On-Site Sampling and Generalized Count Data Models

MAKI MOMMA

Contents

Abstract

The aim of this paper is to derive distributions of count variables based on generalized count data models, when inference is based on an on-site sample. On-site sampling is a method where data are collected from subjects that are engaged in an activity of interest (on-site population) at the time of sampling. While the method inevitably implies selection bias, it is in general easier to implement than random sampling. Furthermore, when a high frequency of zero values is expected in the whole population, on-site sampling makes it possible to draw inferences based on a relatively small sample size. After introducing various forms of generalized count data models, distribution of an on-site population corresponding to each model is derived and their properties studied. Estimation based on an on-site sample is also discussed briefly.

1. Introduction

Count data refer to data taken on the number of events in a specified time interval. In many microeconomic applications, we are interested in the dependence of a count variable on other quantitative or qualitative variables, called regressors. Although count variables are discrete by nature, there is little loss of information when their distributions are approximated by continuous distributions such as the normal, provided the data consist mostly of large values. If this is the case, classical econometric models may be employed for analysis. In contrast, when the data include a number of small values, as is common with microeconomic data, it is essential to derive discrete models for the counts. Such models are called count data models.

While interest in count data model is relatively new in econometrics, its role is becoming increasingly important with the proliferation of microeconomic data. The most basic regression model for count data is the Poisson model, where a count variable follows a Poisson distribution with mean parameter that is a deterministic function of the regressors. Empirical findings suggest however, that the Poisson assumption is not consistent with some features of real-life data, and for this reason, various generalized count data models have been proposed. Some of these features include heterogeneity of the population, observation of excess zeros, and dependence between occurrence times of events. A brief survey of generalized count data models is given in Sections 2 and 3, with emphasis on models based on flexible assumptions for the count distribution.

Sampling method plays an essential role in the analysis of a count variable, since in many cases data exhibit a high frequency of zeros. If random sampling is employed under such circumstances, a large sample size is required to perform reliable analysis. When data are collected only from items taking non-zero values, such inefficiency can be avoided. One such method is to employ on-site sampling, where random samples are taken from a population of subjects that are engaged in an activity of interest (referred to as on-site population) at the time of sampling. Although the method inevitably contains sampling bias, it is in general easier to imply. Furthermore, a smaller sample size is required to perform inference based on on-site samples, since zero values are precluded in the sample. On-site sampling is discussed in Section 4, where distributions of on-site populations corresponding to models of Sections 2 and 3 are derived. Estimation methods are discussed briefly for each case.

It is assumed throughout the paper that data are taken in cross-section form (single observation on the number of counts per each individual) unless noted otherwise.

2. Parametric count data models

Count data models are used extensively in the area of reliability analysis, bio-statistics and demography, where various models as well as estimation techniques have been developed. A special feature of economic data is that the only information available is the number of event counts over a specified time interval. Such data are sometimes called current status data in the statistical literature. In contrast, in areas such as reliability analysis, it is common to assume that event times are observed as well. Observation of current status data is a natural assumption when data are collected in surveys. In general, it implies a loss of information compared to the case where the event times are also observed, the exception being the baseline Poisson model described bellow.

The most basic model of a count variable, sometimes called the baseline model, is the simple Poisson model where the number of counts Y in a given time period (which is standardized to be 1) follows a Poisson distribution:

$$P(Y=y|\lambda)=P(y|\lambda)=\frac{e^{-\lambda}\lambda^{y}}{y!}. \tag{2.1}$$

Here, $\lambda=E(Y)$ is the mean parameter. In case of count data regression, it is customary to assume that λ depends on the regressors through the relation $log\,\lambda=x'\beta$, where x denotes a k dimensional column vector of regressors $(x_1,\dots,x_k)'$ whose values are observed and β, a $k\times 1$ parameter vector $(\beta_1,\dots,\beta_k)'$ to be estimated from data. Then, $E(Y)=\lambda=e^{-x'\beta}$, so the above specification ensures positivity of the expected value of the counts. The log-likelihood of the Poisson model (2.1) based on observations (x_i,y_i) $i=1,\dots,n$ is given by

$$log\,L=-\sum_{i=1}^{n}e^{-x_i'\beta}+\sum_{i=1}^{n}y_i x_i'\,\beta+\sum_{i=1}^{n}log\,y_i!. \tag{2.2}$$

When event counts follow the Poisson model, there is indeed no shortage in the amount of information we obtain through current status data, since in this case, inter-event times follow an exponential distribution which has the famous memory-less property.

A serious drawback of the Poisson model is the restriction that expected value of the counts must equal its variance. This follows since Poisson distribution is a one-parameter distribution. Evidence from empirical data suggest however, that variance usually exceeds the mean (a case called over-dispersion). In some cases, the data also exhibit an excess number of zeros compared to those expected by a Poisson distribution. Two types of models are commonly used to account for excess zeros. The zero-inflated Poisson model, where zero is assumed to come from two different sources, and the hurdle model, where the model consists of a two-part decision process and zero is generated by an independent data generating process. For further details on these models, see for example, Cameron, A. C. and Trivedi, P. K. (1998).

A second and equally serious limitation of the Poisson model is that it does not allow heterogeneity within the population. An assumption of a homogeneous population is not likely to hold in practice. To accommodate heterogeneity, it is often assumed that an unobservable heterogeneity factor affects the expected value of the counts in a multiplicative form. More specifically, for every observation i, it is assumed that $E(Y_i)=\tilde{\lambda}_i=\lambda_i v_i=exp(x_i'\beta)v_i$, where v_i is an unobservable heterogeneity factor with $E(v_i)=1$. Since heterogeneity is unobservable, it needs to be integrated out of the distribution function to obtain the conditional distribution of Y given x. Letting g denote the density of v, marginal density of the counts with multiplicative heterogeneity is then seen to be

$$P(y\,|\,x)=\int\frac{e^{-\lambda v}(\lambda v)^y}{y!}g(v)dv\,, \tag{2.3}$$

the mixed Poisson distribution.

