الجمهورية الجزائرية الديمقراطية الشعبية

Ministère de L'Enseignement Supérieur et de la Recherche Scientifique
UNIVERSITE SAAD DAHLEB DE BLIDA
Faculté des sciences
Département de Mathématiques

MEMOIRE DE MASTER
En Mathématiques

Option : Modèlisation Stochastique et Statistique

THÈME :

Régression des données de comptage et applications sur l'analyse des risques

Réalisé par
BEN ABDELLAZIZ Faiza & BENHASSINE Rayane

Soutenu devant le Jury :

TAMI Omar	Université Blida 1	Président
FRIHI Redhouane	Université Blida 1	Examinateur
RASSOUL Abdelaziz	ENSH de Blida	Promoteur

Juillet 2022

REMERCIEMENTS

Nous tenons à remercier Dieu de nous avoir donné le courage et pour la volonté, la santé, et la foi pour pouvoir réaliser ce travail.
Le travail de mémoire décrit dans ce manuscrit a été réalisé sous la direction du Dr **RASSOUL Abdelaziz**, nous tenons à lui exprimer nous plus vifs remerciements et nous profond reconnaissance pour leurs aide précieuse et pour nous avoir facilité la réalisation de notre mémoire.
Nous adressons également nos remerciements à tous les membres des jury, nous sommes très reconnaissant à leurs remarques et commentaires qui nous aidions beaucoup pour mieux présenter ce document.
Enfin, un grand merci aux membres de ma famille et à toutes les personnes ayant contribués de prés ou de loin à la réalisation de ce modeste travail.

DÉDICACES

je dédie ce modeste travail
A mes chères parents pour leurs soutien inconditionnel , leurs patience et son encouragement au long de ces années , et pour sacrifices leur amour, et leurs prières tout au long de mes études .
A mon promoteur **A.Rassoul** pour ses conseil.
A ma chère tante **Fatma**, mon estime pour eux est immense, je vous remercie pour tout ce que vous avez fait pour moi .
A mes adorables sœurs **Hadil** et **Meriem**, et mon frère **Yousef** .
A mes grands-parents, que Dieu vous préserve une longue vie heureuse .
A ma grande famille ,mes amis et collègues.
A mon binôme **Faiza** pour son soutien moral, sa patience et sa compréhension tout au long de ce projet.

Rayane

DÉDICACES

Je dédie ce modeste travail accompagné d'un profond amour :
A **mes très chers parents** pour leur tendresse, leur amour inestimable, leur soutien et l'encouragement tout le long de mes études, quoi que je fasse ou que je dise, je ne saurai point vous remercie comme il doit. Je vous aime et j'espère que vous êtes fières de moi.
A mes sœurs **Hassiba** , **Nadira**, **Salima** et **Warda** et
mes frères **Mohamed**,**Abde slam** et**Abdelhak** et **AbdelWahab**
pour l'amour qu'ils me réservent, le soutien moral et leur disponibilité à mes côtés.
A tous les membres de ma famille et toutes les personnes qui se tenaient à côté de moi.
Mes chères copines : **Imen**, **Wafa**, **Fella**, et
Et à la fin, je remercie mon binôme, **Rayane** , qui a contribué à la réalisation de ce travail.

Faiza

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

الملخص:

تعد نمذجة بيانات العد مشكلة واسعة الانتشار في مجالات مختلفة مثل البنوك والاقتصاد القياسي. كما تم استكشاف طرق النمذجة التي تم تكييفها مع هذا النوع من البيانات على نطاق واسع في الأدبيات. إن تراجع بواسون هو الملاذ المعتاد في هذا النوع من الحالات، ومع ذلك فإن العديد من التطبيقات على الحالات الحقيقية قد أبرزت الحاجة إلى إيجاد حلول بديلة تسمح بإدارة المشكلات المتعلقة بالتشتت المفرط والإفراط في الأصفار الناجم عن آليات الظاهرة المدروسة. ومن هنا جاءت فكرة استخدام نموذج عد بديل يعتمد على توزيع ثنائي الحدود سالب.

Résumé :

La modélisation de données de comptage est une problématique très répandue dans divers domaines comme la banque,l'économétrie. Aussi les méthodes de modélisation adaptées à ce type de données ont été largement explorées dans la littérature. La régression de Poisson est le recours standard dans ce genre de situation, cependant de nombreuses applications à des cas réels ont mis en évidence la nécessité de trouver des solutions alternatives permettant de gérer les problèmes sur-dispersion et les excès de zéros induits par les mécanismes du phénomène étudié. D'où l'idée d'utiliser un modèle de comptage alternatif basé sur une distribution binomiale négative.

Abstract :

The modeling of counting data is a widespread problem in various fields such as banking, econometrics. Modelling methods adapted to this type of data have also been extensively explored in the literature. The Poisson regression is the standard recourse in this kind of situation, however many applications to real cases have highlighted the need to find alternative solutions allowing to manage the problems over-dispersion and excess of zeros induced by the mechanisms of the phenomenon studied. Hence the idea of using an alternative counting model based on a negative binomial distribution.

ABREVIATION ET NOTATIONS

— $E(.)$: espérance mathématique.
— Var : variance.
— F : fonction de répartition.
— f : densité de probabilité.
— $i.i.d$: indépendantes et identiquement distribuées.
— $v.a$: variable aléatoire.
— $X_1,..,X_n$: échantillon de taille n.
— $P(A)$: probabilité de l'événement A.
— $\mathbb{N}$: ensemble des entiers naturels.
— $\mathbb{N}*$: ensemble des entiers naturels non nuls.
— $\mathbb{R}$: ensemble des réels.
— $\mathbb{R}_+$: ensemble des nombres réels positifs ou nuls.
— $\mathbb{Z}$:ensemble des entiers relatifs.
— $\binom{n}{k}$:le coefficient binomial.
— $N(t)$:un processus stochastique.
— MCO :la méthode des Moindres Carrés Ordinaires.
— β_i :les coefficients associé à la i^{me} variable explicative x_i.
— $\hat{\beta}_i$: les coefficients du modèle estimet.
— SCT :somme des carées total.
— SCE :somme des carées expliquée.
— SCR :somme des carées residuelle.
— ϵ_t=erreurs de specification(inconnue).
— R^2 :le coefficient de détermination.
— GLMs : modèles linéaires généralisés.
— EMV : estimateur du maximum de vraisemblance.
— $L(\beta; y_1,...,y_n,x_1,...,x_n)$:la fonction vraisemblable.
— $Negbin$:binomiale négative.

— Std. Error : l'erreur type de l'estimation du coefficient

INTRODUCTION GÉNÉRALE

Plusieurs auteurs ont récemment proposé des modèles marginaux et des méthodes pour estimer les données de comptage. Cependant, l'étude mathématique de ces modèles n'est pas toujours infaillible et repose sur des hypothèses mathématiques simplificatrices. L'objectif de ce mémoire est d'étudier ces modèles qui reposent encore sur des hypothèses mathématiques simplificatrices. Ainsi, le mémoire couvre plusieurs aspects : la modélisation statistique, une étude mathématique des modèles proposés, l'élaboration d'inférences statistiques d'accompagnement, et des études de simulation numérique dans le but de valider les modèles proposés, afin de déterminer leur domaine de validité et de comparer leurs performances avec les performances des modèles existants et leur application sur des données réelles.
Dans la première partie, nous présentons un résumé sur la probabilité de base,
les variables aléatoires et les distributions de probabilité associées, en particulier les distributions des données de comptage, la distribution de Poisson et les distributions binomiales négatives, qui sont considérées comme une alternative majeure à la distribution de Poisson.
Étant donné que les distributions de probabilité du recensement n'ont pas été entièrement normalisées dans la littérature économétrique, leurs propriétés sont explorées en détail au chapitre 1.
Les distributions peuvent servir de blocs de construction pour améliorer les modèles de régression pour le calcul des données. La deuxième partie de ce mémoire traite de la notion de modèles linéaires et de modèles linéaires généralisés utilisant la méthode des moindres carrés. Il commence ensuite par une discussion détaillée du modèle de régression de Poisson, compris une comparaison avec le modèle linéaire. Deux questions particulièrement pertinentes pour le praticien sont l'interprétation correcte des coefficients de régression, compris l'inférence basée sur les erreurs types appropriées.Les techniques d'estimation de base et les caractéristiques des estimateurs sont
discutées dérivé d'hypothèses de probabilité maximale ou proche du maximum. Une

section consacrée à l'identification de l'erreur potentielle d'un modèle de régression de Poisson : ses origines, ses conséquences et comment détecter une erreur de spécification grâce à des procédures de test appropriées. Pratiquement dans la régression de Poisson en raison de l'abondance de valeurs nulles et de la présence de quelques valeurs aberrantes, la variance est supérieure à la moyenne. Dans ce cas, on parle d'une diffusion excessive de la variable. Cette position implique une sous-estimation des écarts types et une surestimation des tests statistiques et nous rejetons souvent l'hypothèse nulle selon laquelle les coefficients du modèle ne sont pas significatifs. D'où l'idée d'utiliser un modèle de comptage alternatif basé sur une distribution binomiale négative.
Il s'agit d'une généralisation de la régression de Poisson qui supprime l'hypothèse finie selon laquelle la variance est égale à la moyenne définie par le modèle de Poisson. Régression négative conventionnelle , basée sur la distribution gamma du mélange de Poisson. Cette formule est populaire car elle permet de modéliser l'hétérogénéité des poissons à l'aide de la distribution gamma.
Nous étudierons un modèle de régression où la variable énoncée dans la loi binomiale négative suit ses caractéristiques et les méthodes d'inférence et de diagnostic appropriées. Enfin, une étude de simulation complète de tailles d'échantillons limitées est réalisée pour évaluer la cohérence de nos résultats.
Dans le troisième chapitre, nous intéressons à la pérennité des modèles de comptage. Une application de données réelles a été présentée en trouvant un modèle de prédiction de l'ozone approprié et il a été appliqué pour calculer les données sur l'épilepsie. Avec les deux exemples, nous avons constaté que les modèles linéaires classiques ne sont pas adaptés à l'analyse de variables à l'aide d'une interprétation de type « comptage » (ou réponses). Le nombre est distribué non pas selon une loi ordinaire, mais selon la loi de Poisson et une loi binomiale négative.

CHAPITRE 1

VARIABLES ALÉATOIRES ET DISTRIBUTIONS DE PROBABILITÉ

1.1 Introduction

Dans ce chapitre, nous faisons un rappel sur les probabilités de base, les variables aléatoires, leur fonctions de répartition et les lois de probabilités associées .Ensuit nous allons présenter la variables aléatoires discrètes, après la définition de cette notion nous étudions les principales lois de probabilité discrètes et n donné la définition de la variable aléatoire en détail les principales lois de probabilité continues. Ensuite on a présenté deux distributions discrètes pour les donnés de comptage : la distribution de Poisson et la distribution binomiale négative.

1.2 Les variables aléatoires

1.2.1 Définition

De manière très générale, il s'agit d'un caractère déterminé par l'état d'un système résultant d'une expérience aléatoire, c'est-à-dire d'une propriété X du système (ayant couramment la forme d'une grandeur) qui dépend du résultat obtenu.
En langage mathématique, Soit (Ω, A, P) un espace probabilisé ; on appelle variable aléatoire sur (Ω, A, P), à valeurs dans l'ensemble E (fini ou infini dénombrable) toute application X de Ω dans E telle que : $(\forall x \in E), \omega \in \Omega,\ X(\omega) = x \in A$
Si $E = \mathbb{Z}^*$ on dit que X est entière.
Si $E = \mathbb{N}$ on dit que X est entière positive.
Nous distinguons deux types de variables aléatoires : discrètes et continues.

1.3 Lois de probabilité usuelles

Les lois de probabilité permettent de décrire les variables aléatoires sous la forme d'une «expérience type» puis d'analyser cette expérience en détail pour pouvoir déduire les principales caractéristiques de toutes les expériences aléatoires qui sont du même type. Le travail est fait une seule fois mais il sert à toutes les expériences semblables. L'évaluation de la loi de probabilité et des caractéristiques étant effectuée, l'utilisateur n'a plus à "construire" les probabilités mais simplement à identifier le modèle et à utiliser les résultats connus sur le modèle. On s'intéressera ici à quelques lois qui sont très fréquentes.
Soit Ω un univers muni d'une probabilité P, et soit X une variable aléatoire. On appelle loi de probabilité de X, notée P_X, l'application qui à toute partie A de $\mathbb{R}$ associe

$$P_X(A) = P(\{\omega \in \Omega : X(\omega) \in A\})$$

1.4 Fonction de Répartition

1.4.1 Définition

On appelle fonction de répartition d'une variable aléatoire X, la fonction notée F_X, définie sur $\mathbb{R}$ a valeurs dans $[0,1]$ par :

$$F_X(x) = P(X \leq x) = P_X(]-\infty, x]), \forall x \in \mathbb{R}.$$

La fonction F_X est croissante, continue à droite et a pour limite 0 en $-\infty$ et 1 en $+\infty$. Inversement, si on se donne une fonction F ayant ces propriétés, on a vu dans le cours d'intégration qu'il existe une (unique) mesure de probabilité μ telle que $\mu(]-\infty, x]) = F(x)$, pour tout $t \in \mathbb{R}$. Cela montre qu'on peut interpréter F comme la fonction de répartition d'une variable aléatoire réelle.
Il découle des résultats du cours d'intégration que F_X caractérise la loi P_X de X. On a, en particulier

$$P(a \leq X \leq b) = \begin{cases} F_X(b) - F_X(-a) & si \quad a \leq b \\ F_X(-b) - F_X(a) & si \quad a < b \end{cases},$$

et les sauts de F_X correspondent aux atomes de P_X .

1.5 Variables aléatoires discrètes

On appelle variable aléatoire discrète définie sur $(\Omega, \mathcal{A}, \mathbf{P})$ une application telle que $X(\Omega)$ est dénombrable en général $X(\Omega)$ est fini, c'est à dire le résultat d'une épreuve. Si X prend n valeurs $x_1, x_2, ..., x_n$, $n \in \mathbb{N}$ telle que :

$$\sum_{n\in\mathbb{N}} P(\omega \in \Omega : X(\omega) = x_n) = \sum_{n\in\mathbb{N}} P_X(x_n) = 1 \tag{1.1}$$

On appelle espérance de la variable réelle X à valeurs dans un ensemble fini des nombres réels, le nombre réel :

$$m = E(X) = \sum_{i=1}^{n} x_i P(X = x_i) \tag{1.2}$$

La variance de X est définie comme la moyenne quadratique des écarts à la moyenne.

$$Var(X) = \sum_{i=1}^{n} (x_i - m)^2 P(X = x_i) = E(X^2) - [E(X)]^2 \tag{1.3}$$

1.6 Variables aléatoires continues

1.6.1 Définition

Une variable aléatoire est dite continue si elle peut prendre toutes les valeurs dans un intervalle donné (borné ou non borné). En règle générale, toutes les variables qui résultent d'une mesure sont de type continu.
La fonction de densité de probabilité d'une variable aléatoire continue a les propriétés suivantes :
Si la fonction de densité de probabilité $f_X(x)$ est continue en un point x , alors la fonction de distribution cumulative est dérivée, et sa dérivée :

$$dF_X(x) = f_X(x)$$

Étant donné que la valeur de la variable aléatoire X ne dépend que de l'intégrale de la fonction de densité de probabilité, la valeur de la fonction de densité de probabilité à des points individuels n'affecte pas les performances de la variable aléatoire. Plus précisément si une fonction n'a qu'un nombre fini et dénombrable infini de points qui diffèrent de la fonction de densité de probabilité de X, ou a une mesure de 0 par rapport à l'axe réel entier (il s'agit d'un ensemble de mesure nulle), alors cette fonction est aussi peut être la fonction de densité de probabilité de X.

Une variable aléatoire continue a une probabilité de 0 en tout point. En corollaire la probabilité qu'une variable aléatoire continue prenne une valeur dans un intervalle n'a rien à voir avec le fait que l'intervalle soit ouvert ou fermé.
Notez que la probabilité $P(x = a) = 0$ mais $X = a$ n'est pas un événement impossible.

1.6.2 Propriétés d'une variable aléatoire continue

Il s'agit de variables aléatoires continues unidimensionnelles et les variables continues multidimensionnelles sont similaires.
Fonction de densité de probabilité pour les données aléatoires : elle représente la probabilité que l'amplitude instantanée tombe dans une plage spécifiée et est donc une fonction de l' amplitude . Elle varie avec l'ampleur de la gamme prise.
La fonction de densité $f(x)$ a les propriétés suivantes :

- $f(x) \geq 0$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $P(a \leq X \leq b) = \int_a^b f_X(x)dx.$

1.7 Fonction caractéristique

La fonction caractéristique peut être obtenue par transformée de Fourier de la fonction de densité de probabilité .

$$\phi_x(j\omega) = \int_{-\infty}^{+\infty} f(x)e^{j\omega x}dx$$

La fonction propre a une relation biunivoque avec la fonction de densité de probabilité. Par conséquent, connaître la fonction caractéristique d'une distribution équivaut à connaître la fonction de densité de probabilité d'une distribution.
Le nième moment d'une variable aléatoire X est l' espérance mathématique de X élevée à la nième puissance à savoir

$$E[x^n] = \int_{-\infty}^{+\infty} x^n f_X(x)dx$$

On a l'espérance mathématique d'une variable continue X est le nombre réel :

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x)dx.$$

La variance mathématique d'une variable continue X définie par :

$$\sigma^2 = E(X^2) - E(X)^2 = \int_{-\infty}^{+\infty} x^2 f_X(x)dx - E(X)^2.$$

1.8 Distributions pour les données de comptage

1.8.1 Distribution de Poisson

Soit X une variable aléatoire discrète définie sur $\mathbb{N} \cup \{0\} = \{0, 1, 2, ...\}$. X suit une distribution de Poisson de paramètre λ, s'écrit $X \sim Poisson(\lambda)$ si et seulement si la fonction de probabilité est la suivante :

$$P(X = k) = \frac{e^{-\lambda}\lambda^K}{k!}, \lambda \in \mathbb{R}^+, \quad k = 0, 1, 2, ... \tag{1.4}$$

La fonction génératrice de probabilité de la distribution de Poisson est donnée par :

$$G(s) = \sum_{k=0}^{\infty} s^k P(X = K) = \sum_{k=0}^{\infty} s^k \frac{e^{-\lambda}\lambda^K}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda s)^k}{k!} = e^{\lambda_s - \lambda} \tag{1.5}$$

La fonction de probabilité de Poisson est obtenue comme suit :

$$P(X = k) = (k!)^{-1} \frac{d^k P}{ds^k} \mid_{s=0} = \frac{e^{-\lambda}\lambda^k}{k!} \tag{1.6}$$

La distribution de Poisson a une valeur prévue

$$E(X) = P'(1) = \lambda \tag{1.7}$$

et une variance

$$Var(X) = P''(1) + P'(1) - [P'(1)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda \tag{1.8}$$

sinon la valeur attendue peut être dérivée directement en utilisant la probabilité fonction :

$$E(X) = \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} = \lambda \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda \tag{1.9}$$

L'égalité de la moyenne et de la variance est caractéristique de la distribution de Poisson. Il joue un rôle crucial dans la suite de la discussion et sera appelé équi-dispersion. Les écarts par rapport à l'équi-dispersion peuvent être des sur-dispersions (la variance est supérieure à la moyenne) ou sous-dispersion (la variance est plus faible) que la moyenne). Contrairement à d'autres distributions multi-paramètres, tels que la distribution normale, une violation de l'hypothèse de variance est suffisante pour violation de l'hypothèse de Poisson.

