Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Statistical Planning and Inference 136 (2006) 3173–3186
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Robust Poisson regression

Tsung-Shan Tsou*

Institute of Statistics, National Central University, Chung-Li 32054, Taiwan

Received 17 October 2002; accepted 23 December 2004
Available online 14 March 2005

Abstract

Count data are very often analyzed under the assumption of a Poisson model [(Agresti, A., 1996. An Introduction to Categorical Data Analysis. Wiley, New York; Generalized Linear Models, second ed. Chapman & Hall, New York)]. However, the derived inference is generally erroneous if the underlying distribution is not Poisson (Biometrika 70, 269–274).

A parametric robust regression approach is proposed for the analysis of count data. More specifically it will be demonstrated that the Poisson regression model could be properly adjusted to become asymptotically valid for inference about regression parameters, even if the Poisson assumption fails. With large samples the novel robust methodology provides legitimate likelihood functions for regression parameters, so long as the true underlying distributions have finite second moments. Adjustments that robustify the Poisson regression will be given, respectively, under log link and identity link functions. Simulation studies will be used to demonstrate the efficacy of the robust Poisson regression model.

MSC: 62J02; 62G35

Keywords: Robust profile likelihood; Poisson regression; Quasilikelihood; Robust Poisson regression; Likelihood ratio test

1. Introduction

Count data are encountered in many statistical applications. For example, the number of outpatients of a hospital every day, the number of car thefts in year 2000 of New York city,

* Tel.: +886 3 426 7220; fax: +886 3 425 8602.
E-mail address: tsou@stat.ncu.edu.tw.

0378-3758/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.12.008

the number of sister chromatid exchange of a chromosome, etc. The Poisson distribution is frequently assumed to be the sampling model for count data (Frome, 1983; Holford, 1983). For example, the Poisson regression is a popular generalized linear model (GLM) that relates the mean of the count to explanatory variables. Let μ be the expected value of a Poisson variate Y, and let x denote an explanatory variable. The Poisson linear model has the form $\mu = \alpha + \beta x$, and the Poisson loglinear model assumes that $\log \mu = \alpha + \beta x$. Many GLM softwares have implemented Poisson regression models for the analysis of count data with choices of link functions including identity and log links (SAS, proc genmod). However, the ease of the analysis does not guarantee the validity of the analysis.

It has been demonstrated that the application of Poisson regression models on count data that display extra-Poisson variation generates fallacious inference (Paul and Plackett, 1978; Cox, 1983). Certain types of negative-binomial regression models have been proposed for counts that exhibit overdispersion (Engel, 1984; Lawless, 1987). However the correctness of the variance–mean relationship is prerequisite for the validity of these parametric approaches. Less model-dependent methods such as the quasilikelihood methodology (Wedderburn, 1974; McCullagh and Nelder, 1989) have gained popularity over the years. However, the lack of an appropriate likelihood function prevents a proper representation of the likelihood of the interested parameter. As a result a full likelihood inference is unattainable.

Gould and Lawless (1988) showed that regression coefficients estimates in linear location-scale models remain consistent even if the error distribution is misspecified. Valid variance estimates could also be derived by the variance estimator proposed by Royall (1986). In this article robust inference for count data is the focus. The robust Poisson regression model (RPR) is proposed for the inference about regression parameters for more general count data, so that one need not worry about the correctness of the Poisson assumption. Unlike any other semi-parametric approaches the RPR model supplies an asymptotically legitimate likelihood function for the regression parameter. Therefore statistical tools that require a full specification of a likelihood function could be derived on the basis of the proposed robust likelihood function. Here the meaning of model misspecification is distinct from, for example, that of Lin and Wei (1989) in which they studied consistency and variance estimation of regression parameter estimates when some covariates are omitted in the systematic component in the Cox regression setting.

This paper is organized as follows. First the concept of robust profile likelihood functions that motivates the RPR model is briefly described in Section 2. In Section 3 the adjustment that makes the Poisson working model robust for the population mean parameter is given for independent and identically distributed (iid) count data. RPR models for the identity link and for the log link functions are presented in Sections 4 and 5, respectively. Section 6 includes simulation studies that demonstrate the effectiveness of the RPR model for data actually generated from non-Poisson distributions. Concluding remarks are given in Section 7.

2. Robust likelihood

Consider observations $y_1, y_2, \ldots, y_n$ which are regarded as realizations of independent random variables $Y_1, Y_2, \ldots, Y_n$. In addition, we have a working model for the prob-

ability distributions of Y_i's, $\{f_i(\bullet;\psi) = f(\bullet;\eta_i(\psi)), i = 1,\ldots,n, \psi \in \Psi\}$, where Ψ is a fixed-dimensional parameter space. For example, under normal regression settings, $\eta_i(\psi) = (x_i'\gamma, \sigma^2)$, $\psi = (\gamma, \sigma^2)$ and $f_i = f_i(y_i;\psi) = \exp\{-(y_i - x_i'\gamma)^2/2\sigma^2\}/\sqrt{2\pi}\sigma$. Here, x_i represents covariates that are specific to y_i. Note that this model is a collection of probability distributions, each of which is identified by an unique value of ψ.

In many cases, our interest lies in an one-dimensional (1D) characteristic of the distribution, such as the mean, or regression coefficient. In such cases, we will parameterize such problems in terms of the 1D parameter of interest, say, θ, and a finite-dimensional nuisance parameter, ϕ, so that $\psi' = (\theta, \phi')$. In order to make inference about θ, we need to eliminate the nuisance parameter ϕ from the likelihood function $L(\theta,\phi) = \prod_i f_i(y_i;\theta,\phi)$. Under special circumstances, we might have marginal, conditional, or partial likelihood available. Another approach is to use the profile likelihood that is more applicable.

