldots\ldots, n$$

où λ est le paramètre de la distribution de Poisson tel que

$$E(Y_i) = Var(Y_i) = \lambda_i$$

Ce paramètre est lié à p variables exogènes par la forme log-linéaire,

$$\log \lambda_i = xie : i = 1,2,...,n$$

Où xi est un vecteur $(1, p)$ associé au vecteur de paramètre $\beta(p,1)$. Le choix de la spécification log-linéaire s'explique essentiellement par la nécessité d'avoir de paramètres λ_i positifs. Les avantages de cette forme fonctionnelle sont analogues à ceux du modèle économétrique habituel de la régression, en particulier $E(Y_i / x_i)_i = \lambda_i = e^{x_i\beta}$

$$\Leftrightarrow \log E(Y_i / x_i) = x_i\beta$$

β s'interprète donc comme étant une élasticité lorsque les variables exogènes sont en logarithme. Toutefois, contrairement aux modèles log-linéaires traditionnels β n'est pas l'élasticité de la variable endogène mais de son espérance Mathématique. Pour un n échantillon, le modèle de comptage de Poisson peut apriori être estimé par la méthode des moindres carrées non linéaires ou par la méthode du Maximum de vraisemblance. La log-vraisemblance de cette spécification est :

$$\log L = \sum_{i=1}^{n} \left\{ e^{x_i\beta} + y_i\beta - \log(y_i!) \right\}$$

Les équations de vraisemblance sont : $\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} x_i \left(y_i - e^{x_i\beta} \right)$.

L'estimateur du maximum de vraisemblance $\hat{\beta}ML$ de β est solution des équations suivantes des moments empiriques (Lee ; 1986)

$$\sum_{i=1}^{n} x_i' y_i = \sum_{i=1}^{n} x_i' e^{-x_i\beta}$$

Le Hessien est donné par : $\frac{\partial^2 \log L}{\partial\beta\partial\beta'} = -\sum_{i=1}^{n} \lambda_i . x_i' x_i$

Le Hessien est défini négatif $\forall x$ et $\forall \beta$. La méthode de Newton est alors pour ce modèle un algorithme simple qui converge rapidement.

III.1.2. Les données : origine des données

Elles proviennent d'une étude de la caisse centrale de coopération économique (aujourd'hui ; Agence Française de développement) réalisée par Leen Hart, l'hériteau et vama Tanka (1991). Cette étude d'une part à fourni des éléments de comparaison des principales variables macro-économiques des pays d'Afrique Subsaharienne, anglophones et francophones, membres ou non de la zone Franc et d'autre part à faire ressortir les effets de la politique d'ajustement sur les performances économiques vingt-huit pays avaient été retenus et pour lesquels on disposait des données fiables couvrant la période 1980-1989.

Parmi les vingt-huit pays, douze font partie de la zone Franc : Bénin, Burkina-Faso, Côte d'Ivoire, Mali, Niger, Sénégal, Togo, Comores, Cameroun, Centrafrique, Congo, Gabon. Pour les autres pays Africain, il s'agit de 12 pays : Burundi, Gambie, Ghana, Kenya, Madagascar, Malawi, Mauritanie, Maurice, Nigéria, Sierra Léone, Soudan, Somalie, Tanzanie, Zaïre, Zambie et Zimbabwe.

Ainsi, pour ces vingt-huit pays et pour la période considérée, on dispose des données sur les variables suivantes :

-Indice des prix à la consommation ;
- valeur en Droits de Tirage Spéciaux (DTS) des exportations et des importations ;
- solde des paiements courants en % du Produit Intérieur Brut (PIB) ;
- épargne budgétaire (soit recettes définitives hors dons moins dépenses courantes intérêts compris) en % des recettes ;
- croissance du PIB en volume.

Selon les auteurs ci-dessus cités, ces variables avaient été choisies comme représentatives des objectifs de l'ajustement : rétablir les grands équilibres (balance des paiements et finances publiques) et ralentir l'inflation tout en atteignant la meilleure croissance possible. Par ailleurs, elles sont comparables entre elles, quelle que soit la dimension des pays et leur unité monétaire puisqu'elles sont exprimées soit en DTS, soit en ratios ou en taux de croissance. On dispose également sur la période sous réserve des informations sur les pays ayant reçu des prêts d'ajustement structurel.

III.2. Modélisation du risque en assurance automobile [8]

Nous avons entretenu la théorie et la méthodologie à suivre dans le cadre des modèles linéaires généralisés. Nous allons à présent nous intéresser plus en détail à la modélisation du risque automobile. Nous évoquerons ainsi dans un premier temps les notions de coût moyen et de fréquence qu'il est usage de modéliser séparément sous l'hypothèse à contrôler l'indépendance de ces deux facteurs puis nous nous intéressons plus particulièrement à la régression de poisson et aux phénomènes de sur dispersion. Enfin nous reviendrons sur l'objectif de la modélisation à savoir l'analyse des segments sur ou sous tarifés et une éventuelle évolution des tarifs.