1.8.1.1 Quelques propriétés de la distribution de Poisson

1. Prendre la première dérivée de la fonction de probabilité de Poisson avec respect au paramètre λ, on obtient :

$$\frac{\partial P_k}{\partial \lambda} = \frac{e^{-\lambda}.(-1)\lambda^k}{k!} + \frac{e^{-\lambda}k\lambda^{k-1}}{k!} = P_k\left(\frac{k}{\lambda} - 1\right) \tag{1.10}$$

   Par conséquent, les probabilités P_k diminuent avec une augmentation de λ .

2. Considér les résultats dichotomiques $P(X = 0)$ et $P(X > 0)$. Les capacités de la sonde sont données par

$$P_0 = e^{-\lambda} \quad et \quad P_+ = 1 - e^{-\lambda}$$

1.8.1.2 Sommes des variables aléatoires de Poisson

Supposons que $X \sim Poisson(\lambda)$ et $Y \sim Poisson(\mu), \lambda, \mu \in \mathbb{R}_+$ et que X et Y sont indépendants. La variable aléatoire $Z = X + Y$ est Poisson distribué $Poisson(\lambda + \mu)$. Ce résultat découle directement de la définition des fonctions génératrices de probabilités où sous indépendance.

$$E(s^{X+Y}) = E(s^X)E(s^Y).$$

En outre

$$P^{(Z)} = E(s^{X+Y}) = e^{(\lambda+\mu)+(\lambda+\mu)s} \tag{1.11}$$

Qui est exactement la fonction génératrice de probabilité d'un Poisson distribué variable aléatoire avec paramètre $(\lambda + \mu)$. Par conséquent, $Z \sim \text{Poisson}(\lambda + \mu)$.

$$\begin{aligned} P(Z = k) &= \sum_{i=0}^{k} P(X = k - i)P(Y = i) = \sum_{i=0}^{k} \frac{e^{-\lambda}\lambda^{k-i}}{(k-i)!} . \frac{e^{-\mu}\mu^i}{i!} \\ &= \frac{e^{-\lambda-\mu}}{k!} \sum_{i=0}^{k} \frac{k!}{(k-i)!i!} \lambda^{k-i}\mu^i = \frac{e^{-\lambda-\mu}}{k!} \end{aligned} \tag{1.12}$$

Où la dernière égalité découle de la définition des coefficients binomiaux.

1.8.2 Processus de Poisson

Le processus de Poisson est un cas particulier d'un processus de comptage qui à son tour est un cas particulier d'un processus stochastique. Par conséquent, certaines définitions générales seront, d'abord avant que les propriétés du processus de Poisson

soient présentées.
Un processus stochastique $X(t), t \in T$ est un ensemble de variables aléatoires (sur un certain espace de probabilité) indexé par le temps. $X(t)$ est une variable aléatoire qui marque l'occurrence d'un événement à l'instant t. L'expérience sous-jacente elle-même reste non expérimentée et les définitions et les arguments sont encadrés exclusivement en termes de $X(t)$. Si l'index T est un intervalle sur la droite réelle, le processus stochastique est dit être un processus stochastique continu de temps. Si le nombre cardinal de T est égal au cardinal nombre de N il est appelé un processus stochastique temps discret.
Un processus stochastique $N(t), t \geq 0$ est dit être un processus de comptage si $N(t)$ représente le nombre total d'événements qui se sont produits avant t.
Les propriétés suivantes sont valides :

1. $N(t) \geq 0$
2. $N(t)$ est un entier évalué
3. Pour $s < t, N(s) \leq N(t)$
4. Pour $s < t, N(t) - N(s)$ donne le nombre d'événements qui se sont produits dans l'intervalle (s, t).

Un processus de comptage est appelé stationnaire si la distribution du nombre d'événements dans n'importe quel intervalle de temps dépend seulement de la longueur de l'intervalle :

$$(\forall s > 0) \quad N(t_2 + s) - N(t_1 + s) \overset{iid}{\sim} N(t_2) - N(t_1)$$

Un processus de comptage a des incréments indépendants si le nombre d'événements qui se produisent dans des intervalles de temps disjoints sont indépendants.Le processus de Poisson est un processus continu de comptage de temps avec stationnaire et des incréments indépendants. En d'autres termes, il suppose que l'événement d'un événement aléatoire à un moment donné est indépendant du temps et de le nombre d'événements qui ont déjà eu lieu. Soit $N(t, t + \Delta)$ nombre d'événements qui se sont produits entre (t) et $(t + \Delta), t > 0$.

1.8.3 Distribution de Poisson comme limite binomiale

Considérons une expérience dont tous les résultats peuvent être catégorisés sans ambiguïté comme succès (S) ou échec (F). Par exemple, en jetant une pièce de monnaie nous pouvons appeler face un succès et pile un échec. Alternativement en tirant d'une urne qui ne contient que des boules rouges et bleues, nous pouvons appeler rouge un succès et bleu un échec. En général la survenue d'un événement est un succès et la non-occurrence est un échec. la probabilité d'un succès peut être notée par p. Alors

$0 < p < 1$ et la probabilité d'un échec est donnée par $q = 1 - p$.
Supposons maintenant que l'expérience soit répétée un certain nombre de fois,dire n fois. Puisque chaque expérience se traduit par un (F) ou un (S) répéter l'expérience produit une série de S et de F. Ainsi, dans trois dessins d'une urne le résultat rouge, bleu, rouge, dans cet ordre, peut être indiqué par (SFS). L'ordre peut représenter un temps discret. Ainsi, la première expérience est faite au temps $t = 1$, le second au temps $t = 2$, et le troisième au temps $t = 3$. Ainsi la séquence des résultats peut être interprétée comme un processus stochastique de temps discret. La séquence de dessin de l'urne avec remplacement est l'exemple classique et processus temporel discret stationnaire : les résultats des expériences à différents points dans le temps sont indépendants, et la probabilité p d'un succès est constante au fil du temps et égale à la proportion de boules rouges dans l'urne. Dans cette situation, toutes les permutations de la séquence ont la même probabilité.
Définissons une variable X comme le nombre total de succès obtenus dans n répétitions de l'expérience. X est appelé variable de comptage et n constitue une limite supérieure pour le nombre de comptages. Selon les hypothèses d'indépendance et de stationnarité X a une fonction de distribution binomiale avec une fonction génératrice de probabilité

$$P(s) = [q + ps]^n \tag{1.13}$$

n était interprété comme le nombre de répétitions d'une expérience donnée.
Pour introduire explicitement une dimension de temps, considérons un intervalle de temps fixe $(0, T)$ et divisons le en n intervalles de longueur égale, p est maintenant la probabilité de succès dans l'intervalle. La longueur de l'intervalle est donnée par T/n, où T peut être normalisé sans perte de généralité à 1. Dénotons le facteur de proportionnalité par λ Puis $p_n = \lambda/n$ c-à-d $p_n = \lambda$, une constante donnée. De plus, soit $q_n = 1 - \lambda/n$. En substituant ces expressions àP_n et q_n dans 1.13 et en prenant des limites, on obtient

$$\lim_{n\to\infty} P(s) = \lim_{n\to\infty}\left[1 - \frac{\lambda}{n} + \frac{\lambda}{n}s\right]^n = \lim_{n\to\infty}\left[1 + \frac{\lambda(s-1)}{n}\right]^n = e^{\lambda(s-1)} \tag{1.14}$$

Mais 1.14 est précisément la fonction génératrice de probabilité de la distribution de Poisson. Divisons la période de temps fixe en intervalles de plus en plus courts la distribution binomiale converge vers la distribution de Poisson. Ce résultat est «théorème de Poisson» (Feller, 1968)[1], (Johnson et Kotz, 1969)[2]. La limite supérieure du nombre de comptes implicites dans une distribution binomiale disparaît. Le processus de Poisson stochastique à temps discret suppose une dépendance et une stationnarité c-à-d «aléatoire» des essais successifs de Bernoulli.

1.8.4 Distribution binomiale négative

Une variable aléatoire X a une distribution binomiale négative avec des paramètres $\alpha \geq 0$ et $\theta \geq 0$ notées $X \sim Negbin(\alpha, \theta)$ si la fonction de probabilité est donnée par

$$P(X = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)(k+1)} \left(\frac{1}{1+\theta}\right)^{\alpha} \left(\frac{\theta}{1+\theta}\right)^{k} \quad k = 0, 1, 2, ... \tag{1.15}$$

$\Gamma(.)$ désigne la fonction gamma telle que $\Gamma(s) = \int_0^{\infty} Z^{s-1} e^{-z} dz$ pour s > 0. Cette distribution à deux paramètres a une fonction génératrice de probabilité.

$$P(s) = [1 + \theta(1-s)]^{-\alpha} \tag{1.16}$$

La moyenne et la variance sont données par

$$E(X) = \alpha\theta \tag{1.17}$$

et

$$Var(X) = \alpha\theta(1+\theta) = E(X)(1+\theta) \tag{1.18}$$

Pour $\theta \geq 0$ la variance de la distribution binomiale négative dépasse généralement sa moyenne "sur-dispersion "la sur-dispersion disparaît pour $\theta \to 0$. La distribution binomiale négative se présente sous différentes paramétrisations. D'un point de vue économétrique, les considérations suivantes s'appliquent. Afin de pouvoir utiliser la distribution binomiale négative pour l'analyse de régression la première étape consiste à convertir le modèle en paramétrage moyen, par exemple :

$$\lambda = \alpha\theta \tag{1.19}$$

Où λ est la valeur attendue l'inspection de 1.19 montre qu'il y a deux des façons simples de le faire.

1. $\alpha = \lambda/\theta$, Dans ce cas la variance prend la forme

$$Var(X) = \lambda(1+\theta)$$

   Par conséquent,la variance est une fonction linéaire de la moyenne. Ce modèle est appelé " Negbin I " (Cameron et Trivedi,1986) [3].

2. $\theta = \lambda/\alpha$,Dans ce cas la fonction variance prend la forme

$$Var(X) = \lambda + \alpha^{-1}\lambda^2$$

Ce modèle est appelé "Negbin II ". Les fonctions de probabilité associées aux deux modèles sont les suivantes :

◊ **Negbin I :**

$$P(X=k)=\frac{\Gamma(\lambda/\theta+k)}{\Gamma(\lambda/\theta)\Gamma(k+1)}\left(\frac{1}{1+\theta}\right)^{\lambda/\theta}\left(\frac{\theta}{1+\theta}\right)^{k} \tag{1.20}$$

◊ **Negbin II :**

$$P(X=k)=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\Gamma(k+1)}\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(\frac{\lambda}{\alpha+\lambda}\right)^{k} \tag{1.21}$$

Bien que ces deux types soient les paramétrages les plus utilisés en pratique d'autres sont possibles. Par exemple :

$$\alpha=\sigma^{-2}\lambda^{1-k}$$

et

$$\theta=\sigma^{2}\lambda^{k}$$

Comme précédemment, $E(X)=\lambda$. La substitution de α et θ en 1.18donne

$$Var(X)=\lambda(1+\sigma^{2}\lambda^{k})$$

Ainsi, pour $k=0$ cette paramétrisation se réduit à la distribution binomiale négative avec fonction de variance linéaire alors que pour $k=1$, une variance quadratique fonction est obtenue. Winkelmann et Zimmermann (1995)[4] se réfèrent à ce modèle comme « Negbin$_k$ ».
On trouve souvent une autre paramétrisation dans la littérature statistique (voir par exemple DeGroot, 1986)[5]où dans l'expression générale 1.15, $1/(1+\theta)$ est remplacé par p et $\theta/(1+\theta)$ est remplacé par q. Si α est un entier disons n la distribution est appelée distribution Pascal et elle a l'interprétation d'une distribution du nombre d'échecs qui se produiront avant qu'exactement n succès se soient produits dans une séquence infinie d'essais de Bernoulli avec une probabilité de succès p. Pour $n=1$, cette distribution se réduit à la distribution géométrique.

$$P(X=k)=pq^{k-1},\quad k=1,2,3,... \tag{1.22}$$

En résumé, le principal avantage de la distribution Binomiale négative sur la distribution de Poisson est que le paramètre supplémentaire introduit une flexibilité substantielle dans la modélisation de la fonction de variance et donc l'hétéroscédasticité . En particulier il introduit la surdispersion une forme plus générale l'hétéroscédasticité que l'égalité moyenne-variance impliquée par la distribution de Poisson.

1.8.5 Lien avec les autres distributions

La distribution binomiale négative implique la distribution de Poisson. Pour $X\sim Negbin(\alpha,\theta)$, laisser $\theta\to 0$ et $\alpha\to\infty$ de sorte que $\theta\alpha=\lambda$ une constante. La distri-

bution négative binomiale converge vers la distribution de Poisson avec le paramètre λ. Pour une preuve considérer la fonction génératrice de probabilité de la distribution binomiale négative remplacer θ par λ/α, et prendre des limites.

$$\begin{aligned}\lim_{\substack{\alpha\to\infty\\ \theta\alpha\to\lambda}} P(s) &= \lim_{\substack{\alpha\to\infty\\ \theta\alpha\to\lambda}} [1+\theta(1-s)]^{-\alpha} \\ &= \lim_{\alpha\to\infty}\left[1+\frac{\lambda(1-s)}{\alpha}\right]^{-\alpha} \\ &= e^{-\lambda(1-s)} \end{aligned} \tag{1.23}$$

Mais c'est exactement la fonction génératrice de probabilité d'une distribution de Poisson avec le paramètre λ. Une dérivation alternative et un peu plus lourde de ce résultat peut être basée directement sur la fonction de distribution de probabilité

$$\begin{aligned}\lim_{\substack{\alpha\to\infty\\ \theta\alpha\to\lambda}} P(X=k) &= \lim_{\substack{\alpha\to\infty\\ \theta\alpha\to\lambda}} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}\left(\frac{1}{1+\theta}\right)^{\alpha}\left(\frac{\theta}{1+\theta}\right)^{k} \\ &= \lim_{\alpha\to\infty} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\left(\frac{\lambda}{\alpha+\lambda}\right)^{k} \\ &= \lim_{\alpha\to\infty}\left(\prod_{j=1}^{k}\frac{\alpha+j-1}{\alpha+\lambda}\right)\left(\frac{\alpha}{\alpha+\lambda}\right)^{\alpha}\frac{\lambda^k}{k!} \\ &= \lim_{\alpha\to\infty}\left(\prod_{j=1}^{k}\frac{1+(j-1)/\alpha}{1+\frac{\lambda}{\alpha}}\right)\left(\frac{1}{1+\frac{\lambda}{\alpha}}\right)^{\alpha}\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\frac{\lambda^k}{k!}\end{aligned}$$

Où l'on a utilisé l'expression du produit pour le rapport des fonctions gamma et le fait que

$$(\alpha+\lambda)^{-k} = \prod_{j=1}^{k}(\alpha+\lambda)^{-1}.$$

1.8.6 Caractérisation supplémentaire du binôme négatif

La distribution Binomiale négative apparaît de plusieurs façons. C'est une distribution finie d'une série d'expériences de Bernoulli non indépendantes. Elle se présente également comme une distribution mixte et répartition composée.. Pour le mélange, supposons que $X \sim$Poisson(λ) et que λ a une distribution gamma. La distribution marginale de X est alors la distribution binomiale négative. Pour la composition, supposons qu'une distribution de Poisson est composée d'une distribution logarithmique. La distribution composée est alors la distribution binomiale négative.