Let $l(\theta,\phi) = \log L(\theta,\phi)$ be the log likelihood and $l(\theta,\widehat{\phi}(\theta))$ be the log profile likelihood of θ (Kalbfleisch and Sprott, 1970), where $\widehat{\phi}(\theta)$ is the maximum likelihood estimate (MLE) of ϕ for fixed θ. Under certain conditions the profile log likelihood may be used just like an ordinary log likelihood. For example the θ-value that maximizes $l(\theta,\widehat{\phi}(\theta))$ is the overall MLE, and an approximate confidence interval for θ could be obtained from the likelihood ratio test (LRT) based on the profile likelihood (McCullagh and Nelder, 1989). However, the failure of its derivative $\partial l(\theta,\widehat{\phi}(\theta))/\partial\theta$ to have zero expectation is considered the major reason for the MLE to be inconsistent, if the dimension of ϕ is a substantial fraction of n. Many authors have proposed adjustments or corrections to the profile likelihood to accommodate nuisance parameters to ensure the consistency of the MLE and the asymptotic normality (see, for example, Barndorff-Nielsen, 1983; Cox and Reid, 1987; McCullagh and Tibshirani, 1990).

The focus of the previously cited work is mainly on the reduction of the impact of nuisance parameters, *in the context of the assumed model*. Royall and Tsou (2003) proposed an adjustment to the working profile likelihood, so that the adjusted profile likelihood is an asymptotically valid representation of the likelihood of the interested parameter, even if the working model assumption fails. An important condition is required for the working model to be adjustable, in addition to the familiar regularity conditions that ensure the consistency and asymptotic normality of the MLE. In short it is assumed that the MLE, $\widehat{\theta}$, based on f, remains a consistent estimate of the parameter of interest under model misspecification. It is thus the working likelihood function that is to be adjusted, not the working model based MLE $\widehat{\theta}$ of the interested parameter.

Let θ_0 and ϕ_0 be the limiting values of the MLEs, $\widehat{\theta}$ and $\widehat{\phi}$, based on the working model f, when, in fact $y_1,\ldots,y_n$ are actually generated from, say, $\{h_i = h(\bullet;\tau_i(\theta,\lambda)), i = 1,\ldots,n\}$, where λ is the nuisance parameter (or fixed-dimensional nuisance parameters) under h. Here, it is implicitly assumed that the interpretation of θ under f remains the same as in h. This condition is what Royall and Tsou (2003) referred to as the first condition for robustness—the parameter of inference remains the parameter of interest even if the working model is incorrect. This condition is satisfied by the Poisson regression model which makes this working model adjustable.

Let l_θ and l_ϕ be the first derivatives of $l(\theta,\phi)$ with respect to θ and ϕ, respectively, whose derivatives with respect to ϕ are correspondingly denoted by $l_{\theta\phi}$ and $l_{\phi\phi}$. Now let $I_{h\theta\phi}$ and $I_{h\phi\phi}$ be the limiting values of $-l_{\theta\phi}/n$ and $-l_{\phi\phi}/n$, respectively, under h and the limits of

$-l_{\theta\theta}/n$ and $-l_{\phi\theta}/n$ are correspondingly denoted by $I_{h\theta\theta}$ and $I_{h\phi\theta}$, where these limiting values are all evaluated at θ_0 and ϕ_0. In this section regularity conditions on f and h are also assumed, such as $E_h l_\theta(\theta_0, \phi_0) = 0$ and those found in Huber (1981) or White (1982) that ensure the consistency and the asymptotic normality of the MLEs.

The robust log profile likelihood function proposed by Royall and Tsou (2003) takes the form

$$\frac{A}{B} l(\theta, \widehat{\phi}(\theta)), \tag{1}$$

where

$$A = I_{h\theta\theta} - I_{h\theta\phi} I_{h\phi\phi}^{-1} I_{h\phi\theta} \tag{2}$$

and

$$B = V_{h\theta\theta} - 2 I_{h\theta\phi} I_{h\phi\phi}^{-1} V_{h\phi\theta} + I_{h\theta\phi} I_{h\phi\phi}^{-1} V_{h\phi\phi} I_{h\phi\phi}^{-1} I_{h\phi\theta}. \tag{3}$$

Here $V_{h\theta\theta} = \lim_{n\to\infty} E_h[l_\theta(\theta_0, \phi_0)^2/n]$, $V_{h\theta\phi} = \lim_{n\to\infty} E_h[l_\theta(\theta_0, \phi_0) l_\phi(\theta_0, \phi_0)/n]$, and $V_{h\phi\phi} = \lim_{n\to\infty} E_h[l_\phi(\theta_0, \phi_0) l_\phi(\theta_0, \phi_0)'/n]$, where E_h indicates the expectation evaluated under h.

