



UNIVERSITY OF ABOMEY- CALAVI

FACULTY OF AGRONOMIC SCIENCES



Laboratoire de Biomathématiques &
d'Estimations Forestières

Master Program in Statistics, Major Biostatistics

1st batch

Generalized linear models with Poisson family: applications in ecology

A thesis submitted to the Faculty of Agronomic Sciences
in partial fulfillment of the requirements for the degree
of the Master of Sciences in Biostatistics

Presented by:

LOKONON Enagnon Bruno

Supervisor: Pr Romain L. GLELE KAKAI,

Professor of Biostatistics and Forest estimation

Academic year: 2014-2015



UNIVERSITE D'ABOMEY- CALAVI

FACULTE DES SCIENCES AGRONOMIQUES



Laboratoire de Biomathématiques &
d'Estimations Forestières

Programme de Master en Biostatistiques

1^{ère} Promotion

**Modèles linéaires généralisés de la famille
de Poisson : applications en écologie**

**Mémoire soumis à la Faculté des Sciences Agronomiques
pour obtenir le Diplôme de Master recherche en Biostatistiques**

Présenté par:

LOKONON Enagnon Bruno

Superviseur: Pr Romain L. GLELE KAKAÏ,

Professeur titulaire de Biostatistiques et
estimation forestière

Année académique: 2014-2015



Certification

I certify that this work has been achieved by **LOKONON E. Bruno** under my entire supervision at the University of Abomey-Calavi (Benin) in order to obtain his Master of Science degree in Biostatistics.

Pr Romain L. GLELE KAKAÏ

Professor of Biostatistics and Forest estimation

Acknowledgements

This research was supported by WAAPP/PPAAO-BENIN (West African Agricultural Productivity Program/ Programme de Productivité Agricole en Afrique de l'Ouest).

This dissertation could only have been possible through the generous contributions of many people. First and foremost, I am grateful to my supervisor Pr Romain L. GLELE KAKAI, Professor of Biostatistics and Forest estimation who tirelessly played key role in orientation, scientific writing and mentoring during this research. In particular, I thank him for his prompt availability whenever needed.

I would kindly like to thank all the lecturers involved in this training for their useful - teaching and guidance which helped in improving this thesis. I am deeply grateful to all my colleagues, in particular TCHANDAO MANGAMANA Essomanda for the inspiring working atmosphere that they fostered and the wonderful time we had together.

I am grateful to my family for their encouragement.

Finally, I would like to thank the Lord Jesus Christ for his Love, the Holy Spirit for his Light and the Blessed Virgin for her Support.

Abstract

Ecological data are often discrete and do not follow the assumptions of the General linear model and its variants (linear regressions, ANOVA, etc.). Discrete response variables, such as count data, often contain many zero observations and are unlikely to have a normally distributed error structure even if transformed. To solve these problems, Generalized Linear Models (GLM) have been more recently developed. The basic GLM for count data is the Poisson model with log link. Frequently, count data are often overdispersed (variance of the response variable greater than the mean) and invalidating the use of the Poisson distribution. In these conditions, some extensions of Poisson model are usually used to deal with overdispersion, including the Negative binomial, Quasi-Poisson, zero-inflated Poisson (ZIP) models and Zero Inflated Negative Binomial (ZINB). The main objective of this study was to empirically assess the robustness of Poisson model and its extensions to overdispersion situations in ecological count data. The simulation plan considered took into account the overdispersion k ($k=2, 4, 8, 10, 12$ and 20), the sample size, n ($n=25, 50, 100, 500$ and 1000) and the proportion of zeros in the sample p ($p=0.20, 0.40, 0.60$ and 0.80). Two models have been considered: simple model (one explanatory variable) and 2-variables model. The comparison criteria were the mean bias (B), the mean relative error (RE) and the root mean-squared error (RMSE) of the slopes, Akaike Information Criterion (AIC) and Vuong statistic. Results obtained showed that no model perform better in all situations but Negative binomial and Zero Inflated Poisson models recorded overall good performances. Applications of these results in ecology revealed that the number of wilted plants is overdispersed because of the preponderance of zeros in the data set. The results proved that zero inflated models performed better on the number of wilted plants within pineapple cultivars in Benin.

Key words: Poisson models and its extensions, overdispersion, simulation, ecological data.

Résumé

En écologie, les données sont souvent discrètes et ne respectent pas les conditions d'application du modèle linéaire général et ses variantes (régression linéaire, ANOVA, etc.). Les variables discrètes telles que les données de comptage par ailleurs contiennent souvent beaucoup de zéros et ne suivent pas une distribution normale même après l'application d'une transformation. Pour résoudre ces problèmes, les modèles linéaires généralisés ont été récemment développés. Le modèle linéaire généralisé de base pour les données de comptage est le modèle de Poisson avec log comme fonction de lien. La principale hypothèse du modèle de Poisson est l'égalité entre la moyenne et la variance. Fréquemment cependant, les données de comptage sont souvent surdispersées présentant une variance supérieure à la moyenne empêchant ainsi l'utilisation du modèle de Poisson. Dans ces conditions, les extensions du modèle de Poisson sont souvent proposées parmi lesquelles les modèles Négatif Binomial, Quasi-Poisson et des modèles de Poisson à inflation de zéros (zero-inflated Poisson models). Le principal objectif de cette étude est d'évaluer de façon empirique la robustesse de la régression de Poisson et ses extensions dans la résolution des problèmes de surdispersion rencontrés au niveau des données de comptage en écologie. Le plan de simulation considéré prend en compte la paramètre de surdispersion k ($k=2, 4, 8, 10, 12$ et 20), la taille de l'échantillon, n ($n=25, 50, 100, 500$ et 1000) et la proportion p de zéros au sein des échantillons ($p=0,20, 0,40, 0,60$ et $0,80$). Deux modèles ont été considérés, le modèle avec une variable indépendante et le modèle avec deux variables indépendantes. Les critères de comparaison utilisés étaient le biais moyen, l'erreur moyenne relative et l'erreur quadratique moyenne d'une part, Akaike Information Criterion (AIC) et la statistique de Vuong d'autre part. Les résultats obtenus ont montré qu'aucun modèle n'est meilleur dans toutes les situations mais d'une façon globale les modèles ZIP et Négatif binomial présentent de bonnes performances. En appliquant ces résultats en écologie, il a été révélé que le nombre de plants d'ananas attaqués par la maladie de wilt était surdispersé à cause du nombre élevé de zéros dans les données. Les résultats ont montré que les modèles de Poisson à inflation de zéros sont plus performants pour modéliser le nombre de plants d'ananas attaqués par la maladie de wilt.

Mots clés: Modèles de Poisson et ses extensions, surdispersion, simulation, données écologiques.

Table of contents

	Pages
Certification.....	i
Acknowledgements	ii
Abstract	iii
Table of contents	v
List of tables	vii
List of figures	vii
1. Introduction	1
1.1 Problematic and objectives.....	1
1.2 Presentation of the case study	3
2. Principles of the GLMs	5
2.1. From General linear model to Generalized linear model	5
2.2. The exponential family distribution	5
2.2.1. <i>The Normal distribution</i>	6
2.2.2. <i>The Binomial distribution</i>	6
2.2.3. <i>The Gamma distribution</i>	7
2.2.4. <i>The Poisson distribution</i>	7
2.3. Theoretical principles under Poisson model	7
2.3.1. <i>Definition and properties</i>	8
2.3.2. <i>Maximum likelihood estimation of β</i>	8
2.3.3. <i>Goodness of fit test</i>	9
2.3.4. <i>Hypothesis test</i>	10
2.3.5. <i>Overdispersion</i>	11
2.4. Poisson model extensions.....	11
2.4.1. <i>Quasi-Poisson model</i>	11
2.4.2. <i>Negative Binomial model</i>	13
2.4.3. <i>Zero-inflated Poisson (ZIP) model</i>	14
2.4.4. <i>Zero inflated Negative Binomial (ZINB) Regression</i>	17
3. Material and methods	19
3.1. Simulation plan and comparison criteria.....	19
3.1.1. <i>Poisson models and its extensions considered</i>	19
3.1.2. <i>Sample size considered and values of overdispersion parameter</i>	19

3.1.3. Proportion of zeros.....	20
3.1.4. Generation of the populations.....	20
3.1.5. Comparison criteria	21
3.2. Application: identification of the best model to analyze prevalence of wilt disease on pineapple cultivars.....	21
4. Results and discussion.....	23
4.1. Results from the Monte Carlo study	23
4.1.1. Relative efficiency of the models considered for 1 covariate	23
4.1.2. Relative efficiency of the models considered for 2 covariates	28
4.1.3. Relative efficiency of the zero inflated models with 1 covariate	33
4.1.4. Relative efficiency of the zero inflated models with 2 covariates	36
4.2. Application on wilt disease data.....	39
5. Conclusions and perspectives for Future Research.....	48
References	49
Appendix	52

List of tables

Pages

Table 1. Estimated values for Poisson model and its extensions ($\beta_0 = 0.14$ and $\beta_1 = 0.063$)...	24
Table 2. Median ranks of Poisson model and its extensions according to the RMSE values.	27
Table 3 . Estimated values for Poisson model and its extensions ($\beta_0 = 0.14$, $\beta_1 = 0.063$ and $\beta_2 = -0.15$).	29
Table 4. Median ranks of Poisson model and its extensions according to the RMSE values (case of 2 covariates).....	33
Table 5. Estimated values for Poisson, Negative binomial, ZIP and ZINB models	34
Table 6. Vuong statistic's values for ZIP and ZINB, case of 1 covariate	35
Table 7. Rank of Poisson, Negative binomial, ZIP and ZINB models, case of 1 covariate	36
Table 8. Estimated values for Poisson, Negative binomial, ZIP and ZINB models ($\beta_0 = 0.14$, $\beta_1 = 0.063$, $\beta_2 = 1.25$).	37
Table 9. Vuong statistic's values for ZIP and ZINB, case of 2 covariates	37
Table 10. Rank of Poisson, Negative binomial, ZIP and ZINB models, case of 2 covariates.	38
Table 11. Comparison of the models: Poisson, Quasi-Poisson, NB, ZIP and ZINB.	47

List of figures

Pages

Figure 1. Boxplot of mean bias for Poisson model and its extensions.....	25
Figure 2. Boxplot of relative errors for Poisson model and its extensions	25
Figure 3. Plot of mean bias against sample size for all models	26
Figure 4. Boxplot of mean bias for Poisson model and its extensions in case of two	30
Figure 5. Boxplot of relative errors for Poisson model and its extensions in case of	30
Figure 6. Plot of mean bias against sample size for all models (case of slope 1)	31
Figure 7. Plot of mean bias against sample size for all models (case of slope 2)	32
Figure 8. Frequency of number of wilted plants per plot.....	39
Figure 9 : mean versus variance	43

Chapter 1:

Introduction

1.1 Problematic and objectives

1.2 Presentation of the case study

1. Introduction

1.1 Problematic and objectives

The first step in data analysis is to understand types of data considered before choosing a modeling approach (Yaacob *et al.*, 2010). Ecological data are often discrete (O'Hara and Kotze, 2010) and in the context of modeling this non-negative integer nature count of the response variable, the use of least square regression models suffer several methodological limitations and statistical properties (Neter *et al.*, 1996). To solve these problems, Generalized Linear Models (GLM) have been more recently developed (McCullagh and Nelder, 1989). A GLM is a generalization of linear regression that includes modifications allowing a loosening of the regression assumption of normally distributed residuals with constant variances (Czado and Sikora, 2002). It is a maximum likelihood based method that provides a systematic framework by which parameters are estimated when the model error structure belongs to the exponential family. This includes distributions such as Poisson, Gamma and Binomial (Jiao *et al.*, 2004). Given the nature of discrete, non-negative integer value with quite low number of events of the count data, Poisson distribution has been checked to be the best distribution to describe these kinds of data (Ver Hoef and Boveng, 2007, O'Hara and Kotze, 2010). Poisson regression model is a generalized linear model with the logarithm, as canonical link function, and response variable assumed to follow the Poisson distribution (Tsoumanis, 2010). This method finds its application in several areas and situations from which we address some motivated examples below.

Example 1. School administrators study the attendance behavior of high school juniors at two schools. Predictors of the number of days of absence include gender of the student and standardized test scores in math and language arts. The number of days of absence (dependent variable) is a count response that can have values 0, 3, 7, etc.

Example 2. This example refers to data from a study of crabs nesting (Agresti, 1996). Each female crab in the study has been investigated on the number of male crabs, called satellites, attached to her in her nest. The aim is to evaluate effect of some factors on the number of satellites residing close to the females. Explanatory variables that are thought to affect this included the female crab's color and weight. The response outcome for each female crab is her number of satellites. In this example, the response variable is again a count response and Poisson regression is the most appropriate approach to model the data.

Example 3. A scientist would like to model effect of four factors (Site, tick species, body part of cattle and livestock farming system) on ticks dynamics on cattle. The number of ticks is a count response that can also have values 0, 1, 5, etc. and Poisson regression is the suitable method to model these data.

Though Poisson model works well for count data, it still suffers one potential problem. This relates to the assumption of the equality between the variance and the mean of the response variable (Ver hoef and Boveng, 2007). When this assumption is violated, overdispersion occurs, meaning that data often exhibit more variance than the mean. The consequence is that standard errors of estimations are biased and thus lead to incorrect test statistics (Hilbe, 2011), resulting in a poor fit to the data (Harrison, 2014). Moreover, besides the problem of overdispersion, it is also common in real life count data to exhibit many 'zero' observations (Sileshi *et al.*, 2009) and Poisson model is not convenient.

Failing to take overdispersion and excess zeros into account can lead researchers to erroneously conclude that variables have a meaningful effect when in fact they do not (Zuur *et al.*, 2009). The necessity for accurate biological inference therefore demands that we employ tools to both identify and adequately deal with these problems to minimize the risk of Type I error (Hilbe, 2011).

Since overdispersion and excess zeros are so common, several models have been developed for these data, including the Negative binomial, Quasi-Poisson (Wedderburn, 1974), and zero-inflated (Lambert, 1992) models. The issue is linked to the choice of suitable model for a data sample with overdispersion or excess zeros. There is surprisingly little guidance in the statistical literature (Ver hoef and Boveng, 2007).

The main objective of this research work was to assess the robustness of Poisson model and its extensions to overcome problems of overdispersion and excess zeros in ecological count data. Specific objectives of the study were:

- to empirically compare different Poisson model and its extensions under a scenario of increasing degree of overdispersion;
- to empirically compare zero inflated models with Poisson and Negative binomial models under varying proportions of zeros in count data;
- to apply Poisson extension models to identify the best one that well fit the prevalence of wilt disease on the pineapple cultivars in Benin.

1.2 Presentation of the case study

A study aimed to evaluate the effect of climatic zones and cultivars of pineapple (*Ananas comosus* (L.) Merr.) on the prevalence of wilt disease in Benin. Benin is characterized by three climatic zones: Sudanian zone (mean annual rainfall is less than 1000 mm with unimodal regime), Sudano-Guinean zone (mean annual rainfall is 1000-1110 mm with unimodal regime) and Guinean zone (rainfall regime is bimodal, with a mean annual rainfall of 1200 mm) (Assogbadjo et al., 2006). Three cultivars of pineapple cultivated in Benin are concerned by the study: “Smooth Cayenne”, “sugarloaf” and a “local cultivar”. The pineapple is a tropical plant with edible multiple fruit consisting of coalesced berries and the most economically significant plant in the Bromeliaceae family. Pineapples are subject to a variety of diseases, the most serious of which is wilt disease vectored by mealybugs typically found on the surface of pineapples, but possibly in the closed blossom cups (Picture 1). Symptoms include yellowing and wilting of the leaves, resulting in the death of infected plants, especially when accompanied by drought and high temperatures.

Local cultivar (picture 2) is characterized by a small fruit (0.5-1 kg) and the flesh is of a crunchy, fibrous texture with high sugar content and moderate acidity. It has a high tolerance to stress, plagues and pathogens.

The Smooth Cayenne or “cayenne lisse” (picture 3) is characterized by: fruit 2.5–3 kg, pale yellow to yellow flesh, cylindrical in shape, high sugar and acid content, well-adapted to canning and processing, leaves without spines. Smooth Cayenne is very vulnerable to the wilt disease.