1.8.7 Sommes des variables aléatoires binomiales négatives

Supposons que X et Y suivent indépendamment des distributions binomiales négatives avec $X \sim$ Negbin I (λ, θ) et $Y \sim$ Negbin I (μ, θ). Il s'ensuit que la variable aléatoire $Z = X + Y$ est de distribution binomiale négative Negbin I$(\lambda + \mu, \theta)$. Pour preuve rappelons que la fonction génératrice de probabilité générique de la distribution binomiale négative est donnée par $P(s) = [1 + \theta(1 - s)]^{-\alpha}$. En paramétrisation Negbin I on obtient

$$P(s)^{(X)} = [1 + \theta(1 - s)]^{-\lambda/\theta}$$

et

$$P(s)^{(Y)} = [1 + \theta(1 - s)]^{-\mu/\theta}$$

ainsi

$$P(s)^{(Z)} = [1 + \theta(1 - s)]^{-\lambda/\theta}[1 + \theta(1 - s)]^{-\mu/\theta} = [1 + \theta(1 - s)]^{-(\lambda+\mu)/\theta} \tag{1.24}$$

Ainsi, les distributions binomiales négatives du type spécifié ci-dessus sont fermées par convolution.
Ce résultat dépend essentiellement de deux hypothèses : premièrement la spécification NegbinI avec fonction de variance linéaire doit être adoptée. Deuxièmement X et Y doivent partager un paramètre de variance commun θ. En d'autres termes la somme de deux distributions binomiales négatives spécifiées arbitrairement n'est en général pas distribuée en binomiale négative.
Encore une autre paramétrisation est souvent trouvée dans la littérature statistique (voir par exemple DeGroot, 1986)[5], où dans l'expression générale 1.15, $1/(1 + \theta)$ est remplacé par p et $\theta/(1 + \theta)$ est remplacé par q . Si α est un entier, disons n, la distribution est appelée distribution de Pascal, et elle a l'interprétation d'une distribution du nombre d'échecs qui se produiront avant qu' exactement n succès se soient produits dans une séquence infinie d'essais de Bernoulli avec probabilité de succès p. Pour $n = 1$, cette distribution se réduit à la distribution géométrique.

$$P(X = k) = pq^{k-1}, \quad k = 1, 2, 3, ..$$

Pour résumer, le principal avantage de la distribution binomiale négative par rapport à la distribution de Poisson est que le paramètre supplémentaire introduit une flexibilité substantielle dans la modélisation de la fonction de variance et donc de l'hétéroscédasticité. En particulieril introduit une sur-dispersion, une forme plus générale d'hétéroscédasticité que l'égalité moyenne-variance impliquée par la distribution de Poisson.

1.8.8 Distribution binomiale

Une variable aléatoire X suit une loi binomiale de paramètres $n \in \mathbb{N}$, et $p \in (0,1)$, noté $X \sim B(n,p)$), si

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,...,n. \tag{1.25}$$

La fonction génératrice de probabilité est donnée par :

$$P(s) = \sum_{k=0}^{n} s^k \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (ps)^k q^{n-k} = (q+ps)^n \tag{1.26}$$

et la moyenne et la variance sont :

$$E(X) = np$$

et

$$Var(X) = np(1-p)$$

Dans les problèmes d'estimation le paramètre binomial n est généralement traité comme donné. Parfois cependant on estime n en fonction des données également. Sous maximum de vraisemblance, il y a deux possibilités. Première on peut respecter la nature entière du paramètre et maximiser au moyen d'une recherche par grille. L'estimateur résultant n'aura pas les propriétés standard d'un estimateur du maximum de vraisemblance. Alternativement, on peut traiter n comme un paramètre continu. Dans ce cas des dérivées peuvent être prises. Depuis

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)}$$

Où $\Gamma(.)$ désigne la fonction gamma et $\Gamma(n+1) = n!$ si n est un entier cela implique le calcul de la fonction digamma. Alternativement, la différenciation directe peut être basée sur une approximation de la représentation factorielle en utilisant la formule de Stirling

$$k! \approx (2\pi)^{1/2} k^{k+1/2} \exp(-k)\{1 + \frac{1}{12k}\}$$

Dans les deux cas une difficulté logique surgit en ce qui concerne l'espace d'échantillonnage possible de la variable aléatoire sous-jacente X si n est un paramètre continu non négatif.

1.8.9 Distribution logarithmique

La variable aléatoire X a une distribution logarithmique si (Johnson et Kotz, 1969, p. 166)[6]

$$P(X=k)=\alpha\frac{\theta^k}{k},\quad k=1,2,...,0<\theta<1 \tag{1.27}$$

Où $\alpha=-[\log(1-\theta)]^{-1}$, La fonction génératrice de probabilité est donnée par

$$P(s)=\sum_{k=1}^{\infty}s^k\alpha\frac{\theta^k}{k}=\sum_{k=1}^{\infty}\alpha\frac{(\theta s)^k}{k}=-\alpha\ln(1-\theta s) \tag{1.28}$$

Où la dernière égalité découle d'un développement en série de Taylor de $\ln(1-x)$ au voisinage de 0 :

$$\ln(1-x)=-\sum_{k=1}^{\infty}\frac{x^k}{k} \tag{1.29}$$

Alternativement, la fonction génératrice de probabilité peut être écrite en utilisant l'expression explicite de la constante de normalisation α comme

$$P(s)=\frac{\log(1-\theta s)}{\log(1-\theta)}. \tag{1.30}$$

La moyenne et la variance sont données par

$$E(X)=\alpha\frac{\theta}{(1-\theta)}, \tag{1.31}$$

et

$$Var(X)=\alpha\theta\frac{(1-\alpha\theta)}{(1-\theta)^2}. \tag{1.32}$$

La distribution affiche une sur-dispersion pour $0<\alpha<1$ (c'est-à-dire $\theta>1-e^{-1}$) et une sous-dispersion pour $\alpha>1$ (c'est-à-dire $\theta<1-e^{-1}$). Contrairement aux distributions précédentes, l'espace échantillonnage de la distribution logarithmique est donné par l'ensemble des entiers positifs. Et en fait il peut être obtenu comme une distribution limite de la distribution binomiale négative tronquée à zéro (Kocherlakota et Kocherlakota,1992)[7]. La raison probable pour laquelle la distribution logarithmique est un concurrent inefficace des distributions de Poisson ou binomiales négatives doit être vue dans sa fonction moyenne compliquée qui en fait mais pas formellement interdit l'utilisation de la distribution dans un cadre de régression. Par exemple, Chatfield, Ehrenberg et Goodhardt (1966)[8] utilisent la distribution logarithmique pour modéliser le nombre d'éléments d'un produit acheté par un acheteur dans une période de temps spécifiée mais ils n'incluent pas de co-variables c'est-à-dire qu'ils ne spécifient aucune régression.

CHAPITRE 2

RÉGRESSION DES DONNÉES DE COMPTAGE

2.1 Introduction

La régression est une méthode de prévision mathématique très utilisée en économie. À partir d'un ensemble des valeurs expérimentales, représentées par des points sur un graphique. La régression est donc l'opération qui consiste à faire passer une droite ou tout autre courbe mathématique le plus prés possible d'un certain nombre de points obtenus d'une manière expérimentale.
i.e, On cherche essentiellement à déterminer la variation de l'espérance mathématique d'une v.a. Y en fonction des variables explicatives X pouvant prendre toute valeur dans un intervalle I de $\mathbb{R}$ différentes valeurs de X dans I correspondent par hypothèse des variables aléatoires distinctes et on est x donc en fait en présence d'une famille de variable aléatoire $\{Y(x)|x \in I\}$. Admettant que pour tout x l'espérance mathématique existe alors $E(Y(x))$ est la fonction $g(x)$ qu'il s'agit de rechercher cette fonction mettant en évidence l'évolution moyenne de l'entité Y à expliquer en fonction de x et appelée fonction de régression.Dans cette approche on considère naturellement que l'incertitude de la prédiction de Y pour le «niveau» x de X se manifeste par une variable aléatoire $\epsilon(x)$ venant s'ajouter à la composante déterministe $g(x)$ dans sa forme la plus générale un modèle de régression simple s'écrit donc :

$$Y(x) = g(x) + \epsilon(x).$$

Puisque $E(Y(x)) = g(x)$ on a nécessairement $E(\epsilon(x)) = 0$ quel que soit x la variable aléatoire $\epsilon(x)$ est appelée erreur ou aléa (d'où la notation habituelle du «e» grec) dans la plupart des modèles on suppose que l'erreur est de même loi quel que soit x ce qui permet d'écrire $Y(x) = g(x) + \epsilon$ (on écrit même parfois simplement $Y = g(x) + \epsilon$ en omettant d'indiquer que la variable aléatoire Y est assujettie à la valeur x)[9] .

2.2 Régression linéaire

2.2.1 Définition

Dans les modèles statistiques, les modèles dits linéaires sont largement dominants. Que les variables prédictives au sein de ces modèles soient numériques, catégorielles (recodées en indicatrices) ou les deux en même temps, on peut exprimer dans une formulation unifiée les méthodes bien connues que sont la régression linéaire, l'analyse de la variance et l'analyse de la covariance. Ce cadre simple déjà assez intégrateur est classiquement nommé modèle linéaire général. Il fait l'hypothèse d'une distribution gaussienne sur la variable dépendante, conditionnellement aux prédicteurs et d'un lien linéaire ou de proportionnalité entre variables explicatives et à expliquer. Or ces modèles ne sont pas toujours adéquats à tous problèmes statistiques, certains sont assez compliqués pour qu'ils soient modélisés par ces modèles simples.
En 1972 de nouveaux modèles ont été formulés par John Nelder et Robert Wedderburn [10] appelés modèles linéaires généralisés comme une généralisation souple de la régression linéaire. Le GLM "General Linéaire Modelés " généralise la régression linéaire en permettant au modèle linéaire d'être relié à la variable réponse via une fonction lien. Aussi comme un moyen d'unifier les autres modèles statistiques y compris la régression linéaire, la régression logistique et la régression de Poisson[11].

2.2.2 Modèle

On note Y la variable aléatoire réelle à expliquer et X la variable explicative (déterministe) ou effet fixe ou facteur contrôle. Le modèle revient à supposer qu'en moyenne, $E(Y)$ est une fonction affine de X [12].

— Dans le cas où X est déterministe, le modèle s'écrit :

$$E(Y) = f(X) = a_0 + a_1 X$$

— Dans le cas où X est aléatoire, le modèle s'écrit alors conditionnellement aux observations de X :

$$E(Y|X = x) = a_0 + a_1 X$$

Il conduit aux mêmes estimations pour une séquence d'observations aléatoires identiquement distribuées (y_t, x_t); $t = 1,...,n$ ($n > 2$ et les x_t non tous égaux) le modèle s'écrit avec les observations :

$$y_t = a_0 + a_1 x_t + \epsilon; \quad t = 1,...,n$$

2.2.3 Forme matricielle

Le modèle s'écrit sous la forme suivantes : [13]

$$Y = Xa + \epsilon \Leftrightarrow \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{2.1}$$

Sous les hypothèses de base suivantes :

- H1 : le modèle est linéaire en X_t (ou en n'importe quelle transformation de X_t).
- H2 : les valeurs X_t sont observées sans erreur (X_t non aléatoire).
- H3 : $E(\epsilon_t) = 0$, l'espérance mathématique de l'erreur est nulle : en moyenne le modèle est bien spécifié et donc l'erreur moyenne est nulle.
- H4 : $E(\epsilon_t^2) = \sigma_t^2$, la variance de l'erreur est constante : le risque de l'amplitude de l'erreur est le même quelle que soit la période.
- H5 : $E(\epsilon_t \in [t, t']) = 0$ si $t \neq t'$, les erreurs sont non corrélées (ou encore indépendantes) : une erreur à l'instant t n'a pas d'influence sur les erreurs suivantes.
- H6 : $Cov(X_t, \epsilon_t) = 0$, l'erreur est indépendante de la variable explicative.

2.2.4 Estimation des paramètres par la méthode des Moindres Carrés Ordinaires (MCO)

La technique des moindres carrés ordinaire (MCO) apporte une réponse au problème posé.
On doit estimer a_1 et a_0 de façon à minimiser la distance au carré entre chaque point observé Y_t et chaque point $\hat{Y}_t$ donné par la droite $\hat{Y}_t = \hat{a}_0 + \hat{a}_1 x_t + e_t$.
Soit $e = Y_t - \hat{Y}_t$, l'écart entre ces deux mesures, la méthode ou la technique des MCO consiste à rechercher les valeurs de a_0 et a_1 de façon à minimiser la quantité suivante :

$$Min \sum_{t=1}^{n} e_t^2 = min \sum_{t=1}^{n} (Y_t - \hat{Y})^2 = min \sum_{t=1}^{n} S^2$$

pour que cette fonction ait un minimum , il faut que les dérivées par-rapport à a_0 et a_1 soient nulles.

$$\frac{\partial S}{\partial a_0} = 0 \iff 2 \sum_{t=1}^{n} (Y_t - a_0 - a_1 X_t)(-1) = 0 \longrightarrow \sum_{t=1}^{n} Y_t = na_0 + a_1 \sum t = 1^n X_t \tag{2.2}$$

$$\frac{\partial S}{\partial a_1} = 0 \iff \sum_{t=1}^{n} (Y_t - a_0 - a_1 X_t)(-X_t) = 0 \longrightarrow \sum_{t=1}^{n} Y_t X_t = a_0 \sum_{t=1}^{n} X_t + \sum_{t=1}^{n} X_t^2 \tag{2.3}$$

En notant $\hat{a_0}$ et $\hat{a_1}$ les solutions des équations (2.2) et (2.3), on obtient d'après l'équation (2.2) :

$$\hat{a_0} = \frac{\sum_{t=1}^{n} Y_t}{n} - \hat{a_1}\frac{\sum_{t=1}^{n} X_t}{n}$$

ou bien :

$$\hat{a_0} = \bar{Y} - \hat{a_1}\bar{X}$$

Puisque

$$\frac{\sum_{t=1}^{n} Y_t}{n} = \bar{Y} \quad et \quad \frac{\sum_{t=1}^{n} X_t}{n} = \bar{X}.$$

En remplaçant la valeur de $\hat{a_1}$ dans l'équation (2.3), on obtient :

$$\sum_{t=1}^{n} Y_t X_t - \bar{Y}\sum_{t=1}^{n} X_t = \hat{a_1}\left(\sum_{t=1}^{n} X_t^2 - \bar{X}\sum_{t=1}^{n} X_t\right).$$

D'où

$$\hat{a_1} = \frac{\sum_{t=1}^{n} X_t Y_t - \bar{Y}\sum_{t=1}^{n} X_t}{\sum_{t=1}^{n} X_t^2 - \bar{X}\sum_{t=1}^{n} X_t} = \frac{\sum_{t=1}^{n} X_t Y_t - n\bar{Y}\bar{X}}{\sum_{t=1}^{n} X_t^2 - n\bar{X}^2} = \frac{\sum_{t=1}^{n}(Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

Les estimateurs des MCO du modéle de régression linéair simple
$Y_t = a_0 + a_1 + \epsilon_t$ sont :

$$\boxed{\hat{a_0} = \bar{Y} - \hat{a_1}\bar{X}}$$

et

$$\boxed{\hat{a_1} = \frac{\sum_{t=1}^{n} X_t Y_t - n\bar{Y}\bar{X}}{\sum_{t=1}^{n} X_t^2 - n\bar{X}^2} = \frac{\sum_{t=1}^{n}(Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1}^{n}(X_t - \bar{X})^2} = \frac{cov(Y_t, X_t)}{{\sigma_{X_t}}^2}}$$

2.2.5 Equation d'analyse de la variance et le coefficient de détermination

Pour calculer le coefficient de détermination, nous démontrons d'abord les deux relations :

$$\sum_{t=1}^{n} e_t = 0$$

la somme des résidus est nulle (la droite de régression passe par le point moyen cela est valable uniquement pour les modèles contenant le terme constant.

$$\sum_{t=1}^{n} Y_t = \sum_{t=1}^{n} \hat{Y}_t,$$

L'égalité entre la moyenne de la série à expliquer et la moyenne de la série ajustée. On démontre d'abord que :

$$\sum_{t=1}^{n} e_i = \sum_{t=1}^{n} (Y_t - \hat{Y}_t) = 0$$

On sait que :

$$Y_t = \hat{Y}_t + e_t = \hat{a_0} + \hat{a_1} X_t + e_t \iff \sum_{t=1}^{n} Y_t = n\hat{a_0} + \hat{a_1} + \sum_{t=1}^{n} e_i$$

$$\sum_{t=1}^{n} e_i = n\bar{Y} - n\hat{a_0} - \hat{a_1} n\bar{X}$$

On remplace $\hat{a_0}$ par sa valeur on obtient alors :

$$\sum_{t=1}^{n} e_i = n\bar{Y} - n(\bar{Y} - \hat{a_1}\bar{X}) - \hat{a_1} n\bar{X}$$

D'ou :

$$\sum_{t=1}^{n} e_i = n\bar{Y} - n(\bar{Y}) - \hat{a_1} n\bar{X}) + \hat{a_1} n\bar{X} - \hat{a_1} n\bar{X}$$

Donc

$$\boxed{\sum_{t=1}^{n} e_t = 0}$$

Puisque $\sum_{t=1}^{n} e_t = 0$ on déduit alors :

$$\sum_{t=1}^{n} e_t = \sum_{t=1}^{n} (Y_i - \hat{Y}) = 0 \iff \sum_{t=1}^{n} Y_i - \sum_{t=1}^{n} \hat{Y}_i = 0$$

On conclut :

$$\boxed{\sum_{t=1}^{n} Y_i = \sum_{t=1}^{n} \hat{Y}_i}$$

A Partir de ces deux équations nous pourrons déduire la fondamentale d'analyse de la variance. On a :

$$Y_t - \hat{Y}_t = e_t \iff Y_t = \hat{Y}_t + e_t$$

d'où :

$$Y_t - \bar{Y} = \hat{Y}_t + e_t - \bar{Y} \longrightarrow (Y_t - \bar{Y})^2 = (\hat{Y}_t - \bar{Y})^2 + e_t^2 + 2(\hat{Y}_t - \bar{Y})e_t$$

Passant aux sommes on trouve :

$$\sum_{t=1}^{n}(Y_t-\bar{Y})^2=\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2+\sum_{t=1}^{n}e_t^2+2\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})e_t$$

Comme :

$$\sum_{t=1}^{n}e_t=0$$

$$\sum_{t=1}^{n}Y_t=\sum_{t=1}^{n}\hat{Y}_t$$

On déduit alors :

$$\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})e_t=0$$

Il résulte :

$$\sum_{t=1}^{n}(Y_t-\bar{Y})^2=\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2+\sum_{t=1}^{n}e_t^2$$

Qu'on peut ecrire comme suit

$$\sum_{t=1}^{n}(Y_t-\bar{Y})^2=\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2+\sum_{t=1}^{n}e_t^2 \tag{2.4}$$

$$SCT=SCE+SCR$$

(2.4) : appelée équation d'analyse de la variance avec :

$$SCT=\sum_{t=1}^{n}(Y_t-\bar{Y})^2$$

est la somme des carrés totaux.Elle indique la variabilité totale de Y c'est t-a dire l'information disponible dans les données.

$$SCE=\sum_{t=1}^{n}(\hat{Y}_t-\bar{Y})^2$$

est la somme des carrées expliquée . Elle indique la variabilité expliquée par le modèle c'est t-a dire la variation de Y expliquée par X.

$$SCR=\sum_{t=1}^{n}e_t^2$$

est la somme des carrées résiduels .Elle indique la variabilité non-expliqué (résiduelle) par le modèle c'est t-a dire d'écart entre les valeur observées de Y et celles par le modèle.