Now suppose that $\widehat{A}$ and $\widehat{B}$ are the consistent estimates of A and B, respectively. Then the adjusted log profile likelihood function $(\widehat{A}/\widehat{B}) l(\theta, \widehat{\phi}(\theta))$ is an asymptotically valid likelihood of θ. This robust log profile likelihood has two key important performance properties possessed by a correct likelihood function (Royall and Tsou, 2003). The first property is that the object of inference θ under the working model remains the object of interest, which ensures the consistency of the MLE of θ under the misspecified working model. The second property pertains to the same large sample approximation of the probabilities of misleading evidence, $P\{L_p(\theta)/L_p(\theta_0) \geqslant k\}$, $k > 1$, where L_p denotes the profile likelihood function based on the correct model, and that on the basis of the robust profile likelihood fucntion. Let Φ denote the standard normal distribution function and c be proportional to the distance between θ and θ_0, then the two probabilities of misleading evidence both approximate $\Phi\{-c/2 - \log(k)/c\}$.Therefore legitimate relative support between two hypothesized parameter values and support intervals for the interested parameter (Royall, 1997) could be obtained in large samples, so long as h has a finite second moment. It is emphasized that A and B are functions of θ_0, ϕ_0 and h. Nevertheless it will be demonstrated later that the two quantities could be estimated without knowing h.

In terms of hypothesis testing, for instance, the adjusted LRT statistic

$$2\frac{\widehat{A}}{\widehat{B}}[l(\widehat{\theta}, \widehat{\phi}(\widehat{\theta})) - l(\theta_0, \widehat{\phi}(\theta_0)]$$

is asymptotically χ_1^2 distributed under $\mathrm{H}_0 : \theta = \theta_0$, under model misspecification. In contrast the naive LRT $2[l(\widehat{\theta}, \widehat{\phi}(\widehat{\theta})) - l(\theta_0, \widehat{\phi}(\theta_0)]$ has an asymptotic χ_1^2 distribution only if the working model is correctly specified, namely, $f = h$. A simple application of Taylor's expansion indicates that $(A^2/B)^{-1}$ is the correct asymptotic variance of $\sqrt{n}\widehat{\theta}$ even if the Poisson assumption is wrong. Note, however, that the adjustment A/B is certainly less

useful if θ_0 does not carry the same interpretation of θ under f. In general the parameter of inference θ_0 is not the parameter of interest if $f \neq h$ (White, 1982).

The same adjustment was also derived by Stafford (1996) whose motivation, however, was to correct Bartlett's second identity under model misspecification. Similar adjustments were developed by Viraswami and Reid (1998) under the iid situations. The distributional properties of the adjusted LRT were studied in details therein. On the other hand the adjusted profile likelihood function by Royall and Tsou (2003) was developed in the context of the robust representation of statistical evidence, see also Royall (2000). Interested readers could consult Royall and Tsou (2003) for details along the line of robust representation of statistical evidence, and Stafford (1996) and Viraswami and Reid (1998) for the purpose of adjusting the chi-squared approximations.

3. IID count data

Consider iid observations $y_1, y_2, \ldots, y_n$ generated from distribution h. The likelihood function based on the Poisson model is

$$L = \theta^{n\overline{y}} \exp(-n\theta).$$

It is obvious that the MLE of θ based on the working model is simply the sample mean, $\overline{y}$, which is an unbiased and a consistent estimate of the population mean for general h. Despite the consistency of the model-based MLE, the function L is not a valid likelihood function for θ.

According to (2) and (3) factors A and B of the ratio A/B that makes the Poisson likelihood function robust are $A = 1/\theta_0$ and $B = \mathrm{Var}_h(Y)/\theta_0^2$. Here $\mathrm{Var}_h(Y)$ is the variance of Y under h. It is easy to see that A/B equals unity if the Poisson model is correct, since for Poisson data $\mathrm{Var}_h(Y) = \theta$, so that no correction is needed.

It is immediate to see that the adjustment A/B could be consistently estimated by replacing A and B with their empirical versions, $\widehat{A} = 1/\overline{y}$ and $\widehat{B} = \sum_i^n (y_i - \overline{y})^2/n\overline{y}^2$, so that

$$\widehat{A}/\widehat{B} = n\overline{y}/\sum_i^n (y_i - \overline{y})^2.$$

Obviously no knowledge regarding h is necessary in order to estimate $\widehat{A}$ and $\widehat{B}$, except for the assumption that h has a finite second moment. The adjusted robust Poisson likelihood is then

$$\{\theta^{n\overline{y}} \exp(-n\theta)\}^{n\overline{y}/\sum_i^n (y_i-\overline{y})^2}.$$

The resulting adjusted likelihood function is an asymptotically legitimate likelihood for θ, not only for Poisson data but for general h that has a finite variance. Later simulation studies will be given to demonstrate that, with large samples, the adjusted LRT retains the nominal type I error probability and provides confidence intervals with the nominal coverage probability, even when h is not Poisson.

4. Robust Poisson regression—identity link

Consider count data $y_1, y_2, \ldots, y_n$ which are regarded as realizations of independent random variables $Y_1, Y_2, \ldots, Y_n$. Under multiple linear regression settings each Y_i is assumed linearly related to its own covariate vector $x_i = (x_{i0}, x_{i1}.x_{i2}, \ldots, x_{ip-1})'$, $x_{i0} = 1$, so that

$$E(Y_i) = \mu_i = \gamma_0 x_{i0} + \gamma_1 x_{i1} + \cdots + \gamma_{p-2} x_{ip-2} + \beta x_{ip-1}.$$

Suppose β is the parameter of interest and other parameters are considered nuisance parameters. For convenience let $(\gamma_0, \ldots, \gamma_{p-2})'$ be denoted by ϕ.

