DESIGNING AN OFFSET POISSON-GAMMA MIXTURE REGRESSION MODEL

[1]Mbe, E. Nja[*], [2]E.C. Nduka and [3]P.U. Ogoke

[1] Department of Mathematics, Federal University Lafia, Nasarawa State, Nigeria
[2] Department of Mathematics & Statistics, University of Port-Harcourt, Rivers State, Nigeria
[3] Department of Mathematics & Statistics, University of Port-Harcourt, Rivers State, Nigeria
[*]Corresponding E-mail: mbe_nja@yahoo.com

Date Manuscript Received: 04/03/15 Accepted: 18/09/15 Published: December 2015

ABSTRACT

The problem of over-dispersion often encountered by Poisson distribution is what gave rise to the development of the Poisson-Gamma mixture distribution. This distribution is a mixture of a family of Poisson distributions with Gamma mixing weights. The log likelihood function of the Poisson-Gamma mixture distribution has been exponentiated in order to extract the canonical link of the mixture distribution. This link function is re-expressed as a sum of the loglinear link and an offset. By implication, modeling for the Poisson-Gamma mixture model amounts to modeling for the Poisson loglinear model and adding the offset term to the result. This model can be used to model HIV/AIDS infection rate for multiple sex partners. It is theoretically shown that the mixture distribution reduces to the Poisson distribution and that in large samples, the Poisson-Gamma mixture is approximated by the Poisson distribution.

KEYWORDS: *Over-dispersion, Poisson distribution, Gamma mixing weights, Loglinear*

INTRODUCTION

The need to use a Poisson-Gamma mixture distribution arises when the Poisson distribution exhibits over-dispersion. For the Poisson model, the mean and the variance are equal. If the variance of a Poisson model exceeds its mean, over-dispersion is said to occur. Over-dispersion is indicated if Pearson dispersion is greater than 1 Hibe, (2007). The problem with over-dispersion, common to most Poisson models is that it renders the parameter estimates biased. Where over-dispersion is a concern, the alternative models are:

(1) Quasi-Poisson

(2) Negative Binomial regression

MATERIALS AND METHODS

In this paper, we shall consider only the Negative Binomial regression as an alternative to the over-dispersed Poisson model. In designing an offset Poisson-Gamma mixture model, it is logical to use the order listed below:

*(i)*Poisson distribution

*(ii)*Gamma distribution

*(iii)*Poisson-Gamma distribution (Negative Binomial distribution)

(iv) Offset Poisson-Gamma mixture regression model

RESULTS AND DISCUSSION

THE POISSON REGRESSION

The Poisson regression is derived from the poisson probability mass function given Johnson, (2004), Freund, (1992) as

$P(X = x) = \lambda^x e^{-x} / x!$ and (Hilbe 2007) as

$$f(x_i\,;\,\lambda_i) = e^{-t_i\lambda_i}(t_i\lambda_i)^{y_i} / x_i!\ ,$$

$$x = 0, 1, 2, \ldots, \lambda > 0 \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

where x_i is the count response, λ_i is the predicted count or rate parameter, t_i is the area or time in which counts enter the model

(1) reduces to

$$f(x_i, \lambda_i) = e^{-\lambda_i}(\lambda_i)^{x_i} / x_i!$$

where λ_i applies to individual counts devoid of time or size considerations, t_i is set equal to 1. When time or size or space is a concern, $t_i > 1$ and the resulting model is call offset Poisson model.

$$f = P(X = x) = \lambda^x e^{-x} / x_i! \ \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

is the probability of observing any specific count *y*

The Likelihood function is

$$L(\mu_i\,, y_i) = \sum_{i=1}^{n}\{y_i \ln(\mu_i) - \mu_i - \ln(y_i!)\} \ldots\ldots\ldots\ldots (4)$$

where μ_i is the predicted count.

(3) is used when the Poisson model is estimated by a Generalized Linear Model.

When estimation employs a full maximum likelihood update, we use

$$\mu_i = \exp(x_i\beta)$$

and equation (4) becomes Hilbe, (2007)

$$L(\beta_i; y_i) = \sum\{y_i(x_i\beta) - \exp(x_i\beta) - \ln(y_i!)\} \ldots\ldots\ldots\ldots (5)$$

From *(iii)*

$$explogf = exp\{-\mu + xlog\mu - logx!\}$$

Thus the canonical link is

$$\theta = \log(\mu) = X'\beta \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Equation (6) is called the log-linear model.