La variabilité totale (SCT) est égale à la variabilité expliquée (SCE) plus La variabilité des résidus (SCR). Cette équation va nous permettre de juger la qualité de l'ajustement d'un modèle, en effet, plus la variance expliqué est "proche" de la variance totale, l'ajustement global du modèle sera meilleur. C'est pourquoi nous calculons le rapport SCE sur SCT :

$$R^2 = \frac{\sum_{t=1}^{n}(\hat{Y}_t - \bar{Y})^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} = 1 - \frac{\sum_{t=1}^{n} e_t^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2} \tag{2.5}$$

Ou bien :

$$R^2 = \frac{SCE}{SCR} = 1 - \frac{SCR}{SCT}$$

R^2 est appelé le coefficient de détermination, il indique la proportion de variance de Y expliquée par le modèle .
Plus il sera proche de la valeur 1, le modèle sera meilleur, la connaissance des valeur de X permet de deviner avec précision celle de Y .
lorsque R^2 est proche de 0, cela veut dire que X n'apporte pas d'informations utiles (intéressantes) sur Y la connaissance des valeur de X ne nous dit rien sur celle de Y.

2.3 Régression de Poisson

2.3.1 Modèle de Poisson

Le modèle de régression de Poisson est le modèle de référence pour les données de comptage de la même manière que le modèle linéaire normal est la référence pour les données continues à valeurs réelles. Les premières références en économétrie incluent Gilbert (1982) [14], Hausman, Hall et Griliches (1984) [15] et Cameron et Trivedi (1986) [3]. Le modèle de Poisson est simple et il est robuste. Si le seul intérêt de l'analyse réside dans l'estimation des paramètres d'une fonction moyenne log-linéaire, il n'y a guère de raison (hormis l'efficacité) d'envisager jamais autre chose que le modèle de régression de Poisson. En fait, son applicabilité s'étend bien au-delà du domaine traditionnel des données de comptage.

2.3.2 Hypothèses du modèle de régression de Poisson

Le modèle standard pour les données de comptage est le modèle de régression de Poisson, qui est un modèle de régression non linéaire. Ce modèle de régression est dérivé de la distribution de Poisson en permettant au paramètre d'intensité λ de dépendre de co-variables (la régression)[16].
Le modèle de régression de Poisson de base relie la fonction de probabilité d'une variable dépendante y_i (aussi appelée endogène ou variable dépendante) à un vecteur de

variables indépendantes x_i (aussi appelée exogène ou variable indépendante). Soit k le nombre de régression (compris, généralement, une constante). Enfin, n est le nombre d'observations dans l'échantillon.

Le modèle de régression de Poisson uni-varié standard suit trois hypothèses :
Hypothèse 1 :

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, ...$$

Où $f(y|\lambda)$ est la fonction de probabilité conditionnelle de y donnée λ,et il doit tenir que $\lambda > 0$.
Hypothèse 2 :

$$\lambda = \exp(x'\beta)$$

Où β est un vecteur $(k \times 1)$ de paramètres, et x' est un vecteur de colonne de dimension $(k \times 1)$.
Hypothèse 3 :
Les paires d'observations $(y_i, x_i), i = 1, ..., n$ sont distribuée indépendamment.

2.3.2.1 Discussion

Les hypothèses 1 et 2 peuvent être combinées pour obtenir la condition suivante : fonction de probabilité :

$$f(y|x) = \frac{\exp(-\exp(x'\beta))\exp(yx'\beta)}{y!}, \quad y = 0, 1, 2, ... \tag{2.6}$$

La distribution de Poisson contient un seul paramètre qui définit simultanément la moyenne et la variance conditionnelle.
La forme telle que définie par les hypothèses ci-dessus désigne une fonction exponentielle (ou logarithme linéaire),

$$E(y|x) = \lambda = exp(x'\beta) \tag{2.7}$$

Et la fonction de covariance conditionnelle exponentielle

$$Var(y|x) = \lambda = \exp(x'\beta) \tag{2.8}$$

Le fait que la moyenne conditionnelle et la variance conditionnelle soient égales en le modèle de régression de Poisson est une caractéristique spéciale même la dispersion qui le fera dans le contexte d'un modèle de régression, les variables linéaires passées affectent la variable dépendante (le nombre d'événements comptés dans une période de temps) à travers l'intensité (ou le taux de fréquence instantané) du processus. L'hétérogénéité de ces dernières est modélisée comme une fonction dissuasive miniature

des variables explicatives. Cela signifie que, contrairement à un modèle de régression linéaire ordinaire, le caractère aléatoire du modèle de Poisson est intrinsèque et non dû à une erreur aléatoire additive qui représente en outre l'hétérogénéité. Si le processus stochastique de base n'affiche pas le caractère aléatoire souhaité, ou s'il n'y a même pas de processus stochastique sous-jacent significatif car je pense que le modèle de régression de Poisson peut toujours être une approximation valide pour le processus de génération de données réelles ainsi qu'un descriptif utile outil. En combinaison avec les hypothèses 1 et 2, l'hypothèse 3 permet de application directe de la méthode du maximum de probabilité pour estimer les paramètres du modèle. L'estimation du maximum de vraisemblance est discutée ci-dessous.

2.3.3 Moindres carrés ordinaires et autres alternatives

Les avantages et inconvénients d'un modèle de régression de Poisson en répondant à la question du praticien "Quand et pourquoi utiliser un modèle de régression de Poisson ?" la première réponse naturelle semblerait que la variable dépendante devrait être dénombrable. Mais cette condition n'est ni nécessaire ni suffisante. Ce n'est pas nécessaire, car un modèle de régression de Poisson a été montré utile également pour les variables non comptées. Un exemple est le modèle de régression exponentielle avec contrôle correct qui émerge constamment de la modélisation de la durée, qui peut être estimée par la régression de Poisson. Un autre exemple plus important est l'estimation de toute élasticité fixe modèle de régression de Poisson.
Évidemment, le fait que la variable dépendante soit un compte n'est pas non plus suffisant.
Premièrement, il existe d'autres modèles de données de comptage qui prennent la nature de la variable dépendante à considérer qui peut être meilleure que le modèle de Poisson. Souvent, ces modèles généralisés permettront un ensemble beaucoup plus riche d'inférences, en particulier en ce qui concerne la probabilité de résultats uniques (par exemple "zéro") et avec elle dans le processus sous-jacent de génération de données structurelles. Spécification possible de modèles de données de comptage alternatifs et le bon choix ce sont des sujets importants que nous aborderons plus tard.
Deuxièmement, on ne sait pas pourquoi on ne peut pas complètement ignorer la nature particulière de la variable dépendante et appliquer des modèles de régression standard tels que le modèle linéaire normal

$$y = x'\beta + e, \quad e|x \sim \mathcal{N}(0, \sigma^2). \tag{2.9}$$

Plusieurs objections peuvent être faites à une telle approche. 2.9 Il ignore la nature discrète de la variable dépendante. Sous modèle linéaire normal, la probabilité d'un résultat donné est nulle. Ainsi, il n'y a pas de conclusions sur un résultat possible.

De plus, le modèle 2.9 permet des résultats négatifs alors que les nombres sont négatifs. En continuant, le modèle viole l'hypothèse 2 selon laquelle la moyenne la fonction est le log -linéaire. Ainsi, 2.9 donnera un estimateur incohérent pour si le processus de génération de données réelles suit un modèle de régression de Poisson. Enfin, il elle ignore l'élasticité hétérodyne inhérente aux données de comptage (voir l'équation 2.8). Le preuve de cette approche n'apparaît que si les nombres sont très grands. Poissons la distribution, par exemple, peut être approchée par une distribution normale, et l'approximation est généralement considérée comme satisfaisante pour $\lambda > 20$.

2.3.3.1 Modèle log-linéaire

Ces problèmes peuvent être partiellement résolus par des méthodes traditionnelles. De puis fonction moyenne.Nous pouvons préciser

$$\ln(y) = x'\beta + \mu, \quad \mu \sim \mathcal{N}(0, \sigma^2) \tag{2.10}$$

où "ln" désigne le logarithme népérien. Dans ce modèle

$$y = \exp(x'\beta + \mu) \tag{2.11}$$

Distribution log-normale avec espérance conditionnelle

$$E(y|x) = \exp(x'\beta + 1/2\sigma^2)$$

similaire aux modèles de régression de Poisson (jusqu'au facteur d'échelle $\exp(1/2\sigma^2)$), et les valeurs de y sont limitées aux traits pleins non négatifs. Si seulement le modèle a une constante globale, on peut redéfinir

$$\tilde{\beta_0} = \beta_0 - 1/2\sigma^2$$

et deux les modèles ont essentiellement la même fonction moyenne.
Cependant, la distribution log-normale implique une fonction de variance différente. En particulier, il considère

$$Var(y|x) = \phi[E(y|x)]^2, \quad \text{où} \quad \phi = \exp(\sigma^2) - 1.$$

En général, l'erreur type log-normale de l'estimation est non comparable aux modèles de Poisson, l'hétéroscédasticité comprend Les erreurs types doivent être calculées.
Les deux problèmes fondamentaux de la méthode log-normale sont les comptages "zéro" ne sont pas autorisés car les logarithmes ne sont définis que comme positifs résultat, et le problème de "retransformer" se pose : si la condition la variance de y

n'est pas quadratique dans l'espérance conditionnelle.Des solutions temporaires au problème zéro ont été proposées, telles que l'élimination tous les résultats sont nuls, ou ajoutez une constante à chaque comptage, comme 0,1 et 0,5 (Voir King, 1988)[17]. Dans ce cas le modèle s'écrit :

$$\log(y + c) = x'\beta + \mu$$

King (1988)rapporte les résultats d'une analyse de Monte-Carlo où le modèle log-linéaire ajusté est appliqué à des données artificielles d'un modèle de régression de Poisson. Il trouve un biais substantiel pour les estimations des paramètres lorsque le modèle log-linéaire est utilisé au lieu du modèle de régression de Poisson. Le biais ne disparaît pas avec l'augmentation de la taille de l'échantillon. Le modèle log-linéaire a tendance à surestimer les paramètres de pente lorsqu'ils sont positifs et à sous-estimer les paramètres de pente lorsqu'ils sont négatifs, c'est-à-dire que les paramètres sont biaisés par rapport à zéro. En introduisant des biais et en ignorant le caractère discret des données, ce modèle est assez insatisfaisant et son utilisation ne peut être recommandée. De même, bien sûr supprimer tous les zéros n'est pas non plus une bonne idée, car cela conduira à des problèmes de sélection d'échantillon endogène similaires à ceux connus du modèle linéaire (Heckman,1979)[18].

2.3.3.2 Moindres carrés non linéaires

Une partie du problème vient du fait que nous considérons un modèle avec une erreur multiplicative $\epsilon = \exp(\mu)$ (voir l'équation 2.11). Envisager des alternatives modèle

$$y = \exp(x'\beta) + v \quad v \quad i.i.d \sim N(0, \sigma^2) \tag{2.12}$$

Ce modèle a la même fonction moyenne que le modèle de régression de Poisson.Le logarithme de y n'a pas besoin d'être pris, et le problème que le nombre de dis est égal à zéro se produit. En fait, du point de vue de l'estimation, le modèle 2.12, s'il estime par maximum de vraisemblance (identique aux moindres carrés non linéaires dans ce cas), si β le monde réel est Poisson. Il n'est pas efficace car il ignore hétéroscédasticité inhérente dans les modèles de régression de Poisson. Cependant le modèle peut être modifié comme $\sigma^2 = \exp(x'\beta)$, dans ce cas itérer les moindres carrés non linéaires pondérés donneront les mêmes résultats que le maximum de vraisemblance maximum de Poisson. Le principal problème de ce modèle est que, bien qu'il n'y ait aucun avantage à faciliter l'estimation ou l'interprétation des paramètres, il non prise en compte du caractère non négatif et entier de la variable dépendante. Le modèle ne peut pas être utilisé pour prédire la probabilité d'occurrence résultat unique.

2.3.4 Estimation des paramètres

La forme exponentielle de la fonction moyenne implique l'augmentation nécessaire en $x'\beta = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$ augmente $E(y|x)$ d'une unité plus il est petit, plus il est éloigné de zéro. en d'autres termes, changement de niveau de $x'\beta$ requis pour maintenir un pourcentage de changement donné dans $E(y|x)$ continu. Par conséquent, la dérivée partielle de $E(y|x)$ tout élément de x dépend de la valeur de $x'\beta$:

$$\frac{\partial(E(y|x))}{\partial x_j} = \exp(x'\beta)\beta_j = E(y|x)\beta_j \quad j = 1,...,k \tag{2.13}$$

Ces effets marginaux diffèrent sensiblement d'un individu à l'autre. Parfois ça Peut être utile pour calculer les effets marginaux pour des individus représentatifs, tels que les moyennes d'échantillon pour les variables explicatives. Ou alors, les effets marginaux attendus (ou moyens) peuvent être calculés rédécesseur

$$E_x\left[\frac{\partial E(y|x)}{\partial x_j}\right] = \beta_j E[\exp(x'\beta)]$$

peut être estimé de manière cohérente par prédécesseur

$$\widehat{E_x}\left[\frac{\partial E(\widehat{y|x})}{\partial x_j}\right] = \frac{1}{n}\sum_{i=1}^{n}(\exp(x'\hat{\beta})\hat{B_j})$$

Cependant, il est plus courant et plus simple de considérer les changements relatifs est associé à de petits changements de x_j dans $E(y|x)$ car celui-ci est constant et égal à β_j :

$$\frac{\partial(y|x)/E(y|x)}{\partial x_j} = \beta_j \tag{2.14}$$

Si x est sous forme logarithmique, alors β_j a une interprétation élastique, donnant la variation en pourcentage de $E(y|x)$ pour chaque variation en pourcentage de x_j. Parfois, nous sommes intéressés à évaluer l'effet d'unités (discrètes) la variation de x_j par rapport à la valeur attendue de y. C'est-à-dire que nous voulons comparer la valeur attendue de y pour x_j et $x_j + 1$. Dans ce cas, la méthode de calcul ne donne qu'une approximation du changement relatif. Définition $\tilde{x} = (1, x_2, ..., x_{j+1}, ..., x_k)'$ .alors le changement relatif exact est

$$\frac{E(y|\tilde{x}'\beta) - E(y|x'\beta)}{E(y|x'\beta)} = \frac{\exp(x'\beta + \beta_j) - \exp(x'\beta)}{\exp(x'\beta)} = \exp(\beta_j) - 1 \tag{2.15}$$

L'exemple le plus typique est une variable fictive qui prend la valeur 0 ou 1. Donc, l'effet relatif de la variable fictive sur les nombres attendus est $\exp(\beta_j) - 1$ autour de

$\beta_j^0 = 0$ donne

$$\exp(\beta_j) - 1 \approx [\exp(\beta_j^0) - 1] + \exp(\beta_j^0)(\beta_j - \beta_j^0)|_{\beta_j^0=0} = \beta_j$$

Ainsi, β_j est l'approximation de premier ordre de l'impact relatif d'une variable fictive pour un petit β_j, et l'approximation linéaire est d'autant meilleure que β_j .
Ces résultats sont similaires à ceux rencontrés dans le modèle log-linéaire standard.
Cependant, il existe une différence conceptuelle qui lève une certaine ambiguïté dans l'interprétation des paramètres de Poisson, ambiguïté qui a d'abord été relevée par Goldberger (1968)[19] pour le modèle log-linéaire (voir aussi Winkelmann, 2001)[20]. Là $E(\log y|x) = x'\beta$, d'où il ne s'ensuit pas que $E(y|x) = \exp(x'\beta)$. Ce n'est que sous certaines hypothèses supplémentaires qu'une expression telle que $\exp(\beta_j) - 1$ identifie correctement le changement relatif de $E(y|x)$ en raison d'un changement d'unité dans x_j .
La situation dans le modèle de régression de Poisson est beaucoup plus simple. Cependant, l'estimation reste un problème. Comme indiqué par Goldberger (1968) pour le modèle log-linéaire, estimant $\exp(\beta_j) - 1$ par $\exp(b_j) - 1$, où b_j est l'estimateur du maximum de vraisemblance, bien que cohérent, introduit un biais de petit échantillon. Un estimateur amélioré a été proposé par Goldberger (1968) et Kennedy (1981)[21].