The log likelihood function contributed by y_i, under a Poisson working model, is

$$\begin{aligned} l_i(\beta, \phi) = {} & y_i \log(\gamma_0 + \cdots + \gamma_{p-2} x_{ip-2} + \beta x_{ip-1}) \\ & - (\gamma_0 + \cdots + \gamma_{p-2} x_{ip-2} + \beta x_{ip-1}). \end{aligned}$$

The MLEs of the regression parameters are solutions of the system of estimating equations

$$\sum_{i=1}^{n} x_{ij-1} \left(\frac{y_i - \mu_i}{\mu_i} \right) = 0, \quad j = 1, \ldots, p. \tag{4}$$

The solutions of (4) are the maximum quasilikelihood (MQL) estimators (Wedderburn, 1974; McCullagh and Nelder, 1989; Heyde, 1997) or M-estimators (Huber, 1981) when the Y_i's are not necessarily generated from Poisson distributions. McCullagh (1983) showed that under mild regularity conditions, the consistency of the MQL estimators under model misspecification depends only upon the correct specification of the regression (see also Fahrmeir, 1990). In other words, the Poisson regression model provides consistent estimates of regression parameters for more general count data. Hence the first condition of robustness of Royall and Tsou (2003) is satisfied. Thus the Poisson working likelihood function could be adjusted to become asymptotically legitimate for regression parameters.

After lengthy derivation the elements A and B of A/B, denoted by A_e and B_e, that makes the Poisson likelihood robust, under the identity link function, are

$$A_e = \lim_{n\to\infty} \sum_{i}^{n} \frac{1}{\mu_{i,0}} \left(x_{ip-1} - \sum_{j=1}^{p-1} \frac{|\Delta_{ej}|}{|\Delta_e|} x_{ij-1} \right)^2 \Bigg/ n$$

and

$$B_e = \lim_{n\to\infty} \sum_{i}^{n} \frac{\mathrm{Var}_h(Y_i)}{\mu_{i,0}^2} \left(x_{ip-1} - \sum_{j=1}^{p-1} \frac{|\Delta_{ej}|}{|\Delta_e|} x_{ij-1} \right)^2 \Bigg/ n,$$

respectively, where Δ_e is the $(p-1) \times (p-1)$ matrix with

$$\left(\sum_{i}^{n} x_{i0} x_{ij-1}/\mu_{i,0}, \sum_{i}^{n} x_{i1} x_{ij-1}/\mu_{i,0}, \ldots, \sum_{i}^{n} x_{ip-3} x_{ij-1}/\mu_{i,0}, \sum_{i}^{n} x_{ip-2} x_{ij-1}/\mu_{i,0} \right)'$$

as the jth column, $j = 1, \ldots, p-1$, and Δ_{ej}, $j = 1, \ldots, p-1$, is the $(p-1) \times (p-1)$ matrix obtained from Δ_e by replacing the jth column of Δ_e, namely,

$$\left(\sum_i^n x_{i0} x_{ij-1}/\mu_{i,0}, \sum_i^n x_{i1} x_{ij-1}/\mu_{i,0}, \ldots, \sum_i^n x_{ip-3} x_{ij-1}/\mu_{i,0}, \sum_i^n x_{ip-2} x_{ij-1}/\mu_{i,0} \right)'$$

with

$$\left(\sum_i^n x_{i0} x_{ip-1}/\mu_{i,0}, \sum_i^n x_{i1} x_{p-1}/\mu_{i,0}, \ldots, \sum_i^n x_{ip-3} x_{ip-1}/\mu_{i,0}, \sum_i^n x_{ip-2} x_{ip-1}/\mu_{i,0} \right)'.$$

Here $\mu_{i,0}$ is the true mean value of Y_i under h. Notice that Δ_{ej} and Δ_e can be estimated by replacing $\mu_{i,0}$ by its MLE $\widehat{\mu}_i$, and consistent estimates $\widehat{A}_e$ and $\widehat{B}_e$ for A_e and B_e can then be obtained by further estimating $\mathrm{Var}_h(Y_i)$ by $(y_i - \widehat{\mu}_i)^2$. It should be stressed here that the knowledge of h is not required in order to estimate A and B. The consistent estimate $\widehat{\mu}_i$ and the resultant estimate $(y_i - \widehat{\mu}_i)^2$ of $\mathrm{Var}_h(Y_i)$,in general, warrant the consistency of both $\widehat{A}_e$ and $\widehat{B}_e$.

For the special case of a simple regression model $\mu_i = \gamma_0 + \beta x_i$,

$$A_e = \lim_{n\to\infty} \sum_i^n \frac{1}{\mu_{i,0}} (x_i - \widehat{x})^2/n,$$

$$B_e = \lim_{n\to\infty} \sum_i^n \frac{\mathrm{Var}_h(Y_i)}{\mu_{i,0}^2} (x_i - \widehat{x})^2/n,$$

where $\widehat{x} = \left(\sum_i^n x_i/\mu_{i,0}\right) / \left(\sum_i^n 1/\mu_{i,0}\right)$.

5. Robust Poisson regression—log link

With the log link function $\log \mu_i = \gamma_0 + \cdots + \gamma_{p-2} x_{ip-2} + \beta x_{ip-1}$, the log likelihood function based on the Poisson regression model is

$$l_i(\beta, \phi) = y_i(\gamma_0 + \cdots + \gamma_{p-2} x_{ip-2} + \beta x_{ip-1}) - \exp(\gamma_0 + \cdots + \gamma_{p-2} x_{ip-2} + \beta x_{ip-1}).$$

Now the MLEs of the regression parameters are solutions of the system of estimating equations

$$\frac{\partial \sum_i^n l_i}{\partial \gamma_{j-1}} = \sum_i^n x_{ij-1}(y_i - \mu_i). \tag{5}$$

Similarly, the solutions of (5) are MQL estimators as well. The consistency of the MLEs under model misspecification is, thus, preserved.