If time or space or size is a concern and we use a rate as $\frac{\mu}{t}$ instead of μ, (6) becomes

$$\log(\mu/t) = X'\beta - \log(t) \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

where log *(t)* is called an offset. Equation (7) is then termed an offset Poisson regression model.

GAMMA DISTRIBUTION

This distribution models the waiting time between Poisson distributed events. The probability of waiting time until the *nth* poisson (λ) event is a gamma probability density expressed as

$$P(x) = \lambda(\lambda x)^{n-1} / (n-1)!\, e^{-\lambda x}$$

where λ is the rate at which time changes.

If $X \sim$ Gamma $(K; \theta)$ the pdf becomes

$$P(x) = x^{k-1}e^{-x/\theta} / \Gamma(x)\theta^k$$

for $x > 0,\ k,\ \theta > 0$

k = Shape parameter = no of occurrences of an event

θ = Scale parameter

By letting $\alpha = k$ (shape parameter) and $\beta = 1/\theta$ (rate parameter)

The pdf when $X \sim \Gamma(\alpha, \beta)$ becomes $(x, \alpha, \beta) = \beta^\alpha X^{\alpha-1}e^{-X\beta} / \Gamma(x)$ for $x \geq 0$ & $\alpha, \beta > 0$

$= X^{\alpha-1}e^{-X/\beta} / \beta^\alpha\Gamma(\alpha)$, for $x > 0,\ \alpha > 0,\ \beta > 0$

If $\beta = 1/\alpha$

α = Poisson rate parameter

$g(x, \alpha, \beta) = 0$ elsewhere

$$g(x, \alpha, \beta) = logX^{\alpha-1}e^{-\alpha/\beta} - log\beta^\alpha \int_0^\alpha y^{\alpha-1}e^{-y}dy$$

$$g(x, \alpha, \beta) = exp\{logf(x, \alpha, \beta)\}$$

$$= exp\left\{-1/\beta\, X - \alpha log\beta + log\Gamma(\alpha) + (\alpha - 1)logx\right\}$$

The canonical link

$$\theta = {}^{-1}/_{\beta}$$

$$-{}^{1}/_{\beta} = E(X) = X'\beta$$

$$\text{But } \beta = {}^{1}/_{\alpha}$$

$$\Rightarrow -\frac{1}{\beta} = -\alpha$$

$$\text{but} - {}^{1}/_{\beta} = \alpha'\beta$$

$$-\alpha = \alpha'\beta$$

Thus Gamma regression for rates yields an offset Gamma regression as

$$\alpha/_{t} = \alpha'\beta$$

$$-\alpha = t\alpha'\beta$$

t = time or size or shape.

POISSON-GAMMA MIXTURE DISTRIBUTION

The variance of the Poisson-Gamma mixture distribution differs from that of Poisson by $\mu^2 k$, the extra component arising from mixing the Poisson distribution with a gamma distribution. Thus this variance is given as

$$var(X) = \mu + \mu^2 k$$

The mean, $E(X) = \mu$ (same as poisson)

The distribution is also termed, Negative Binomial. This is because the negative binomial distribution is a mixture of a family of poisson distributions with gamma mixing weights. In this case, the poisson parameter is a random variable distributed as gamma. The negative binomial distribution with r number of failures, k number of successes and having the probability of success p is given as

$$f(k;r;p) = P(X = x) = \binom{k-1}{k-r}(1-p)^r p^{k-r}, \quad k = r, r+1, \ldots$$

By letting $P \to 0$ or $r \to \infty$, the Negative Binomial tends to the poisson distribution. where the stopping parameter $r \to 0$ in a Negative Binomial distribution and the probability of success in each trial $p \to 0$, the mean parameter λ, is kept constant.