III.2.1. Fréquence et coût moyen

Comme nous l'avons fait remarquer plus haut, il est d'usage en assurance automobile de Modéliser séparément le coût moyen de sinistre et la fréquence de sinistre. La prime pure est ensuite calculée en multipliant le cout moyen pour la fréquence. L'hypothèse sous-jacente à cette Méthodologie est indépendance entre ces deux notions. Cette indépendance est une règle générale admise, mais il est tout de même préférable de la contrôler. Pour ce faire nous pourront utiliser un test d'indépendance basé sur le coefficient de corrélation de Pearson, sur le taux de Kendall ou encore sur le Rhô de Pearson.

En règle générale, les montants de sinistres seront modalisées à partir d'une loi Gamma. En effet, ces derniers correspondent bien à une distribution avec la moyenne. En pratique, nous observons souvent une distinction entre les sinistres matériels et les sinistres corporels l'échelle de valeurs associées à ces deux types de sinistre étant trop différente d'une catégorie à l'autre. De même, la modélisation des sinistres d'un montant exceptionnel, fait souvent l'objet d'une attention particulière et utilise la théorie des valeurs extrêmes. Dans le cadre de cette étude, les sinistres dits graves feront l'objet d'une modélisation spécifique, d'une part de leur montant moyen (avec une loi gamma) et d'autre part probabilité d'occurrence (avec une régression logistique). En ce qui concerne la fréquence de sinistres, elle fait en règle générale l'objet d'une modélisation semblable à une régression de poisson. En effet, nous observons bien un processus de comptage pour lequel nous désirons modéliser une proportion (fréquence).

Cependant, la modélisation de la fréquence est généralement relativement complexe, d'une part parce que le nombre d'observations sans sinistres est très important et d'autre part car l'hypothèse (sous-jacente à l'utilisation d'une loi de poisson selon laquelle la variance est égale à la moyenne est rarement vérifiée. Dans ce dernier cas, on parle alors soit de sous-dispersion soit de surdispersion

et l'on est souvent amené à effecteur des modélisations plus complexes afin de corriger ce phénomène.

III.2.2. Description de la régression

En montant Y la variable à expliquer et X les variables explicatives, nous cherchons à maximiser la log-vraisemblance que l'on peut écrire facilement. La loi conditionnelle de Y sachant l'observation i étant une loi de poisson (de paramètre Mi nous pouvons écrire,

$$\alpha(\beta) = \ln\left(\prod_{i=1}^{n} \exp(-\mu_i)\frac{\mu_i^{y_i}}{y_i!}\right)$$

$$= \sum_{i=1}^{n} \ln\left[\left(-\mu_i - \log(y_i!) + y_i \ln \mu_i\right)\right]$$

Or nous savons que dans le cadre des modèles linéaires généralisées nous avons la relation $g(\mu) = n(x)$ *avec* $g(\mu)$ la fonction de lien. En choisissant la fonction de bien canonique pour la régression de poissons, nous savons ainsi que $log(\mu) = n(x_i)$ Par concavité de la fonction de vraisemblance en β, il suffit alors de regarder les dérivées du premier ordre, qui s'écrivent $\frac{\gamma\alpha(\beta)}{\gamma\beta_j} = \sum_{i=1}^{n} x_{ij}(Y_i - \mu_i)$

Ceci nous permet alors de remarquer que si l'on considère une catégorie de risque définie par l'occurrence d'une variable qualitative, alors le nombre de sinistres observés associés à ce niveau de risque est égal à son homologue théorique. En effet, si l'on ne sélectionne qu'une catégorie de risque particulière (les hommes par exemple) alors nous avons la relation : $\sum Y_i = \sum \mu_i$

Ceci nous indique donc que les « primes fréquences » attribuées aux différentes catégories de risque composent exact égaux à 1) de plus, le modèle reconstitue sans erreur le nombre total de sinistres observés (pour autant qu'il existe un intercept) nous ferons alors remarquer qu'un règle générale, nous posséderons

liée à chaque observation. Bien entendu, cette information est importante et joue un rôle capital dans la modélisation de la fréquence. Nous utilisons cette information, mais sans estimer de coefficient associé (que nous fixons alors à 1). En notant ti la durée d'observation associée à la i^{ème} observation, nous utiliserons alors la relation suivante : $\ln(\mu_i) = n(x_i) + \ln(t_i)$

Ainsi, lorsque nous désirerons obtenir une fréquence annuelle associé à une observation, il nous suffira de calculer non pas $\hat{\mu}_i = t_i \exp(\hat{\eta}_i)$ mais : $\hat{\mu}_i = 365 \times \exp(\hat{\eta}_i)$ en supposant ici que la durée d'exposition est exprimée en nombre de jours.