Sugarloaf (picture 4), fruit 2.5–3 kg, white flesh with no woodiness in the center, cylindrical in shape, it has a high sugar content but no acid, an unusually sweet fruit. Sugarloaf is resistant to the wilt disease.

The experimentation has consisted in installing plots of 25 m² in randomly selected sites in each zone. In Guinean zone 5 plots were considered per cultivar. In each plot 117 plants were cultivated. In Sudano-Guinean zone, 6 plots were installed per cultivar and in each plot 117 plants were cultivated. In Sudanian zone, 4 plots were installed per cultivar and 117 plants were cultivated per plot. A total of 45 plots were installed in the three zones. At one old age, the number of plants affected by the wilt disease was counted per plot (see data in appendix).

The data were extracted from the data set of a PhD student involved in the WAAPP (West African Agricultural Productivity Program) which has supported this research.



Picture 1. Pineapple subject to Wilt disease



Picture 2. Local cultivar



Picture 3. Smooth Cayenne



Picture 4. Sugarloaf

Based on these data, Poisson extension models were applied to identify the best one that well fit the prevalence of wilt disease on the pineapple cultivars in Benin after a simulation study.

Chapter 2:

Principles of the GLMs

- 2.1. From General linear model to Generalized linear model**
- 2.2. The exponential family distribution**
- 2.3. Theorectical principles under Poisson model**
- 2.4. Poisson model extensions**

2. Principles of the GLMs

2.1. From General linear model to Generalized linear model

A simple linear model that describes the relationship between a single covariate x and a continuous response variable y is written as:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (2.1.1)$$

where $i=1,2,\dots,n$ (number of observations), α the intercept (constant), β the regression coefficient for x and ε_i the error term.

It is also possible to write the same model without explicitly specifying ε_i which uses the expected value of y_i considered x .

$$E(y_i | x_i) = \mu_i = \alpha + \beta x_i. \quad (2.1.2)$$

This linear model can be transformed to a generalized linear model (Olsson, 2002) by replacing μ_i by $g(\mu_i)$

$$g(\mu_i) = \alpha + \beta x_i = \eta_i \quad (2.1.3)$$

η_i is a linear combination of the covariates and $g()$ defines the relationship between μ_i and η_i . $g()$ is called a link function.

GLM is a generalization of general linear model in the following ways:

1. Specification of the distribution of the response variable that belongs to the exponential family of distributions;
2. Specification of the link function $g()$ accordingly;
3. Specification of the linear predictor η_i .

Many commonly used distributions in the exponential family are the Normal, Binomial, Poisson, Gamma and Inverse Gaussian distributions. In addition, several other distributions are in the exponential family and they include the Beta, Multinomial, Dirichlet, and Pareto. There are other several distributions which are not in the exponential family but are used for statistical modeling and they include the Student's t and Uniform distributions.

2.2. The exponential family distribution

Suppose that we have a set of independent random response variables, y and that the probability (density) function can be written in the form (Olsson, 2002):

$$f(y; \theta, \phi) = \exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \quad (2.2.1)$$

Any distribution that can be written in this way is a member of the exponential family where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions depending on the exponential distribution considered.

ϕ is the dispersion parameter, θ is the canonical parameter with $\theta = g(\mu)$.

Standard theory for this type of distribution gives expressions for the mean and variance in the form:

- a) $\mu = E(y) = b'(\theta)$, where $E(y)$ is the first derivative of $b(\theta)$;
- b) $\text{Var}(y) = a(\phi) b''(\theta)$, where $b''(\theta)$ is the second derivative of $b(\theta)$;
- c) The variance expressed as a function of the mean is:

$\text{Var}(y) = a(\phi) V(\mu)$, where $V(\mu)$ is the *variance function*.

In the following lines some common distributions belong to exponential family and used in GLM are presented.

2.2.1. The Normal distribution

The probability density function of Normal distribution is (Dobson, 2002):

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left\{\frac{\mu y - \frac{1}{2}\mu^2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2}\text{Log}(2\pi\sigma^2)\right\} \end{aligned} \quad (2.2.2)$$

This is an exponential family distribution with:

$$\theta = \mu, \quad b(\theta) = \frac{1}{2}\mu^2, \quad a(\phi) = \sigma^2$$

Thus, $E(y) = \mu$ and $\text{Var}(y) = \sigma^2$

The link function of the normal distribution is the identity, $\theta = g(\mu) = \mu$.

2.2.2. The Binomial distribution

Consider a serial of binary events, called ‘trials’, each with only two possible outcomes: ‘success’ or ‘failure’. Let the random variable y be the number of ‘success’ in n independent trials in which the probability of success, p , is the same in all trials. Then y has the binomial distribution with probability density function (Olsson, 2002):

$$\begin{aligned} f(y; p) &= \binom{n}{y} p^y (1-p)^{n-y} \\ &= \exp\left[y \log\left(\frac{p}{1-p}\right) + n \log(1-p) + \log\left(\binom{n}{y}\right)\right] \end{aligned} \quad (2.2.3)$$

Let's consider $\theta = \log\left(\frac{p}{1-p}\right)$ or $p = \frac{\exp(\theta)}{1 + \exp(\theta)}$, and then the equation (2.2.3) is,

$$f(y; p) = \exp\left[y\theta + n \log\left(\frac{1}{1 + \exp(\theta)}\right) + \log\binom{n}{y}\right]$$

It follows that the Binomial distribution is an exponential family distribution with,

$$\theta = \log\left(\frac{p}{1-p}\right), b(\theta) = n \log[1 + \exp(\theta)], c(y, \varphi) = \log\binom{n}{y} \text{ and } a(\varphi) = 1$$

The link function of the binomial distribution is the logit, $\theta = g(p) = \log\left(\frac{p}{1-p}\right) = \text{logit}(p)$.

2.2.3. The Gamma distribution

The probability density function of gamma distribution is (Olsson, 2002):

$$f(y_i; \mu_i, v) = \left(\frac{v}{\mu_i}\right)^v \frac{y_i^{v-1} e^{-\frac{vy_i}{\mu_i}}}{\Gamma(v)}$$

$$= \exp\{[-y_i / \mu_i - \log(\mu_i)]v + (v-1)\log(y_i) + v\log(v) - \log[\Gamma(v)]\} \quad (2.2.4)$$

where $\theta_i = -1/\mu_i$, $b(\theta_i) = -\log(\theta_i)$, $a_i(\varphi) = 1/v$,

and $c(y_i, \varphi) = (v-1)\log(y_i) + v\log(v) - \log[\Gamma(v)]$

v is the shape parameter and μ_i is the scale parameter.

The link function of Gamma distribution is the reciprocal, $\theta = g(\mu) = -1/\mu$.

2.2.4. The Poisson distribution

The Poisson distribution can be written as a special case of an exponential family distribution. It has probability function (Olsson, 2002):

$$P(Y = y; \mu) = \frac{\mu^y e^{-\mu}}{y!}$$

$$= \exp\left[\frac{[\log(\mu)]y - \mu}{1} - \text{Log}(y!)\right] \quad (2.2.5)$$

where $\theta = \text{Log}(\mu)$, $b(\theta) = \mu$, $a(\varphi) = 1$ and $E(y) = \text{Var}(y) = \mu$.

The link function of the Poisson distribution is the log $\theta = g(\mu) = \log(\mu)$.

2.3. Theoretical principles under Poisson model

Poisson regression is a GLM where the response variable is modeled as having a Poisson distribution. Poisson distribution models random variables with non-negative integer

values called count data. It is a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time. Count data are data in which the observations can take only the non-negative integer value {0, 1, 2, 3, etc.}.

2.3.1. Definition and properties

For any $\lambda > 0$, a random variable Y has a Poisson distribution with parameter (or mean) λ if (Mouatassim and Ezzahid, 2012):

$$\Pr[Y = y_i] = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, y_i = 0, 1, 2, \dots \quad (2.3.1)$$

where, Y is the random variable representing the number of occurrences, λ_i is the parameter that represents the expected value of the count i . y_i represents the observed number of cases. The effect of explanatory variables on the response variable Y is modeled through the parameter λ . An important property of the Poisson distribution is that the sum of the independent Poisson random variables is a further Poisson random variable with mean equal to the sum of the individual means. It arises naturally in many contexts in particular:

- as the count of the total number of events (e.g. accidents) occurring in any given period of time, when the occurrence of an event at any one time does not influence (is independent of) the occurrences of events at other times;
- as an approximation for the binomial distribution: if n is large and p is small, thus the Binomial distribution with parameters n and p , $(B(n; p))$ is well-approximated by Poisson distribution with parameter np ;
- The mean and variance of this distribution can be shown to be equal. Since the mean is equal to the variance, any factor that affects one will also affect the other.

Since the logarithmic function is the natural link function for Poisson distribution, a log linear model is considered:

$$\ln(\lambda_i) = X_i' \beta \quad (2.3.2)$$

Where $X_i' \beta$ is the usual linear combination of predictors for case i , β is a vector of parameters associated with the vector of covariates, X_i . The regression parameter β is estimated by the maximum likelihood.

2.3.2. Maximum likelihood estimation of β

The likelihood function for n independent Poisson observations is a product of probabilities given by:

$$L(y, \lambda) = \prod_{i=1}^n \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} \quad (2.3.3)$$

where $\frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$ is the probability function of Poisson distribution.

Taking logarithm of likelihood function, we find that the log-likelihood function is:

$$\ln L(y, X, \beta) = \sum_{i=1}^n (y_i X_i' \beta - e^{X_i' \beta} - \ln(y_i!)) \quad (2.3.4)$$

The parameter β of this equation can be estimated using maximum likelihood method:

$$g_i = \frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^n (y_i - \lambda_i) X_i' = 0 \quad (2.3.5)$$

The Hessian of this function is:

$$H = \frac{\partial^2 \ln L}{\partial \beta \partial \beta'} = - \sum_{i=1}^n (X_i' X_i e^{X_i' \beta}) \quad (2.3.6)$$

Newton-Raphson Algorithm is used to estimate the parameter β as follow (Mc Cullagh and Nelder, 1989):

$$\beta_{(i+1)} = \beta_{(i)} - H^{-1}(i) g_i \quad (2.3.7)$$

And then the mean is:

$$\hat{\lambda}_i = e^{X_i' \hat{\beta}} \quad (2.3.8)$$

The main hypothesis of Poisson model is:

$$E[y_i / x_i] = \hat{\lambda}_i = e^{X_i' \hat{\beta}} = \text{var}[y_i / x_i] \quad (2.3.9)$$

After fitting the model, this hypothesis must be tested. The model is:

Overdispersed if $E[y_i / x_i] < \text{var}[y_i / x_i]$ and Underdispersed if $E[y_i / x_i] > \text{var}[y_i / x_i]$.

2.3.3. Goodness of fit test

In order to assess the adequacy of Poisson regression model, one should first look at the basic descriptive statistics for the event count data. If the count mean and variance are significantly different (equivalent in a Poisson distribution), thus the model is likely to be overdispersed or underdispersed. The model analysis option gives a scale parameter as a measure of overdispersion; this is equal to the Pearson chi-square statistic divided by the number of observations minus the number of parameters. Underdispersion is very uncommon to various forms of count data (O'Hara and Kotze, 2010).

Test of Cameron and Trivedi (1990)

The Poisson model fits well under the null hypothesis:

$$H_0 : \text{Var}(y_i) = \lambda_i ; H_1 : \text{Var}(y_i) = \lambda_i + \alpha g(\lambda_i) \quad (2.3.10)$$

Deviance D

The deviance function for the Poisson distribution is (Dobson, 2002):

$$D = 2 \sum_i^n \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right], \quad (2.3.11)$$

Where y is the number of events, $i=1 \dots n$, n is the number of observations and $\hat{\mu}_i$ is the fitted Poisson mean.

When the model is adequate these statistics follow chi-squared distribution with $n-q$ degrees of freedom, q is the number of parameters.

Pearson's chi-squared statistic

The Pearson goodness of fit test statistic is (Dobson, 2002):

$$\chi^2 = \sum_i^n (y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i). \quad (2.3.12)$$

When the model is adequate, this statistic follows chi-square distribution with $n-q$ degrees of freedom.

Akaike Information Criterion (AIC)

Akaike Information Criterion is a way of selecting a model from a set of models. It is defined as (Akaike, 1973):

$$AIC = -2(\ln(\text{likelihood})) + 2q \quad (2.3.13)$$

where $\ln$ is the natural logarithm, likelihood is the probability of the data given a model.

2.3.4. Hypothesis test

Wald's test

The Wald's test represents for GLM what the t test represents for linear model. The test is displayed to compare β_i against 0, $i=1 \dots k$, k is the number of parameters.

The hypotheses are: $H_0: \beta_1 = \dots = \beta_k = 0$ vs H_1 : at least one $\beta_i \neq 0$

The statistic is (Olsson, 2002):

$$\text{Wald} = \frac{\hat{\beta}_i^2}{S^2(\hat{\beta})} \quad (2.3.14)$$

and one reject H_0 if $\text{Wald} \geq \chi_{1-\alpha}^2$.

Likelihood ratio test

A simple test on the overall fit of the model, as an analogous to the F -test in the classical regression model is a Likelihood Ratio test on the "slopes" (Olsson, 2002). The

Likelihood ratio test is based on the following principle. Denote with L_1 the likelihood function maximized over the full parameter space, and denote with L_0 the likelihood function maximized over the model with only the intercept.

The likelihood ratio statistic is:

$$\text{LRT} = -2 \log (L_0/L_1) = -2 [\log (L_0) - \log (L_1)] \quad (2.3.15)$$

When the model is adequate, it can be shown that the distribution of the likelihood ratio statistic approaches a χ^2 distribution.

2.3.5. Overdispersion

As stated above, the Poisson distribution assumes that the variance is equal to the mean. In real situations, this can fail when there is positive dependence between the observations, or incomplete information about all relevant covariates. Thus, it can often be observed that the variance is larger than the mean. In these cases, the data are said to be overdispersed. Parameter estimates in Poisson regression models on overdispersed data have standard errors and p-values that are too small (Cameron and Trivedi, 2005). The implication of overdispersion has similar consequences as the failure of the assumption of homocedasticity in the linear regression model.

Overdispersion occurs primarily for two reasons (Harrison, 2014); ‘apparent overdispersion’ (Hilbe, 2011) arises when models have been poorly specified, for example by failing to include important predictors, interactions between predictors that have already been measured, or by specifying the incorrect link function (Hilbe, 2011). Conversely, ‘real overdispersion’ can arise when there is clustering in the count data, meaning observations are not truly independent of one another (Hilbe, 2011), when there is an excess number of zeros in the data (zero-inflation) (Zuur *et al.*, 2009), or when the variance of the response is truly greater than the mean (i.e., is not accurately described by a Poisson process).

Evidence of overdispersion indicates inadequate fit of the Poisson model. Several alternative models have emerged to correct for overdispersion. Among them, the Quasi-Poisson estimation, Negative Binomial regression model, Zero Inflated Poisson model and Zero Inflated Negative Binomial model (Miller, 2007).

2.4. Poisson model extensions

2.4.1. Quasi-Poisson model

Quasi-Poisson is defined as a modification of Poisson model allowing to take into account the overdispersion by introducing a dispersion parameter k into the Poisson model

(Allain and Brenac, 2001), so that the variance of the response is now $\text{Var}(y) = k\mu$, μ is the mean. If $k > 1$, the model is overdispersed. There is no exponential family corresponding to this specification, and the resulting GLM does not imply a specific probability distribution for the response variable. Rather, the model specifies the conditional mean and variance of the response y directly. Since the model does not give a probability distribution for y , it cannot be estimated by maximum likelihood. Nevertheless, the usual procedure for maximum likelihood estimation of a GLM yields the so-called Quasi-likelihood estimators of the regression coefficients, which share many of the properties of maximum likelihood (ML) estimators. As it turns out, Quasi-likelihood estimates of the regression coefficients are identical to ML estimates for the Poisson model. The estimated coefficient standard errors differ, however (Tsoumanis, 2010).