2.3.5 Période à risque

Les données de comptage mesurent le nombre de fois qu'un certain événement se produit pendant un intervalle de temps donné. La durée de cet intervalle est parfois appelée «période de risque» ou «exposition».
Dans le modèle de Poisson standard, on suppose que la période de risque est la même pour toutes les observations. Sous cette hypothèse, il peut être normalisé à l'unité sans perte de généralité, et $\exp(x'\beta)$ est la valeur attendue de y par intervalle de temps (comme l'année, le mois ou la semaine). Cependant, dans d'autres cas, la période de risque varie selon les observations. Par exemple, Mc Cullagh et Nelder (1989)[22] analysent le nombre d'incidents signalés par type de navire.
Le nombre total de mois de service varie de 45 mois pour un type de navire à 44,882 mois pour un autre. De toute évidence, on s'attendrait à ce que le nombre d'incidents augmente avec le nombre total de mois de service. Dans le même ordre d'idées, Diggle, Liang et Zeger (1994)[23] utilisent les données d'une expérience randomisée pour comparer le nombre de crises d'épilepsie au cours d'une période de 8 semaines période d'observation pré-traitement avec le nombre de crises d'épilepsie pendant la période d'observation post-traitement de 2 semaines.
Enfin, Barmby, Nolan et Winkelmann (2001)[24] analyse le nombre de jours d'absence du travail pour un échantillon de travailleurs dont certains sont sous contrat à 4 jours

de travail par semaine tandis que d'autres sont sous contrat à 5 jours de travail par semaine.
Il n'est pas nécessaire de limiter les différences d'exposition au temps civil. Par exemple, Rose (1990)[25] analyse les déterminants des incidents de la circulation aérienne. Dans son cas, la taille différente de l'exploitation entre les différents transporteurs est exprimée par le nombre de départs réguliers par année (en milliers). Bauer et al. (1998)[26] s'intéressent au nombre d'accidents du travail, en utilisant les données au niveau de l'entreprise pour l'Allemagne.
Encore une fois, on s'attendrait à ce que le nombre d'accidents augmente avec la taille du groupe à risque, ici le nombre de travailleurs. Le cas de référence pour le traitement de l'exposition est d'assumer la proportionnalité. Dans les exemples ci-dessus, Mc Cullagh et Nelder (1989) modélisent le nombre prévu d'incidents causant des dommages aux navires par mois d'exploitation, tandis que Rose (1990)[25] modélise le nombre prévu d'incidents de la circulation aérienne par 1000 départs. Si nous indiquons le niveau d'exposition individuel par t, nous pouvons écrire :

$$E(y|x) = t\exp(x'\beta) = \exp(x'\beta + \log t) \tag{2.16}$$

Ainsi, en proportionnalité un doublement du temps d'exposition double le nombre attendu. À la deuxième ligne du paragraphe 2.16, le $logt$ est parfois appelé « décalage logarithmique ».
Sinon on pourrait vouloir donner l'hypothèse de proportionnalité libre pour l'essai. Une simple possibilité est d'inclure le $logt$ comme un régresseur sans limiter son coefficient à l'unité :

$$E(y|x,t) = \exp(x'\beta + \gamma \log t) \tag{2.17}$$

La restriction $H_0 : y = 1$ peut alors être testée avec des méthodes standard. Alternativement, on peut reparametériser en utilisant $\theta = \gamma - 1$. La fonction moyenne lit alors

$$E(y|x,t) = \exp(x'\beta + \theta \log t + \log t) \tag{2.18}$$

Le temps logarithmique d'exposition est inclus deux fois, à la fois comme décalage et comme régression, et le test de proportionnalité simplifie maintenant le test $H_0 : \theta = 0$.
Il existe une autre variante de ce test si le temps d'exposition ne peut prendre que deux valeurs, t_1 et t_2. Ce cas est présenté dans Barmby, Nolan et Winkelmann (2001). Définissez une variable fictive $D = 1$ si $t = t_1$ et $D = 0$ si $t = t_2$. Ensuite, dans le modèle de régression

$$E(y|x,D) = \exp(x'\beta + \delta D) \tag{2.19}$$

le test de proportionnalité à l'exposition est réduit à $H_0 : \delta = \log(t_1/t_2)$. Pour établir l'équivalence, à noter que

$$\log t = \log t_2 + (\log t_1 - \log t_2)D$$

Ainsi, en stricte proportionnalité

$$E(y|x, D) = \exp(\log t_2 + x'\beta + (\log t_1 - \log t_2)D),$$

$$\delta = \log t_1 - \log t_2 = \log(t_1/t_2)(\log t_2)$$

.

2.3.6 Estimateur Maximum de Vraisemblance

2.3.6.1 Définition

Cette section traite du problème d'estimation de β, le vecteur $(k \times 1)$ des coefficients de régression dans le modèle de régression de Poisson. La majeure partie de ce travail traite de la méthode du maximum de vraisemblance, car il s'agit de la méthode la plus courante pour estimer les modèles de données de comptage. Le principe du maximum de vraisemblance stipule que le paramètre doit être choisi de manière à maximiser la probabilité que le modèle spécifié ait généré l'échantillon observé. De nombreuses bonnes références économétriques aux principes généraux d'estimation du maximum de vraisemblance sont disponibles, notamment Amemiya, (1985) [27] et Cramer (1989)[28].

2.3.6.2 Fonction de vraisemblance et maximisation

Étant donné un échantillon indépendant de n paires d'observations(y_i, x_i), la distribution de probabilité conjointe des échantillons est le produit des distributions conditionnelles individuelles distributions de probabilité :

$$f(y_1, ..., y_n|x_1, ..., x_n; \beta) = \prod_{i=1}^{n} f(y_i|x_i; \beta) \tag{2.20}$$

D'après la compréhension paramétrique, 2.20 est appelée la fonction de vraisemblance, nous écrivons :

$$L = L(\beta; y_1, ..., y_n, x_1, ..., x_n) \tag{2.21}$$

L'estimateur du maximum de vraisemblance est défini comme :

$$\hat{\beta} = \arg_\beta \max L(\beta; y_1, ..., y_n, x_1, ..., x_n)$$

Comme le logarithme est une transformation monotone, la maximisation de la fonction de vraisemblance est équivalente à la maximisation de la fonction logarithmique ou log de vraisemblance $ell = \log L$. En général, cette transformation simplifie les choses considérablement, car il remplace les produits par des sommes. De plus, il permet l'utilisation de le théorème central limite lors de l'étude des propriétés du maximum comme estimateur de probabilité. La fonction log-vraisemblance pour le modèle de régression de Poisson prend la forme :

$$\begin{aligned} \ell(\beta; Y; X) &= \log \prod_{i=1}^{n} f(y_i|x_i;\beta) \\ &= \sum_{i=1}^{n} \log f(y_i|x_i;\beta) \\ &= \sum_{i=1}^{n} -\exp(x_i'\beta) + y_i x_i'\beta - \log(y_i!) \end{aligned} \tag{2.22}$$

La valeur de maximisation de β, notée $\hat{\beta}$ est trouvée en calculant la première dérivées de la fonction log-vraisemblance et en les fixant à zéro. Dans le modèle de régression de Poisson, il existe k telles dérivées, par rapport à β_1, β_2 et ainsi de suite. Le vecteur (colonne) qui collecte ces k premières dérivées est alternativement désigné comme vecteur de gradient ou comme vecteur de score. Ce dernier terme est utilisé dans ce qui suit. Nous écrivons :

$$s_n(\beta, y, x) = \frac{\partial \ell(\beta, y, x)}{\partial \beta} = \sum_{i=1}^{n} [y_i - \exp(x_i'\beta)]x_i$$

Nous utilisons l'indice "n" pour rappeler que le score dépend de la taille de l'échantillon. L'estimateur du maximum de vraisemblance $\hat{\beta}$ est la valeur de β qui résout les conditions du premier ordre pour un maximum

$$s_n(\beta, y, x) = 0 \tag{2.23}$$

Notez que tant que x_i inclut une constante, les conditions du premier ordre impliquent que $\sum_{i=1}^{n} \hat{u}_i = 0$, où $\hat{u}_i$ est un résidu implicite défini comme

$$\hat{u}_i = y_i - \hat{E}(y_i|x_i) = y_i - \exp(x_i'\hat{\beta})$$

Pour les régresseurs non constants, 2.23 peut être interprété comme un ensemble d'orthogonalité conditions :

$$\sum_{i=1}^{n} \hat{u}_i x_j = 0, j = 2, ..., k$$

L'équation 2.23 donne les conditions nécessaires pour un maximum. Si en plus la matrice des dérivées secondes, la matrice Hessienne, est définie négative pour tout valeurs de β, la solution de 2.23 est appelée estimateur du maximum de vraisemblance. La matrice Hessienne de la fonction de log-vraisemblance de Poisson est donnée par :

$$H_n(\beta; y, x) = \frac{\partial^2 \ell(\beta; y, x)}{\partial^2 \beta}$$

$$= -\sum_{i=1}^{n} \exp(x_i'\beta) x_i x_i' \tag{2.24}$$

H_n est définie négative, la fonction log-vraisemblance de la régression de Poisson modèle est globalement concave, et l'ensemble des paramètres résolvant le premier ordre les conditions sont les estimateurs uniques du maximum de vraisemblance.

2.3.6.3 Algorithme de Newton-Raphson

Puisque2.23 est non linéaire en β, le système de k équations doit être résolu en utilisant un algorithme itératif. Un choix courant qui fonctionne bien pour les fonctions objectives concaves est la méthode de Newton-Raphson. Il peut être motivé comme suit. Étant donné toute estimation de paramètre initiale, disons $\hat{\beta}$, nous pouvons obtenir une approximation de second ordre de $\ell(\beta)$ autour de $\hat{\beta}^0$:

$$\ell^*(\beta) = \ell(\hat{\beta}^0) + s_n(\hat{\beta}^0)'(\beta - \hat{\beta}^0) + \frac{1}{2}(\beta - \hat{\beta}^0)' H_n(\hat{\beta}^0)(\beta - \hat{\beta}^0) \approx \ell(\beta)$$

Maintenant, nous pouvons maximiser $\ell^*(\beta)$ (plutôt que $\ell(\beta)$ par rapport à β, donnant un nouvelle valeur de paramètre que nous appelons $\hat{\beta}^1$. La condition de premier ordre de ce plus simple le problème est

$$s_n(\hat{\beta}^0) + H_n(\hat{\beta}^0)(\hat{\beta}^1 - \hat{\beta}^0) = 0$$

ou

$$\hat{\beta}^1 = \hat{\beta}^0 - [H_n(\hat{\beta}^0)]^{-1} s_n(\hat{\beta}^0)$$

Ainsi, pour une valeur de départ arbitraire $\hat{\beta}^0$, la règle de mise à jour de Newton-Raphson est donné par :

$$\hat{\beta}^{t+1} = \hat{\beta}^t - [H_n(\hat{\beta}^t)]^{-1} s_n(\hat{\beta}^t) \quad t = 0, 1, ... \tag{2.25}$$

où $s(.)$ désigne le score et $H(.)$ la hessienne de la fonction log-vraisemblance de Poisson. Si on évalue le membre de droite à l'estimateur du maximum de vraisemblance, on observe que $s(\hat{\beta}^t) = 0$ et donc $\hat{\beta}^{t+1} = \hat{\beta}^t$. La procédure itérative se termine lors-

qu'un critère de convergence prédéfini est satisfait. Les critères possibles incluent le changement de la valeur de l'estimation $\hat{\beta}^{t+1} - \hat{\beta}^t$, la variation de la log-vraisemblance $\ell(\hat{\beta}^{t+1}) - \ell(\hat{\beta}^t)$ ou la valeur de la score à l'estimation $s(\hat{\beta}^t)$. La convergence se produit lorsque l'une de ces valeurs, ou une combinaison de celles-ci, est proche de zéro (par exemple inférieure à 10^{-5} en valeur absolue).

2.3.6.4 Propriétés de l'estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance $\beta = argmaxL(\beta)$ est en général une fonction non linéaire de la variable dépendante. Par conséquent, les résultats analytiques sur les propriétés de petit échantillon de la distribution d'échantillonnage de $\hat{\beta}$ ne sont pas disponibles. Si un certain nombre de conditions de régularité sont satisfaites, on peut montrer que l'estimateur du maximum de vraisemblance est :

— asymptotiquement sans biais
— asymptotiquement normal
— efficacité asymptotique

Ces trois observations sont résumées dans le résultat de convergence suivant Cramer(1989)[28] Amemiya(1985)[27]

$$\sqrt{n}(\hat{\beta}_{ML} - \beta_0) \underrightarrow{d} \mathcal{N}(0, I(\beta_0)^{-1}) \tag{2.26}$$

où $\underrightarrow{d}$ signifie « converge en distribution », et on a la matrice d'information de Fisher $I(\beta_0)$ donne par :

$$I(\beta_0) = -E\left[\frac{\partial^2 \ell(\beta; y_i, x_i)}{\partial \beta \partial \beta'}\right]_{\beta_0} \tag{2.27}$$

L'estimateur du maximum de vraisemblance est asymptotiquement sans biais puisque la distribution vers laquelle elle converge est centrée sur la valeur réelle du paramètre β_0. C'est asymptotiquement efficace, puisque sa variance est égale à l'inverse de l'information de Fisher, la borne inférieure de Cramer-Rao pour tout estimateur sans biais. Alors que ces propriétés asymptotiques au sens strict ne tiennent que dans la limite de taille d'échantillon infinie, dans la pratique, ils sont souvent supposés être d'environ valide, surtout lorsque la taille de l'échantillon n'est pas si petite. Le approximatif distribution de $\hat{\beta}$ est alors donnée par :

$$\hat{\beta}_{ML} \sim \mathcal{N}(\beta_0, [nI(\beta_0)]^{-1}) \tag{2.28}$$

Ce résultat nécessite, en général, que le modèle soit correctement spécifié. Laisse le densité vraie (conditionnelle) soit notée $f_0(y_i|x_i)$. Il doit existerβ_0 un tel que :

$$\prod_{i=1}^{n} f(y_i|x_i; \beta_0) = \prod_{i=1}^{n} f_0(y_i|x_i) \tag{2.29}$$

Les propriétés de l'estimation du maximum de vraisemblance dans les modèles mal spécifiés sont discutées . En dehors des spécifications correctes, quelques autres des conditions de régularité sont requises, essentiellement pour assurer l'interchangeabilité des opérations de différenciation et de prise d'anticipations.
Les dérivées première et seconde de la fonction log-vraisemblance doivent être définies, et la matrice d'information de Fisher doit être non nulle .

2.4 Régression binomiale négative :

2.4.1 Le modèle

La distribution binomiale négative est l'alternative la plus couramment utilisée au modèle de Poisson quand il est douteux que les exigences strictes d'indépendance du processus sous-jacent et l'inclusion de tous les régresseurs pertinents soient satisfaites. En particulier, le modèle binomial négatif (Negbin) est approprié lorsque la distribution conditionnelle de $y|\tilde{\lambda}$ est distribuée par Poisson et que $\tilde{\lambda}$ est distribuée indépendamment par gamma. Ainsi, le modèle Negbin a l'interprétation d'un modèle de mélange de Poisson qui rend compte d'une manière spécifique pour le caractère aléatoire du paramètre de Poisson $\tilde{\lambda}$. Par ailleurs, le modèle Negbin apparaît lorsque le processus de comptage sous-jacent n'est pas indépendant et lorsque la dépendance peut être décrite par un type spécifique de véritable contagion. D'autres références au modèle Negbin incluent Cameron et Trivedi (1986), Lawless (1987b)[29] et Hausman, Hall et Griliches (1984).
Pour faire le pas vers le modèle de régression Negbin, les paramètres α et λ sont spécifiés en termes de variables exogènes. Le modèle Negbin II est obtenu pour $\alpha = \sigma^{-2}$ et $\lambda = \exp(x'\beta)$. Dans ce cas, la fonction d'attente conditionnelle est

$$E(y|x) = \exp(x'\beta) \tag{2.30}$$

alors que la fonction de variance conditionnelle est donnée par

$$Var(y|x) = \exp(x'\beta) + \sigma^2[\exp(x'\beta)]^2 \tag{2.31}$$

La variance conditionnelle est toujours supérieure à la moyenne conditionnelle : modèle binomial négatif est un modèle de surdispersion. Le modèle Negbin I est obtenu en laissant α varier entre les individus de sorte que $\alpha = \sigma^-2exp(x'\beta)$ et $\lambda = exp(x'\beta)$. Ce paramétrage produit une variance qui est une fonction linéaire de la moyenne :

$$Var(y|x) = (1 + \sigma^2)\exp(x'\beta) \tag{2.32}$$

Une autre façon de caractériser la différence entre le Nègre I et le modèle Negbin II sont en termes de fonction de dispersion ϕ, de sorte que $Var(y|x) = \phi E(y|x)$. Pour le modèle Negbin I, $\phi = (1+\sigma 2)$, une fonction constante, alors que pour le modèle Negbin II, $\phi = 1 + \sigma^2 \exp(x'\beta)$.

2.4.2 Modèle Negbin II

La fonction de probabilité (conditionnelle) du modèle Negbin II peut être écrite comme

$$f(y|.) = \frac{\Gamma(\sigma^{-2}+y)}{\Gamma(\sigma^{-2})\Gamma(y+1)}\left(\frac{\sigma^{-2}}{\exp(x'\beta)+\sigma^{-2}}\right)^2\left(\frac{\exp(x'\beta)}{\exp(x'\beta)+\sigma^{-2}}\right)^y.$$

Pour $\sigma^2 \to 0$, ce modèle converge vers le modèle de régression de Poisson. Depuis $\sigma^2 \leq 0$ le modèle de Poisson est obtenu à la limite de la espace de paramètres. Ceci doit être gardé à l'esprit lors de l'évaluation du modèle : un test de rapport de vraisemblance modifié doit être utilisé pour tester $H_0 : f$ est Poisson contre $H_1 : f$ est binomial négatif. Le problème du dépistage des restrictions au la limite de l'espace des paramètres a été discutée. En supposant un échantillon indépendant, la fonction de log-vraisemblance du modèle Negbin II est donnée par

$$\ell(\beta,\sigma^2) = \sum_{i=1}^{n}\left[\sum_{i=1}^{y_i}\log(\sigma^{-2}+j-1) - \log y_i! - (y_i+\sigma^{-2})\log(1+\sigma^2 exp(x_i'\beta)) + y_i\log\sigma^2 + y_i x_i'\beta\right] \tag{2.33}$$

où le rapport des fonctions gamma dans simplifié .
Les estimateurs de vraisemblance maximale $\hat{\beta}$ et $\hat{\sigma^2}$ de Negbin II sont obtenus comme solutions aux conditions de premier ordre

$$\sum_{i=1}^{n}\frac{y_i - exp(x_i'\beta)}{1+\sigma^2 exp(x_i'\beta)}x_i = 0 \tag{2.34}$$

et

$$\sum_{i=1}^{n}\left[\frac{1}{\sigma^4}\left(log(1+\sigma^2\exp(x_i'\beta)) - \sum_{j=1}^{y_i}\frac{1}{\sigma^{-2}+j-1}\right) - \frac{(y_i+\sigma^{-2})\exp(x_i'\beta)}{1+\sigma^{-2}\exp(x_i'\beta)} + \frac{y_i}{\sigma^2}\right] = 0 \tag{2.35}$$

De plus, on peut montrer (voir Lawless, 1987b) que la matrice d'information est en bloc diagonal. Par conséquent, $\hat{\sigma^2}$ et $\hat{\beta}$ sont asymptôtiquement indépendants.
La variance de $\hat{\beta}$ est donnée par

$$Var(\hat{\beta}) = \left(\sum_{i=1}^{n}\frac{\exp(x_i'\beta)}{1+\sigma^2\exp(x_i'\beta)}x_i x_i'\right)^{-1}. \tag{2.36}$$

2.4.3 Modèle Negbin I

Le modèle Negbin I a une fonction de log-vraisemblance

$$\ell(\beta,\sigma^2)=\sum_{i=1}^{n}{}'\left[\left(\sum_{j=1}^{y_i}\log(\sigma^{-2}\exp(x'\beta)+j-1)\right)-\log y_i!-(y_i+\sigma^{-2}\exp(x_i'\beta))\log(1+\sigma^2)+y_i\log\sigma^2\right] \tag{2.37}$$

Avec conditions de premier ordre pour $\hat{\beta}$:

$$\sum_{i=1}^{n}\left[\left(\sum_{i=n}^{y_i}\frac{\sigma^{-2}\exp(x_i'\beta)}{\sigma^{-2}\exp(x_i'\beta)+j-1}\right)x_i+\sigma^{-2}exp(x_i'\beta)x_i\right]=0 \tag{2.38}$$

Contrairement au modèle Negbin II, les conditions de premier ordre du Negin I modèle ne sont pas de la forme $\sum(y_i-\mu_i)f(\mu_i)=0$. Le modèle Negbin I ne appartiennent à la classe des familles exponentielles linéaires, et les résultats de robustesse ne s'appliquent donc pas dans ce cas. En fait, le Negbin le modèle II est le seul modèle de cette famille. Par rapport, c'est aussi le seul modèle Negbin avec matrice d'information en bloc-diagonal.