III.2.3. Surdispersion : définition, causes et détection

Comme nous l'avons vu précédemment, l'utilisation de la régression de poisson repose sur l'hypothèse forte d'égalité entre la variance et l'espérance de la variable à appliquer (on parle alors d'équidispersion) .En pratique cette équidispersion est rarement vérifiée ce qui peut remettre en doute l'utilisation de la régression de poisson. Si la variance est supérieur à la moyenne, nous parlerons alors de surdispersion inversement si la variance est inférieur à la moyenne nous parlerons de sous-dispersion.

Ce phénomène est généralement dû à l'omission de variables explicatives, pas toujours connues ou accessibles. Une interprétation simple de cette relation de cause à effet peut être mise en avant si nous considérons deux classes du risque C_1 et C_2 de poids P_1 et P_2 sans effet de surdispersion.

$\left(\hat{\sigma}_1^{\,2} = \hat{m}_1 \quad et \quad \hat{\sigma}_2^{\,2} = \hat{m}_2\right)$ mais que nous aurons omis de séparer l'espérance de la classe $C_1 U\, C_2$ correspondrait à la somme pondérée de m_1 et m_2 Tandis que la variance voudrait : $\hat{\sigma}^2 = P_1\hat{\sigma}_1^{\,2} + P_2\hat{\sigma}_2^{\,2} + P_1(\hat{m}_1 - \hat{m}_2)^2 + P_2(\hat{m}_2 - m)^2 \geq \hat{m}$

Nous constatons donc bien une surdispersion, l'égalité n'étant possible que dans l'hypothèse où les classes du risque C1 et C2 ne sont pas différentiables et ainsi $\hat{m}_1 = \hat{m}_2 = \hat{m}$. Il est ainsi possible de contrôler la présence d'une sur ou sous dispersions en représentant pour chaque classe de risque la variance empirique en fonction de la moyenne empirique. Si les points sont autour de la première bissectrice, nous pourront alors valider l'hypothèse d'équidispersion. Dans le cas contraire le phénomène de surdispersion sera celui le plus observée en pratique et mis en évidence par une variance plus élevées en règle générale.

La présence de la sur ou sous dispersion dans les données relativement à un modèle de poisson peut également être mise en évidence en estimant un paramètre de surdispersion. Une estimation de ce paramètre est donnée en effectuant le rapport du Khi généralisés de Pearson sur $n-p$ avec n le nombre d'observations et p le nombre de variables explicatives. Cette estimation peut également être effectuée en faisant le rapport est proche de 1, l'hypothèse d'équidispersion peut être retenue.

A l'inverse, si ce rapport est supérieur à 1 (respectivement inférieur à 1) : Nous sommes en présence de sur dispersion (respectivement de sous dispersion)

Ne doit laisser paraître aucune tendance sous l'hypothèse d'équidispersion et une droite de régression de ce ration devrait correspondre à une droite horizontale d'ordonnée 1. En pratique, cette analyse graphique permet d'obtenir une indication quant à la modélisation à mettre en œuvre en cas de sur ou sous dispersion.

III.2.4. Solutions d'amélioration

Approche préalable, la quasi vraisemblance

Avant de mettre en avant les possibilités d'améliorations en présence de sur dispersion, nous allons définir et mettre en avant l'unité de ce que l'on appelle la quasi vraisemblance. Nous nous plaçons dans le cadre classique des modèles linéaires généralisés, avec Y une variable à expliquer, X une matrice des variables explicatives, β un vecteur des vecteurs de coefficient à estimer , ϕ un paramètre de dispersion et g () une fonction de lien. Nous définissons alors la quasi-vraisemblance comme suit en considérant Y un vecteur d'observations (individus) de moyenne μ et de fonction de variance $V(\mu)$

$$\mathbb{Q}(Y,M)=\int_{y}^{a}\frac{y-t}{\phi v(t)}dt$$

Nous pouvons alors vérifier que cette fonction possède trois propriétés communes avec la log- vraisemblance d'une loi de la famille exponentielle utilisée en Generalised Linear Method (GLM) , à savoir ;

$$E\left(\frac{\partial\mathbb{Q}}{\partial M}\right)=0$$

$$E\left(\frac{\partial^2\mathbb{Q}}{\partial M^2}\right)=-\frac{1}{\phi v(1)}$$

$$Var\left(\frac{\partial\mathbb{Q}}{\partial M}\right)=\frac{1}{\phi v(1)}=E\left(\frac{\partial^2\mathbb{Q}}{\partial^2 M^2}\right)$$