On the other hand, AIC cannot be computed because the likelihood is not defined, and also the residual deviance is the same for the Poisson and Quasi-Poisson models (Miller, 2007).

Estimation of coefficient β and adequacy of the model

In frame of Quasi-Poisson model, the coefficient β is estimated by using Quasi-likelihood function. Quasi-likelihood function is:

$$\begin{aligned}
 Q(\mu; y) &= \sum_{i=1}^n Q_i(\mu_i; y_i) \quad \text{with} \quad (2.4.1) \\
 Q_i(\mu_i; y_i) &= \int_{\mu_i} \frac{y_i - \mu_i}{V(\mu_i)} d\mu_i + f(y_i) \\
 &= \int_{\mu_i} \frac{y_i - \mu_i}{k\mu_i} d\mu_i + f(y_i) \\
 &= \frac{1}{k} \int_{\mu_i} \frac{y_i - \mu_i}{\mu_i} d\mu_i + f(y_i) \\
 &= \frac{1}{k} (y_i \log \mu_i - \mu_i + C)
 \end{aligned}$$

In this expression k and c are constants, $i=1, 2, \dots, n$, n =number of observations.

The maximization of $Q(\mu, y)$ is similar to the likelihood maximization in Poisson model (Allain and Brenac, 2001) when $\log(\mu_i)$ is replaced by $x_i'\beta$.

To assess the adequacy of the model, the similar statistics are used such as Deviance, Pearson χ^2 statistic as in the case of Poisson model.

2.4.2. Negative Binomial model

Another extension of the Poisson model that can be used to model overdispersion is the Negative Binomial (NB) model. Several authors (Miaou and Lum, 1993; Kumara and Chin, 1995; Shankar *et al.*, 1995) have used the NB model to relax for overdispersion. This was done by introducing a stochastic component into the relationship between observed count and covariates. NB model addresses the issue of overdispersion by introducing a dispersion parameter k to accommodate for unobserved heterogeneity in count data. This model is a generalization of Poisson regression which assumes that the mean μ_i of y_i (the response variable) is not only determined by X_i (the covariates) but also a heterogeneity component of ε_i (error term) unrelated to x_i . The formulation can be expressed as (Yaacob, 2010):

$$\hat{\mu}_i = \exp(X_i' \beta + \varepsilon_i) = \exp(X_i' \beta) \exp(\varepsilon_i) \quad (2.4.2)$$

where $\exp(\varepsilon_i) \sim \text{Gamma}(k, k)$. As a result, the density function of y_i can be derived (Yaacob, 2010):

$$f(y_i / X_i) = \frac{\Gamma(y_i + k)}{\Gamma(y_i + 1) \Gamma(k)} \left(\frac{k}{k + \mu_i} \right)^k \left(\frac{\mu_i}{k + \mu_i} \right)^{y_i} \quad (2.4.3)$$

We denote a random variable y having a negative binomial distribution as $y \sim \text{NB}(\mu, k)$ with mean $E[y] = \mu$ and variance $\text{Var}[y] = V_{\text{NB}}(\mu) = \mu + \frac{\mu^2}{k}$, where $\mu > 0$ and $k > 0$.

Here, the overdispersion is the multiplicative factor $1 + \frac{\mu}{k}$, which depends on μ (in contrast to the Quasi-Poisson). Clearly, NB distribution approaches a Poisson distribution when k tends to infinity.

ML estimation can be computed by iterating estimation of β and k , the two parameters of NB using Newton-Raphson algorithm.

Estimation of β and k and adequacy of the model

In the case of Negative Binomial model, coefficients β and k are estimated using likelihood function. The log-likelihood function for NB regression is:

$$\ln L(\beta, k, y) = \sum_{i=1}^n \left[\sum_{t=0}^{y_i-1} \ln\left(1 + \frac{t}{k}\right) + y_i \ln \mu_i - (y_i + k) \ln\left(1 + \frac{\mu_i}{k}\right) \right]. \quad (2.4.4)$$

By equating the first derivatives of the log-likelihood with respect to β and k to zero, ML estimates can be written, respectively, as:

$$\frac{\partial \ln L(\beta, k, y)}{\partial \beta} = \sum_i^n \frac{y_i - \mu_i}{\mu_i \left(1 + \frac{\mu_i}{k}\right)} \left(\frac{\partial \mu_i}{\partial \beta_i} \right) = 0; \quad (2.4.5)$$

$$\frac{\partial \ln L(\beta, k, y)}{\partial k} = \sum_i^n \left\{ \sum_{t=0}^{y_i-1} \left(\frac{t}{1 + \frac{t}{k}} \right) + k^2 \ln \left(1 + \frac{\mu_i}{k} \right) - \frac{(y_i + k)}{1 + \frac{\mu_i}{k}} \mu_i \right\} = 0. \quad (2.4.6)$$

Newton-Raphson algorithm steps:

Let us denote the following formulas:

$$g_k = \sum \frac{\partial l_i}{\partial k^*} \quad G_k = \sum \frac{\partial^2 l_i}{\partial k^{*2}} \quad (2.4.7)$$

$$g_\beta = \sum \frac{\partial l_i}{\partial \beta} \quad G_\beta = \sum \frac{\partial^2 l_i}{\partial \beta \partial \beta}, \quad g_{\beta k} = \begin{bmatrix} g_\beta \\ g_k \end{bmatrix} \quad G_{\beta k} = \begin{bmatrix} G_\beta & \sum \frac{\partial^2 l_i}{\partial k^* \partial \beta} \\ \left(\sum \frac{\partial^2 l_i}{\partial k^* \partial \beta} \right)' & G_{k^*} \end{bmatrix} \quad (2.4.8)$$

where l_i is likelihood of NB regression and $k^* = \ln(k)$. The Newton-Raphson algorithm steps are:

Step 1: Set $k^*=0$ and iterate to obtain estimate of β : $\tilde{\beta}^{(i)} = \tilde{\beta}^{(i-1)} - [G_\beta]^{-1} g_\beta$;

Step 2: Set $\beta' = [1 \ 0 \ 0 \ 0]$ and iterate to obtain estimate of k^* : $\tilde{k}^{*(i)} = \tilde{k}^{*(i-1)} - [G_k]^{-1} g_k$;

Step 3: Use results from steps 1 and 2 as starting values to obtain estimates of k^* and β ;

Step 4: Back-transform k^* to get estimate of k : $k = \exp(k^*)$;

Step 5: repeat the steps 2 and 3 until convergence.

To assess the adequacy of the model, the similar statistics are used such as Deviance, Pearson χ^2 statistic like in the case of Poisson model.

2.4.3. Zero-inflated Poisson (ZIP) model

Poisson model is inadequate in the case of excess of zeros in the sample. In such a situation, the zero-inflated Poisson regression model is more appropriated. ZIP model is a modification of Poisson regression model that allows for an over-abundance of zero counts in the data (Mouatassim and Ezzahid, 2012). For instance, in counting disease lesions on plants, a plant may have no lesions either because it is resistant to the disease, or simply because no disease spores have landed on it. This is the distinction between structural zeros, which are inevitable, and sampling zeros, which occur by chance. The essential idea is that the data

comes from two regimes. In one regime (R_I), the outcome is always a zero count, while in the other regime (R_{II}) the counts follow a standard Poisson process.

Suppose that,

$$\Pr[y_i \in R_I] = \omega_i, \quad (2.4.9)$$

and,

$$\Pr[y_i \in R_{II}] = (1 - \omega_i); i=1,2,\dots,n. \quad (2.4.10)$$

Thus, the probability density of ZIP can be formulated as follows (Mouatassim and Ezzahid, 2012):

$$P(Y = y_i | x_i, z_i) = \begin{cases} \omega_i(z_i) + (1 - \omega_i(z_i))\text{Pois}(\mu_i; 0 | x_i) & \text{if } y_i = 0 \\ (1 - \omega_i)\text{Pois}(\mu_i; y_i | x_i) & \text{if } y_i > 0 \end{cases} \quad (2.4.11)$$

where z_i is a vector of covariates defining the probability ω_i .

With $\text{Pois}(\mu_i; 0 | x_i) = \exp(-\mu_i)$ and $\text{Pois}(\mu_i; y_i | x_i) = \exp(-\mu_i) \mu_i^{y_i} / y_i!$

Covariates enter the model through the conditional mean, μ_i of the Poisson distribution:

$\mu_i = \exp(x_i' \beta)$. The mean and the variance of ZIP are $E[y_i | x_i] = (1 - \omega_i) \mu_i$ and

$$\text{Var}[y_i | x_i] = (1 - \omega_i)(\mu_i + \omega_i \mu_i^2).$$

One can observe clearly that ZIP reduces to the classical Poisson model when $\omega_i = 0$. Otherwise (if $\omega_i > 0$), ZIP is overdispersed because the variance exceeds the mean. Following Lambert (1992), it is common, and convenient, to model ω_i using a Logit model, so: $\omega_i = [\exp(z_i' \gamma)] / [1 + \exp(z_i' \gamma)]$, where z_i is a (1xp) vector of covariates defining the probability ω_i , and γ is a (px1) vector of its corresponding parameters.

Estimation of β and γ

If we have n independent observations in our sample, we can define the likelihood function of Y_i as follow (Lambert, 1992):

$$L(\beta, \gamma) = \prod_{y_i=0} [\omega_i + (1 - \omega_i) \exp(-\mu_i)] \prod_{y_i \neq 0} [(1 - \omega_i) \exp(-\mu_i) \mu_i^{y_i} / y_i!]. \quad (2.4.12)$$

The log-likelihood function may be written as:

$$\begin{aligned} \log L(\beta, \gamma) = & \sum_{y_i=0} \log [\exp(z_i' \gamma) + \exp(-\exp(x_i' \beta))] \\ & + \sum_{y_i \neq 0} [y_i x_i' \beta - \exp(x_i' \beta) - \log(y_i!)] - \sum_{i=1}^n \log [1 + \exp(z_i' \gamma)] \end{aligned} \quad (2.4.13)$$

Let D_i be an indicative function defined as follow:

$$D_i = \begin{cases} 1 & \text{if } y_i = 0 \\ 0 & \text{otherwise} \end{cases} \quad (2.4.14)$$

The log likelihood function becomes:

$$\begin{aligned} \log L(\beta, \gamma) &= \sum_{i=1}^n D_i \log[\exp(z_i' \gamma) + \exp(-\exp(x_i' \beta))] \\ &+ \sum_{i=1}^n (1 - D_i) [y_i x_i' \beta - \exp(x_i' \beta) - \log(y_i!)] - n \log[1 + \exp(z_i' \gamma)] \end{aligned} \quad (2.4.15)$$

The log likelihood function can be maximized using EM algorithm (Hall, 2000). The EM algorithm estimates the expectations of the i -th missing observation at each iteration and uses these expectations to estimate the parameters, and iterates until convergence. The missing data of this problem are defined as indicator variables $\delta = (\delta_1, \dots, \delta_n)'$ where $\delta_i = 1$ when y_i is from the zero state and $\delta_i = 0$ when y_i is from the Poisson state. Lambert (1992) defined the complete-data as follow:

$$\log L_c(\beta, \gamma, y, \delta) = \log L_c(\gamma, y, \delta) + \log L_c(\beta, y, \delta) + \sum_{i=1}^n (1 + \delta_i) \log(y_i!) \quad (2.4.16)$$

where $\log L_c(\gamma, y, \delta) = f(\delta_i | \gamma) = \sum_{i=1}^n [\delta_i z_i' \gamma - \log(1 + \exp(z_i' \gamma))]$ and

$$\log L_c(\beta, y, \delta) = f(y_i | \delta_i, \beta) = \sum_{i=1}^n (1 - \delta_i) [y_i x_i' \beta - \exp(x_i' \beta)].$$

This log-likelihood can be easily maximized because $\log L_c(\gamma, y, \delta)$ and $\log L_c(\beta, y, \delta)$ can be maximized separately. The EM proceeds iteratively via three steps: E step, M step for β and M step for γ . For the $(k+1)$ iteration, the three steps are defined as follows (Lambert, 1992):

E step: estimate δ_i by its conditional expectation $\delta_i^{(k)}$ given by the current estimates $\beta^{(k)}$ and $\gamma^{(k)}$:

$$\delta_i^{(k)} = \begin{cases} [1 + \exp(z_i' \gamma^{(k)} - \exp(x_i' \beta^{(k)}))]^{-1} & \text{if } y_i > 0 \\ 0 & \text{if } y_i = 0 \end{cases}. \quad (2.4.17)$$

M step for β : it consists on finding $\beta^{(k+1)}$ by maximizing $\log L_c(\beta, y, \delta^{(k)})$. This can be accomplished by fitting a weighted log linear Poisson regression of y on the covariate matrix x with weights $1 - \delta^{(k)}$.

M step for γ : it consists on finding $\gamma^{(k+1)}$ by maximizing $\log L_c(\gamma, y, \delta^{(k)})$ as function of γ :

$$\log L_c(\gamma, y, \delta^{(k)}) = \sum_{y_i=0} \delta_i^{(k)} z_i' \gamma - \sum_{y_i=0} \delta_i^{(k)} \log(1 + \exp(z_i' \gamma)) - \sum_{i=1}^n (1 - \delta_i^{(k)}) \log(1 + \exp(z_i' \gamma)) \quad (2.4.18)$$

Note that $\gamma^{(k+1)}$ can be found by fitting a weighted logistic regression of the response y on the covariate matrix z with weights $(1-\delta^{(k)})$ (Lambert, 1992).

Adequacy of ZIP regression model

The adequacy of ZIP is assessed by comparing it to Poisson using Vuong's statistic V (Vuong, 1989). Vuong (1989) has introduced, in frame of maximum likelihood estimations, a test which is a well suited method to compare non-nested models: ZIP (or ZINB) against Poisson (or Negative binomial). Let f_1 be the density function of the model 1 (ZIP or ZINB) and f_2 the density function of the model 2 (Poisson or Negative binomial).

Let $m_i = \text{Log}\left(\frac{f_1(y_i)}{f_2(y_i)}\right)$ where $i=1, \dots, n$, n =number of observations. Hence, the Vuong's statistic for the hypothesis $E(m_i) = 0$ is given by:

$$V = \frac{\sqrt{n}\left(\frac{1}{n} \sum_{i=1}^n m_i\right)}{\sqrt{\frac{1}{n} \sum_{i=1}^n (m_i - \bar{m})^2}} \quad (2.4.19)$$

Under the null hypothesis, the Vuong's statistic is asymptotically normally distributed. At 5 % significance level, the first model (ZIP or ZINB) is preferred if $V > 1.96$; if $V < -1.96$, then the second one (Poisson or Negative binomial) is preferred and the two models are equivalent when $|V| < 1.96$.

2.4.4. Zero inflated Negative Binomial (ZINB) Regression

ZINB distribution is a mixture distribution, similar to ZIP distribution, where the probability p for excess zeros and with probability $(1-p)$ the rest of the counts followed Negative binomial distribution. ZINB distribution is given by:

$$P(Y = y) = \begin{cases} p + (1-p)\left(1 + \frac{\mu}{k}\right)^{-k} & \text{if } y = 0 \\ (1-p) \frac{\Gamma(y+k)}{y! \Gamma(k)} \left(1 + \frac{\mu}{k}\right)^{-k} \left(1 + \frac{\mu}{k}\right)^{-y} & \text{if } y = 1, 2, \dots \end{cases} \quad (2.4.20)$$

k is a shape parameter which quantifies the amount of overdispersion, and Y is the response variable of interest.

The mean and variance of ZINB distribution are:

$E(y) = (1-p)k$ and $V(y) = (1-p)\mu(1-p\mu+\mu/k)$, respectively. It can be noted that this distribution approaches the ZIP distribution and the negative binomial distribution as k tends to 1 and p tends to 0, respectively.

If both $1/k$ and p tends to 0 then ZINB distribution reduces to Poisson distribution. ZINB regression model relates p and k to covariate matrix x and z with regression parameters β and γ as: $\log(\mu_i) = x\beta$ and $\logit(p_i) = z\gamma$ $i=1, 2, \dots, n$.