2.4.4 Modèle Negbin$_k$

Malgré ces avantages du modèle Negbin II, on pourrait néanmoins souhaiter de se lancer dans la recherche d'estimateurs alternatifs asymptotiquement efficaces s'ils sont correctement spécifiés. Un de ces modèles est le binôme négatif généralisé modèle de Winkelmann et Zimmermann (1991, 1995) [30, 4]. Un modèle similaire a été employé indépendamment par Ruser (1991)[31]. Ce modèle a été redécouvert par Saha et Dong (1997) [32] qui ignoraient apparemment la littérature précédente.
Soit $\alpha=\sigma^{-2}\lambda^{1-k}$ et $\lambda=\exp(x'\beta)$. k, est une non-linéarité .Par rapport au modèle de Poisson, deux paramètres supplémentaires doivent être estimés et ce modèle a été appelé Negbin $_k$.
Le Negbin$_k$ peut être interprété comme un hyper-modèle pour les modèles Negbin I et Negbin II non imbriqués. En particulier, le Negbin$_k$ niche le Negbin II et Negbin I à travers les restrictions paramétriques $k=1$ et $k=0$ respectivement.
Ainsi, un test entre les deux sous-modèles non emboîtés peut se dérouler (voir Ozuna et Gomez (1995)[33] pour un certain nombre d'autres approches de test entre les modèles Negbin I et Negbin II).
Une représentation possible de la fonction de probabilité du modèle de $Negbin_k$ utilise la notation suivante :

$$\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha}=(\alpha^{-1}\lambda+1)^{-\alpha}=(\sigma^2\lambda^k+1)^{-\lambda^{1-k}/\sigma^2} \tag{2.39}$$

De plus

$$\begin{aligned}\left(\frac{\lambda}{\lambda+\alpha}\right)^{y} &= (1+\alpha\lambda^{-1})^{-y} \\ &= \prod_{i=1}^{y}\frac{1}{1+\sigma^{-2}\lambda^{-k}}\end{aligned}$$

Enfin, en utilisant la propriété récursive de la fonction gamma :

$$\frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y)} = \prod_{j=1}^{y}\frac{\sigma^{-2}\lambda^{1-k}+j-1}{j}$$

En mettant tout ensemble, la fonction de probabilité du Negbin$_k$ peut être exprimé comme :

$$f(y|\lambda,\sigma^2,k) = C \times \begin{cases}\prod_{j=1}^{y}\frac{\sigma^{-2}\lambda^{1-k}+j-1}{(1+\sigma^{-2}\lambda^{-k})j}, & \text{si y=1,2,...}\\ 1, & \text{si y=0}\end{cases} \tag{2.40}$$

avec

$$C = (\sigma^2\lambda^k+1)^{-\lambda^{1-k}}/\sigma^2$$

$$\lambda = \exp(x'\beta), \sigma^2 \geq 0$$

Étant donné un échantillon indépendant d'observations, la log-vraisemblance de l'échantillon est la somme du logarithme des probabilités $f(y)$.

2.4.5 Modèle Negbin$_X$

Encore une autre paramétrisation de la distribution binomiale négative est proposée par Santos Silva et Windmeijer (2001) [34], le modèle binomial négatif peut être représenté comme une somme arrêtée (ou composée) distribution, où

$$Y = R_1 + R_2 + ... + R_S = \sum_{i=j}^{S} R_j$$

où $S = 0, 1, 2, ...$ est une distribution de Poisson, et les composantes $R_j = 1, 2, ...$ ont une distribution logarithmique identique et indépendante. Le logarithmique distribution a un seul paramètre θ avec $0 < \theta < 1$. Il est donc naturel de permettre pour les covariables en laissant

$$\theta = \frac{\exp(x'y)}{1+\exp(x'y)}$$

par conséquent

$$\frac{\theta}{1-\theta} = \exp(x'y)$$

D'après les propriétés de la distribution logarithmique, il s'ensuit que le nombre attendu de comptages dans chaque composante est

$$E(R_j) = \frac{\exp(x'y)}{\log[1+\exp(x'y)]}.$$

Si on laisse pour la partie de Poisson en plus $\lambda = E(S) = \exp(x'y)$ comme d'habitude, alors il s'ensuit que Y est un binôme négatif distribué avec des paramètres

$$\alpha = \frac{\exp(x'\beta)}{\log[1+\exp(x'\gamma)]} \quad \text{et} \quad \lambda = \exp(x'\beta)$$

En remplaçant ces expressions dans la fonction de probabilité binomiale négative , et après quelques simplifications supplémentaires, on obtient la fonction de probabilité Negbin_X

$$f(y|x) = \frac{\Gamma\left(y + \frac{\exp(x'\beta)}{\log[1+exp(x'y)]}\right)\exp(-\exp(-x'\beta))}{\Gamma(y+1)\Gamma\left(\frac{\exp(x'\beta}{\log[1+\exp(x'y)]}\right)(1+\exp(-x'y))^y} \tag{2.41}$$

avec

$$E(y|x) = \frac{\exp(x'\beta + x'y)}{\log[1+\exp(x'y)]}.$$

Bien sûr, on peut encore modifier le modèle en incluant différents ensembles de régression z et x dans les différentes parties du modèle. Habituellement, il y aura cependant peu de raisons a priori pour justifier une telle sélection, et le modèle inclure deux coefficients pour chaque co-variable disponible. Le côté intéressant du modèle est l'interprétation du processus sous-jacent de génération de données. L'effet global d'un régression sur le nombre total de comptages est la somme de deux effets distincts.
Premièrement, une variable peut affecter le nombre de composants S.
Deuxièmement, une variable peut affecter le nombre de comptages dans chaque composante R_j . Cette séparation peut avoir des analogies avec les processus de la vie réelle. Santos Silva et Windmeier[34] motive son modèle par la demande de visites chez le médecin. Ici le nombre total de visites peut dépendre du nombre total de périodes de maladie une période plus le nombre de visites au cours de chaque période.

CHAPITRE 3
APPLICATION

3.1 Application sur des données de la qualité de l'air

3.1.1 Trouver le modèle linéaire approprié pour la prévision de l'ozone

On a un ensemble de données sur la qualité de l'air pour expliquer comment les modèles linéaires sont interprétés. Nous avons commencé par un modèle linéaire de base, et à partir de là, nous avons essayé de trouver un modèle linéaire avec un meilleur ajustement.

3.1.2 Le jeu de données sur la qualité de l'air

L'ensemble de données sur la qualité de l'air contient 154 mesures des quatre paramètres de qualité de l'air suivants, tels qu'obtenus à New York :

— **Ozone :** niveau moyen d'ozone en parties par milliard.

— **Solar.R :** Rayonnement solaire à langleys.

— **Vent(wind)** : vitesse moyenne du Lent en miles par heure.

— **Tempe :** Température quotidienne maximale en degrés Fahrenheit.

3.1.3 Exploration et préparation des données

La tâche de prédiction est la suivante : pouvons-nous prédire le niveau d'ozone compte tenu du rayonnement solaire, de la vitesse du vent et de la température ? Pour

voir si les hypothèses du modèle linéaire sont appropriées pour les données disponibles, nous calculerons la corrélation entre les variables(tableau3.1)

	Ozone	Solar.R	vent(wind)	Temp
Ozone	1.0000000	0.3483417	-0.6124966	0.6985414
Solar.R	0.3483417	1.00000000	-0.1271835	0.2940876
vent(Wind)	-0.6124966	-0.1271835	1.0000000	-0.4971897
Temp	0.6985414	0.2940876	-0.4971897	1.0000000

TABLE 3.1 – Corrélations de variables par paires

D'après tableau3.1 on remarque
-L'ozone a une relation positive avec la température.
-L'ozone a une relation négative avec le vent.
Cela suggère qu'il devrait être possible de former un modèle linéaire prédisant le niveau d'ozone en utilisant les caractéristiques restantes.

3.1.4 Pré traitement des données

Étant donné que l'ensemble de données sur la qualité de l'air contient certaines valeurs manquantes, nous les supprimerons avant de commencer à ajuster les modèles et sélectionnerons 70%des échantillons pour la formation et utiliserons le reste pour les tests.

3.1.5 Etude du modèle linéaire

Le modèle s'écrit :

$$Y_t = \beta_0 + \beta_1 Solar.R + \beta_2 Temp + \beta_3 Wind + e_t$$

L'estimateur des coefficients β_i est obtenu en minimisant la distance au carré entre chaque observation et la droite, d'où le nom d'estimateur des moindres carrés ordinaires (MCO).

3.1.5.1 Modèle des moindres carrés ordinaires

On applique la modèle des Moindres Carrés Ordinaires dont l'écriture sous R on trouve (tableau 3.2).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-74.18929	28.65268	-2.589	0.0116
Solar.R	0.04507	0.02844	1.585	0.1173
Temp	1.84999	0.31696	5.837	1.32e-07
vent(Wind)	-3.34510	0.80164	-4.173	8.10e-05

TABLE 3.2 – Les coefficients du modèle

— **Std. Error :** est l'erreur type de l'estimation du coefficient

— **t value :** indique la valeur du coefficient en fonction de l'erreur type

— **Pr(>|t|) :** est la valeur de p pour le test t, qui indique la signification de la statistique de test

Les résidus :
$SCR = \sum_{t=1}^{n} e_t^2$ où le résidus est défini comme la différence entre les valeurs observées et prédites,tableau 3.3.

Min	1Q	Median	3Q	Max
-29.913	-16.014	-3.949	7.818	92.987

TABLE 3.3 – Les résidus

La valeur médiane résiduelle suggère que le modèle prédit généralement des valeurs d'ozone légèrement plus élevées que celles observées. La grande valeur maximale, cependant, indique que certaines prédictions aberrantes sont également beaucoup trop faibles.
Le coefficient de détermination R^2 :
Le multiple R au carré indique le coefficient de détermination. Il est défini comme le carré de la corrélation entre les estimations et les résultats observés. et on a :
R^2=0.60619815
Après de définir le modèle, nous tracer les modèles linéaires.
Intervalles de confiance :
Construction des intervalles de confiance des estimations de caractéristiques(tableau3.4).

	2.5%	97.5%
Intercept	-131.28099447	-170.9758115
Solar.R	-0.01159323	0.1017332
Temp	1.21843797	2.4815478
vent(Wind)	-4.94241076	-1.7477984

TABLE 3.4 – Les intervalles de confiance

Nous voyons que le modèle ne semble pas être aussi sûr du réglage de l'interception. Voyons si le modèle fonctionne toujours bien :

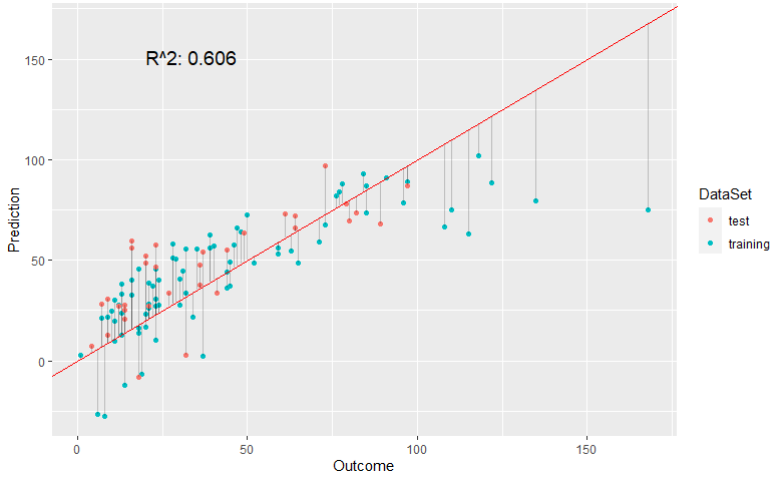

FIGURE 3.1 – Modèle des moindres carrés ordinaires

En ce qui concerne l'ajustement du modèle, il y a deux observations principales :

— Les niveaux élevés d'ozone sont sous-estimés.
— Des niveaux d'ozone négatifs sont prévus.

Examinons ces deux questions plus en détail dans ce qui suit.

3.1.5.2 Estimation de niveaux élevés d'ozone

En regardant le graphique, on remarque que le modèle linéaire s'adapte bien au résultat lorsque l'ozone se situe dans la plage [0,100]. Cependant, lorsque la concentration d'ozone réellement observée est supérieure à 100, le modèle sous-estime considérablement la valeur.
Une question que nous devrions nous poser est de savoir si ces niveaux élevés d'ozone ne pourraient pas être le résultat d'erreurs de mesure.
Compte tenu des niveaux d'ozone typiques, les valeurs mesurées semblent raisonnables. Le niveau maximal d'ozone est de 168 ppm (parties par milliard), et 150 à 510 ppm sont des concentrations maximales typiques pour les villes américaines. Cela signifie que nous devrions en effet nous préoccuper des valeurs aberrantes.
Sous-estimer les niveaux élevés d'ozone serait particulièrement préjudiciable car des niveaux élevés peuvent être dangereux pour la santé. Examinons les données pour identifier pourquoi le modèle a des problèmes avec ces valeurs aberrantes.

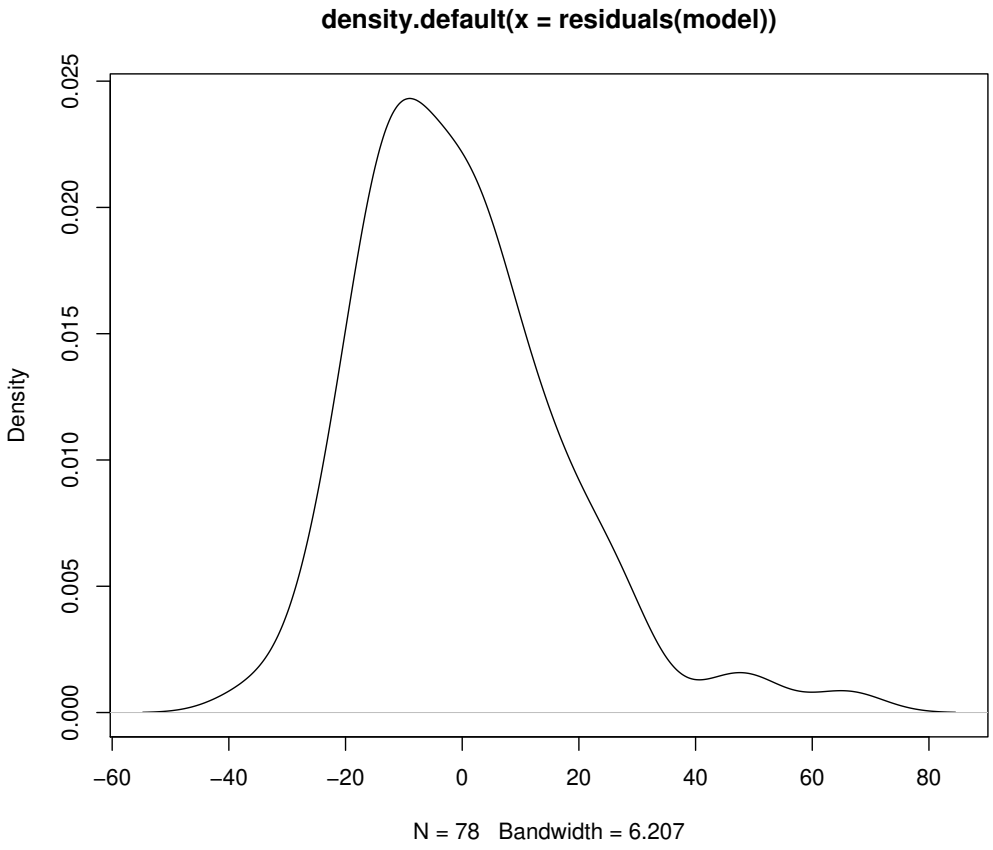

FIGURE 3.2 – Densité des résidus

L'histogramme indique que les valeurs à l'extrémité droite de la distribution résiduelle sont en effet problématiques. Étant donné que les résidus ne sont pas vraiment distribués normalement, un modèle linéaire n'est pas le meilleur modèle. En effet, les résidus semblent plutôt suivre une certaine forme de distribution de Poisson. Pour savoir pourquoi l'ajustement du modèle des moindres carrés est si mauvais pour les valeurs aberrantes, nous examinons à nouveau les données.