Note that when regressors are observed with error (the errors-in-variables case) and no heterogeneity is assumed, the resulting distribution of Y has exactly the same form as above. To show this, let $z_i'\beta=(x_i'+u_i')\beta$ where z_i is a vector of observed variables and u_i, a vector of observation errors. Assuming as in the Poisson model, $E(Y_i)=exp(x_i'\beta)$, define $\xi=exp(-u_i'\beta)$ and let $\overline{g}$ be the density function of ξ. Then

$$P(y\,|\,z)=\int\frac{e^{-\overline{\lambda}\xi}(\overline{\lambda}\xi)^y}{y!}\overline{g}(\xi)d\xi\,, \tag{2.4}$$

where $\overline{\lambda}=exp(z_i'\beta)$. It is not possible to identify whether mixing is due to heterogeneity or errors-in-variables or both, unless there is additional information. In the following discussion, it will be assumed that the model implies heterogeneity. This is done mainly for expository purposes. It should be kept in mind that the same argument applies for the errors-in-variables case as well.

Multiplicative heterogeneity does not change the expected value of Y, but changes its variance and causes over-dispersion. As a result, zeros as well as large values are more frequently observed than in the simple Poisson model. Regardless of the form of g, it can be shown that

$$Var(Y) = \lambda^2 Var(v) + \lambda \tag{2.5}$$

provided $E(v) = 1$, a standardization employed for identification purposes. When the distribution of the count variable belongs to an exponential family, Shaked (1980) has derived a more general result referred to as the Two Crossings Theorem. The theorem states that mixed distribution always have heavier tails than the original distribution.

When the true model is the mixed Poisson distribution, consistency of MLE based on the simple Poisson model is still valid. This follows since from (2.2), the first order condition for maximum likelihood estimation of the Poisson model is seen to be

$$\sum_{i=1}^{n} (y_i - e^{x_i'\beta}) x_i = 0 \tag{2.6}$$

which holds as long as the relation between the mean of the counts and the regressors is valid. A straightforward approach to estimating a mixed model then is to use the Poisson MLE and adjust for the variance. A common method is to describe the variance as a function of the mean, the most popular being $Var(Y \mid \lambda) = \lambda + \alpha\lambda^p$, where α is a scalar parameter and p is some specified value, usually 1 or 2. This is the method of pseudo maximum likelihood. Using this approach, no assumption is necessary regarding the form of the heterogeneity distribution.

A second approach to estimating a mixed model is to assume a parametric distribution for heterogeneity. The form of the mixed distribution depends on the form of g, so in order to estimate the model parametrically, it is necessary to specify the distribution of the unobserved heterogeneity factor v. The most popular choice for the form of g is the Gamma distribution $\Gamma(\alpha,\alpha)$, which results in a Negative Binomial for the distribution of the counts. The shape and scale parameter of the Gamma distribution are set equal to accommodate the assumption that $E(v) = 1$. The corresponding distribution of the counts is seen to be

$$P(y \mid x, \lambda, \alpha) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y+1)} \left(\frac{\alpha}{\lambda + \alpha} \right)^{\alpha} \left(\frac{\lambda}{\lambda + \alpha} \right)^{y}, \tag{2.7}$$

which is the Negative Binomial distribution with mean $\lambda = e^{-x'\beta}$ and variance $\lambda\left(1 + \frac{\lambda}{\alpha}\right)$. Parameter estimates are obtained by maximum likelihood method. It is to be noted that there are other possible parameterization of the Gamma distribution, which will also lead to the Negative Binomial for the marginal distribution of Y, but with slightly different parameterization.

In many empirical cases, the Negative Binomial model seems to fit the data fairly well.

This does not necessarily imply that it is indeed the correct model. In fact, it may simply be the result that the count variable follows an over-dispersed distribution that is a member of the linear exponential family with two parameters, and that the functional form of variance (2.5) is known. See Gourieroux, C., A. Monfort and A. Trognon (1984a,b) for details. Other choices for heterogeneity distribution include inverse Gamma distribution by Dean, Lawless and Willmot (1989), and lognormal distribution by Hinde (1982). Fully parametric methods may produce biased results when the assumption on the form of the distribution does not hold, and often there is no strong foundation for the assumption on the heterogeneity distribution.

When data are obtained in panels, a more elaborate model can be employed. See for example, Hausman, Hall and Griliches (1984) for a detailed discussion on parametric estimation of panel data. A time series data of the counts present a different type of difficulty, since the data collected will typically be dependent of one another. One of the popular models used in this instance is the binomial-thinning model, where counts from a previous time period are thinned down while new independent counts occur within a given time period. For a detailed account of the binomial thinning model, see Al-Osh, M. A. and Alzaid, A. A. (1987).

3. Generalized count data models

Assuming a parametric distribution for heterogeneity is somewhat arbitrary and mainly for computational ease. With the advancement of computer technology, estimating models requiring computer intensive methods have become less inhibitive. For this reason, various models in a more general framework imposing less restriction on the distributional form have been proposed. Several of these models are presented in this section.