Under the log-linear specification the elements A and B of A/B, denoted by A_l and B_l, that makes the Poisson likelihood robust are

$$A_l = \lim_{n\to\infty} \sum_i^n \mu_{i,0} \left(x_{ip-1} - \sum_{j=1}^{p-1} \frac{|\Delta_{lj}|}{|\Delta_l|} x_{ij-1} \right)^2 \Big/ n$$

and

$$B_l = \lim_{n\to\infty} \sum_i^n \mathrm{Var}_h(Y_i) \left(x_{ip-1} - \sum_{j=1}^{p-1} \frac{|\Delta_{lj}|}{|\Delta_l|} x_{ij-1} \right)^2 \Big/ n,$$

where Δ_l is the $(p-1)\times(p-1)$ matrix with

$$\left(\sum_i^n x_{i0}x_{ij-1}\mu_{i,0}, \sum_i^n x_{i1}x_{ij-1}\mu_{i,0}, \ldots, \sum_i^n x_{ip-3}x_{ij-1}\mu_{i,0}, \sum_i^n x_{ip-2}x_{ij-1}\mu_{i,0} \right)',$$

as the jth column, $j = 1, \ldots, p-1$, and Δ_{lj}, $j = 1, \ldots, p-1$, is the $(p-1)\times(p-1)$ matrix obtained from Δ_l by replacing the jth column of Δ_l, namely,

$$\left(\sum_i^n x_{i0}x_{ij-1}\mu_{i,0}, \sum_i^n x_{i1}x_{ij-1}\mu_{i,0}, \ldots, \sum_i^n x_{ip-3}x_{ij-1}\mu_{i,0}, \sum_i^n x_{ip-2}x_{ij-1}\mu_{i,0} \right)'$$

with

$$\left(\sum_i^n x_{i0}x_{ip-1}\mu_{i,0}, \sum_i^n x_{i1}x_{p-1}\mu_{i,0}, \ldots, \sum_i^n x_{ip-3}x_{ip-1}\mu_{i,0}, \sum_i^n x_{ip-2}x_{ip-1}\mu_{i,0} \right)'.$$

Likewise Δ_{lj} and Δ_l can be estimated by replacing $\mu_{i,0}$ with its MLE $\widehat{\mu}_i$, and consistent estimates $\widehat{A}_l$, $\widehat{B}_l$ for A_l, B_l can then be obtained by further substituting $(y_i - \widehat{\mu}_i)^2$ for $\mathrm{Var}_h(Y_i)$.

Under a simple regression model $\log \mu_i = \gamma_0 + \beta x_i$

$$A_l = \lim_{n\to\infty} \sum_i^n \mu_{i,0}(x_i - \widetilde{x})^2/n,$$

$$B_l = \lim_{n\to\infty} \sum_i^n \mathrm{Var}_h(Y_i)(x_i - \widetilde{x})^2/n,$$

where $\widetilde{x} = \left(\sum_i^n x_i\mu_{i,0}\right) / \left(\sum_i^n \mu_{i,0}\right)$.

6. Simulation studies

Simulation studies are carried out to examine the effectiveness of the adjusted RPR model. The performance of the adjusted LRT is specifically considered. The iid case is

considered first. Two types of iid count data are generated. First, consider data generated from a negative-binomial distribution. Let p be the success probability and y be the total number of failures before r successes are observed, then $E(Y) = \theta = r(1-p)/p$. The parameter r is set as 2 and mean values considered include $H_0 : \theta = 6$ and $H_1 : \theta = 7$. The combination of $r = 2$ and $\theta = 6$ gives the variance of Y the value of 24, and 31.5 when $\theta = 7$. Another type of count data are generated by first letting the interarrival failure time follow a Weibull distribution. Then the number of failures observed within the time period (0, 10) is recorded as the count. Two sets of parameter values are chosen for the Weibull distribution, and the corresponding theoretical mean counts are 6.16 and 6.46 that are designated as H_0 and H_1, respectively. The corresponding variance is 0.90 under H_0 and is 0.85 under H_1. Sample sizes $n = 50$ and 100 are considered separately for two sources of count data stated above.

Two thousand data sets are generated for each sample size. The empirical type I error probability, α, is calculated as the proportion of rejecting the null hypothesis H_0, if twice the adjusted log likelihood ratio is greater than 3.84, so that the nominal type I error probability is 0.05. The empirical power of the test is defined as the proportion of rejecting H_0 when data are actually generated under H_1. The coverage probability, denoted by cp, of the adjusted LRT-based confidence interval with 95% nominal coverage probability is also computed.

Tables 1 and 2 display the average of $\widehat{\theta}$ (m), the median of $\widehat{\theta}$ (med), and the sample variance of $\widehat{\theta}(s^2)$. Also included is the average of the 2000 variance estimates of $\widehat{\theta}$ (adjvar), which is based on $(\widehat{A}^2/\widehat{B})^{-1}/n$.