	Ozone	Solar.R	Vent(Wind)	Temp
min	110.0	207.0	2.300	79.00
1st Qu	115.8	223.5	3.550	81.75
Median	120.0	231.5	4.050	86.50
Mean	128.0	236.2	4.583	86.17
3rd Qu	131.8	250.8	5.300	89.75
Max	168.0	269.0	8.00	94.00
	Ozone	Solar.R	vent(Wind)	Temp
min	1.0	7.0	2.30	57.00
1 st Qu	18.0	113.5	7.40	71.00
Median	31.0	207.0	9.70	79.00
Mean	42.1	184.8	9.94	77.79
3rd Qu	62.0	255.5	11.50	84.50
Max	168.0	334.0	20.70	97.00

TABLE 3.5 – Comparer les observations avec un niveau élevé d'ozone avec l'ensemble de données (seuil à 95% quantile).

En regardant les distributions des deux groupes d'observations, nous ne pouvons pas voir une grande différence entre les observations d'ozone élevé et les autres échantillons. Nous pouvons cependant trouver le coupable en utilisant le tracé des prédictions du modèle ci-dessus. Dans le graphique, nous voyons que la plupart des points de données sont centrés autour de la plage d'ozone [0, 50]. Pour bien s'adapter à ces observations, l'ordonnée à l'origine a une grande valeur négative de -74,19, c'est pourquoi le modèle sous-estime les niveaux d'ozone pour des valeurs d'ozone plus grandes, qui sont sous-représentées dans les données de formation.

3.1.5.3 Le modèle des niveaux d'ozone négatifs

Si la concentration d'ozone observée est proche de 0, le modèle prédit souvent des niveaux d'ozone négatifs. Bien sûr, cela ne peut pas être dû au fait que les concentrations ne peuvent pas descendre en dessous de 0. Encore une fois, nous étudions les données pour découvrir pourquoi le modèle fait toujours ces prédictions.
À cette fin, nous sélectionnerons toutes les observations dont le niveau d'ozone se situe dans le 5% centile et étudierons leurs valeurs caractéristiques :

	Ozone	Solar.R	Vent(Wind)	Temp
Min	1.0	8.00	9.70	57.0
1st Qu	4.5	20.50	9.85	59.5
Median	6.5	36.50	12.30	61.0
Mean	5.5	37.83	13.75	64.5
3rd Qu	7.0	48.75	17.38	67.0
Max	8.0	78.00	20.10	80.0

	Ozone	Solar.R	Vent(Wind)	Temp
Min	1.0	7.0	2.30	57.00
1st Qu	18.0	113.5	7.40	71.00
Median	31.0	207.0	9.70	79.00
Mean	42.1	184.8	9.94	77.79
3rd Qu	62.0	255.5	11.50	84.50
Max	168.0	334.0	20.70	97.00

TABLE 3.6 – Comparer les observations avec un faible niveau d'ozone avec l'ensemble de données (seuil à 5% de quantile).

Ce que nous constatons, c'est que, pour de faibles niveaux d'ozone, le rayonnement solaire moyen est beaucoup plus faible, tandis que la vitesse moyenne du vent(Wind) est beaucoup plus élevée. Pour comprendre pourquoi nous avons des prédictions négatives, regardons les coefficients du modèle : (tableau 3.2).
Ainsi, pour les faibles niveaux d'ozone, le coefficient positif de **Solar.R** ne peut pas compenser l'attraction négative de l'ordonnée à l'origine et du coefficient de **Wind** car, pour les faibles niveaux d'ozone, les valeurs de **Solar.R** sont faibles, tandis que les valeurs de **Wind** sont élevées.

3.1.5.4 Traiter les prévisions négatives concernant les niveaux d'ozone

Abordons d'abord le problème de la prévision des niveaux d'ozone négatifs.

3.1.5.5 Modèle des moindres carrés prdinaires tronqués

Une approche simple pour traiter les prédictions négatives consiste à les remplacer par la valeur minimale possible à la place. De cette façon, si nous remettions notre modèle à un client, il ne commencerait pas à soupçonner que quelque chose ne va pas avec le modèle.
Vérifions comment cela améliorerait nos prédictions sur les données de test. Se souvenir du R^2 du modèle initial était était 0,604.

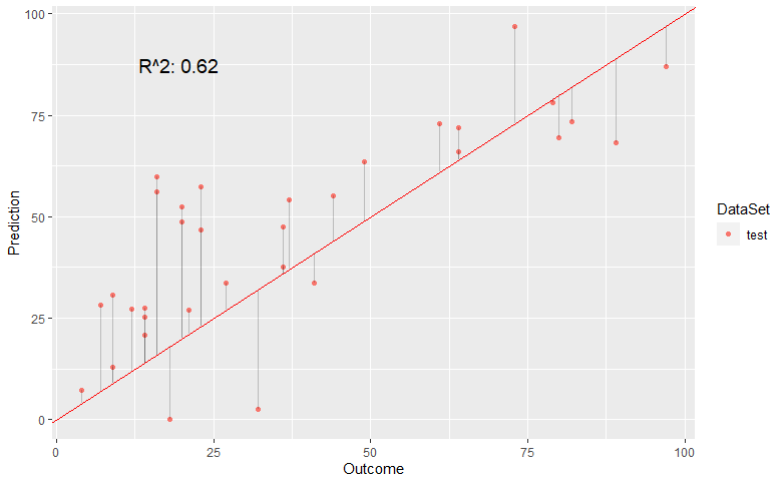

FIGURE 3.3 – Modèle des moindres carrés ordinaires tronqués

Comme on le voit, ce hack limite le problème et augmente R^2 pour 0,646. Cependant, corriger les valeurs négatives de cette manière ne change rien au fait que notre modèle est erroné car la procédure d'ajustement n'a pas pris en compte le fait que les valeurs négatives devraient être impossibles.

3.1.6 Régression de Poisson

Pour éviter les estimations négatives, nous pouvons utiliser un modèle linéaire généralisé (GLM) qui suppose une distribution de Poisson plutôt qu'une distribution normale :

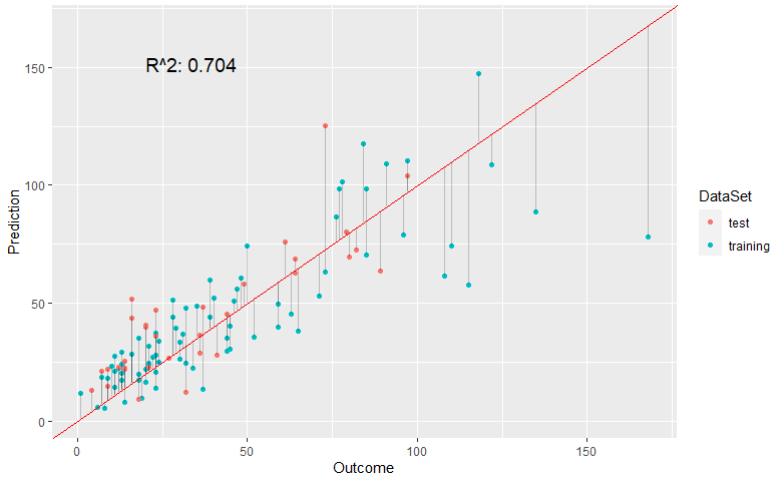

FIGURE 3.4 – Le modèle de Poisson

R^2 la valeur de 0,616 indique que la régression de Poisson est légèrement supérieure aux Moindres carrés ordinaires (0,604). Cependant, ses performances ne sont pas supérieures au modèle qui tronque les valeurs négatives à (0,646). C'est probablement parce que la variance du niveau d'ozone est beaucoup plus grande que ce que

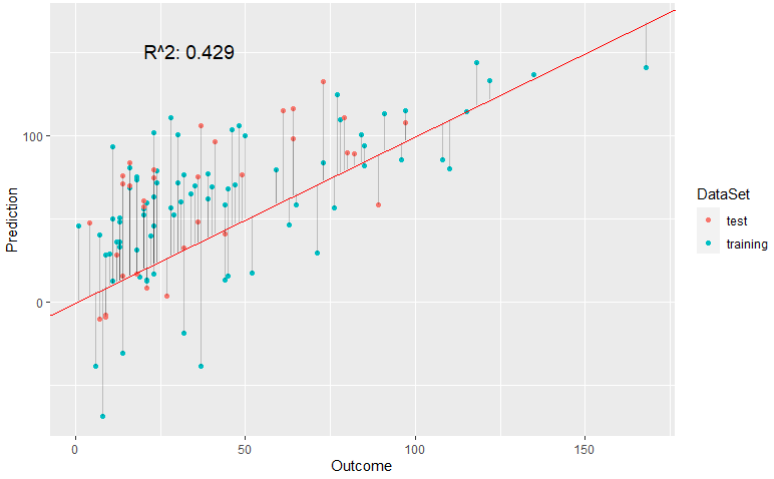

FIGURE 3.5 – Le modèle pondérée

suppose le modèle de Poisson, la moyenne et la variance doivent être les mêmes pour la distribution de Poisson,dans ce cas la moyenne et la variance ne sont pas égaux.
E(x) : 42.0991
Var :1107.29

3.1.6.1 Transformation logarithmique

Une autre approche pour traiter les prédictions négatives consiste à prendre le logarithme du résultat : 0.704. Notez que bien que le résultat soit identique au résultat via la régression de Poisson, les deux méthodes ne sont pas identiques en général.

3.1.6.2 Faire face à la sous-estimation des niveaux élevés d'ozone :

Idéalement, nous aurions un meilleur échantillonnage des mesures avec des niveaux d'ozone élevés. Cependant, comme nous ne pouvons pas collecter plus de données, nous devons nous contenter de ce que nous avons. Une façon de traiter la sous-estimation des niveaux élevés d'ozone consiste à ajuster la fonction de perte.

3.1.7 Régression pondérée

En utilisant la régression pondérée, nous pouvons influencer l'impact des résidus des valeurs aberrantes. À cette fin, nous calculerons les scores z des niveaux d'ozone, puis nous utiliserons leur exponentielle comme poids pour le modèle de sorte que l'impact des valeurs aberrantes soit augmenté. Ce modèle est nettement plus approprié que le modèle ordinaire des moindres carrés car il traite mieux les valeurs aberrantes.

3.1.7.1 Échantillonnage

Échantillonnons à partir des données d'entraînement de sorte que les niveaux élevés d'ozone ne soient plus sous-représentés. Cela revient à faire une régression pondérée. Cependant, plutôt que d'attribuer de petits poids aux observations avec de faibles niveaux d'ozone, nous avons implicitement défini leurs poids sur 0.
"N (rame avant)=78"
"N(rame après)=51"
Construisons maintenant un nouveau modèle basé sur les données échantillonnées.
Comme on peut le voir, le modèle basé sur les données échantillonnées n'est pas plus performant que celui utilisant les poids.

3.1.7.2 Combinaison les preuves

Ayant vu que la régression de Poisson est utile pour éviter les estimations négatives et que la pondération est une stratégie efficace pour améliorer la prédiction des valeurs aberrantes, nous devrions essayer de combiner les deux approches, ce qui conduit à une régression de Poisson pondérée.

3.1.8 Régression de Poisson pondérée

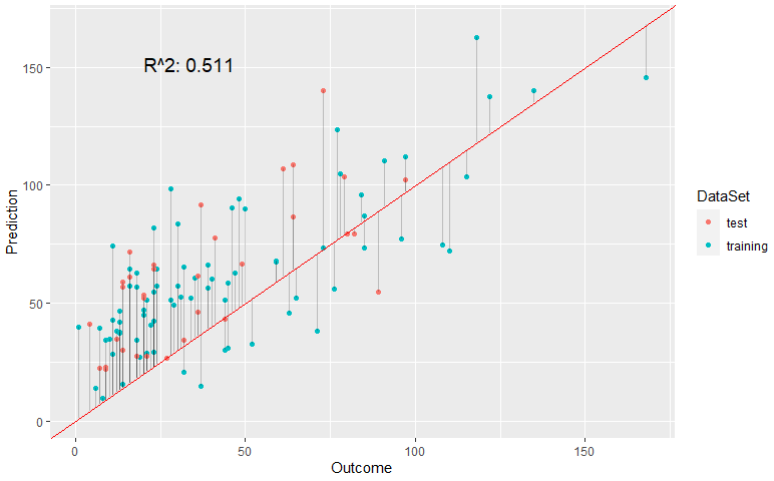

FIGURE 3.6 – Le modèle de Poisson pondérée

Comme on le voit, ce modèle combine les avantages de l'utilisation de la régression de Poisson (prédictions non négatives) avec l'utilisation de poids (sous-estimation des valeurs aberrantes). En effet, le R^2 de ce modèle est le plus bas à ce jour (0,652 contre 0,646 du modèle linéaire tronqué). Examinons les coefficients du modèle :

(Intercept)	Solar.R	Temp	vent(Wind)
5.074448971	0.001987798	-0.001202156	-0.137611701

TABLE 3.7 – Les coefficients du modèle de Poisson pondérée

Le modèle est toujours dominé par l'ordonnée à l'origine mais maintenant il est positif. Ainsi, si toutes les autres caractéristiques ont une valeur de 0, la prédiction du modèle sera toujours positive. Cependant, qu'en est-il de l'hypothèse selon laquelle la moyenne devrait être égale à la variance pour la régression de Poisson ?
Var : 1107.29
moy : 42.1
L'hypothèse du modèle n'est définitivement pas satisfaite et nous souffrons d'une sur-dispersion puisque la variance est supérieure à celle supposée par le modèle.

3.1.9 Modèle binomial négatif pondéré

Nous devrions donc essayer de choisir un modèle plus adapté à la sur-dispersion, tel que le modèle binomial négatif

FIGURE 3.7 – Le modèle binomial négatif pondéré

Ainsi, en termes de performances sur l'ensemble de test, le modèle binomial négatif pondéré n'est pas meilleur que le modèle de Poisson pondéré. Cependant, en ce qui concerne l'inférence, la valeur doit être supérieure car ses hypothèses ne sont pas brisées. En regardant les deux modèles, il est évident que leurs valeurs de p diffèrent considérablement :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.074448971	0.1668873283	30.4064366	4.515679e-203
Solar.R	0.001987798	0.0002630893	7.5556013	4.169292e-14
Temp	-0.001202156	0.0017067486	-0.7043545	4.812120e-01
vent(Wind)	-0.137611701	0.0042937826	-32.0490610	2.262476e-22

TABLE 3.8 – Les coefficients du modèle de Poisson pondéré

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.651242295	0.4874400886	7.490648	6.853421e-14
Solar.R	0.002591735	0.0006855076	3.780752	1.563551e-04
Temp	0.013312402	0.0052097134	2.555304	1.060950e-02
vent(Wind)	-0.127395396	0.0113987431	-11.176267	5.328447e-29

TABLE 3.9 – Les coefficients du modèle de binomial négatif pondérée

Alors que le modèle de Poisson prétend que tous les coefficients sont hautement significatifs, le modèle binomial négatif démontre que l'ordonnée à l'origine n'est pas significative, les bandes de confiance pour un binomial négatif peuvent être trouvées. Et en utilisant le cadre de données construit contenant les valeurs des caractéristiques dans l'ensemble de test ainsi que les prédictions avec leurs bandes de confiance, nous pouvons tracer comment les estimations fluctuent en fonction des variables indépendantes :

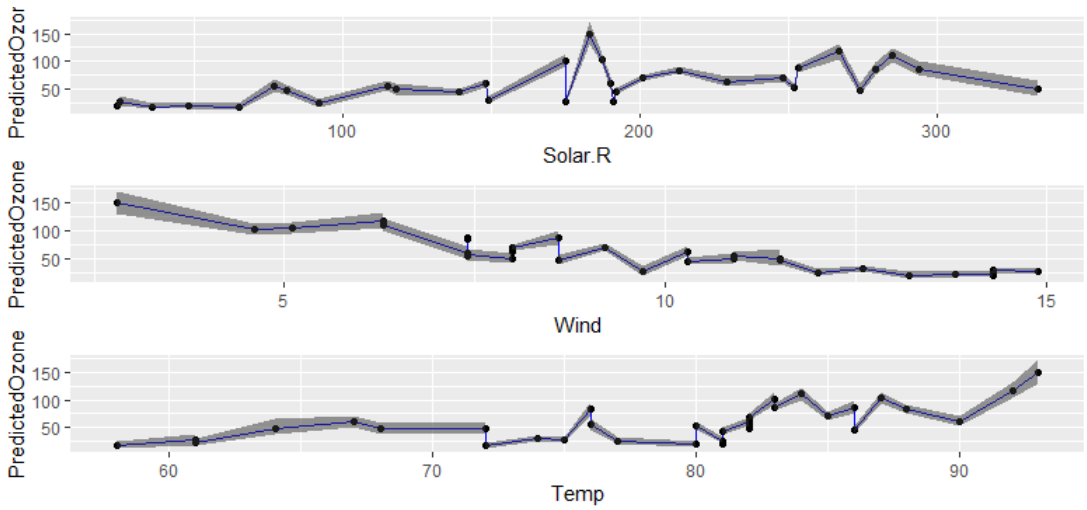

FIGURE 3.8 – Estimations de tracé par rapport aux caractéristiques individuelles dans l'ensemble de test

Ces graphiques montrent deux choses : Il existe des relations linéaires claires pour **Wind** et **Temp**. Les niveaux d'ozone estimés chutent lorsqu'ils **Wind** augmentent, tandis que les niveaux d'ozone estimés augmentent lorsqu'ils **Temp** augmentent. Le modèle est plus fiable pour les faibles niveaux d'ozone, mais moins fiable pour les niveaux d'ozone élevés.

3.2 Application sur les données de comptage de crises épileptiques

Le jeu de données décrit par **Thall** et Vail (1990) provient d'une étude de "Leppik (1985)" sur l'effet du médicament Progabide contre des crises partielles d'épilepsie. D'après les travaux de Thall et Vail (1990), il s'agissait d'un essai à double insu avec permutation. Les patients étaient divisés en deux groupes ; un groupe recevant le médicament Progabide et un groupe placebo. De plus, cette étude a été faite parallèlement au traitement standard de chimiothérapie. Par ailleurs, ils ont récupéré des données d'une période de 8 semaines avant la prise du médicament/placebo. Cet instant de collecte de données va-t-on appeler (visite zéro).