3.1. Series models

Gurmu, Rilstone, and Stern (1999) developed a semi-parametric model (referred to as the GRS-model) of the counts based on a series expansion for the distribution of unobserved heterogeneity. Their model assumes that conditional distribution of the counts given the value of heterogeneity follows a Poisson distribution, and the distribution of the heterogeneity factor ν is approximated by an orthonormal polynomial expansion. More specifically, the distribution of ν is given by

$$g(v_i)=\frac{1}{\phi}w(v_i)[P_K(v_i)]^2, \qquad (3.1)$$

where $w(v)$ is the baseline density of heterogeneity, $P_K(v_i)$ denotes a polynomial of degree K, and $\phi=\int w(v_i)[P_K(v_i)]^2 dv_i$ is the constant of proportionality. $P_K(v_i)$ is squared to ensure positivity of the density of the counts. In particular, Gamma distribution is employed as the baseline distribution of heterogeneity,

$$w(v_i)=\frac{v_i^{\alpha-1}\lambda^{\alpha}}{\Gamma(\alpha)}e^{-\lambda v_i} \qquad (3.2)$$

and a generalized Laguerre polynomial is employed for $P_K(v_i)$,

$$L_K^{\alpha}(v_i)=\sum_{l=0}^{K}\binom{K}{l}\frac{\Gamma(K+\alpha)}{\Gamma(\alpha)\Gamma(K+1)}\lambda^l(-v_i)^l. \qquad (3.3)$$

The model nests the Negative Binomial and the geometric count model as special cases. Hence, it is a more flexible form of specification. Provided the density $g(v)$ has finite-order moments, it gives consistent estimators regardless of the form of g. The estimating equation is quite complex and computer intensive methods are necessary to implement. The model can also be extended to incorporate truncations or excessive zeros. For details, see Gurmu, Rilstone and Stern (1999).

Cameron and Johansson (1997) developed a model (CJ-model) where the distribution of the count variable itself is depicted using a series expansion. This model is attractive in that it allows the case of under-dispersion (mean exceeding variance) as well as over-dispersion. It remains to be seen whether it is possible to approximate the distribution of an arbitrary discrete variable using a series expansion, and the model may not always prove to be parsimonious since it requires quite a few polynomial terms to deviate significantly from the baseline distribution. Their model assumes the following distribution for the counts:

$$P_p(y\mid\lambda,a)=f(y\mid\lambda)\frac{h_p^2(y\mid a)}{\eta_p(\lambda,a)}. \qquad (3.4)$$

Here, $f(y\mid\lambda)$ is the baseline density, $h_p(y\mid a)=\sum_{k=0}^{p}a_k y^k$ is the pth order polynomial, $a=(a_0,a_1,\ldots,a_p)'$ is the vector of parameters, and $\eta_p=\sum_{k=0}^{p}\sum_{l=0}^{p}a_k a_l m_{k+l}$ is a normalizing constant with m_k denoting kth non-central moment of the baseline density $f(y\mid\lambda)$. The polynomial

$h_p(y \mid a)$ is again squared to ensure positivity. The appropriate order of expansion is determined via a model selection criterion such as AIC or BIC.

A reasonable choice for $f(y \mid \lambda)$ is the Poisson distribution, in which case the distribution of the count variable is specified as follows:

$$P_p(y \mid \lambda, a) = \frac{e^{-\lambda}\lambda^y}{y!}\frac{h_p^2(y \mid a)}{\eta_p(\lambda, a)} = \frac{\lambda^y}{y!}\sum_{j=0}^{\infty}\frac{(-1)^j\lambda^j}{j!}\frac{h_p^2(y \mid a)}{\eta_p(\lambda, a)}. \tag{3.5}$$

When the baseline distribution is Negative Binomial, the corresponding distribution of the counts is seen to be

$$P_p(y \mid \lambda, a, \alpha) = \frac{\lambda^y}{y!}\frac{h_p^2(y \mid a)}{\eta_p(\lambda, a)}\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha}\left[\prod_{i=1}^{y}\left(\frac{y+\alpha-i}{\lambda+\alpha}\right)\right]. \tag{3.6}$$

Cameron and Trivedi (1998) have shown that in general, Negative Binomial baseline model fits the empirical data better. It is more flexible compared to the Poisson baseline model (3.5) with the cost of estimating one additional parameter α. The Negative Binomial Baseline model (3.6) corresponds to the GRS-model model with gamma baseline distribution for heterogeneity. Empirical comparison by Cameron and Trivedi of a CJ-model with Negative Binomial baseline distribution and GRS-model with Gamma baseline distribution for heterogeneity suggests that performances of both models are compatible.

CJ-model is not derived as an approximation to the distribution of unobserved heterogeneity. To interpret their model from this point of view, rewrite the mixed Poisson density assuming exchangeability of integration and addition, as

$$f(y \mid \lambda) = \frac{\lambda^y}{y!}\sum_{j=0}^{\infty}\frac{(-1)^j}{j!}\lambda^j\mu_{j+y}, \tag{3.7}$$

where as before, v is the unobserved heterogeneity factor with $E(v)=1$, and $\mu_{j+y} = E\left(v^{j+y}\right)$ is the $(j+y)$th non-central moment of heterogeneity v. Comparing equations (3.5) and (3.7), CJ-model with baseline Poisson distribution can be interpreted as estimating the "weighted average" of the non-central moments of heterogeneity by a finite polynomial of the observed number of counts. It should be noted that CJ-model with Negative Binomial distribution as the baseline distribution corresponds to estimating the non-central moment of heterogeneity in (3.7) using higher order polynomial of the count variable.