ZINB log-likelihood given the observed data is (Yesilova *et al.*, 2012):

$$\begin{aligned} l(\beta, \gamma, k; y, z, x) = & \sum_{i=1}^n \log(1 + \exp(z_i \gamma)) - \sum_{i=1: y_i=0}^n \log \left(\exp(z_i \gamma) + \left(\frac{\exp(x_i \beta) + k}{k} \right)^{-k} \right) \\ & + \sum_{i=1: y_i > 0}^n \left(k \log \left(\frac{\exp(x_i \beta) + k}{k} \right) + y_i \log(1 + \exp(-x_i \beta) k) \right) \\ & + \sum_{i=1: y_i > 0}^n (\log \Gamma(k) + \log \Gamma(1 + y_i) - \log \Gamma(k + y_i)) \end{aligned} \quad (2.4.21)$$

ML estimations for β , k and γ can be obtained by using the EM algorithm like in the case of ZIP.

Chapter 3:

Material and methods

3.1. Simulation plan and comparison criteria

**3.2. Application: identification of the best model to analyze
prevalence of wilt disease on pineapple cultivars**

3. Material and methods

Monte Carlo simulation method was used in the frame of this study. Paxton *et al.* (2001) provide a succinct explanation of the method as follow: The researcher begins by creating a model with known population parameters (i.e., the values are set by the researcher). The analyst then draws repeated samples of size N from that population and, for each sample, estimates the parameters of interest. Next, a sampling distribution is estimated for each population parameter by collecting the parameter estimates from all the samples. The properties of that sampling distribution, such as its mean or variance, come from this estimated sampling distribution. Similarly, Mooney (1997) explains that Monte Carlo simulation offers an alternative to analytical mathematics for understanding a statistic's sampling distribution and evaluating its behavior in random samples. Monte Carlo simulation does this empirically using random sample from known populations of simulated data to track a statistic's behavior.

3.1. Simulation plan and comparison criteria

3.1.1. Poisson models and its extensions considered

Two Poisson generalized linear models with a log link were considered and defined by:

First model: $\log(E(Y_i)) = \log(\mu) = \beta_0 + \beta_1(X_{1i})$

Second model: $\log(E(Y_i)) = \log(\mu) = \beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i})$,

where Y is the response variable; X1 and X2 are the explanatory variables; $i=1 \dots n$, n =number of observations and μ is parameter (or mean) of Poisson distribution.

The vector $\beta = (\beta_0, \beta_1, \beta_2)$ used in the simulation study were prespecified as: $\beta_0=0.14$; $\beta_1=0.063$ and $\beta_2= -0.15$. These values were chosen in order to obtain the mean of populations to be $\mu \approx 1$. Moreover, the Poisson extension models such as Quasi-Poisson model, Negative binomial model, Zero Inflated Poisson model were considered in the simulations as well as the linear model applying $\log_{10}$ on the response variable (Y+1).

3.1.2. Sample size considered and values of overdispersion parameter

From each population of size $N=10000$, five different samples of size n were generated ($n= 25, 50, 100, 500, 1000$). The overdispersion parameter k is the ratio between the variance and the mean. If k is greater than 1 the model is overdispersed. The

underdispersion ($k < 1$) situation is very scarce in ecology (O'Hara and Kotze, 2010). For this reason we defined 6 values of k such as, $k = 2, 4, 8, 10, 12$ and 20 .

3.1.3. Proportion of zeros

Previous research displayed zero-inflation ranging from 0.20 (20 % of zeros) (Mullahy, 1986) to 0.96 (96 % of zeros) (Zorn, 1996). To reflect a range including both zero-deflation and zero-inflation, four populations were established differing in the proportion of zeros present in the count outcome variable. The populations contained 0.20, 0.40, 0.60, 0.80 proportions of zeros.

3.1.4. Generation of the populations

Populations of size $N=10000$ were simulated. In each population, the covariates X_1 and X_2 were defined from the standard Normal distribution. The outcome variable, Y , was randomly established from Poisson distribution of parameter $\mu \approx 1$. Since for a Poisson distribution, a linear transformation $y=kY$, changes the equality between the mean and the variance, $\text{mean}(y)=k\text{mean}(Y)$ and $\text{var}(y)=k^2\text{var}(Y)$. The algorithm of simulation is presented in the following steps:

Step 1. Define values for the coefficients $\beta_0, \beta_1, \beta_2$ (in the case of 2 explanatory variables);

Step 2. Generate the explanatory variables X_1 and X_2 from the standard Normal distribution;

Step 3. Calculate the mean such that $E(Y_i) = \mu = \exp[\beta_0 + \beta_1(X_{1i}) + \beta_2(X_{2i})]$;

Step 4. Generate the outcome variable, Y from Poisson distribution of parameter μ with sample size N ;

Step 5. Apply the overdispersion parameter k to the Poisson distribution considered to obtain an overdispersed distribution, through a linear transformation $y=kY$.

Step 6. Consider a multivariate bootstrap sample (3 variables) of size n ($n=25, 50, 100, 500, 1000$) taken from the populations generated at the previous steps.

Step 7. For each combination of n and k , repeat step 6 S times ($S=1,000$) and for each bootstrap sample, run the following models a) Poisson model ; b) Quasi-Poisson model ; c) Negative binomial model; d) Zero Inflated Poisson model; d) Linear model $\log_{10}$ transformation of the response variable.

The MASS and pscl packages were used to run Negative binomial model and Zero Inflated Poisson model, respectively.

The algorithm used for data sample with zeros is the same as the one described above except at the step 4 where the outcome variable, Y is generated from the Zero Inflated Poisson

distribution using the function `rzipois ( )` of R software specifying the proportion of excess zeros. We used the VGAM package of R software to generate the random variates from a zero-inflated Poisson distribution.

3.1.5. Comparison criteria

For each of the 30 combinations of factors levels considered (n and k), the 5 different models were compared by calculating the mean bias (B), the mean relative error (RE) and root mean-squared error (RMSE) of the slopes; for the first slope β_1 , we have:

$$B = \frac{1}{S} \sum_{j=1}^S (\beta_1 - \hat{\beta}_{1j}); \text{RAE} = \frac{1}{S} \sum_{j=1}^S \frac{|\beta_1 - \hat{\beta}_{1j}|}{\beta_1}; \text{RMSE} = \sqrt{\frac{1}{S} \sum_{j=1}^S (\beta_1 - \hat{\beta}_{1j})^2}; \quad (3.1.1)$$

where $\hat{\beta}_{1j}$ is the estimated parameter, β_1 is the true value and $j=1 \dots S$, S is the number of the bootstrap samples considered (S=1000). The same formulas above were also applied to β_2 (for two covariates models).

At each iteration, the model showing the low values of these statistics is the best. Estimated mean biases (B) and Relative errors (RE) were plotted and analyzed. As far as root mean-squared error (RMSE) was concerned, a rank was carried out for each combination of n and k from 1 (the best model) to 5 (the less accurate model).

Moreover, in the case of Zero Inflated Poisson simulation, for the 20 combinations of n and p, ZIP model was compared to Poisson and Zero Inflated Negative binomial model (ZINB) to Negative binomial using Vuong's statistic V (equation (2.4.19)). ZIP or ZINB is preferred to Poisson or Negative binomial respectively if $V > 1.96$. ZIP is equivalent to Poisson and ZINB is equivalent to Negative binomial model when $|V| < 1.96$.

To choose between Poisson, negative binomial and zero inflated models for each combination of n and p, Akaike's Information Criterion (equation (2.3.13)). The model with small value is the better. A rank was carried out for each combination of n and p from 1 (the best model) to 4 (the model of lowest performance).

3.2. Application: identification of the best model to analyze prevalence of wilt disease on pineapple cultivars

In a first step, we used the simulation study to select the best model based on the overdispersion of the data, the proportion of zeros, the sample size and the number of covariates. In a second step, we applied GLM on real life data collected on prevalence of wilt disease on the pineapple cultivars in Benin. We started by applying Poisson model. Since Poisson model did not fit well, we used the Poisson extension models (Quasi-Poisson,

Negative binomial, ZIP, ZINB), and on the basis of appropriate statistics (AIC, Log-likelihood, etc.), we have chosen the best model.

All data were analyzed using R software.

Chapter 4:

Results and discussion

4.1. Results from the Monte Carlo study

4.2. Application on wilt disease data

4. Results and discussion

4.1. Results from the Monte Carlo study

4.1.1. Relative efficiency of the models considered for 1 covariate

Table 1 presents the estimated values of Poisson model and its extensions for the combinations n and k . Generally, the estimated slopes of Poisson, Quasi-Poisson and Negative binomial are close to the true value (0.063) for high sample size ($n=500$ and $n=1000$). For sample size less than 500, these values are less or greater than the true value. In all five models and for every value of n , the value of the intercept increases with the value of k . Moreover, the estimated slope of ZIP is less than the true value for all combinations of n and k . As far as Linear Model (LM) is concerned, for each sample, the estimated slope is less than the true value when k is equal to 2 and 4 but greater than the actual value in other cases. The result shows also that Poisson and Quasi-Poisson models gave the same values of parameter estimates.

Figures 1 and 2 present boxplots of the mean bias and the relative error of the 5 models. These figures show the best behavior of the Negative binomial, Quasi-Poisson and Poisson. Nevertheless, the Negative binomial obtained the lowest median values of bias and relative error for all combinations of n and k . Moreover, the dispersion around the median values of bias and relative error is less pronounced in the case of Quasi-Poisson and Poisson models. These two models yielded same coefficient estimates and therefore same bias and relative error but standard errors of the coefficient estimates model are always different and are adjusted for overdispersion in the case of Quasi-Poisson (Agresti, 1990). The equality of the parameters of the two models come from the form of likelihood (Poisson model) and Quasi-likelihood (Quasi-Poisson) used in the estimation (Allain and Brenac, 2001).

Table 1. Estimated values for Poisson model and its extensions ($\beta_0=0.14$ and $\beta_1=0.063$).

n	k	Poisson		Quasi-Poisson		Negative binomial		ZIP		LM	
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
25	2	0.792	0.069	0.792	0.069	0.791	0.071	1.144	0.033	0.957	0.051
25	4	1.487	0.057	1.487	0.057	1.486	0.059	1.888	0.027	1.334	0.053
25	8	2.190	0.067	2.190	0.067	2.189	0.070	2.591	0.033	1.762	0.080
25	10	2.410	0.070	2.410	0.070	2.410	0.070	2.812	0.033	1.901	0.084
25	12	2.592	0.063	2.592	0.063	2.591	0.064	2.989	0.027	2.022	0.086
25	20	3.102	0.063	3.102	0.063	3.102	0.064	3.502	0.029	2.351	0.095
50	2	0.819	0.058	0.819	0.058	0.819	0.059	1.166	0.025	0.962	0.044
50	4	1.507	0.064	1.507	0.064	1.507	0.066	1.894	0.031	1.340	0.060
50	8	2.203	0.069	2.203	0.069	2.203	0.070	2.597	0.031	1.757	0.080
50	10	2.428	0.054	2.428	0.054	2.428	0.054	2.818	0.024	1.902	0.072
50	12	2.603	0.061	2.603	0.061	2.603	0.061	2.998	0.027	2.007	0.082
50	20	3.120	0.062	3.120	0.062	3.120	0.062	3.510	0.029	2.351	0.093
100	2	0.832	0.064	0.832	0.064	0.832	0.064	1.171	0.035	0.967	0.046
100	4	1.519	0.065	1.519	0.065	1.519	0.066	1.902	0.030	1.340	0.062
100	8	2.214	0.061	2.214	0.061	2.214	0.061	2.598	0.027	1.761	0.074
100	10	2.441	0.066	2.441	0.066	2.441	0.066	2.820	0.033	1.911	0.081
100	12	2.611	0.062	2.611	0.062	2.611	0.062	2.997	0.031	2.013	0.080
100	20	3.129	0.069	3.129	0.069	3.129	0.069	3.514	0.032	2.353	0.107
500	2	0.831	0.063	0.831	0.063	0.831	0.063	1.174	0.034	0.960	0.046
500	4	1.525	0.061	1.525	0.061	1.525	0.062	1.906	0.028	1.338	0.059
500	8	2.218	0.064	2.218	0.064	2.218	0.064	2.601	0.031	1.758	0.076
500	10	2.441	0.063	2.441	0.063	2.441	0.063	2.823	0.029	1.899	0.080
500	12	2.620	0.063	2.620	0.063	2.620	0.063	3.005	0.029	2.009	0.084
500	20	3.134	0.063	3.134	0.063	3.134	0.063	3.517	0.030	2.347	0.094
1000	2	0.833	0.062	0.833	0.062	0.833	0.062	1.172	0.034	0.962	0.045
1000	4	1.524	0.063	1.524	0.063	1.524	0.063	1.905	0.029	1.338	0.060
1000	8	2.219	0.063	2.219	0.063	2.219	0.063	2.601	0.030	1.759	0.074
1000	10	2.440	0.062	2.440	0.062	2.440	0.062	2.823	0.029	1.897	0.078
1000	12	2.623	0.063	2.623	0.063	2.623	0.063	3.006	0.030	2.013	0.084
1000	20	3.135	0.063	3.135	0.063	3.135	0.063	3.517	0.028	2.349	0.097

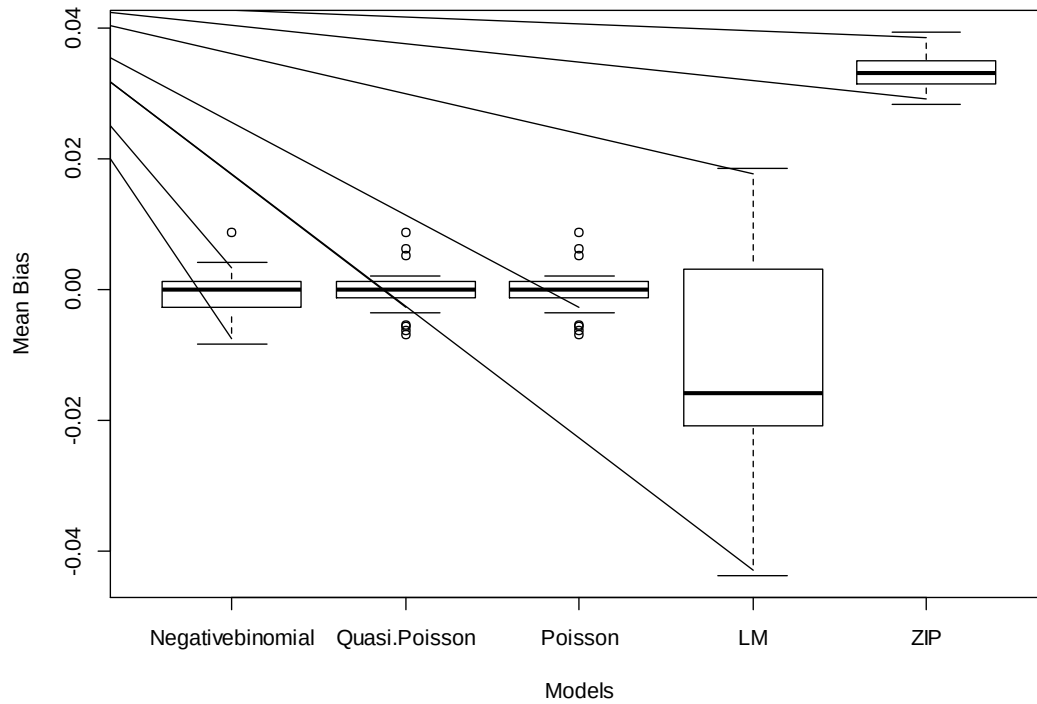


Figure 1. Boxplot of mean bias for Poisson model and its extensions

LM, followed by ZIP are the less performers since they yielded large median values of bias and relative error as depicted in Figure 1 and Figure 2 respectively. Moreover, the dispersion around the median values of bias and relative error is largely pronounced in case of LM.