In the regression settings, without loss of generality a simple regression model is considered. Independent samples are generated, respectively, from negative binomial distributions. The negative binomial datum y_i is generated in the following manner. Recall that if p_i is the success probability and y_i is the total number of failures before three successes are observed, then $E(Y_i) = 3(1-p_i)/p_i$. When the mean value of Y_i is linearly related to the covariate x_i, that is, $E(Y_i) = \gamma + \beta x_i$, the parameter value of the intercept γ is set as 1, and $\beta = 2$ and 3 are used as H_0 and H_1, respectively. Under the log-linear specification $\log E(Y_i) = \gamma + \beta x_i$, the value of γ is set as 1, and $\beta = 1$ and 1.3 are selected as H_0 and H_1, respectively. The covariate x is generated from an uniform distribution on the interval (1, 4).

Table 1
Negative binomial ($r = 2$)

		$m(\widehat{\theta})$	$\text{med}(\widehat{\theta})$	$s^2(\widehat{\theta})$	$\text{adjvar}(\widehat{\theta})$	α	cp
$n = 50$	$H_0 : \theta_0 = 6$	6.017	6.000	0.484	0.474	0.068	0.937
	$H_1 : \theta_0 = 7$	6.999	6.980	0.646	0.620	0.275	0.941
$n = 100$	$H_0 : \theta_0 = 6$	5.984	5.980	0.240	0.238	0.053	0.953
	$H_1 : \theta_0 = 7$	7.024	7.020	0.305	0.314	0.486	0.949

Table 2
Interarrival time = Weibull

		$m(\widehat{\theta})$	med$(\widehat{\theta})$	$s^2(\widehat{\theta})$	adjvar$(\widehat{\theta})$	α	cp
$n=50$	$H_0:\theta_0=6.160$	6.164	6.160	0.016	0.016	0.049	0.970
	$H_1:\theta_0=6.470$	6.462	6.460	0.017	0.017	0.652	0.960
$n=100$	$H_0:\theta_0=6.160$	6.167	6.170	0.009	0.008	0.052	0.963
	$H_1:\theta_0=6.460$	6.463	6.460	0.009	0.009	0.915	0.966

Table 3
y ~negative binomial, $E(y)=1+\beta x$

		$m(\widehat{\beta})$	med$(\widehat{\beta})$	$s^2(\widehat{\beta})$	adjvar$(\widehat{\beta})$	α	cp
$n=50$	$H_0:\beta=2$	2.016	2.014	0.621	0.584	0.072	0.932
	$H_1:\beta=3$	2.978	2.991	1.054	1.021	0.154	0.928
$n=100$	$H_0:\beta=2$	1.979	1.978	0.284	0.332	0.077	0.936
	$H_1:\beta=3$	2.994	3.008	0.559	0.555	0.267	0.947
$n=500$	$H_0:\beta=2$	1.996	2.004	0.064	0.063	0.055	0.958
	$H_1:\beta=3$	2.994	2.997	0.110	0.109	0.849	0.956

Table 4
y ~negative binomial, $E(y)=\exp(1+\beta x)$

		$m(\widehat{\beta})$	med$(\widehat{\beta})$	$s^2(\widehat{\beta})$	adjvar$(\widehat{\beta})$	α	cp
$n=50$	$H_0:\beta=1$	0.992	0.996	0.027	0.022	0.095	0.926
	$H_1:\beta=1.3$	1.298	1.296	0.029	0.025	0.534	0.921
$n=100$	$H_0:\beta=1$	1.000	0.992	0.013	0.012	0.076	0.949
	$H_1:\beta=1.3$	1.299	1.298	0.018	0.017	0.681	0.939
$n=500$	$H_0:\beta=1$	1.001	0.999	0.002	0.002	0.066	0.966
	$H_1:\beta=1.3$	1.299	1.296	0.004	0.003	0.999	0.967

Two thousand data sets are generated for sample sizes $n=50$, 100 and 500, respectively. Empirical probabilities α and cp are similarly defined as in the iid case. Tables 3 and 4 also display the average of $\widehat{\beta}$ (m), the median of $\widehat{\beta}$ (med), and the sample variance of $\widehat{\beta}$ (s^2). Also included is the average of the 2000 variance estimates of $\widehat{\beta}$ (adjvar), which is based on $(\widehat{A}^2/\widehat{B})^{-1}/n$.

It is not surprising to see that the distribution of $\widehat{\theta}$ is symmetric, by the similarity between $m(\widehat{\theta})$ and med$(\widehat{\theta})$, that is attested by the central limit theorem. The variance of $\widehat{\theta}=\bar{y}$ is correctly estimated by utilizing the adjusted Poisson likelihood function. Without a proper

correction, it would be overestimated more than 10 times higher for negative-binomial data, and more than 100 times higher if the interarrival time is Weibull (results not shown). The simulation with negative-binomial data also shows that the empirical type I error probability derived from the naive LRT is 0.311 for $n=50$, and 0.335 for $n=100$. When the interarrival time is Weibull the corresponding empirical type I error probabilities are all 0.000 for $n=50$ and for $n=100$. The biases of employing the unadjusted likelihood are partly reflected in the biases of the results from the naive LRT. The biases are obviously dominated by the factor A/B. So long as there is extra-Poisson variation, the bias persists regardless of the sample size. The adjusted likelihood, on the other hand, has successfuly reduced the biases to the order of $O(n^{-1/2})$ instead of $O(1)$ (see Viraswami and Reid (1998) for details). Obviously, the adjusted Poisson likelihood provides valid inference for count data not necessarily generated from a Poisson distribution.