Les comptages des crises épileptiques ont été enregistrés tous les deux semaines durant 8 semaines où les patients ont pris le médicament/placebo. Ces données constituent les comptages avant la permutation, le but principal de l'étude est de déterminer si le médicament antiépileptique Progabide a un effet réducteur sur le nombre de crises épileptiques ou non.
On va traiter l'article de Breslow (1996) qui a utilisé une régression de Poisson et une binomiale négative avec la méthode d'estimation des moindres carrées re-pondérés itérativement. La réponse est la somme des quatre comptages de crises épileptiques par patient.

3.2.1 Régression de Poisson

Le premier modèle (breslow-poiss) est le suivant :

$$\ln E[Y_i|prdicteurs] = \beta_0 + \beta_1 1_{(X_i^{(1)}=1)} + \beta_2 X_i^{(2)} + \beta_3 X_i^{(3)} + \beta_4 1_{(X_i^{(1)}=1)} X_i^{(3)} + \beta_5 X_i^{(2)} X_i^{(3)}$$

On a les notations suivantes avec i=1, . . ., 58 :

— Y_i=Ysum="nombre total de crises épileptiques du patient i"
— $X_i^{(1)}$=Trt="Traitement du i^{ime} patient "=$\begin{cases} 0 & \text{pour placebo} \\ 1 & \text{pour progabid} \end{cases}$
— $X_i^{(2)}$=log(Base4) =logarithme népérien du comptage sur 8 semaines avant l'étude.
— $X_i^{(3)}$=Age10="Ages des patients dévisés par 10"

On applique une régression de Poisson simple dont l'écriture sous R on trouve :

Min	1Q	Median	3Q	Max
-6.1377	-1.5986	-0.5294	1.2061	8.9098

Table 3.10 – Deviance résiduelle

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.19409	0.45583	4.813	1.48e-06
log(base)	0.29035	0.20065	1.447	0.147890
Age10	-0.20307	0.15284	-1.329	0.183953
Trtprogabide	-1.52397	0.17032	-8.948	<2e-16
log(Base4) :Age10	0.22468	0.06698	3.355	0.000795
lod(Base4) :Trtprogabide	0.66239	0.07323	9.045	<2e-16

Table 3.11 – Coefficients de régression de Poisson

◊ déviance nulle : 1284.72 sur 57 les degrés de liberté
◊ déviance résiduelle : 408.41 sur 52 les degrés de liberté
◊ AIC : 696.12

Ce modèle s'obtient également par élimination pas-à-pas descendante basée sur les p-valeurs. Si on choisit par contre la famille quasi-Poisson, il n'y a presque aucun effet significatif et par la méthode pas-à-pas descendante, on retient le modèle à 2 prédicteurs : $\log(Base4)$ et $Age10$.
En reproduisant les calculs de Breslow (1996), on obtient les mêmes résultats et on observe que $\log(Base4)$ n'a pas d'effet significatif sur la réponse. Pourtant, on ne pouvait pas l'exclure du modèle parce que ses interactions avec $Age10$ et Trt sont statistiquement significatives.
Par défaut, la fonction glm(.) estime les coefficients du modèle par la méthode des moindres carrées re-pondérés itérativement .
De plus, en prenant le nombre total de crises épileptiques, on s'est occupé du problème de l'indépendance des réponses. Néanmoins, en comparant la déviance résiduelle avec son degré de liberté, on constate toujours un problème de sur-dispersion. Il doit donc exister encore une autre source qui entraîne ce problème. Testant alors si la présence de sur-dispersion est statistiquement significative en appliquant un test de Dean (1992) la statistique vaut cette fois-ci :

$$T = \frac{\sum_{j=1}^{59}[(y_j - \hat{\mu}_i)^2 - y_j]}{\sqrt{2\sum_{j=1}^{59}\hat{\mu}_j}}$$

Ou $\hat{\mu}_j = \exp(X_j\beta)$ (Cas poissonien sous H_0) T vaut approximativement $35.9(p-valeur < 2.210-16)$. L'hypothèse nulle est donc rejetée et on peut affirmer qu'on est très probablement en présence de sur-dispersion. La valeur estimée du paramètre de sur-dispersion $\hat{\phi}$ vaut approximativement 8.78. Un modèle de Poisson pour ces données n'est donc probablement pas un très bon choix. En ce qui concerne l'effet du traitement, on produit à nouveau le meme genre de graphique déja utilisé plusieurs fois avant.
On constate que l'équation nous donne essentiellement des valeurs négatives ce qui est une bonne nouvelle pour l'effet du traitement. Ainsi pour quelqu'un qui a un comptage de référence de 30 crises épileptiques sur une période de huit semaines($4\exp(2) \cong 30$), le nombre de crises épileptiques d'un patient du groupe Progabide équivaut en moyenne a 0.82 fois($\exp(-0.61042 + 0.20446) \cong 0.82$) le nombre de crises d'un patient prenant un placebo. Progabide aurait donc un effet réducteur de crises. On voit également à l'aide des deux graphiques, presque tous les patients vont expérimenter une réduction. A partir de 80 comptages de référence, Progabide ne semble plus être efficace contre les crises épileptiques. 2 des 58 patients inclus dans notre étude (environ 3.4% de notre échantillon) se trouvent alors dans ce cas où le traitement n'est plus effectif. Si on regarde les résidus de Pearson studentisés tracés en fonction des variables continues transformées ou non, on remarque que la transformation log népérienne a

amélioré l'aspect de la courbe. Néanmoins, on constate qu'on se trouve en présence de quelques outliers. De plus, on voit que la transformation de la variable correspondant à l'age n'entraine presque pas de changement d'où elle peut être considérée comme étant inutile. En ce qui concerne les résultats de la régression faite avec le modèle sans transformations, tous les prédicteurs ont été jugés comme ayant un effet significatif sur la réponse. Par contre le modèle avec les transformations est mieux adapté aux données ce qu'on déduit de la valeur du AIC qui est plus petit par rapport à celui du modèle sans transformations ($696.12 contre 784.96$).

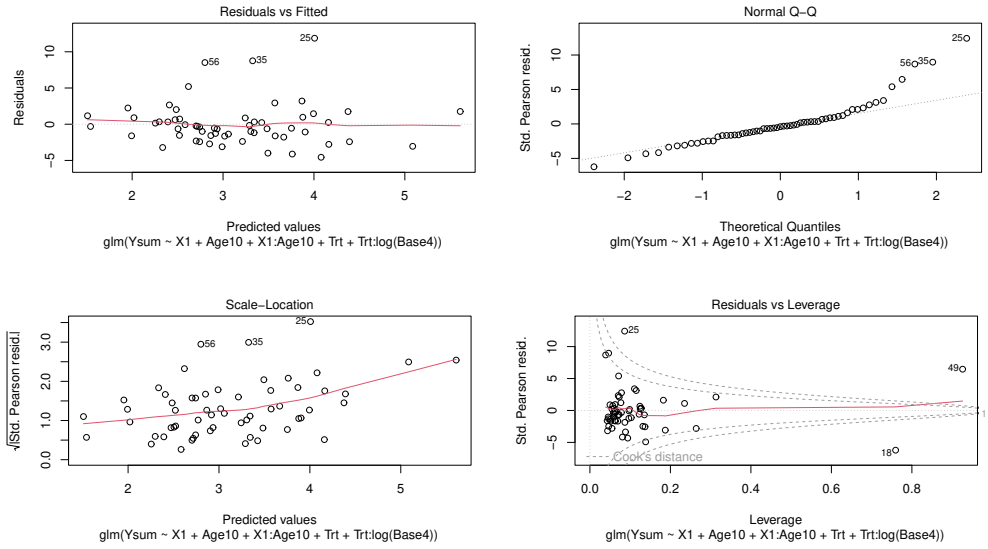

FIGURE 3.9 – Résidus de Pearson studentisés en fonction des variables continues (transformées ou non) pour poisson

3.3 Binomiale négative

Le deuxième modèle (breslow-nb-glm) proposé par Breslow exige une loi binomiale négative avec l'idée de se charger de la sur-dispersion à l'aide du paramètre supplémentaire de cette loi. On applique exactement le même modèle qu'auparavant à l'exception du fait que suit maintenant une loi binomiale négative au lieu d'une Poisson. La fonction glm.nb est basée sur la théorie décrite dans Venables W. N. et Ripley, B. D.(2002) Modern Applied Statistics with S. Fourth edition. Springer. Et se trouve dans le package R MASS.Cette fonction nous fournit également le paramètre theta qui est en fait l'inverse du paramètre de dispersion. Les valeurs de Breslow pour ce paramètre ne coïncident pas avec les nôtres. On conclut alors que glm.nb n'estime pas le paramètre de dispersion par la méthode des moments. Sur la page d'aide de R on nous dit que le paramètre est estimé par des itérations de score et d'information et est retenu quand les deux convergent. Evidemment, comme on a trouvé un autre paramètre de

dispersion, nos estimations et erreurs-standards varient par rapport à celles de Breslow. Dans le package MASS de R, on peut trouver une fonction theta.mm qui estime le theta par la méthode des moments. Il s'agit effectivement de la valeur 3.303, dont l'inverse vaut 0.3027. Comme on connait la Valeur de theta, on peut maintenant faire appel à la fonction glm.
Ce modèle n'est quand même pas préservé si on procède par la méthode pas-à-pas descendante basée sur la p-valeur à partir du modèle complet à interactions d'ordre 2.
Le modèle final obtenu par cette méthode est constitué de deux prédicteurs sans aucune interaction : log(Base4) et Trt tous les deux significatif. L'intercept lui aussi est significatif.On se concentre néanmoins sur le modèle de Breslow et on obtient quasi les mêmes résultats

Min	1Q	Median	3Q	Max
-3.0495	-0.5571	-0.1094	0.2227	2.0977

Table 3.12 – Déviance résiduelle

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.880157	1.098469	2.622	0.0114
log(base)	0.002866	0.572380	0.005	0.9960
Age10	-0.399131	0.369311	-1.081	0.2848
Trtprogabide	-0.702344	0.452637	-1.552	0.1268
log(Base4) :Age10	0.302876	0.191303	1.583	0.1194
lod(Base4) :Trtprogabide	0.248120	0.24397	1.032	0.3068

Table 3.13 – Coefficients de la régression avec un modèle Binomiale négative

- ◊ déviance nulle : 138.439 sur 57 les degrés de liberté
- ◊ déviance résiduelle : 51.728 sur 52 les degrés de liberté
- ◊ AIC : 455.29

On voit bien que le rapport de la déviance résiduelle et son degré de liberté vaut presque 1 ce qui nous suggère que la loi binomiale négative a bien géré le problème de sur-dispersion. Un aspect négatif de ce modèle est quand même qu'il trouve qu'aucun prédicteur n'a un effet statistiquement significatif. La situation souhaitable est telle qu'au moins le facteur traitement est jugé d'avoir un effet significatif sur la réponse, ce qui n'est quand même pas le cas ici.
Nos résultats nous suggèrent que la prise de Progabide n'a pas d'effet statistiquement significatif sur le nombre de crises épileptiques. Ce modèle est le mieux adapté aux données ce qui nous indique l'AIC et donc on fait plutôt confiance à ce dernier modèle. Ainsi on maintient notre interprétation : Progabide n'a probablement pas d'effet sur le nombre de crises épileptiques.
le critère AIC nous suggère que le modèle avec les transformations est le meilleur car

son AIC vaut 455.29 contre 460.61 de celui du modèle sans transformations.
Breslow a choisi d'exclure les observations du patient 207. Si on les garde, avec le modèle utilisant une loi binomiale négative, on obtient que l'intercept, Trt et l'interaction de log($Base4$) et Trt sont significatifs. Il y a quand même de forte différence en ce qui concerne l'estimation des paramètres, mais des erreurs standards très similaires. La sur-dispersion semble bien prise en compte avec une déviance résiduelle de 53.471 et 53 degrés de libertés.

CONCLUSION GÉNÉRALE

En dépit du fait que les modèles linéaires généralisés sont désormais des outils classiques en statistiques, ils ne sont encore que trop rarement intégrés. Leurs extensions récentes, avec inclusion d'effets aléatoires ou de classes latentes, rendent ces outils d'une importance considérable pour une étude statistique. Leur maîtrise suppose d'avoir clairement à l'esprit les hypothèses inhérentes à tout modèle de régression, en particulier les concepts de distribution, de moyenne et de variance conditionnelles.
Nous avons tenus dans ce mémoire à mettre en évidence ce type de modèles statistiques ainsi que la démarche à suivre pour leurs réalisation : la récoltes de données, proposition du modèle adéquat, estimations des paramètres etc. Nous terminons ce travail par une application sur les différents modèles, présentés auparavant, sur des données réelles. Enfin, nous souhaitons que notre travail a permit de montrer l'importance de ces modèles statistiques et qu'il servira de référence pour d'éventuelles études statistiques théoriques ou appliquées.

BIBLIOGRAPHIE

[1] William Feller. An extention of the law of the iterated logarithm to variables without variance. *Journal of Mathematics and Mechanics*, 18(4) :343–355, 1968.

[2] Norman L Johnson, S Kotz, and AW Kemp. Negative binomial distribution. *Distributions in Statistics : Discrete Distributions, Wiley, New York*, pages 122–142, 1969.

[3] A Colin Cameron and Pravin K Trivedi. Econometric models based on count data. comparisons and applications of some estimators and tests. *Journal of applied econometrics*, 1(1) :29–53, 1986.

[4] Rainer Winkelmann and Klaus F Zimmermann. Recent developments in count data modelling : theory and application. *Journal of economic surveys*, 9(1) :1–24, 1995.

[5] MH DeGroot. Probability and statistics. 2nd edn reading. *Mass. : Addison-Wesley*, 1986.

[6] Norman L Johnson and Samuel Kotz. Discrete distributions : Distributions in statistics. 1969.

[7] Kocherlakota and Kocherlakota. Bivariate discrete distributions. *Marcel Dekker-New York*, page 191, 1992.

[8] Chatfield C. A.S.C. Ehrenberg and G.J. Goodhardt. Progress on a simplified model of stationary purchasing behaviour. *journal of the Royal Statistical Society A 129*, pages 317–367, 1966.

[9] Michel Lejeune. Régressions linéaire, logistique et non paramétrique. In *Statistique-La thèorie et ses applications*, pages 289–322. Springer, 2010.

[10] John Ashworth Nelder and Robert WM Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society : Series A (General)*, 135(3) :370–384, 1972.

[11] Philippe Besse. Pratique de la modélisation statistique. *Publications du laboratoire de statistiques et probabilités. Université Paul Sabatier, Toulouse. Disponiblea partir de l'URL http ://www-sv. cict. fr/lsp/Besse*, 2003.

[12] Cullagh et Nelder. Introduction au modèle de régression linéaire simple. *université de toulouse.*, 1992.

[13] Régis Bourbonnais et al. *Econométrie*. Dunod, 2021.

[14] C.L Gilbert. Econometric models for discrete (integer valued) economic processes. *in : E.G. Charatsis (ed.) Selected Papers on Contemporary Econometric Problems, The Athens School of Economics and Business science*, 1982.

[15] Jerry A Hausman, Bronwyn H Hall, and Zvi Griliches. Econometric models for count data with an application to the patents-r&d relationship, 1984.

[16] A Colin Cameron and Pravin K Trivedi. *Regression analysis of count data*, volume 53. Cambridge university press, 2013.

[17] Gary King. Statistical models for political science event counts : Bias in conventional procedures and evidence for the exponential poisson regression model. *American Journal of Political Science*, pages 838–863, 1988.

[18] James J Heckman. Sample selection bias as a specification error. *Econometrica : Journal of the econometric society*, pages 153–161, 1979.

[19] Arthur S Goldberger. The interpretation and estimation of cobb-douglas functions. *Econometrica : Journal of the Econometric Society*, pages 464–472, 1968.

[20] R Winkelmann. Correctly interpreting the results from a log-linear regression under heteroskedasticity-methods and an application to the relative wages of immigrants. *JAHRBUCHER FUR NATIONALOKONOMIE UND STATISTIK*, 221(4) :418–431, 2001.

[21] Peter E Kennedy et al. Estimation with correctly interpreted dummy variables in semilogarithmic equations [the interpretation of dummy variables in semilogarithmic equations]. *American Economic Review*, 71(4) :801–801, 1981.

[22] Peter McCullagh and John A Nelder. Generalized linear models. chapman and hall. *London, UK*, 1989.

[23] Peter Diggle, Kung-Yee Liang, and Scott L Zeger. Longitudinal data analysis. *New York : Oxford University Press*, 5 :13, 1994.

[24] Tim Barmby, Michael Nolan, and Rainer Winkelmann. Contracted workdays and absence. *The Manchester School*, 69(3) :269–275, 2001.

[25] Nancy L Rose. Profitability and product quality : Economic determinants of airline safety performance. *Journal of Political Economy*, 98(5, Part 1) :944–964, 1990.

[26] Thomas Bauer, Andreas Million, Ralph Rotte, Klaus F Zimmermann, et al. Immigrant labor and workplace safety. 1998.

[27] Amemiya Takeshi and TAKESHI AUTOR AMEMIYA. Advanced econometrics. 1985.

[28] Jan Salomon Cramer. Econometric applications of maximum likelihood methods. 1989.

[29] Jerald F Lawless. Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 209–225, 1987.

[30] Rainer Winkelmann and Klaus F Zimmermann. A new approach for modeling economic count data. *Economics Letters*, 37(2) :139–143, 1991.

[31] John W Ruser. Workers' compensation and occupational injuries and illnesses. *Journal of Labor Economics*, 9(4) :325–350, 1991.

[32] Atanu Saha and Diansheng Dong. Estimating nested count data models. *Oxford Bulletin of Economics and Statistics*, 59(3) :423–430, 1997.

[33] Teofilo Ozuna and Irma Adriana Gomez. Specification and testing of count data recreation demand functions. *Empirical Economics*, 20(3) :543–550, 1995.

[34] João MC Santos Silva and Frank Windmeijer. Two-part multiple spell models for health care demand. *Journal of Econometrics*, 104(1) :67–89, 2001.