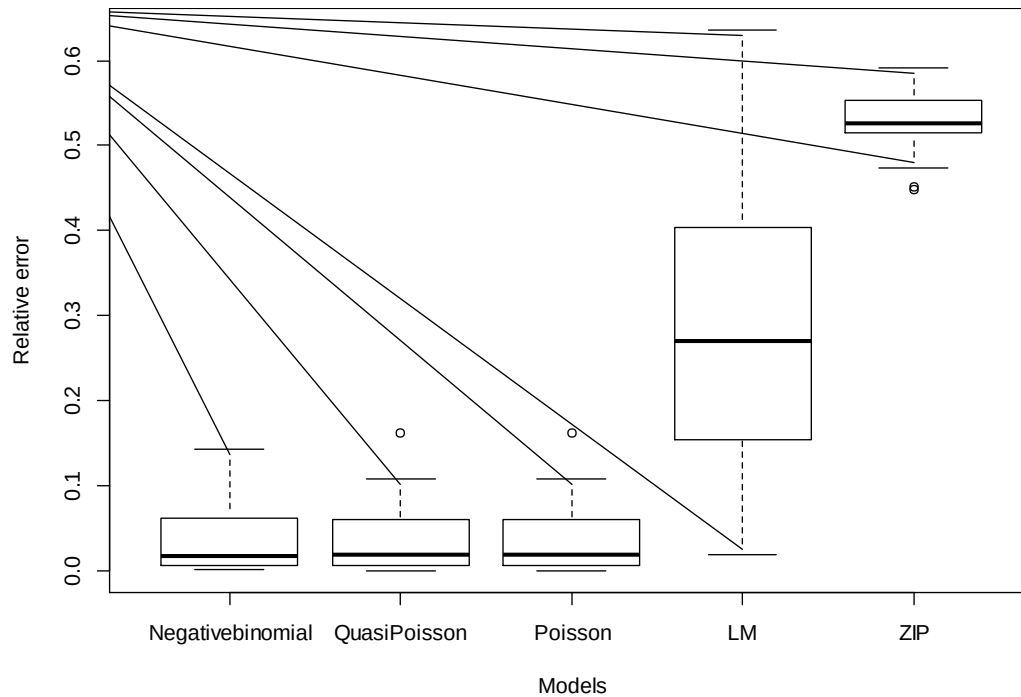


Figure 2. Boxplot of relative errors for Poisson model and its extensions

Effect of overdispersion

Figure 3 shows bias of the 5 models according to the values of the overdispersion k for each sample size. It appears that the mean bias is close to zero for Poisson, Quasi-Poisson and Negative binomial models along of k . ZIP shows relative constant mean bias values between 0.02 and 0.06. As far as LM is concerned, for small k ($k \leq 5$) the mean bias values is close to zero but when k increases, the mean absolute bias values become important (around to -0.06 at $k=20$).

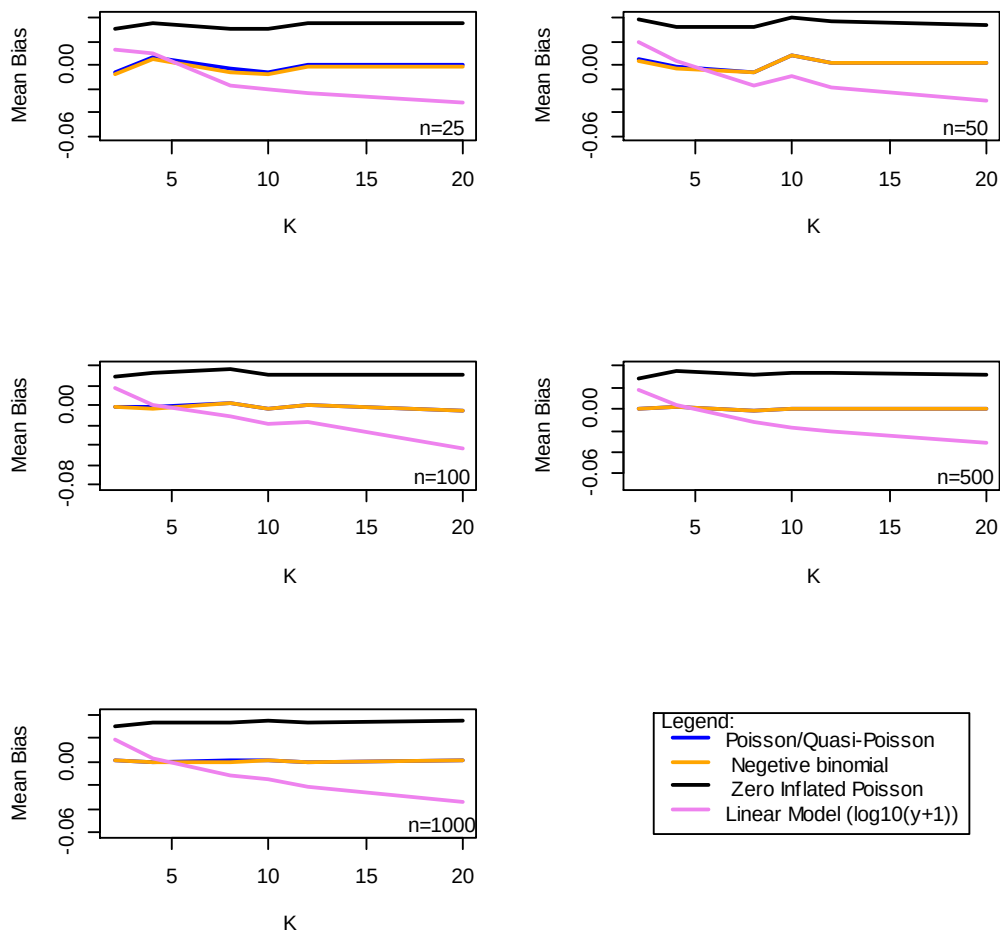


Figure 3. Plot of mean bias against sample size for all models

The Table 2 shows that when the overdispersion parameter k is equal to 2 the LM is the best performer model (rank 1) at small sample size ($n=25$ and 50). At the same value of k , LM is ranged at the second place for the other samples. When k is greater than 4, ZIP is the best performer model (rank 1) at relatively small sample size ($n=25$, 50 and 100). Moreover, for relatively high sample size ($n= 500$ and 1000), NB appears to be the best model. It is also noticed that Poisson and Quasi-Poisson have the same performance (same rank) for each of the combinations of n and k .

Table 2. Median ranks of Poisson model and its extensions according to the RMSE values

n	k	Poisson	Quasi-Poisson	Negative binomial	ZIP	LM
25	2	3	3	5	2	1
25	4	3	3	5	1	2
25	8	2	2	4	1	5
25	10	2	2	4	1	5
25	12	2	2	4	1	5
25	20	2	2	4	1	5
50	2	3	3	5	2	1
50	4	2	2	5	1	4
50	8	2	2	4	1	5
50	10	2	2	4	1	5
50	12	2	2	4	1	5
50	20	2	2	4	1	5
100	2	4	4	1	3	2
100	4	3	3	2	1	5
100	8	3	3	2	1	5
100	10	3	3	2	1	5
100	12	3	3	2	1	5
100	20	3	3	2	1	5
500	2	3	3	1	5	2
500	4	2	2	1	5	4
500	8	2	2	1	4	5
500	10	2	2	1	4	5
500	12	2	2	1	4	5
500	20	2	2	1	4	5
1000	2	3	3	1	5	2
1000	4	2	2	1	5	4
1000	8	2	2	1	4	5
1000	10	2	2	1	4	5
1000	12	2	2	1	4	5
1000	20	2	2	1	4	5

Through these simulation outcomes, it has been shown that even when data are generated from the Poisson model, ZIP and Negative binomial models tend to yield small RMSE. For small sample sizes ($n \leq 100$), ZIP seems to be more robust than the other models. But for high sample sizes, ($n = 500, 1000$), NB is the most robust model. Moreover, for count data, our results suggest that log-transformations in frame of LM perform poorly. Similar results have been found by O'Hara and Kotze (2010) who suggested that an evidence of overdispersion in data demands a correction by using appropriate methods.

4.1.2. Relative efficiency of the models considered for 2 covariates

Table 3 presents the estimated values of Poisson model and its extensions for the combinations of n and k . Results obtained showed that Poisson, Quasi-Poisson and Negative binomial models have the same values of parameter estimates especially from $(n=100, k=12)$ until $(n=1000, k=20)$. Generally, the estimated slopes of Poisson, Quasi-Poisson and Negative binomial are close to the true values (0.063 and - 0.15). In all five models, for each value of n , it is noticed that increase in k increases the value of the intercept. Moreover, the estimated slopes of ZIP and LM are relatively different from the true values.

Table 3 . Estimated values for Poisson model and its extensions ($\beta_0 = 0.14$, $\beta_1 = 0.063$ and $\beta_2 = -0.15$).

n	k	Poisson			Quasi-Poisson			Negative binomial			ZIP			LM		
		$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$	$\tilde{\beta}_0$	$\tilde{\beta}_1$	$\tilde{\beta}_2$
25	2	0.792	0.071	-0.165	0.792	0.071	-0.165	0.797	0.077	-0.172	1.136	0.028	-0.078	0.970	0.048	-0.109
25	4	1.472	0.059	-0.149	1.472	0.059	-0.149	1.480	0.066	-0.161	1.883	0.022	-0.064	1.336	0.053	-0.129
25	8	2.164	0.059	-0.156	2.164	0.059	-0.156	2.170	0.064	-0.165	2.576	0.027	-0.069	1.759	0.065	-0.167
25	10	2.386	0.071	-0.152	2.386	0.071	-0.152	2.388	0.074	-0.155	2.801	0.024	-0.060	1.887	0.086	-0.175
25	12	2.576	0.069	-0.155	2.576	0.069	-0.155	2.579	0.069	-0.158	2.985	0.028	-0.062	2.024	0.085	-0.194
25	20	3.066	0.067	-0.151	3.066	0.067	-0.151	3.065	0.067	-0.150	3.481	0.030	-0.058	2.332	0.092	-0.218
50	2	0.805	0.066	-0.144	0.805	0.066	-0.144	0.807	0.066	-0.148	1.156	0.034	-0.075	0.960	0.043	-0.096
50	4	1.502	0.059	-0.152	1.502	0.059	-0.152	1.507	0.062	-0.159	1.894	0.024	-0.062	1.343	0.053	-0.136
50	8	2.188	0.053	-0.142	2.188	0.053	-0.142	2.191	0.055	-0.146	2.582	0.024	-0.055	1.762	0.058	-0.165
50	10	2.404	0.061	-0.141	2.404	0.061	-0.141	2.405	0.062	-0.142	2.804	0.024	-0.058	1.890	0.073	-0.169
50	12	2.599	0.068	-0.148	2.599	0.068	-0.148	2.600	0.068	-0.149	2.999	0.032	-0.061	2.017	0.083	-0.190
50	20	3.114	0.060	-0.156	3.114	0.060	-0.156	3.113	0.060	-0.155	3.507	0.023	-0.065	2.358	0.089	-0.223
100	2	0.824	0.064	-0.156	0.824	0.064	-0.156	0.827	0.065	-0.159	1.170	0.033	-0.081	0.964	0.042	-0.103
100	4	1.506	0.062	-0.154	1.506	0.062	-0.154	1.509	0.063	-0.157	1.894	0.026	-0.062	1.336	0.055	-0.138
100	8	2.209	0.060	-0.151	2.209	0.060	-0.151	2.211	0.059	-0.153	2.598	0.025	-0.064	1.762	0.067	-0.170
100	10	2.425	0.065	-0.150	2.425	0.065	-0.150	2.425	0.066	-0.151	2.814	0.027	-0.059	1.899	0.078	-0.184
100	12	2.608	0.069	-0.147	2.608	0.069	-0.147	2.608	0.069	-0.147	3.005	0.030	-0.065	2.003	0.087	-0.181
100	20	3.121	0.063	-0.148	3.121	0.063	-0.148	3.121	0.063	-0.148	3.514	0.026	-0.064	2.345	0.094	-0.212
500	2	0.831	0.062	-0.148	0.831	0.062	-0.148	0.831	0.062	-0.148	1.172	0.032	-0.074	0.961	0.041	-0.099
500	4	1.520	0.061	-0.147	1.520	0.061	-0.147	1.521	0.061	-0.149	1.905	0.025	-0.063	1.335	0.055	-0.131
500	8	2.217	0.062	-0.151	2.217	0.062	-0.151	2.217	0.062	-0.151	2.600	0.025	-0.064	1.760	0.071	-0.170
500	10	2.439	0.064	-0.152	2.439	0.064	-0.152	2.438	0.064	-0.152	2.823	0.026	-0.065	1.897	0.077	-0.182
500	12	2.622	0.064	-0.150	2.622	0.064	-0.150	2.622	0.064	-0.150	3.007	0.027	-0.064	2.014	0.080	-0.190
500	20	3.133	0.063	-0.153	3.133	0.063	-0.153	3.133	0.063	-0.153	3.518	0.027	-0.065	2.346	0.090	-0.220
1000	2	0.831	0.062	-0.150	0.831	0.062	-0.150	0.831	0.062	-0.150	1.173	0.030	-0.075	0.961	0.042	-0.100
1000	4	1.523	0.063	-0.149	1.523	0.063	-0.149	1.523	0.064	-0.149	1.905	0.028	-0.063	1.337	0.055	-0.133
1000	8	2.217	0.062	-0.150	2.217	0.062	-0.150	2.217	0.062	-0.150	2.602	0.026	-0.064	1.756	0.070	-0.168
1000	10	2.442	0.062	-0.150	2.442	0.062	-0.150	2.442	0.062	-0.150	2.825	0.027	-0.064	1.900	0.075	-0.180
1000	12	2.623	0.063	-0.148	2.623	0.063	-0.148	2.623	0.063	-0.148	3.006	0.027	-0.063	2.016	0.079	-0.188
1000	20	3.132	0.063	-0.149	3.132	0.063	-0.149	3.132	0.063	-0.149	3.518	0.026	-0.064	2.344	0.092	-0.215

Figure 4 and 5 present the boxplots of the mean bias and the relative error on the two slopes. These figures show the best behavior of the Negative binomial, Quasi-Poisson and Poisson models for the first slope. These three models possess the lower median values of bias and relative error for all combinations of n and k . Though LM presents the lowest median values of bias and relative error for all combinations of n and k for the second slope, the dispersion around the median values is largely more pronounced.

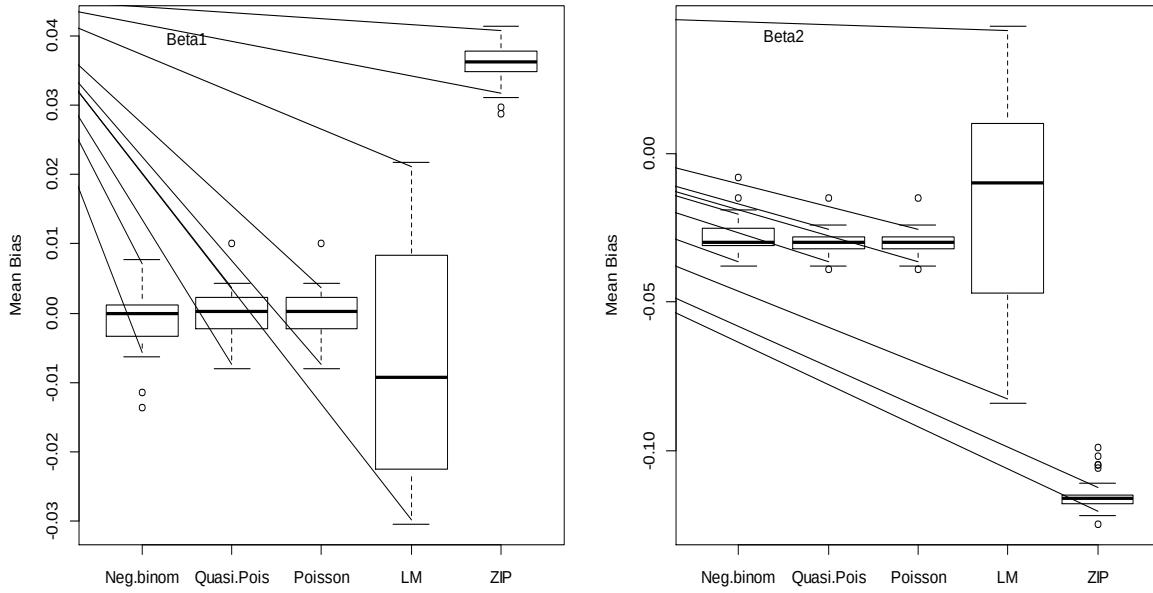


Figure 4. Boxplot of mean bias for Poisson model and its extensions in case of two Covariates

Like in case of one covariate, it is observed that ZIP is the worst performer since they yielded large bias and Relative error as shown in Figure 4 and Figure 5 respectively.