3.2. Finite mixture models

Another approach to modeling heterogeneity is to use finite mixture models. In this approach, the count variable Y is divided into several latent classes, the number c of which is also estimated from data. When Y is generated from c groups each with a Poisson distribution but with different parameters $\beta_j = \left(\beta_{1j}, \beta_{2j} \cdots \beta_{kj}\right)'$ $j = 1,\ldots,c$, the distribution of Y is given by

$$P\left(y_i \mid x_i, \beta, p\right) = \sum_{j=1}^{c} p_j \frac{\lambda_{ij}^{y_i}\, exp(-\lambda_{ij})}{y_i!}, \tag{3.8}$$

where p_j denotes the mixing probabilities $j = 1,\ldots,c$ with $p = (p_1,\ldots,p_c)$, $\beta = (\beta_1,\ldots,\beta_c)$ is a $k \times c$ matrix of parameters to be estimated, and $\lambda_{ij} = exp(x_i'\beta_j)$. For this model, the mean and variance of the count variable Y_i are seen to be

$$E(Y_i) = \sum_{j=1}^{c} p_j \lambda_{ij} \tag{3.9}$$

and

$$Var(Y_i) = \sum_{j=1}^{c} p_j \lambda_{ij} + \sum_{j=1}^{c} p_j \lambda_{ij}^2 - \left\{\sum_{j=1}^{c} p_j \lambda_{ij}\right\}^2 \tag{3.10}$$

respectively, so that $E(Y_i) = Var(Y_i)$ if and only if $\lambda_{i1} = \lambda_{i2} = \cdots \lambda_{ic}$, the case with no heterogeneity. Although the model implies discreteness of the heterogeneity distribution, the approach provides good numerical approximation even when the true mixing distribution is continuous. It is also straightforward to incorporate the case of excess zeros using this model. The approach differs from the semi-parametric approach in Section 3.1 in that it changes the mean-variance relationship, as is seen from (3.9) and (3.10). See Wang, Puterman, Cockburn and Lee (1996) for further details.

3.3. Models based on waiting times

Models discussed in Sections 3.1 and 3.2 focus on the heterogeneity factor to generalize the baseline Poisson model. Since the flip side of the number of event counts is the waiting time between events, an alternative method of generalization is to consider the model in terms of the waiting time distribution. Poisson model implicitly assumes that waiting times between $(k-1)$th and kth event τ_k $(k = 1,\ldots,y)$ are independent and identically distributed with an exponential distribution. In this case, the hazard function $\zeta_t = \dfrac{f(t)}{1-F(t)}$, where $f(t)$ is the density and $F(t)$ the distribution of waiting times τ, remains constant over time. To generalize the model, various waiting time distributions may be employed. A straightforward extension of the exponential waiting time distribution is the Gamma distribution. Based on the assumption that waiting times are independent and identically distributed with density $f(\tau \mid \alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\tau^{\alpha-1}e^{-\beta\tau}$, Winklemann (1995) has shown that the distribution of the counts follows

$$P(y \mid \alpha,\beta)=e^{-\beta}\sum_{i=1}^{\alpha-1}\frac{\beta^{\alpha y+i}}{(\alpha y+i)!}. \tag{3.11}$$

When the waiting times are Gamma distributed, hazard function is either monotone decreasing or monotone increasing. Moreover, negative duration dependence (hazard function is a decreasing function of time) causes asymptotic over-dispersion of the count variable, whereas positive duration dependence causes asymptotic under-dispersion. In order to obtain a Gamma count regression model, it is further assumed that $\frac{\beta}{\alpha}=e^{x_i'\gamma}$ where as before, x denotes a k dimensional column vector of regressors $(x_1,\ldots,x_k)'$. The resulting likelihood function is nonlinear in α and γ, and requires iterative numerical algorithms for estimation.

Gourieroux and Visser (1997) constructed a model based on the assumption that waiting times are influenced by several factors; an observable individual specific factor x_i, unobserved individual specific factor (heterogeneity factor) that is constant through the observation period v_i, and an unobservable individual and spell specific factor η_{ik}, where k denotes the number of events so far. Inclusion of η_{ik} in the model implies that waiting time between the $(k-1)$th and the kth event τ_k depends not only on individual factors but also on the number of events so far. According to their model, the heterogeneity factors satisfy the following:

A1: $v_i,\eta_{i1},\cdots\eta_{ik},\cdots$ are independent of x_i, and v_i's are i.i.d. random variables.

A2: Conditionally on (x,v,η_k), duration times τ_k $k=1,\cdots,$ follow an exponential distribution independently with parameters $\lambda(x,v,\eta_k)$.

A3: The parameter $\lambda(x,v,\eta_k)$ is decomposed as $\lambda(x,v,\eta_k)=\lambda(x,v)(1+\eta_k)^{-1}$.

Observed variables are the number of counts y and values of individual specific factors x_i.

Conditional distribution of the count variable based on the above condition takes a complicated form, from which we need to derive the marginal distribution to pursue estimation. This makes the model unattractive. Instead, Gourieroux and Visser employ a local approximation of the model based on the assumption that unobserved heterogeneity is independent of the regressors, and that individual and spell specific factors η_{ik} are small. Expanding the characteristic function of the waiting time using this assumption, they obtain the local count data model as

$$P(y\,|\,x) \cong \overline{P}_y(\lambda) + \overline{M}_{y+1}\overline{P}_{y+1}(\lambda) - \overline{M}_K \overline{P}_y(\lambda)\,, \tag{3.12}$$

where $\overline{P}_y(\lambda) = \underset{v}{E}\,P(Y=y)$ with $M_y = \sum_{k=1}^{y} \eta_k$, $\overline{M}_y = E(M_y)$, and $M_0 = 0$.