The parameter p becomes

$$P = {}^{\lambda}/_{\lambda + r}, \quad \Rightarrow \lambda = {}^{rp}/_{1-p}$$

Under this parameterization,

$$\text{The } pmf = f(k;r;p) = {}^{\Gamma(K+V)}/_{K!\,\Gamma(r)}\, P^K (1-P)^r$$

$$= {}^{\lambda^K}/_{K!}\, {}^{\Gamma(r+K)}/_{\Gamma(r)(r+\lambda)^K} \cdot {}^{1}/_{(1+\lambda/r)^r} \quad \ldots\ldots\ldots\ldots\ldots (9)$$

Where k is the number of successes in a sequence of independently and identically distributed Bernoulli trials. r is the number of failures.

$$\lim_{r\to\infty} f(K;r;p) = {}^{\lambda^K}/_{K!}\, e^{-\lambda}$$

λ = expected value or mean

Thus the poisson distribution is a limiting Negative Binomial distribution. The Poisson-Gamma mixture pmf, $f(k;r;p)$ can be shown to be equal to that of the negative binomial as follows:

$$f(k;r;p) = \int_0^\infty f_{poisson(\lambda)}(K) \cdot f_{Gamma(r, {}^{1-p}/_{p})}(\lambda)\, d\lambda$$

$$= \int_0^\infty \frac{\lambda^K}{K!} e^{-\lambda} \cdot \lambda^{r-1} e^{-\lambda(1-p)/p} \Big/ \left(\tfrac{p}{1-p}\right)^r \Gamma(r)\, d\lambda$$

$$= {}^{(1-p)^r p^{-r}}/_{K!\,\Gamma(r)} \int_0^\infty \lambda^{r+k-1} e^{-\lambda/p}\, d\lambda$$

$$= {}^{(1-p)^r}/_{K!\,\Gamma(r)}\, p^{r+k}\, \Gamma(r+k)$$

$$= {}^{\Gamma(r+k)}/_{K!\,\Gamma(r)}\, p^k (1-p)^r$$

Hence the Negative Binomial distribution is also the poisson-Gamma mixture distribution.

Under the parameterization $k\alpha$, $E(Y) = \mu = k\alpha$

$$var\,(Y) = K\alpha + K\alpha^2 = \mu + {}^{\mu^2}/_{K}$$

The probability mass function $P(Y = y; \alpha; k)$ becomes

$$P(Y = X; \alpha; K) = {}^{(X+K-1)!}/_{X!\,(K-1)!}\, {}^{\alpha^X}/_{(1+\alpha)^{X+K}}, \quad X = 0,1,2,\ldots$$

by which we obtain the log likelihood as

$$l = X log\left\{{}^{\alpha}/_{(1+\alpha)}\right\} - K log(1+\alpha) + log\binom{X-1}{K-1}$$

The canonical link function is

$$\eta = \log({}^{\alpha}/_{1+\alpha}) = log\left({}^{\mu}/_{\mu+k}\right) \ldots\ldots\ldots\ldots\ldots\ldots (10)$$

Thus the generalized linear regression model arising from (10) is

$$log\left({}^{\mu}/_{\mu+k}\right) = \alpha + X'\beta$$

for a fixed k

Using another parameterization with

$$f(x;k;p) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \qquad x = k, k+1, k+2, \ldots$$

where p = probability of success

k = no of successes before rth failure

$$log f = k log\binom{x-1}{k-1} + k log\left({}^{p}/_{1-p}\right) + x log(1-p)$$

$$exp f = exp\left\{x log(-p) + k log\left({}^{p}/_{1-p}\right) + log\binom{x-1}{k-1}\right\}$$

Canonical link, $g(p) = \log(1-p)$

$$= \log q$$

$$= X\beta$$

The offset model is achieved as

$$log({}^{q}/_{t}) = \log(q) - \log(t) = X\beta$$

$$\Rightarrow \quad \log(q) = X\beta + \log(t)$$

q = probability that a person mates with an infected person

$$t = {}^{1}/_{2} \text{ month}$$

ESTIMATION IN POISSON-GAMMA MIXTURE MODEL

The poisson-Gamma mixture (Negative Binomial) x is a non-negative discrete random variable with probability mass function

$$P(X = x) = \begin{cases} \left({}^{r}/_{r+m}\right)^r {}^{\Gamma(r+k)}/_{k!\,\Gamma(r)} \left({}^{r}/_{r+m}\right)^k & m, r > 0 \\ 0 & elsewhere \end{cases}, X = 0,1,2,\ldots$$

m = *mean* = location parameter
r = dispersion parameter = heterogeneity
The variance $\sigma^2 = m + m^2/r$
when r is known, the Negative binomial distribution with parameter m becomes a member of the exponential family.
$\bar{x}$ = is a minimum variance unbiased estimator for m Al-Khasawneh, (2010).
For the two-parameter poisson-Gamma mixture distribution, both m and r are unknown. This situation is more practical.