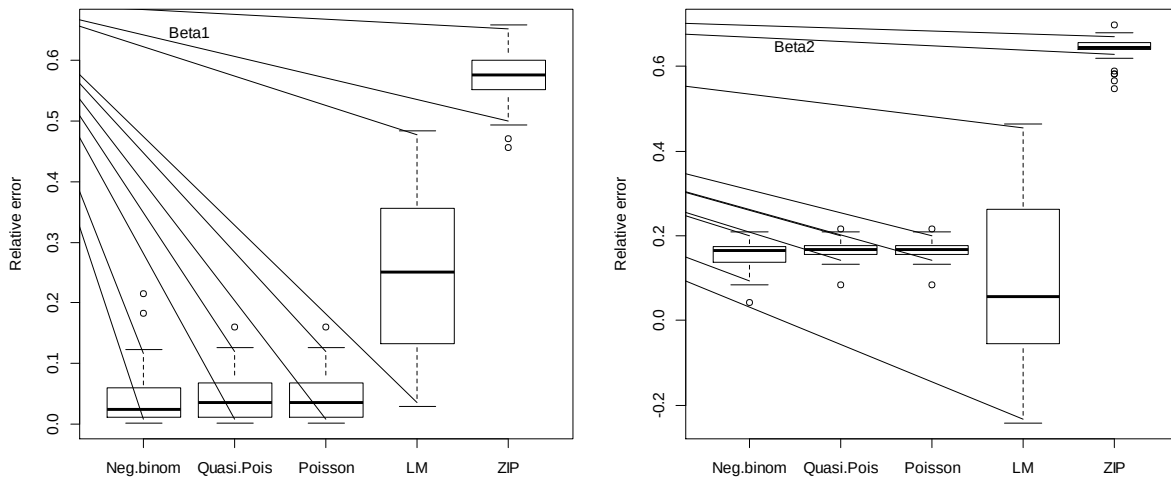


Figure 5. Boxplot of relative errors for Poisson model and its extensions in case of two covariates

Figure 6 shows how the bias relative to the first slope moves with the k at each sample size. These results are similar to those found in case of one covariate. It appears that the mean bias is close to zero for Poisson, Quasi-Poisson and Negative binomial models. ZIP shows the mean bias values between 0.02 and 0.06. As far as the LM is concerned, for small k ($k \leq 5$), the mean bias value is close to zero but when k increases, the mean bias absolute values become important (around to -0.06 at $k=20$).

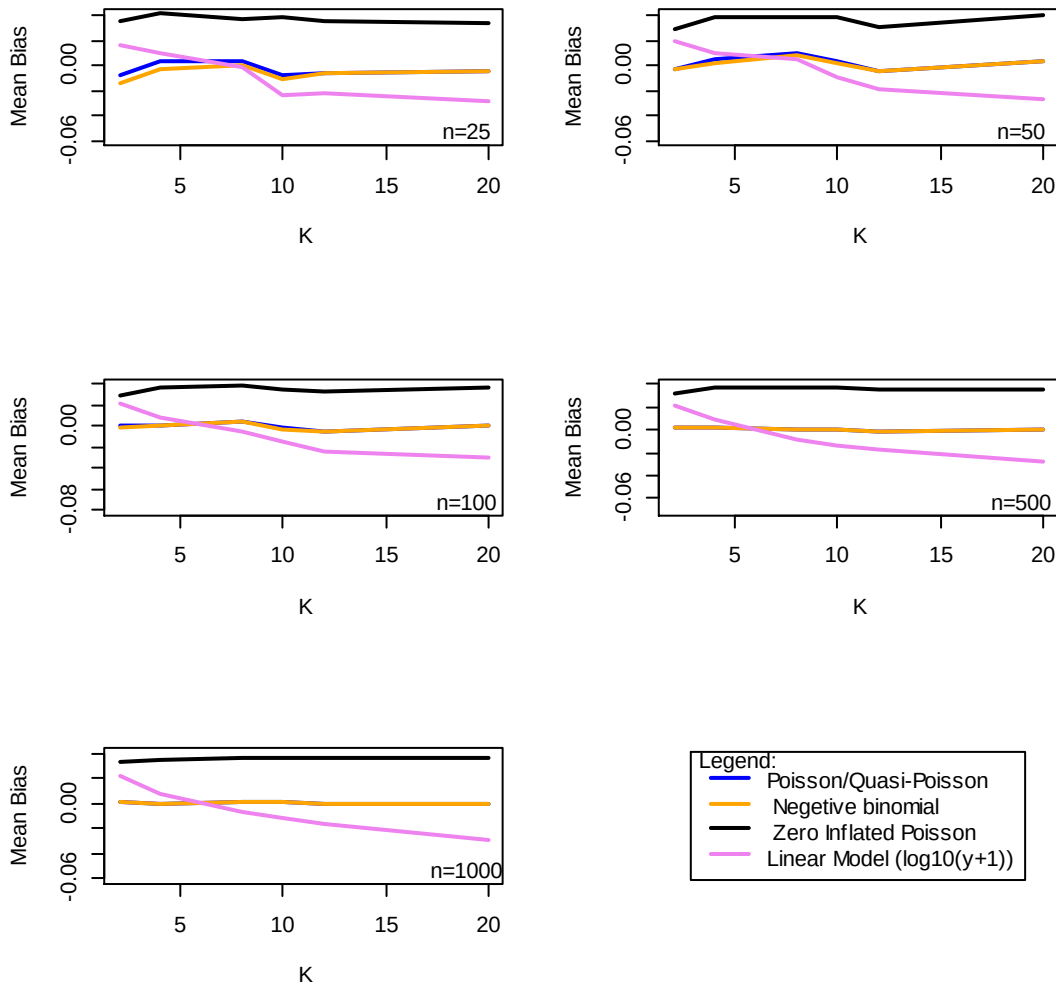


Figure 6. Plot of mean bias against sample size for all models (case of slope 1)

Figure 7 shows how the bias relative to the second slope moves according to k values at each sample size. It appears that the mean bias is close to zero for Poisson, Quasi-Poisson and Negative binomial models. ZIP shows the mean bias values between - 0.10 and - 0.15. As far as the LM is concerned, the mean bias values decrease with increase in k .

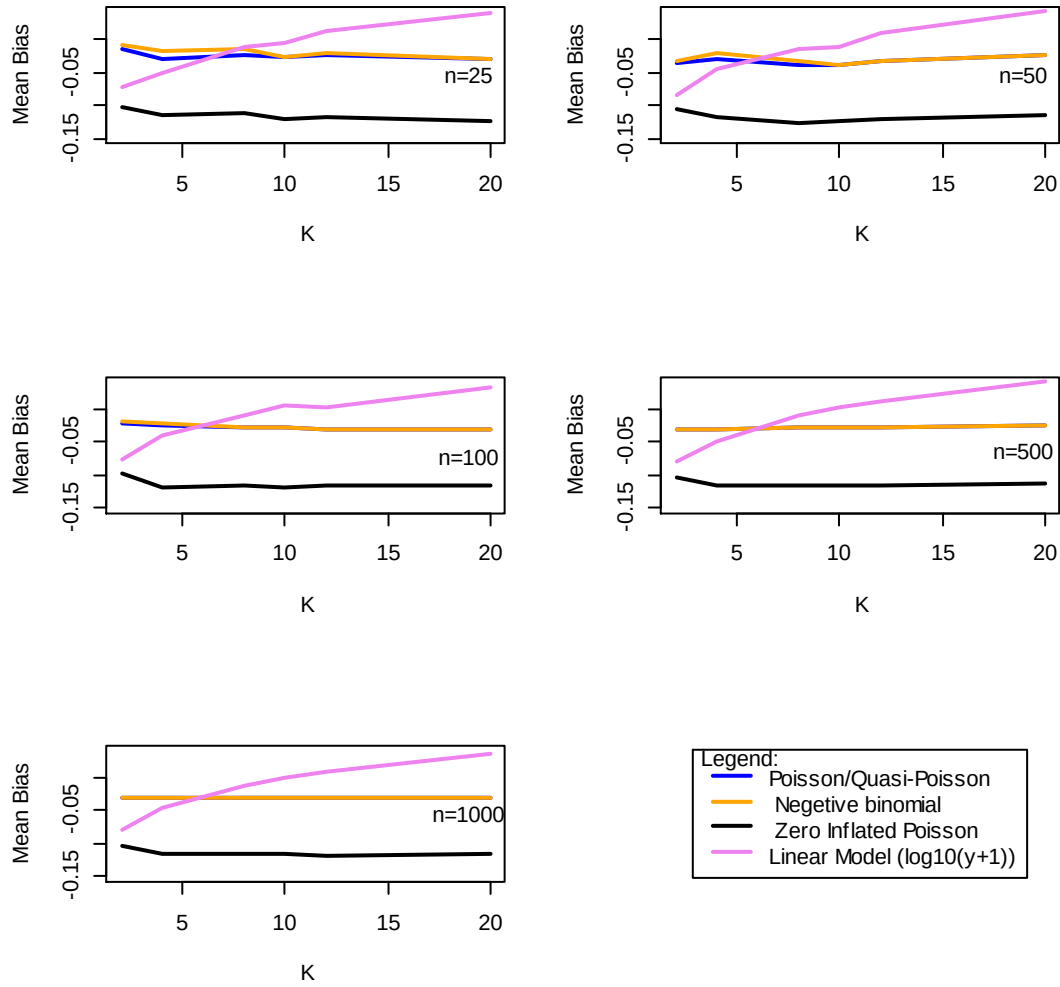


Figure 7. Plot of mean bias against sample size for all models (case of slope 2)

The mean rank of each model for each combination of k and n is presented in Table 4. From the outcomes, it appears that when k is equal to 2, LM is the best model (rank 1) at sample size $n=25$ and $n=50$ but also when k is equal to 4 and $n=50$. For the samples $n=25$ and $n=50$, ZIP is the best performer model (rank 1) at sample for the big values of k while Poisson and Quasi-Poisson are relatively well ranged when the samples are biggest ($n=500, 1000$). Moreover, from the sample size $n=100$ to $n=1000$, the Negative binomial appears to be the best model. It is also noticed that Poisson and Quasi-Poisson have the same performance for each combination of n and k . These results are very close to those obtained in the case of one covariate.

Table 4. Median ranks of Poisson model and its extensions according to the RMSE values (case of 2 covariates)

n	k	Poisson	Quasi-Poisson	Negative binomial	ZIP	LM
25	2	2	2	5	3	1
25	4	3	3	5	1	2
25	8	2	2	4	1	5
25	10	2	2	4	1	5
25	12	2	2	4	1	5
25	20	2	2	3	1	5
50	2	2	2	5	4	1
50	4	2	2	5	2	1
50	8	2	2	4	2	5
50	10	2	2	4	2	5
50	12	2	2	4	1	5
50	20	3	3	2	1	5
100	2	3	3	1	4	2
100	4	3	3	1	3	2
100	8	2	2	1	2	5
100	10	2	2	1	2	5
100	12	2	2	1	2	5
100	20	2	2	1	2	5
500	2	2	2	1	3	3
500	4	2	2	1	5	3
500	8	2	2	1	5	4
500	10	2	2	1	5	4
500	12	2	2	1	5	4
500	20	2	2	1	5	4
1000	2	2	2	1	5	3
1000	4	2	2	1	5	4
1000	8	2	2	1	5	4
1000	10	2	2	1	5	4
1000	12	2	2	1	5	4
1000	20	2	2	1	5	4

4.1.3. Relative efficiency of the zero inflated models with 1 covariate

Table 5 presents the estimated values of Poisson, Negative binomial, ZIP and ZINB model for the combinations of n and p. Generally, estimated values for the slope are close to the true value (0.063) for the high sample size (n=500, 1000). Moreover, for each value of n, the intercept estimated values relatively increase in absolute term when p increases.

Table 5. Estimated values for Poisson, Negative binomial, ZIP and ZINB models ($\beta_0 = 0.14, \beta_1 = 0.063$).

		Poisson		Negative binomial		ZIP		ZINB	
n	p	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
25	0.2	-0.140	0.046	-0.140	0.047	0.096	0.043	0.076	0.044
25	0.4	-0.471	0.068	-0.473	0.068	-0.013	0.058	-0.045	0.068
25	0.6	-0.945	0.023	-0.956	0.024	-0.378	-0.264	-0.396	-0.268
25	0.8	-1.507	0.022	-1.541	-0.015	-0.566	-0.314	-0.846	-0.328
50	0.2	-0.117	0.057	-0.117	0.056	0.104	0.043	0.073	0.044
50	0.4	-0.413	0.068	-0.414	0.069	0.063	0.074	0.015	0.073
50	0.6	-0.862	0.083	-0.863	0.089	-0.017	0.091	-0.113	0.091
50	0.8	-1.595	0.073	-1.603	0.081	-0.476	0.609	-0.601	0.599
100	0.2	-0.102	0.060	-0.102	0.060	0.109	0.063	0.076	0.063
100	0.4	-0.389	0.067	-0.389	0.067	0.101	0.070	0.046	0.071
100	0.6	-0.817	0.058	-0.817	0.059	0.072	0.061	-0.025	0.059
100	0.8	-1.575	0.053	-1.577	0.056	-0.019	0.094	-0.156	0.099
500	0.2	-0.086	0.063	-0.086	0.064	0.133	0.064	0.105	0.065
500	0.4	-0.373	0.063	-0.373	0.063	0.132	0.065	0.099	0.066
500	0.6	-0.785	0.061	-0.785	0.061	0.127	0.060	0.085	0.062
500	0.8	-1.491	0.064	-1.491	0.065	0.109	0.072	0.029	0.072
1000	0.2	-0.085	0.063	-0.085	0.063	0.136	0.064	0.115	0.064
1000	0.4	-0.371	0.064	-0.371	0.064	0.134	0.066	0.109	0.067
1000	0.6	-0.781	0.066	-0.781	0.066	0.134	0.065	0.097	0.066
1000	0.8	-1.481	0.060	-1.481	0.060	0.120	0.063	0.080	0.064

Table 6 shows the Vuong statistic's values comparing ZIP to Poisson and ZINB to Negative Binomial. Results indicate that from a combination ($n=25, p=0.2$) to ($n=100, p=0.4$), ZIP and Poisson models are equivalent while from ($n=25, p=0.2$) to ($n=500, p=0.2$), ZINB and Negative binomial models are equivalent ($V < 1.96$). Furthermore, from the combination ($n=100, p=0.6$), ZIP is preferred to Poisson while from ($n=500, p=0.4$), ZINB is preferred to Negative binomial ($V > 1.96$).

Table 6. Vuong statistic's values for ZIP and ZINB for different combinations of n and p, case of 1 covariate.

n	p	ZIP vs Poisson	ZINB vs Negative binomial
25	0.2	0.535	1.136
25	0.4	1.243	1.059
25	0.6	1.528	1.453
25	0.8	1.659	1.410
50	0.2	0.491	0.792
50	0.4	1.560	0.839
50	0.6	1.735	1.023
50	0.8	1.612	1.493
100	0.2	1.303	1.041
100	0.4	1.675	1.138
100	0.6	2.093	1.148
100	0.8	2.003	1.476
500	0.2	2.242	1.395
500	0.4	3.642	2.134
500	0.6	4.651	1.980
500	0.8	4.096	2.002
1000	0.2	2.736	2.126
1000	0.4	5.080	2.471
1000	0.6	6.290	2.660
1000	0.8	5.749	2.479

The rank of each model for each combination of p and n on the basis of AIC is presented in Table 7. From the outcomes, it appears that when the p is equal to 0.2 Poisson is the best performer model (rank 1) at sample size n=25 and n=50. For the other combinations of n and p, the ZIP is the best model (rank 1) while Poisson is the less performer model.