When $\lambda = E(Y) = e^{x'\beta}$ is independent of v, so that heterogeneity stems only from x_i and η_{ik}, the local distribution of the counts becomes

$$P(y\,|\,x) = \frac{e^{-\lambda}\lambda^y}{y!}\left[1 - \overline{M}_y + \overline{M}_{y+1}\frac{\lambda}{(y+1)}\right]. \tag{3.13}$$

This is the model used by Gourieroux and Visser for estimation. To obtain a model that corresponds to a generalization of the Negative Binomial model, assume that v follows a $\Gamma(\alpha,\alpha)$ distribution. Then, $\overline{P}_y(\lambda)$ is Negative Binomial and the local distribution of the counts is seen to be

$$P(y\,|\,\alpha,\lambda) = P_y(\lambda)\left[1 - \overline{M}_y + \overline{M}_{y+1}\frac{\alpha+y+1}{(y+1)}\frac{\lambda}{\lambda+\alpha}\right]. \tag{3.14}$$

4. On Site Sampling

When the population distribution of a count variable contains a mass at zero, random sampling is likely to produce a sample with many zeros. To pursue reliable inference in such a case, a large sample size is necessary so that enough non-zero values are observed. Instead, on-site sampling takes random samples from an on-site population, that is, from a population of subjects engaged in an activity of interest at the time of sampling. For example, if we want a sample on the number of visits to hospitals during a certain period, an on-site sample will take random samples from patients visiting a hospital on a particular day. This sampling method is in general easier to implement than random sampling of the whole population, and saves a considerable amount of time and cost.

A slightly different form of sampling that is sometimes confused with on-site sampling consists of drawing a random sample from a population of items with positive data values. For example, we could draw a random sample from the owners of registered vehicles, etc. In this case, the sample distribution P_S is simply a conditional distribution of the population distribution, that is

$$P_S(y) = P(y \mid y > 0) = \frac{P(y)}{P(y > 0)} = \frac{P(y)}{1 - P(y = 0)} \ . \tag{4.1}$$

This type of sampling method has limited usage, for information on data values is usually not available prior to sampling.

When discussing on-site sampling, it is essential to consider sampling bias. Note that taking random samples from an on-site population does not correspond to a random sample from the whole population conditioned to take positive values, since the more time a subject spends in an on-site population, the higher its chance to be in the sample. Shaw (1988) has derived the distribution of a count variable on-site when the population distribution is Poisson. His derivation is based on a "hypothesized stratified population". A perhaps simpler interpretation is to assume that a sample is chosen approximately proportional to the number of times a subject engages in the activity of interest. Then we have a familiar case of biased sampling. For cases such as visits to recreational facilities, it is more accurate to assume that subjects are sampled proportional to the average length of time they spend in the facility. This however, will require additional assumption on the distribution of time spent in the facility, and may produce a result that is sensitive to the underlying distributional assumptions, therefore lacking robustness.

A familiar formula for biased sampling is given by

$$f^*(y) = \frac{yf(y)}{E(Y)}, \tag{4.2}$$

where f denotes the density of the whole population, and f^* the biased density, which in this case corresponds to the density of an on-site population. Regardless of the form of f , it can be shown that $E\left(\frac{1}{Y^*}\right) = \frac{1}{\lambda}$, $E\left(Y^*\right) = \frac{E\left(Y^2\right)}{\lambda}$ and $Var(Y)^* = \frac{E\left(Y^3\right)}{\lambda} - \left(\frac{E\left(Y^2\right)}{\lambda}\right)^2$, where Y denotes the variable of interest in the whole population, Y^* the corresponding variable in an on-site population, and $\lambda = E(Y)$. A simple estimate of the mean parameter is given by the reciprocal of $\frac{1}{n}\sum_{i=1}^{n}\frac{1}{Y_i^*}$.

Distribution of an on-site population corresponding to a baseline Poisson model is a dislocated Poisson distribution. So far, distributions corresponding to generalized count data models do not seem to have been studied. I will derive distributions of count variables in an on-site population for models introduced in Sections 2 and 3, and investigate their properties.

4.1. Negative Binomial models

Negative Binomial model is one of the most widely used parametric models for count-data. When the population distribution is Negative Binomial (2.7), distribution of an on-site population is derived using (4.2) as

$$P^*(y\,|\,\alpha,\lambda)=\frac{\Gamma(\alpha+y)}{\Gamma(\alpha+1)\Gamma(y)}\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha+1}\left(\frac{\lambda}{\lambda+\alpha}\right)^{y-1} \tag{4.3}$$

which is a displaced Negative Binomial distribution with mean $E\left(Y^*\right)=\lambda+\frac{\lambda}{\alpha}+1$ and variance $Var\left(Y^*\right)=\frac{\lambda}{\alpha}(\alpha+\lambda)\left(1+\frac{1}{\alpha}\right)$ respectively. The variance of this distribution is larger than that of the whole population, which is $\lambda\left(1+\frac{\lambda}{\alpha}\right)$. Since there has been an increase in the mean value as well, distribution of an on-site population does not always result in over-dispersion. In fact, it is seen that over-dispersion occurs if and only if $\frac{1+\alpha}{\alpha}>\lambda^2$.