Tables 3 and 4 clearly demonstrate that, with large samples the adjusted Poisson likelihood function provides empirical probabilities that are close to the nominal levels. The naive Poisson regression model produces the following results. Simulations for sample sizes $n = 50$, 100 and 500 show that type I error probabilities from the unadjusted model are 0.348, 0.365 and 0.328, respectively, under $H_0 : \beta = 2$ for the identity link function. With the log link function, the unadjusted empirical type I error probabilities are 0.724, 0.698 and 0.721, respectively, under $H_0 : \beta = 1$. It is clearly revealed that the type I error probability based on the naive Poisson regression is seriously inflated if the working model assumption is incorrect.

On the other hand, the RPR model furnishes a likelihood ratio statistic which performs remarkably well on non-Poisson data, with identity link and log link alike. Nominal levels are evidently better approximated when sample size increases.

7. Conclusions

The Poisson regression model has been integrated in many statistical softwares (SAS Institute Inc., 1999; EGRET, 1999) as a standard procedure for the analysis of count data. It makes the analysis fast and easy to carry out. As demonstrated here, the crucial Chi-square approximation of the likelihood ratio test collapses, when data are generated from non-Poisson distributions.

Methods of dealing with extra-Poisson variation in the regression situation have been proposed. For example, McCullagh and Nelder (1989) considered the relationship $\mathrm{Var}(Y_i) \propto E(Y_i)$, and Armitage (1957) focused on the case when $\mathrm{Var}(Y_i) \propto E(Y_i)^b$. Lawless (1987) proposed the negative-binomial model for the mixed Poisson situation. However the mean–variance relationships need to be correctly specified for their inferences to be valid.

The proposed robust Poisson regression approach warrants asymptotically valid inference so long as the true underlying distributions have finite second moments. Unlike aforementioned methods whose validity depends heavily on the correct specification of the variance–mean relationship, the robustness property is achieved by the proposed method more generally. As a tradeoff between efficiency and robustness this novel robust approach might not be as efficient as those mentioned above when the variance and mean relationship is correctly modeled. Yet asymptotically valid likelihood functions for regression param-

eters are provided that none of the semi-parametric approaches could deliver. Breslow (1990) developed an empirical score test for evaluating the significance of added variables in the regression setting for overdispersed count data. With large samples the proposed test statistic that permits the misspecification of the variance function retains nominal probabilities . However as another semi-paramteric alternative, still there is no likelihood function available for the parameter of interest.

The negative-binomial regression model can also be made robust when properly adjusted. The adjustments for multiple linear regression situation can be found in appendix. Under a simple regression scenario,

$$A_b = r_0 \lim_{n\to\infty} \sum_{i=1}^{n} \frac{1}{\mu_{i,0}(\mu_{i,0}+r_0)}(x_i - \widehat{x})^2/n,$$

$$B_b = r_0^2 \lim_{n\to\infty} \sum_{i=1}^{n} \frac{\mathrm{Var}_h(Y_i)}{[\mu_{i,0}(\mu_{i,0}+r_0)]^2}(x_i - \widehat{x})^2/n,$$

where $\widehat{x} = \left(\sum_{i=1}^{n} x_i/[\mu_{i,0}(\mu_{i,0}+r_0)]\right) / \left(\sum_{i=1}^{n} 1/[\mu_{i,0}(\mu_{i,0}+r_0)]\right)$, and r_0 is the limiting value of the MLE of r, the index parameter of negative-binomial model. The adjusted negative-binomial regression model provides asymptotically valid inference for regression parameters, just as the RPR model does. However, the need to estimate the nuisance parameter r make it less desirable compared with the RPR for which only regression parameters are present.

The adjusted Poisson regression model proposed here is a parametric robust approach against model misspecification. The new methodology is a more compelling tool for inference. For example with its parametric nature, a valid likelihood function is available. One could therefore carry out the robust LRT and the robust goodness of fit test that would be difficult for other methodology, such as GEE, to perform.

To generate robust testing procedures for a vector, say θ, of length $q > 1$ of the regression parameters, one would need to calculate $q \times q$ matrices A and B according to (2) and

$$B = V_{h\theta\theta} - I_{h\theta\phi}I_{h\phi\phi}^{-1}V_{h\phi\theta} - V_{h\theta\phi}I_{h\phi\phi}^{-1}I_{h\phi\theta} + I_{h\theta\phi}I_{h\phi\phi}^{-1}V_{h\phi\phi}I_{h\phi\phi}^{-1}I_{h\phi\theta},$$

respectively. A direct adjustment to the likelihood function itself is, however, not feasible. Nonetheless the asymptotic equivalence of the robust LRT, namely, the robust score test and the robust Wald test could be easily obtained. For example, the adjusted correct covariance matrix of the score function is B instead of the naive model-based matrix A. The robust Wald test could be constructed by properly normalizing the MLE of θ with the robust covariance matrix $A^{-1}BA^{-1}$ rather than A^{-1} which is legitimate only when the model assumption is correct.

Acknowledgements

This work is partially supported by Grant 91-H-FA07-1-4 of the Ministry of Education and Grant NSC-89-2118-M-008–003 of the National Science Council, Taiwan, ROC.