ESTIMATION OF THE DISPERSION PARAMETER r

Both the method of the moments and maximum likelihood method individually impose constraints on the estimation of the dispersion parameter. Al-Khasawneh, (2010) combines both the method of moments and maximum quasi-likelihood estimation in a variety of ways using appropriate weights.

Method of moments

This is done by solving simultaneously, the following equations:

$$\hat{m} = \bar{x} \text{ .... (11)}$$
$$s^2 = m + m^2/r \text{ .... (12)}$$

Thus $\hat{r} = \bar{x}^2/(s^2 - \bar{y})$ ....(13)

From (13) it is obvious that if the sample variance s^2 equals the sample mean, $\hat{r}$ cannot be determined.

$$\hat{\mu}_{\dots E} = \bar{X}$$

$\hat{r}_{MLE}$ is solution to

$$n\ln(1 + \bar{x}/r) = n_1(1/r) + n_2(1/r + 1/(r+1)) + n_3(1/r + 1/(r+1) + 1/(r+2)) + \cdots$$

.... (14)

Where n is the size, n_1 is the number of ones in the sample, n_2 is the number of twos in the sample and so on.

Equation (15) can be written as

$\sum\sum 1/(r+t) + n\ln r/(r+\bar{x}) = 0$ ....(15)

Equation (15) is solved by $IWLS$ or Newton-Raphson method to obtain $\hat{r}_{MLE}$.

$\hat{r}_{MLE}$ exist only for over-dispersed samples Levin and Reeds, (1977)

The re-parameterization of $\alpha = 1/r$ was suggested by Piegorsch, (1990) and Anraku and Yanagimoto (1990), where α was estimated by Maximum Likelihood.

Piegorsch, (1990) obtained $\hat{\alpha}_{MLE}$ as follows:

$$P(X = x) = \begin{cases} \dfrac{\Gamma(x + \alpha^{-1})}{x!\,\Gamma(\alpha^{-1})}\left(\dfrac{\alpha m}{1 + \alpha m}\right)^{x}(1 + \alpha m)^{-1/\alpha}, & m, \alpha > 0,\ x = 0,1,2, \\ 0 & elsewhere \end{cases} \quad \text{..... (16)}$$

$$\partial l/\partial m = \sum_{i=1}^{n}\left(X_i/m - 1 + \frac{\alpha X_i}{1 + \alpha m}\right) = 0 \text{ .... (17)}$$

$$\partial l/\partial \alpha = \sum_{i=1}^{n}\left[\sum_{v=0}^{X_i - 1} \frac{1}{\alpha(1 + \alpha v)} - \frac{n}{\alpha^{-2}}\log(1 + \alpha m) + \frac{m\mu(\bar{X} + \alpha^{-1})}{1 + \alpha m}\right] = 0 \text{ .... (18)}$$

$$m_{MLE} = \bar{Y}$$

Solving (18) at $m = \hat{m}$ yields $\hat{\alpha}_{MLE}$

THE OFFSET POISSON-GAMMA MIXTURE MODEL

The canonical link function of the Poisson-Gamma mixture regression model can be expressed as

$$\eta = log(1 - p) \text{ .... (19)}$$

Where $p = m/(m + r)$ is the probability of success in each trial

$$\text{or } \eta = log\left(m/(m + k)\right) \text{ ....(20)}$$

where m is $E(X)$ = mean.