Maximum likelihood method may be employed to obtain parameter estimates for this case, since the model is fully parameterized. Likelihood function corresponding to an on-site sample is given by

$$L=\prod_{i=1}^{n}\frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha+1)\Gamma(y_i)}\left(\frac{\alpha}{\lambda+\alpha}\right)^{\alpha+1}\left(\frac{\lambda}{\lambda+\alpha}\right)^{y_i-1}. \tag{4.4}$$

After some calculation, the log-likelihood is seen to be

$$log\, L\propto\sum_{i=1}^{n}\left[\sum_{k=1}^{y_i-1}log\,(\alpha+k)-(\alpha+y_i)log\,(\lambda+\alpha)+(y_i-1)log\,\lambda\right]+n(\alpha+1)log\,\alpha\,, \tag{4.5}$$

where as before, $\lambda=e^{x_i'\beta}$. From this, first order conditions for the maximization of the likelihood function is derived as

$$\frac{\partial\, log\, L}{\partial\alpha}=n\, log\,\alpha+\frac{n(\alpha+1)}{\alpha}+\sum_{i=1}^{n}\left[\sum_{k=1}^{y_i-1}\frac{1}{\alpha+k}-log\,(\lambda+\alpha)-(\alpha+y_i)\right]=0 \tag{4.6}$$

and

$$\frac{\partial\, log\, L}{\partial\beta}=n\, log\,\alpha+\frac{n(\alpha+1)}{\alpha}+\sum_{i=1}^{n}\left[\sum_{k=1}^{y_i-1}\frac{1}{\alpha+k}-log\,(\lambda+\alpha)-(\alpha+y_i)\right]=0\,, \tag{4.7}$$

respectively. Using the second condition, $\lambda = exp(x_i'\beta)$ is re-expressed as

$$\lambda = \frac{\alpha\left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} x_i\right)}{(1+\alpha)\sum_{i=1}^{n} x_i}.$$

Since the first condition cannot be simplified further, numerical method is necessary to obtain an estimate for α.

Distribution of an on-site population corresponding to a general mixed Poisson model (2.3) is given by

$$P^*(y \mid \lambda) = \int \frac{e^{-\lambda v}(\lambda v)^{y-1}}{(y-1)!} v g(v) dv, \tag{4.8}$$

which can be rewritten as

$$P^*(y \mid \lambda) = \frac{\lambda^{y-1}}{(y-1)!} \sum_{j=0}^{\infty} \frac{(-1)^j \lambda^j}{j!} \mu_{j+y}, \tag{4.9}$$

where $\mu_{j+y} = E\left(v^{j+y}\right)$. The mean and variance of an on-site distribution are seen to be $E(Y*) = E(v) + \lambda E\left(v^2\right) = 1 + \lambda E\left(v^2\right)$ and $Var(Y*) = \lambda E\left(v^2\right) + \lambda^2 E\left(v^3\right) - \lambda^2\left(E\left(v^2\right)\right)^2$ respectively, so for an on-site sample, it is seen that over-dispersion occurs if and only if $E\left(v^3\right) \geq \left(E\left(v^2\right)\right)^2 + \frac{1}{\lambda^2}$.

When the distribution of the original population belongs to a linear exponential family, i.e. when the density of the counts can be expressed as

$$P(y \mid \lambda) = exp\{A(\lambda) + B(y) + C(\lambda)y\}, \tag{4.10}$$

the distribution of an on-site population will also belong to an exponential family given by

$$P(y \mid \lambda) = exp\{A'(\lambda) + B'(y) + C(\lambda)y\}, \tag{4.11}$$

where $A'(\lambda) = A(\lambda) + log\,\lambda$, and $B'(y) = B(y) + log\,y$.

4.2. Series models

The form of GRS-model is quite complicated in the original population. Since CJ-model with Negative Binomial baseline distribution is compatible with the GRS-model with Gamma baseline density, I will derive the distribution of an on-site population for the CJ-model, which has a simpler and more manageable form. Distribution for the CJ-model in an on-site population takes the following form

$$P_p^*(y \mid \lambda,\alpha) = f(y \mid \lambda,\alpha)\frac{\sum_h \sum_j a_h a_j y^{h+j+1}}{\sum_k \sum_l a_k a_l m_{k+l+1}}, \qquad (4.12)$$

where as before, m_k denotes the k th non-central moment of $f(y|\lambda,\alpha)$. From this, the mean and variance of the distribution is seen to be

$$E(Y^*) = \frac{\sum_h \sum_j a_h a_j m_{h+j+2}}{\sum_k \sum_l a_k a_l m_{k+l+1}}$$

and

$$Var(Y^*) = \frac{\sum_h \sum_j a_h a_j m_{h+j+3} \sum_k \sum_l a_k a_l m_{k+l+1} - \left(\sum_h \sum_j a_h a_j m_{h+j+2}\right)^2}{\left(\sum_k \sum_l a_k a_l m_{k+l+1}\right)^2},$$

respectively. The distribution is over-dispersed if and only if

$$\sum_h \sum_j a_h a_j (m_{h+j+3} - m_{h+j+2}) \sum_k \sum_l a_k a_l m_{k+l+1} > \left(\sum_h \sum_j a_h a_j m_{h+j+2}\right)^2.$$

Log-likelihood for an on-site population is seen to be

$$log\, L = \sum\left[log\, f\left(y|\lambda(x_i,\beta)\right) + log\left(\sum_h \sum_j a_h a_j y^{h+j+1}\right) - log\left(\sum_k \sum_l a_k a_l m_{k+l+1} |\lambda(x_i,\beta),\alpha\right)\right].$$

(4.13)

In particular, when the baseline density is Poisson, the corresponding distribution of an on-site population becomes

$$P_p^*(y \mid \lambda,\alpha) = \frac{e^{-\lambda}\lambda^y}{y!}\frac{\sum_h \sum_j a_h a_j y^{h+j+1}}{\sum_k \sum_l a_k a_l m_{k+l+1}}. \qquad (4.14)$$

From this, first order conditions for maximum likelihood estimation based on an on-site sample is derived as

$$\sum_{i=1}^{n}\left[y_i - \frac{\sum_{k}\sum_{l} a_k a_l m_{k+l+2}}{\sum_{h}\sum_{j} a_h a_j m_{j+h+1}} \right] x_i = 0 \qquad (4.15)$$

or equivalently,

$$\sum_{i=1}^{n}\left[y_i - E(Y_i^* \mid x_i) \right] x_i = 0$$

and

$$\sum_{i=1}^{n}\left[\frac{2\sum_{j} a_j y^{l+j+1}}{\sum_{h}\sum_{j} a_h a_j y^{h+j+1}} - \frac{2\sum_{k} a_k m_{k+l+1}}{\sum_{h}\sum_{j} a_h a_j m_{h+j+1}} \right] = 0 . \qquad (4.16)$$

Estimates of parameters are obtained following the procedure for the case of a random sample from the whole population. Cameron and Johansson suggest simulated annealing since the model is non-linear in the parameters.