Appendix

Factors A_b and B_b of the adjustment A_b/B_b for the negative-binomial regression model are

$$A_b = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{r_0}{\mu_{i,0}(r_0+\mu_{i,0})} \left(x_{ip-1} - \sum_{j=1}^{p-1} \frac{|\,\Delta_{bj}|}{|\,\Delta_b|} x_{ij-1} \right)^2 \Big/ n$$

and

$$B_b = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{r_0^2 \operatorname{Var}_h(Y_i)}{\mu_{i,0}^2(r_0+\mu_{i,0})^2} \left(x_{ip-1} - \sum_{j=1}^{p-1} \frac{|\,\Delta_{bj}|}{|\,\Delta_b|} x_{ij-1} \right)^2 \Big/ n,$$

where Δ_b is the $(p-1)\times(p-1)$ matrix with

$$\left(\sum_{i=1}^{n} x_{i0}x_{ij-1}/[\mu_{i,0}(r_0+\mu_{i,0})], \sum_{i=1}^{n} x_{i1}x_{ij-1}/[\mu_{i,0}(r_0+\mu_{i,0})], \ldots, \sum_{i=1}^{n} x_{ip-2}x_{ij-1}/[\mu_{i,0}(r_0+\mu_{i,0})] \right)'$$

as the jth column, $j=1,\ldots,p-1$, and Δ_{bj}, $j=1,\ldots,p-1$, is the $(p-1)\times(p-1)$ matrix obtained from Δ_b by replacing the jth column of Δ_b, namely,

$$\left(\sum_{i=1}^{n} x_{i0}x_{ij-1}/[\mu_{i,0}(r_0+\mu_{i,0})], \sum_{i=1}^{n} x_{i1}x_{ij-1}/[\mu_{i,0}(r_0+\mu_{i,0})], \ldots, \sum_{i=1}^{n} x_{ip-2}x_{ij-1}/[\mu_{i,0}(r_0+\mu_{i,0})] \right)',$$

by

$$\left(\sum_{i=1}^{n} x_{i0}x_{ip-1}/[\mu_{i,0}(r_0+\mu_{i,0})], \sum_{i=1}^{n} x_{i1}x_{p-1}/[\mu_{i,0}(r_0+\mu_{i,0})], \ldots, \sum_{i=1}^{n} x_{ip-2}x_{ip-1}/[\mu_{i,0}(r_0+\mu_{i,0})] \right)'.$$

References

Agresti, A., 1996. An Introduction to Categorical Data Analysis. Wiley, New York.

Armitage, P., 1957. Studies in the variability of Pock counts. J. Hygiene Camb. 55, 564–581.

Barndorff-Nielsen, O., 1983. On a formula for the distribution of the maximum likelihood estimator. Biometrika 70, 343–365.

Breslow, N., 1990. Tests of hypotheses in overdispersed Poisson regression and other quasi-likeihood models. J. Amer. Statist. Assoc. 85, 565–571.

Cox, D.R., 1983. Some remarks on over-dispersion. Biometrika 70, 269–274.
Cox, D.R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. J. Roy. Statist. Soc. B 49, 1–39.
EGRET, 1999. EGRET for Windows version 2.0, Cytel software corporation, USA.
Engel, J., 1984. Models for response data showing extra-Poisson variation. Statist. Neerlandica 38, 159–167.
Fahrmeir, L., 1990. Maximum likelihood estimation in misspecified generalized linear models. Statistics 21, 487–502.
Frome, E.L., 1983. The analysis of rates using Poisson regression models. Biometrics 39, 665–674.
Gould, A., Lawless, J.F., 1988. Consistency and efficiency of regression coefficient estimates in location-scale models. Biometrika 75, 535–540.
Heyde, C.C., 1997. Quasi-likelihood and its Application. Springer, New York.
Holford, T.R., 1983. The estimation of age, period and cohort effects for vital rates. Biometrics 39, 311–324.
Huber, P.J., 1981. Robust Statistics. Wiley, NewYork.
Kalbfleisch, J.D., Sprott, D.A., 1970. Application of likelihood methods to models involving large numbers of parameters (with discussion). J. Roy. Statist. Soc. B 32, 175–208.
Lawless, J.F., 1987. Negative binomial and mixed Poisson regression. Canad. J. Statist. 15, 209–225.
Lin, D.Y., Wei, L.J., 1989. The robust inference for the Cox proportional hazards model. J. Amer. Statist. Assoc. 84, 1074–1078.
McCullagh, P., 1983. Quasi-likelihood functions. Ann. Statist. 11, 59–67.
McCullagh, P., Nelder, J.A., 1989. Generalized Liner Models. second ed. Chapman & Hall, New York.
McCullagh, P., Tibshirani, R., 1990. A simple method for the adjustment of profile likelihoods. J. Roy. Statist. Soc. B 52, 325–344.
Paul, S.R., Plackett, R.L., 1978. Inference sensitivity for Poisson mixtures. Biometrika 65, 591–602.
Royall, R.M., 1986. Model robust confidence intervals using maximum likelihood estimators. Internat. Statist. Rev. 54, 221–226.
Royall, R.M., 1997. Statistical Evidence—a Likelihood Paraduam. Chapman & Hall, New York.
Royall, R.M., 2000. On the probability of observing misleading statistical evidence (with discussion). J. Amer. Statist. Assoc. 95, 760–780.
Royall, R.M., Tsou, T.-S., 2003. Interpreting statistical evidence using imperfect models: robust adjusted likelihood functions. J. Roy. Statist. Soc. B 65, 391–404.
SAS Institute Inc., 1999. SAS Software Release 8.2. SAS Institute, Cary, NC, USA.
Stafford, J.E., 1996. A robust adjustment of the profile likelihood. Ann. Statist. 24, 336–352.
Viraswami, K., Reid, N., 1998. A note on the likelihood-ratio statistic under model misspecification. Canad. J. Statist. 26, 161–168.
Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika 61, 439–447.
White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica 50, 1–25.