Table 7. Rank of Poisson, Negative binomial, ZIP and ZINB models according to the AIC values, case of 1 covariate.

n	p	Poisson	Negative binomial	ZIP	ZINB
25	0.2	1	3	2	4
25	0.4	2	3	1	4
25	0.6	4	2	1	3
25	0.8	4	3	1	2
50	0.2	1	3	2	4
50	0.4	4	2	1	3
50	0.6	4	2	1	3
50	0.8	4	2	1	3
100	0.2	3	2	1	4
100	0.4	4	2	1	3
100	0.6	4	3	1	2
100	0.8	4	2	1	3
500	0.2	4	3	1	2
500	0.4	4	3	1	2
500	0.6	4	3	1	2
500	0.8	4	3	1	2
1000	0.2	4	3	1	2
1000	0.4	4	3	1	2
1000	0.6	4	3	1	2
1000	0.8	4	3	1	2

Böhning *et al.* (1999) compared the Poisson and ZIP models under different proportions of zeros. These results are similar to those of this study. ZIP model is superior to other models except for the small samples ($n=25, 50$) and small p ($p=0.2$). However, this is not surprising since data are generated from ZIP. Al Mamun (2014) also found that when the true percentages of zeros vary from 20 % to 60 %, the zero-inflated models almost perfectly fit the data while the classical models never able to do it.

4.1.4. Relative efficiency of the zero inflated models with 2 covariates

Table 8 shows the estimated values of Poisson, Negative binomial, ZIP and ZINB model for the combinations of n and p . From this one notice that, the estimated values for the slopes are close to the true value (0.063, 1.25) for the high sample size ($n=500, 1000$). Moreover, for each value of n , the intercept estimated values relatively increase in absolute term when p increases.

Table 8. Estimated values for Poisson, Negative binomial, ZIP and ZINB models ($\beta_0 = 0.14$, $\beta_1 = 0.063$, $\beta_2 = 1.25$).

n	p	Poisson			Negative binomial			ZIP			ZINB		
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
25	0.2	-0.097	0.064	1.241	-0.104	0.062	1.252	0.129	0.061	1.253	0.127	0.061	1.254
25	0.4	-0.383	0.049	1.212	-0.363	0.049	1.220	0.125	0.060	1.254	0.124	0.060	1.255
25	0.6	-0.788	0.063	1.178	-0.687	0.065	1.159	0.112	0.061	1.261	0.108	0.061	1.263
25	0.8	-1.465	0.065	1.257	-1.325	0.061	1.160	0.135	0.062	1.252	0.135	0.062	1.252
50	0.2	-0.088	0.068	1.239	-0.088	0.064	1.246	0.142	0.062	1.247	0.141	0.062	1.247
50	0.4	-0.386	0.059	1.212	-0.352	0.062	1.210	0.117	0.059	1.258	0.113	0.059	1.260
50	0.6	-0.788	0.063	1.178	-0.687	0.065	1.159	0.112	0.061	1.261	0.108	0.061	1.263
50	0.8	-1.465	0.065	1.256	-1.326	0.061	1.160	0.135	0.062	1.252	0.135	0.062	1.252
100	0.2	-0.076	0.061	1.239	-0.097	0.062	1.255	0.143	0.060	1.248	0.142	0.060	1.248
100	0.4	-0.368	0.049	1.228	-0.336	0.056	1.219	0.134	0.061	1.252	0.132	0.061	1.253
100	0.6	-0.768	0.065	1.200	-0.685	0.059	1.168	0.128	0.068	1.255	0.126	0.068	1.256
100	0.8	-1.464	0.065	1.227	-1.325	0.061	1.160	0.140	0.064	1.252	0.136	0.064	1.252
500	0.2	-0.081	0.063	1.246	-0.082	0.063	1.248	0.139	0.063	1.250	0.139	0.063	1.250
500	0.4	-0.363	0.061	1.239	-0.358	0.060	1.240	0.136	0.063	1.252	0.135	0.063	1.252
500	0.6	-0.771	0.067	1.235	-0.702	0.065	1.200	0.137	0.065	1.250	0.136	0.065	1.251
500	0.8	-1.464	0.065	1.224	-1.335	0.061	1.161	0.136	0.064	1.252	0.138	0.064	1.252
1000	0.2	-0.086	0.064	1.250	-0.087	0.063	1.251	0.138	0.063	1.251	0.138	0.063	1.251
1000	0.4	-0.363	0.062	1.244	-0.366	0.063	1.247	0.137	0.063	1.251	0.137	0.063	1.251
1000	0.6	-0.781	0.060	1.246	-0.731	0.060	1.222	0.138	0.063	1.251	0.137	0.063	1.251
1000	0.8	-1.464	0.065	1.234	-1.364	0.061	1.182	0.138	0.063	1.251	0.137	0.063	1.251

Vuong statistic's values comparing ZIP to Poisson and ZINB to Negative Binomial (Table 9) indicate that the zero inflated models are better than the other models in all combinations of n and p. These results are similar to those obtained from the table 10 based on the rank of the models according to the AIC values. The Table 10 shows that the zero inflated models are the best with ZIP ranked at the first place.

Table 9. Vuong statistic's values for ZIP and ZINB for different combinations of n and p, case of 2 covariates.

n	p	ZIP vs Poisson	ZINB vs Negative binomial
25	0.2	1.751	2.714
25	0.4	2.271	2.922
25	0.6	3.178	3.484
25	0.8	2.745	3.074
50	0.2	1.775	2.975
50	0.4	3.457	3.937
50	0.6	3.677	3.556
50	0.8	3.566	3.313
100	0.2	2.494	5.776
100	0.4	4.400	6.459

100	0.6	4.584	5.243
100	0.8	4.401	3.667
500	0.2	4.702	10.066
500	0.4	7.489	12.471
500	0.6	10.417	9.637
500	0.8	8.264	8.062
1000	0.2	6.212	15.952
1000	0.4	10.143	18.998
1000	0.6	14.176	15.349
1000	0.8	12.236	11.653

Table 10. Rank of Poisson, Negative binomial, ZIP and ZINB models according to the AIC values, case of 2 covariates.

n	p	Poisson	Negative binomial	ZIP	ZINB
25	0.2	4	3	1	2
25	0.4	4	3	1	2
25	0.6	4	3	1	2
25	0.8	4	3	1	2
50	0.2	4	3	1	2
50	0.4	4	3	1	2
50	0.6	4	3	1	2
50	0.8	4	3	1	2
100	0.2	4	3	1	2
100	0.4	4	3	1	2
100	0.6	4	3	1	2
100	0.8	4	3	1	2
500	0.2	4	3	1	2
500	0.4	4	3	1	2
500	0.6	4	3	1	2
500	0.8	4	3	1	2
1000	0.2	4	3	1	2
1000	0.4	4	3	1	2
1000	0.6	4	3	1	2
1000	0.8	4	3	1	2

4.2. Application on wilt disease data

The summary statistics of all variables are given below. It appears that the number of wilted plants per plot is between 0 and 26 (Box 1).

```
> Data=read.table("clipboard",h=T)
> names(Data)
[1] "ZONES"      "CULTIVARS" "count"
> summary(Data)
      ZONES      CULTIVARS      count
Guinean      :15  cayennelisse:15  Min.   : 0.000
Soudanian     :12    local      :15  1st Qu.: 0.000
SoudanoGuinean:18  sugarloaf   :15  Median : 0.000
                                     Mean    : 5.133
                                     3rd Qu.: 8.000
                                     Max.    :26.000
```

By taking a look at the frequency of the number of wilted plants found in each plot, we obtain Figure 8.

```
> hist(count,breaks=45,col="grey",xlim=c(0,23),xlab=expression(paste("Number of plants")))
```

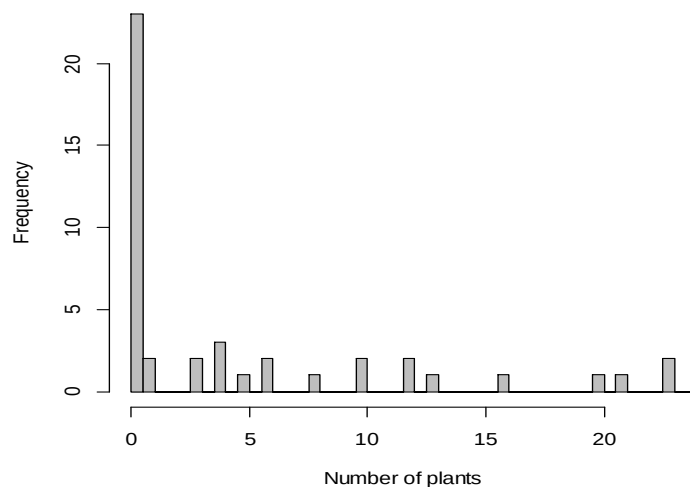


Figure 8. Frequency of number of wilted plants per plot

We can see that there are only positive and integer values and zero has the most important frequency in the data. Given these specificities, the Poisson distribution, seems to be the perfect choice to model this data. Let calculate the mean and variance of the response variable (Box 2).

```
> var(count)
[1] 57.52727
> mean(count)
[1] 5.133333
```

As shown above, the variance of count (57.53) is about eleven times of the mean (5.13). This indicates that there is overdispersion in the data. Further inspection on the data also shows that more than 50 % of the plots do not contain any affected plants, indicating probably zeros are in excess proportion. These data show, in summary, the following traits, the sample size is $n=45$, the proportion of zeros is $p=0.5$ and two covariates have been used. Considering all these traits in the data and according to the results obtain from the simulation study, it appears that the zero inflated models are the best models to fit the data, especially the ZIP model (Table 10).

The GLM is then applied to the data to select the best model. Poisson regression is developed here as a starting point of a count data model, which produces the following outputs using the additive model without interaction.

```
> pois = glm(count~ZONES+CULTIVARS, family="poisson",data=Data)
> summary(pois)

Call:
glm(formula = count ~ ZONES + CULTIVARS, family = "poisson",
    data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6699  -0.8094  -0.0003   0.2264   3.8655

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.07077    0.09144  33.581  < 2e-16 ***
ZONESsoudanian  -20.24374  1309.55887  -0.015    0.988
ZONESsoudanoGuinean -1.11992    0.14632  -7.654 1.95e-14 ***
CULTIVARSlocal    -3.40120    0.45461  -7.482 7.34e-14 ***
CULTIVARSsugarloaf -0.67990    0.14080  -4.829 1.37e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 455.479  on 44  degrees of freedom
Residual deviance:  75.451  on 40  degrees of freedom
AIC: 170.79

Number of Fisher Scoring iterations: 16
```

We test for goodness-of-fit of the model with a chi-square test based on the residual deviance and degrees of freedom. The null hypothesis is H_0 : the model fit well the data (Box 3).

```
> 1 - pchisq(summary(pois)$deviance,summary(pois)$df.residual)
[1] 0.0005913139
```

Poisson model does not fit the data ($p < 0.05$).

We will then use Quasi-Poisson model to fit the data. Quasi-Poisson assumes that the variance is proportional rather than equal to the mean, and estimates the scale parameter k dividing Pearson's chi-squared by its degree of freedom:

```
> qpois = glm(count~ZONES+CULTIVARS, family="quasipoisson",data=Data)
> summary(qpois)
```

Call:

```
glm(formula = count ~ ZONES + CULTIVARS, family = "quasipoisson",
    data = Data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6699	-0.8094	-0.0003	0.2264	3.8655

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.0708	0.1242	24.731	< 2e-16 ***
ZONESSoudanian	-20.2437	1778.1832	-0.011	0.990973
ZONESSoudanoGuinean	-1.1199	0.1987	-5.637	1.53e-06 ***
CULTIVARSlocal	-3.4012	0.6173	-5.510	2.31e-06 ***
CULTIVARssugarloaf	-0.6799	0.1912	-3.556	0.000985 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.843754)

Null deviance: 455.479 on 44 degrees of freedom
 Residual deviance: 75.451 on 40 degrees of freedom
 AIC: NA

Number of Fisher Scoring iterations: 16

We notice that Poisson and Quasi Poisson models give the same estimates but standard error values are greater for Quasi-Poisson than Poisson. In practice, Poisson GLM underestimates the variance in the data.

We can verify this fact easily. First we write a useful function to extract standard errors and the ratio between these values:

```
> se <- function(model) sqrt(diag(vcov(model)))
> round(data.frame(p=coef(pois), q=coef(qpois),se.p=se(pois), se.q=se(qpois), ratio=se(pois)/se(qpois)), 4)
```

	p	q	se.p	se.q	ratio
(Intercept)	3.0708	3.0708	0.0914	0.1242	0.7365
ZONESSoudanian	-20.2437	-20.2437	1309.5589	1778.1832	0.7365
ZONESSoudanoGuinean	-1.1199	-1.1199	0.1463	0.1987	0.7365
CULTIVARSlocal	-3.4012	-3.4012	0.4546	0.6173	0.7365
CULTIVARssugarloaf	-0.6799	-0.6799	0.1408	0.1912	0.7365

We suspect that Quasi-Poisson do not good fit the data. To confirm we will use chi-squared test (Box 4).

```
> 1 - pchisq(summary(qpois)$deviance,summary(qpois)$df.residual)
[1] 0.0005913139
```

Quasi-Poisson model does not fit the data ($p < 0.05$).

We now use negative binomial model. The function “glm.nb” of the MASS library (Ripley *et al.*, 2013) of R 2.15.3 software (R Core Team, 2012) is considered for the analysis.

```
> nb = glm.nb(count~ZONES+CULTIVARS,data=Data)
> summary(nb)

Call:
glm.nb(formula = count ~ ZONES + CULTIVARS, data = Data, init.theta = 4.898697581,
link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.200   -0.652    0.000    0.000    2.013

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.167e+00  1.975e-01  16.036 < 2e-16 ***
ZONESsoudanian -3.728e+01  6.381e+06   0.000  1.00000
ZONESsoudanoGuinean -1.217e+00  2.422e-01  -5.023 5.08e-07 ***
CULTIVARSlocal    -3.478e+00  5.038e-01  -6.902 5.12e-12 ***
CULTIVARSsugarloaf -8.101e-01  2.494e-01  -3.248 0.00116 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(4.8987) family taken to be 1)

Null deviance: 245.393  on 44  degrees of freedom
Residual deviance: 39.999  on 40  degrees of freedom
AIC: 160.22

Number of Fisher Scoring iterations: 1

              Theta:  4.90
             Std. Err.:  2.79

2 x log-likelihood:  -148.224
```

We test for goodness-of-fit of the model with a chi-square test based on the residual deviance and degrees of freedom (Box 5).

```
> 1 - pchisq(summary(nb)$deviance,summary(nb)$df.residual)
[1] 0.4702961
```

We obtain that P-value is 0.47 and is greater than 0.05, so we accept H_0 and we can say that negative binomial model fit well the data.

We can also use variance function to compare Quasi-Poisson and Negative binomial models. Quasi-Poisson and Negative binomial models have different variance functions. One way to check which one may be more appropriate is to create groups based on the linear predictor, compute the mean and variance for each group, and finally plot the mean-variance relationship.

The graph (Figure 9) plots the mean versus the variance and overlays the curves corresponding to Quasi-Poisson model, where the variance is $k\mu$, and the negative binomial model, where the variance is $\mu(1+\mu/k)$.