4.3. Finite mixture models

For finite mixture models with baseline Poisson distribution, distribution of an on-site population is given as follows:

$$P^*(y_i \mid x_i, \beta, p) = \frac{\sum_{j=1}^{c} p_j \frac{\lambda_{ij}^{y_i}\, exp(-\lambda_{ij})}{(y_i - 1)!}}{\sum_{k=1}^{c} p_k \lambda_{ik}} \qquad (4.17)$$

From this, the expected value and variance of an on-site population is seen to be

$E^*(Y_i) = 1 + \frac{\sum_{j=1}^{c} p_j \lambda_{ij}^2}{\sum_{j=1}^{c} p_j \lambda_{ij}}$ and $Var^*(Y_i) = 1 + \frac{\sum_{j=1}^{c} p_j \lambda_{ij}^3}{\sum_{j=1}^{c} p_j \lambda_{ij}} + \frac{\sum_{j=1}^{c} p_j \lambda_{ij}^2}{\sum_{j=1}^{c} p_j \lambda_{ij}} - \frac{\left(\sum_{j=1}^{c} p_j \lambda_{ij}^2\right)^2}{\left(\sum_{j=1}^{c} p_j \lambda_{ij}\right)^2}$ respectively. The distribution is over-dispersed if and only if $\sum_{j=1}^{c} p_j \lambda_{ij} \sum_{j=1}^{c} p_j \lambda_{ij}^3 \geq \left(\sum_{j=1}^{c} p_j \lambda_{ij}^2\right)^2$. Define z_{ij} as an indicator variable that takes the value 1 if observation i belongs to group j, and 0 otherwise. Then, the log likelihood of an on-site population is given by

$$\log L = \sum_{i=1}^{n}\sum_{j=1}^{c}\left[-z_{ij}\log(\sum_{k} p_k\lambda_{ik}) - z_{ij}\log(y_i - 1)! + z_{ij}\log p_j + z_{ij}y_i\log\lambda_{ij} - z_{ij}\lambda_{ij}\right] \tag{4.18}$$

from which the first order conditions are derived as

$$p_j \sum_{i} z_{ij}(\lambda_{ij} - \lambda_{ic}) = (\sum_{k} p_k\lambda_{ik})\sum_{i} z_{ij} \tag{4.19}$$

for $j = 1,\ldots,c-1$ and

$$\sum_{i}\sum_{j} z_{ij}\; p_j\,\lambda_{ij}x_i = (\sum_{k} p_k\lambda_{ik})\sum_{i}\sum_{j}(\lambda_{ij} - y_i)z_{ij}x_i \tag{4.20}$$

Since the model includes unobservable variables or "missing data", EM algorithm is employed to obtain estimates of parameters. In addition, it is necessary to use a numerical method to maximize the expected log-likelihood (M-step), since estimators cannot be obtained in a closed form. Derivation is much more complicated for an on-site sample compared to a random sample of the whole population, since parameters p_j and λ_{ij} are no longer separable in the likelihood function of an on-site population.

4.4. Models based on waiting times

Distributions of an on-site population corresponding to models based on waiting times are complicated. For the Gamma waiting time model, the distribution is seen to be

$$P^*(y \mid \alpha,\beta) = \frac{yG(\alpha y,\beta)}{\sum_{i=1}^{\infty} G(\alpha i,\beta)} - \frac{yG(\alpha y + \alpha,\beta)}{\sum_{i=1}^{\infty} G(\alpha i,\beta)}, \tag{4.21}$$

where

$G(\alpha y,\beta) = \dfrac{1}{\Gamma(\alpha n)}\displaystyle\int_0^{\beta} u^{\alpha n-1}e^{-u}du$ is an incomplete Gamma function. The term cannot be simplified unless $\alpha = 1$, which case corresponds to the simple Poisson model.

Distribution of Gourieroux and Visser's local model in an on-site population is seen to be

$$P^*(y \mid x) = \frac{e^{-\lambda}\lambda^{y-1}}{(y-1)!}\left[1 - \overline{M}_y + \overline{M}_{y+1}\frac{\lambda}{y+1}\right] \tag{4.22}$$

when individual heterogeneity v is absent, and

$$P^*(y \mid x) = \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)\Gamma(y)}\frac{\alpha^{\alpha}\lambda^{y-1}}{(\lambda+\alpha)^{\alpha+y}}\left[1 - \overline{M}_y + \overline{M}_{y+1}\frac{\lambda(y+\alpha+1)}{(y+1)(\lambda+\alpha)}\right] \tag{4.23}$$

when v follows a $\Gamma(\alpha,\alpha)$ distribution. To obtain a maximum likelihood estimator using the local model, it is necessary to transform the model so that (4.22) or (4.23) becomes a proper density function for all values of $\overline{M}_y$.