Equation (19) derives from a probability mass function defined as

$$f(k; r; p) = \frac{\Gamma(r + k)}{k!\,\Gamma(r)}p^k(1 - p)^r$$

while (20) is derived from a *pmf* version of the Poisson-Gamma model expressed as

$$f(x; \alpha; k) = \frac{(X + K - 1)!}{X!\,(K - 1!)}\frac{\alpha^X}{(1 + \alpha)^{X+K}}; \qquad X = 0,1,2, \dots$$

where $E(X) = m = K\alpha$ = mean

$$Var(X) = K\alpha + K\alpha^2 = m + m^2/K$$

It can be shown that expression (19) & (20) are equivalent. Using equation (20), a generalized linear model having the Poisson-Gamma mixture model is expressed as

$$\eta = log(m/(m + K)) = X\beta$$

Note that $m/(m + K) = p$

$$\Rightarrow \log(m) - \log(m + K) = X\beta$$

$$\log(m) = X\beta + \log(m + K) \text{ .... (21)}$$

$\log(m + K)$ is an off-set for the Poisson-Gamma mixture regression model.

Under the condition that $m + k = 1$, the Poisson-Gamma mixture model (21) reduces to a Poisson (m) loglinear model.

$$\hat{\beta}_{MQLE} = (X'WX)^{-}X'WZ$$

Where W is a diagonal weight matrix with Poisson weights.

$$w = m/\phi, \qquad \phi > 1 \quad (\phi \text{ is over-dispersion parameter})$$

β estimates are Maximum Quasi Likelihood Estimates (MQLE).

Over-dispersion occurs when the observed variance is larger than the expected variance. Under the assumption that the logistic model is correct $D \sim \psi^2_{n-p}$ Collet (2003)

Thus Collet,(2003) $D > n - p = E(\phi^2_{n-p})$ can indicate over-dispersion, even though the condition $D > n - p$ can be caused by other reasons such as wrong link function, existence of large outliers, etc.

Illustration for an HIV/AIDS study

We assume a community where multiple sex partners live. Let the meeting with a non-HIV / AIDS member of a community by a multiple sex partner be counted as a success and the meeting with a HIV / AIDS member be counted as a failure.

Let p represent the proportion of non-HIV / AIDS.
Let k = no of times a multiple sex member of the community meets non-HIV / AIDS members before meeting with the *rth* infected person.
Then r = no of times the multiple members meets with infected persons.
k is likely to be higher for higher populations.
$\therefore k$ is a function of the population (size). r is also going to be higher with higher populations.
Let $\alpha = 1/_r$, then α is a measure of Poisson over-dispersion. Recall that for the Poisson-Gamma mixture distribution, $var(X) = m + m^2/_r = m + \alpha m^2$ thus as $r \to \infty$, over-dispersion $\to 0$ and the Poisson-Gamma mixture distribution tends to the Poisson distribution.

Thus in large populations the Poisson-Gamma mixture distribution will be approximated by the Poisson distribution.

CONCLUSION

The offset Poisson-Gamma mixture model has been formulated to show that the Poisson-Gamma mixture link is actually the sum of the Poisson loglinear link and an offset.

Thus modeling a Poisson-Gamma mixture regression model amounts to modeling a Poisson loglinear model and thereafter adding the offset term. It is also shown in this study that by using the fact that the Poisson overdispersion measure is a reciprocal of the number of failures, increasing the sample size reduces Poisson overdispersion.

REFERENCES

Al-Khasawneh, M. F. (2010). Estimating the Negative Binomial Dispersion Parameter. *Asian J. Maths Stat.*, 3, 1-5.

Anraku, K. & Yanagimoto, S. (1990). Estimation for the negative binomial distribution based on the conditional likelihood. Comm. Statist. *Simul. Comput*. 19: 771-786.

Collet, D. (2003). Modelling binary data (2nd edition). London: Chapman and Hall

Hilbe, J. M. (2007). Negative Binomial Regression. 2nd edition. Cambridge University Press. In press
Levin, B. & Reeds, J. (1977). Compound Multinomial likelihood functions are unimodal: Proof of a conjecture of I. *J. Good. Ann. Statist;* 5, 79-87.

Piegorsch, W. W., (1990). Maximum Likelihood estimation for negative binomial dispersion parameter . *Biometrics*, 46, 863-867.

Williams, D. (1982). Extra binomial variation in logistic linear models. *Applied statistics*, 31, 144-148.