Ces propriétés dont la démonstration sera laissée aux soins de l'assuré (aucune complexité majeure), correspondent aux propriétés ainsi que pour les notions de convergence et de normalité asymptotique. Nous pouvons en effet chercher à

maximiser la quasi vraisemblance par rapport à β en calculant en une observation Y_i, la dérivée partielle par rapport à β

$$\frac{\partial \mathbb{Q}(M_i, y_i)}{\partial \beta} = \frac{\partial \mathbb{Q}}{\partial M_i} \frac{\partial M_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta} = \frac{Y_i - M_i}{\phi} xg - 1'(M_i) x X_i^T$$

Nous retrouvons alors exactement le même résultat que pour la maximisation de log-vraisemblance. De plus la variance $\frac{\partial \mathbb{Q}(M_i, Y_i)}{\partial \beta}$ qui est égale à l'espérance de la dérivée seconde se calcule de la même façon que pour la log-vraisemblance et donne résultat. Il est alors possible un algorithme de ficher scoring qui conduit exactement à ce que l'on a appelle IRLS. La normalité asymptotique de l'estimation par « Quasi-maximum de vraisemblance) peut alors être montre et l'on obtient ainsi exactement les mêmes résultats avec la maximisation de la log-vraisemblance. L'utilisation de la quasi-vraisemblance nous permet ainsi de rester dans le cadre des modèles linéaires généralisés avec les résultats identiques. Tout en fixant uniquement deux hypothèses sur l'indépendance des observations ainsi que sur leurs deux premiers moments, alors qu'il était nécessaire de fixer une hypothèse concernant une loi tout entière de la variable à expliquer auparavant. Notons que par analogie, il est possible de définir la quasi –déviance pour une observation y d'espérance M comme

$d(Y, \mu) = -2\phi \mathbb{Q}(Y, \mu)$, la quasi-déviance du modèle saturé étant nulle.

CONCLUSION GENERALE

En définitive, par un cheminement plus ou moins intellectuel effectué par les spécialistes de la notion de régression de Poisson, nous sommes partis des notions de variables aléatoires et des lois les plus utilisées en probabilité et en statistique dans le but de mieux amorcer la régression de Poisson à laquelle on a essayé de donner la formule et la démonstration de ses méthodes. On a donné telle ou telle autre méthode utilisée dans chaque cas qui se présente et les résidus appropriés. On a clôturé avec les exemples d'application illustrant la théorie déjà énoncée.

Ainsi, notre travail n'est pas exhaustif, cependant, il peut servir d'orientation à ceux qui voudront mener une étude approfondie sur la notion de régression de Poisson et ses applications; ce qui permettra de trouver des applications suffisantes dans ce domaine.

REFERENCES BIBLIOGRAPHIQUES

[1] D.YADOLAH et M.GIUSEPPE, *Premier pas en simulation, collection statistique et probabilités appliquées*, Springer-Verlag, Paris, 2008.

[2] G. MILLOT, *Comprendre et réaliser les tests statistiques à l'aide de R*, 1ère Edition de Boeck Université, 2008

[3] G. SAPORTA ; *Probabilité Analyse des données et Statistique*, 2ème édition révisée et augmentée, Mars 2006, Edition TECHNIP.

[4] J.P. LECOUTRE ; *Statistiques et probabilités, Manuel et exercices corrigés*, 3ème et 4ème Edition DUNOD, Paris, 2006 et 2009

[5] N. SAVY ; *Probabilités et Statistiques pour modéliser et décider*, Ellipse Edition Marketing S.A. 2006.

[6] P. MILAN ; *Lois de Probabilité à densité, loi normale*, Terminale S 2013

[7] T. PHAN et J.P. ROWENCZYK, *Statistique et probabilités. Exercices et problèmes*, Dunod, Paris, 2007

Webographies

[8] http://www.ressources-actuarielles.net, consulté le 10/04/2017

[9] http://www.solution stat.ca, consulté> uploads>20, consulté le 13/07/2017

[10] http://google.bi/search?q=la+famille+exponentielle nobelis.eu>photis>, consulté le 7/01/2018

[11] http: //www.archimede.mat.ulaval.ca>L-Veilleux-05-2 pdf, consulté le 15/01/2018

[12] http://www.bdsp ehesp.fr/Base/77 437/ Myriam Khat. (population, French edition), vol. 47, No.4, consulté le 17/09/2017.