```
> xb <- predict(nb)
> g <- cut(xb, breaks=quantile(xb,seq(0,45,5)/45))
> m <- tapply(Data$count, g, mean)
> v <- tapply(Data$count, g, var)
> plot(m, v, xlab="Mean", ylab="Variance", main="Mean-Variance Relationship")
> plot(m, v, xlab="Mean", ylab="Variance", xlim=c(0,15),main="Mean-Variance Relationship")
> mtext("wilt disease data",padj=-0.5)
> pr <- residuals(pois,"pearson")
> phi <- sum(pr^2)/df.residual(pois)
> x <- seq(0.63,15,0.25)
> lines(x, x*phi, lty="dashed")
> lines(x, x*(1+x/nb$theta))
> legend("topleft", lty=c("dashed","solid"),legend=c("Q. Poisson","Neg. Binom."), inset=0.05)
`
```

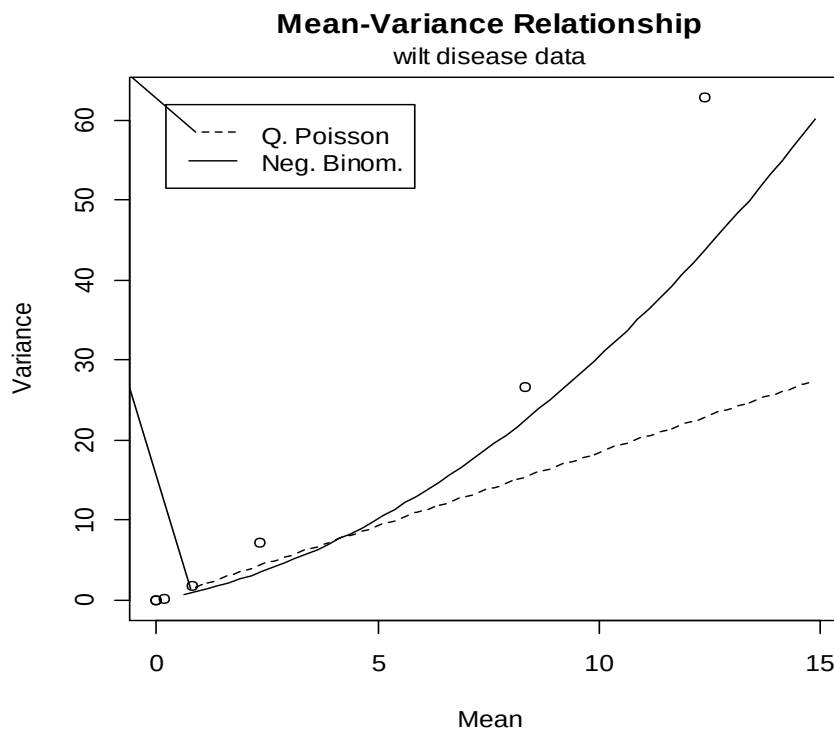


Figure 9 : mean versus variance

Quasi-Poisson variance function does a pretty good job for the bulk of the data, but fails to capture the high variances of the most of the data. Negative binomial variance function is not too different but, being a quadratic, can rise faster and does a better job at the high end. We conclude that the negative binomial model provides a better description of the data than the Quasi-Poisson model.

Since there is an important proportion of zero (more than 50 %) we will use Zero inflated models such as Zero inflated Poisson (ZIP) and Zero inflated Negative binomial (ZINB) to fit the data and compare the result to NB.

These types of models can be fit in R using the `zeroinfl()` function in the `pscl` package.

```
> zip = zeroinfl(count~ZONES+CULTIVARS|1, data = Data)
> summary(zip)

Call:
zeroinfl(formula = count ~ ZONES + CULTIVARS | 1, data = Data)

Pearson residuals:
      Min       1Q   Median       3Q      Max
-1.5799547 -0.5120014 -0.0001684  0.3000062  2.9251251

Count model coefficients (poisson with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.01309    0.09508  31.690 < 2e-16 ***
ZONESsoudanian  -20.24374  2320.55690  -0.009  0.993040
ZONESsoudanoguinean -0.92846    0.15146  -6.130  8.79e-10 ***
CULTIVARSlocal    -3.23272    0.48629  -6.648  2.98e-11 ***
CULTIVARSsugarloaf -0.51382    0.14501  -3.543  0.000395 ***

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8354      0.6539  -2.807  0.00501 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 13
Log-likelihood: -76.36 on 6 Df
```

Zero inflated Negative binomial (ZINB)

```
> zinb = zeroinfl(count~ZONES+CULTIVARS|1, dist="negbin",data = Data)
> summary(zinb)

Call:
zeroinfl(formula = count ~ ZONES + CULTIVARS | 1, data = Data, dist = "negbin")

Pearson residuals:
      Min       1Q   Median       3Q      Max 
-1.3924776 -0.4936973 -0.0001302  0.2401948  2.5370851 

Count model coefficients (negbin with log link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.0459    0.1699  17.928 < 2e-16 ***
ZONESsoudanian -20.2437  2282.8285  -0.009  0.99292
ZONESsoudanoGuinean -0.9854    0.2296  -4.292 1.77e-05 ***
CULTIVARSlocal   -3.2788    0.5107  -6.420 1.36e-10 ***
CULTIVARSsugarloaf -0.5673    0.2338  -2.426  0.01526 *
Log(theta)       2.2140    0.6866   3.225  0.00126 **

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9929    0.7749  -2.572  0.0101 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 9.152
Number of iterations in BFGS optimization: 15
Log-likelihood: -73.02 on 7 Df
```

To choose between the Negative binomial and zero inflated models we need to resort to other criteria. A very simple way to compare models with different numbers of parameters is to compute Akaike's Information Criterion (AIC). We compute it for the three models (Box 6).

```
> AIC(nb);AIC(zip);AIC(zinb)
[1] 160.2245
[1] 164.7233
[1] 160.0382
```

On the basis of the AIC values, Zero inflated Negative binomial is the model that best fit the data.

Table 11 shows the summary of the results indicating that the zero inflated models are the best models one the basis of AIC and Log-likelihood. These results are similar to those from simulation study. However, in the simulation study, the best zero inflated model is ZIP while ZINB is the best according to the Table 11. This relative difference can be explained by nature of the covariates and the fact that the simulation data are generated from ZIP (O'Hara and Kotze, 2010; Al Mamun, 2014).

The table reveals also that all coefficients are significant at 5 % level except the coefficient of Sudanian zone. The five models indicate relatively similar values of the coefficients. The number of affected plants per plot is rare in Sudian zone ($\exp(-20.24) = 2.061154e-09$) compared to the other zones. This outcome can be explained by the fact the Sudanian zone is not an optimal zone of production of pineapple like the other zones where the disease is more present (Tossou *et al.*, 2015). Moreover, local cultivar shows to be more tolerant with wilt disease. Similar report has been done by Medina and García (2005) who found the local cultivar has high tolerance to stress, plagues and pathogens, better than “Smooth Cayenne” cultivar which is very susceptible to fading.

Table 11 . Comparison of the models: Poisson, Quasi-Poisson, NB, ZIP and ZINB.

Variables	Poisson		Quasi-Poisson		Negative binomial		Zero inflated Poisson		Zero inflated Negative binomial	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
intercept	3.071***	0.091	3.071***	0.124	3.167***	0.198	3.013***	0.095	3.046***	0.170
Soudanian	-20.244	1309.559	-20.244	1778.183	-37.280	6.381e+06	-20.244	2320.557	-20.244	2282.829
SoudanoGuinean	-1.120***	0.146	-1.120***	0.199	-1.217***	0.242	-0.928***	0.151	-0.985***	0.230
Local	-3.401***	0.455	-3.401***	0.617	-3.478***	0.504	-3.233***	0.486	-3.279***	0.511
"sugarloaf"	-0.680***	0.141	-0.680***	0.191	-0.810**	0.249	-0.514***	0.145	-0.567*	0.234
Log(theta)									2.214**	0.687
Log-likelihood			-		-148.224		-76.360		-73.020	
AIC	170.790		-		160.220		164.723		160.030	

Chapter 5:
Conclusions

5. Conclusions and perspectives for Future Research

Modeling count variables is common task in ecology. The classical Poisson regression model for this kind of data is often of limited use because empirical count data sets in ecology typically exhibit overdispersion and/or excess number of zeros. Through Monte Carlo simulation approach, this study showed that even when the population data are generated from Poisson distribution, ZIP fits well the data for the small samples. In contrary, when the samples increase, the Negative binomial is the best approach to fit the data. Moreover, for the data generated from ZIP, the Poisson fit well the data for small samples and small proportion of excess zero. When the samples are large, ZIP is the best model to fit the data. As far as the case study is concerned, Zero Inflated Negative Binomial (ZINB) regression performed better the number of wilted plants within pineapple cultivars in Benin. We suggest that count data should not be transformed, but instead, Poisson model and its extensions should be used. The outcomes of this study provide ample suggestions for future research. It is suggested that future research consider other means for Poisson generated populations, other overdispersion parameters and other proportions of zeros. The research should also be extended to incorporate random effects and more covariates. We suggest that these techniques can also be extended to other real application than ecological application such as credit risk, quality control, etc.

References

- Agresti A. 1990. Categorical Data Analysis. John Wiley & Sons, New York, 558 p.
- Agresti A. 1996. An introduction to categorical data analysis. Wiley, New York.
- Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N. and Csàki, F. (eds): Second international symposium on inference theory, Budapest, Akadémiai Kiadó, pp. 267-281.
- Al Mamun, A. 2014. Zero-inflated regression models for count data: an application to under-5 deaths, master of science, Ball State University, Muncie, Indiana.
- Allain E. et Brenac T. 2001. Modèles linéaires généralisés appliqués à l'étude des nombres d'accidents sur des sites routiers : le modèle de Poisson et ses extensions. *Recherche Transports Sécurité* 72, 3-18
- Assogbadjo A. E., Kyndt T., Sinsin B., Gheysen G. and Van Damme P. 2006. Patterns of genetic and morphometric diversity in baobab (*Adansonia digitata*) populations across different climatic zones of Benin (West Africa). *Ann Bot* 97:819–830
- Böhning D., Dietz E., Schlattmann P., Mendonça, L. and Kirchner, P. 1999. The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Association, Series A*, 162, 195-209.
- Cameron AC, and Trivedi PK 1990. Regression-based tests for overdispersion in the Poisson model. *J Econom* 46(3):347–364.
- Cameron, A.C., and Trivedi, P.K. 2005. *Microeconometrics. Methods and Application*. Cambridge University Press.
- Czado C. and Sikora I. 2002. Quantifying overdispersion effects in count regression data. Sonderforschungsbereich 386, paper 289 (200). <http://epub.ub.uni-muenchen.de/>
- De La Cruz Medina J. and García H.S. 2005. PINEAPPLE: Post-harvest Operations. Report, Instituto Tecnológico de Veracruz (<http://www.itver.edu.mx>).
- Dobson A. J. 2002. An introduction to generalized linear models. Second edition 2002. Chapman & Hall/CRC, 221 p.
- Hall, D. B. (2000). Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics* 56, 1030–1039.
- Harrison X. A. 2014. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* 2:e616; DOI 10.7717/peerj.616

- Hilbe J. M. 2011. Negative binomial regression. 2nd edition. Cambridge: Cambridge University Press.
- Jiao Y., Chen Y., Schneider D. and Wroblewski J. 2004. A simulation study of impacts of error structure on modeling stock-recruitment data using generalized linear models. *Canadian Journal of Fisheries and Aquatic Sciences*, 61, 122–133.
- Kumara, S.S.P. and Chin, H.C. 2004. Study of fatal traffic accidents in Asia Pacific countries. *Transportation Research Record*. No.1897. 43-47.
- Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34:1–14.
- Mooney, C. Z. 1997. *Monte Carlo Simulation*. Thousand Oaks, CA: Sage.
- McCullagh P. and Nelder J.A. 1989. Generalized Linear Models, 2nd edn. Chapman&Hall, London.
- Ver Hoef, J.M. & Boveng, P.L. (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88, 2766–2772.
- Miaou, S.P., and Lum, H. 1993. Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis and Prevention* 25(6): 689-709.
- Miller J. M. 2007. Comparing Poisson, hurdle, and zip model fit under varying degrees of skew and zero-inflation. A dissertation presented to the graduate school of the University of Florida in partial fulfillment of the requirements for the degree of doctor of philosophy, USA
- Mouatassim Y. and Ezzahid E., 2012. Poisson regression and Zero-inflated Poisson regression: application to private health insurance data. *Eur. Actuar. J.* DOI 10.1007/s13385-012-0056-2.
- Mullahy, J. 1986. Specification and testing of some modified count data models. *Journal of Econometrics*, 33, 341-365.
- Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. 1996. *Applied Linear Statistical Models, Fourth Edition*, Irwin, Chicago.
- O'Hara RB, Kotze DJ. 2010. Do not log-transform count data. *Methods in Ecology and Evolution* 1:118–122 DOI 10.1111/j.2041-210X.2010.00021.x.
- Olsson U. 2002. Generalized Linear Models: An Applied Approach. Printed in Sweden, Studentlitteratur, Lund Web-address: www.studentlitteratur.se, 243 p.
- Paxton P., Curran P. J., Bollen K. A, Kirby J, and Chen, F. 2001. *Monte Carlo experiments: Design and implementation. Structural Equation Modeling*, 8, 287-312.

- R Core Team, 2012. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>
- Ripley B., Venables B., Hornik K., Gebhardt A. and Fith D., 2013. Package MASS: support functions and datasets for Venables and Ripley's MASS. R package version 7.3-26. Available: <http://cran.rproject.org/web/packages/MASS/index.html>.
- Shankar, V.N, Mannering. F. and Barfield, W. 1995. Effect of Roadway Geometric and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention*. 27(3):371-389.
- Sileshi, G., Hailu, G. & Nyadzi, G.I. 2009. Traditional occupancy-abundance models are inadequate for zero-inflated ecological count data. *Ecological Modelling*, 220, 1764–1775.
- Tossou C. C., Capo-Chichi D. B. E. et Yedomonhan H. 2015. Diversité et caractérisation morphologique des variétés d'ananas (*Ananas comosus* (L.) Merrill) cultivées au Bénin. *Journal of Applied Biosciences* 87:8113– 8120.
- Tsoumanis A. 2010. Regression analysis of mortality with respect to seasonal influenza in Sweden 1993-2010. Master Thesis in Mathematical Statistics, Stockholm University, Sweden.
- Ver Hoef J. M. and Boveng P. L. 2007, "quasi-Poisson vs. Negative binomial regression: how should we model overdispersed count data?" *Ecology*, 88(11), pp. 2766–2772.
- Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61:439–447.
- Yaacob W. F. W. Mohamad A. L. Yap B. W. 2010. A Practical Approach in Modelling Count Data. *Proceedings of the Regional Conference on Statistical Sciences 2010 (RCSS10). June 2010, 176-183*.
- Yesilova A., Kaydan M. B. and Kaya Y. 2012. Modeling insect-egg data with excess zeros using zero-inflated regression models. *Hacetatepe Journal of Mathematics and Statistics*, 39 (2), 273 – 282.
- Zorn, C. J. W. 1996. Evaluating zero-inflated and hurdle Poisson specifications. Midwest Political Science Association, 1-16.
- Zuur A. F, Ieno E. N., Walker N. J, Saveliev A. A and Smith G. M. 2009. Mixed effects models and extensions in ecology with R. New York: Springer.

Appendix

ZONES	CULTIVARS	Number of wilted plants
Guinean	cayennelisse	20
Guinean	sugarloaf	8
Guinean	local	3
Guinean	cayennelisse	21
Guinean	sugarloaf	12
Guinean	local	1
Guinean	cayennelisse	23
Guinean	sugarloaf	26
Guinean	local	0
Guinean	cayennelisse	23
Guinean	sugarloaf	6
Guinean	local	0
Guinean	cayennelisse	13
Guinean	sugarloaf	10
Guinean	local	0
SoudanoGuinean	cayennelisse	12
SoudanoGuinean	sugarloaf	4
SoudanoGuinean	local	0
SoudanoGuinean	cayennelisse	16
SoudanoGuinean	sugarloaf	6
SoudanoGuinean	local	1
SoudanoGuinean	cayennelisse	10
SoudanoGuinean	sugarloaf	4
SoudanoGuinean	local	0
SoudanoGuinean	cayennelisse	3
SoudanoGuinean	sugarloaf	0
SoudanoGuinean	local	0
SoudanoGuinean	cayennelisse	5
SoudanoGuinean	sugarloaf	0
SoudanoGuinean	local	0
SoudanoGuinean	cayennelisse	4
SoudanoGuinean	sugarloaf	0
SoudanoGuinean	local	0
Soudanian	cayennelisse	0
Soudanian	sugarloaf	0
Soudanian	local	0
Soudanian	cayennelisse	0
Soudanian	sugarloaf	0
Soudanian	local	0
Soudanian	cayennelisse	0
Soudanian	sugarloaf	0

Soudanian	local	0
Soudanian	cayennelisse	0
Soudanian	sugarloaf	0
Soudanian	local	0