Models based on waiting time distributions have complicated forms in the original population, and require computer intensive methods for estimation. Taking on-site samples add to the complexity of the model. It remains to be seen whether the model is robust to deviances from the underlying assumptions.

5. Conclusion

On-site sampling is a method that is easier to implement than random sampling. When data are expected to exhibit a number of zeros, on-site sampling makes it possible to infer about the underlying distribution based on a relatively small sample size. I have derived distributions of an on-site population for some generalized count data models. When the distribution of the counts in the whole population is in a regular form, the corresponding distribution of the counts in an on-site population is likely to have a manageable form as well. For the finite mixture model however, on-site sampling adds considerably to the difficulty of the estimation process, since the parameters are no longer separable. A Monte Carlo comparison of the properties of estimators based on an on-site sample and estimators obtained by random sampling is left for further study.

An important estimation method that is not discussed in this paper is the generalized method of moments. Instead, focus is on models with flexible forms for the distribution of the count variable. It is to be noted that when there is a possibility of zeros being generated by an independent process, on-site sampling fails to provide information on the data generating process regarding those zeros. In such cases, it may prove useful to investigate the possibility of stratified sampling, combining random sampling and on-site sampling.

References

1. Al-Osh, M. A. and Alzaid, A. A. (1987), "First-order integer-valued autoregressive (INAR(1)) process", *Journal of Time Series Analysis,* 8, 261-275.

2. Cameron, A. C. and Johansson, P. (1997), "Count data regression using series expansions:

with applications", *Journal of Applied Econometrics*, 12, 203-223.

3. Cameron, A. C. and Trivedi, P. K. (1998), *Regression analysis of count data*, Cambridge University Press.

4. Dean, C Lawless, J. F. and Wilmot G.E. (1989), "A Mixed Poisson-Inverse Gaussian Regression Model", *Canadian Journal of Statistics*, 17, 171-182.

5. Gourieroux, C. and Visser, M. (1997), "A count data model with unobserved heterogeneity", *Journal of Econometrics*, 79, 247-268.

6. Gourieroux, C., Monfort, A. and Trognon, A. (1984a), "Pseudo maximum likelihood methods: theory", *Econometrica*, 52, 681-700.

7. Gourieroux, C., Monfort, A. and Trognon, A. (1984b), "Pseudo maximum likelihood methods: applications to poisson models", *Econometrica*, 52, 701-720.

8. Gurmu, S. (1997), "Semi-parametric estimation of hurdle regression models with an application to medicaid utilization", *Journal of Applied Econometrics*, 12, 225-242.

9. Gurmu, S. (1998), "Generalized hurdle count data regression models", *Economics Letters,* 58, 263-268.

10. Gurmu, S. and Trivedi, P. K. (1992), "Overdispersion tests for truncated Poisson regression models", *Journal of Econometrics*, 54, 347-370.

11. Gurmu, S. and Trivedi, P. K. (1996), "Excess zeros in count models for recreational trips", *Journal of Business & Economic Statistics*, 14,4 69-477.

12. Gurmu, S. Rilstone, P. and Stern, S. (1999), "Semiparametric estimation of count regression models", *Journal of Econometrics*, 88, 123-150.

13. Hausman, J., Hall, B. H. and Griliches, Z. (1984), "Econometric models for count data with an application to the patents-R&D relationship", *Econometrica*, 52, 909-938.

14. Heckman, J. and Singer, B. (1984), "A method for minimizing the impact of distributional assumptions in econometric models for duration data", *Econometrica*, 52, 271-320.

15. Hinde, J. (1982), "Compound Poisson Regression Models", in R. Gilchrist, ed., *GLIM 82; Proceedings of the International Conference on Generalised Linear Models*, New York, Springer-Verlag.

16. Jorgensen, B., Lundbye-Christensen, S. Song, P. X. and Sun, L. (1999) "A state space model for multivariate longitudinal count data" *Biometrika*, 86, 169-181.

17. Laird, N. (1978), "Nonparametric maximum likelihood estimation of a mixing distribution" *Journal of the American Statistical Association*, 73, 805-811.

18. Lambert D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing", *Technometrics*, 34, 1-14.

19. Mullahy, J. (1997), "Heterogeneity, excess zeros, and the structure of count data models", *Journal of Applied Econometrics*, 12, 337-350.

20. Santos Silva, J. M. C. (1997), "Unobservables in count data models for on-site samples", *Economics Letters,* 54, 217-220.

21. Shaw, D. (1988), "On-site samples' regression --problems of non-negative integers, truncation, and endogenous stratification", *Journal of Econometrics*, 37, 211-223.

22. Simar, L. (1976), "Maximum likelihood estimation of a compound Poisson process", *Annals of Statistics*, 4, 1200-1209.

23. Wang, P., Puterman, M. L. Cockburn, I. and Le, N. (1996), "Mixed Poisson regression models with covariate dependent rates", *Biometrics*, 52, 381-400.

24. Wedel, M., Desarbo, W. S., Bult, J. R. and Ramaswamy, V. (1993), "A latent class poisson regression model for heterogeneous count data", *Journal of Applied Econometrics*, 8, 397-411.

25. Winkelmann, R. (1995) "Duration dependence and dispersion in count-data models", *Journal of Business & Economic Statistics*, 13, 467-474.

26. Winkelmann, R. (1997), *Count Data Models:Econometric Theory and Application to Labor Mobility*, Berlin, Springer-Verlag.

27. Winkelmann, R. (1998), "Count data models with selectivity" *Econometric reviews*, 17, 339-359.

28. Winkelmann, R. and Zimmermann, K. F. (1995), "Recent developments in count data modeling: theory and application" *Journal of Economic Surveys*, 9, 1-24.