A Note on Generalized Poisson Regression Models

K. P. Singh[a], **J. T. Wulu**[b] **and A. A. Bartolucci**[c]

[a] *Department of Epidemiology and Biostatistics, School of Public Health, University of North Texas Health Science Center at Fort Worth, Fort Worth, Texas 76107, USA (ksingh@hsc.unt.edu)*

[b] *Department of Health and Human Services, Bureau of Primary Health Care, Bethesda, MD 20814, USA (JWulu@hrsa.gov)*

[c] *Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA (albartol@uab.edu)*

Abstract: Many applications in medicine, public health, engineering, and biological sciences have discrete responses (not necessarily binary) and/or occurrences that are rare or exceptional and follow the Poisson distribution. However, such responses are over-dispersed or under-dispersed. For such situations generalized Poisson regression models (GPR) is more appropriate. In this paper, assumptions and properties of GPR are explored from a public health point. An application of GPR is illustrated and applied to a data set. Poisson regression (PR) and GPR models are used to identify relationships between covariates and the response variable. The Wald t test is used for testing the significance of each regression parameter. Based on the dispersion parameter and the goodness-of-fit measure for each data, the GPR model performs better than the PR model.

Keywords: Poisson Regression; Over-Dispersion; Under-dispersion; Maximum Likelihood Estimation; Dispersion

1. INTRODUCTION

In numerous scientific experiments, frequently, the response or dependent variable is a count generated by processes in which the number of incidents is due to a rare or chance event, and that rare or chance event obeys the principle of randomness. In such cases, the type of data can be fitted by a Poisson model. In theory, data of the Poisson distribution should have its mean equal to its variance. However, in reality, data arising from groups or individuals (e.g., study populations) are statistically dependent within a group, so the observed variance of the data may be larger or smaller than the corresponding mean.

There are a number of approaches to dealing with count data, or data arising from accumulated or aggregated binomial (or multinomial) trials. The more familiar is the Poisson regression (PR) model, which has been commonly used for the analysis of cell frequencies in cross-tables, and somewhat less for "event-count" kinds of studies, and a modification of it to deal with over-dispersion may involve the negative binomial regression model. But, the generalized Poisson regression (GPR) model has shown statistical advantages over standard Poisson regression, negative binomial regression, generalized negative binomial regression, and generalized linear models in the event of fitting count data that may be over-dispersed or under-dispersed or equi-dispersed. The GPR model provides a versatile approach for analyzing count random variables and their relationships to other variables or covariates.

Consul [1989] presented pioneering work on a generalization of Poisson distribution. A number of

studies have also suggested various models to deal with *extra-Poisson* variation data [Manton *et al.*, 1981; Cox, 1983; Efron, 1986; Lawless 1987; Stein and Juritz, 1988; Breslow, 1990; Singh and Famoye, 1993;]. Approaches and models for analyzing over-dispersed Poisson data and Poisson rates include generalized linear models [Nelder and Wedderburn, 1972]; asymmetric maximum likelihood methods and methods using double-exponential families [Efron, 1986]; and Bayesian overdispersed models and quasi-likelihood methods [Lu and Morris, 1994].

Singh and Famoye [1993] used and highly recommended the GPR model instead of the PR model in their analysis of life table and follow-up data. They indicated that the PR model was not appropriate to analyze an extra-Poisson variation survival data set using a Poisson regression model. applied the GPR model to study the relation between chromosome aberrations and radiation dose in human lymphocytes.

2. THE GENERALIZED POISSON REGRESSION MODEL

Let Y_i be a count response variable that follows a generalized Poisson distribution. Then the mean and variance of Y_i are given by:

$$\mathrm{E}(Y_i \mid x_i) = \mu_i , \tag{1}$$

and

$$\mathrm{V}(Y_i \mid x_i) = \mu_i (1 + \alpha\mu_i)^2 \tag{2}$$

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

x_i is a (k-1) dimensional vector of covariates, β is a k-dimensional vector of regression parameters and $y_i = 0,\ 1,\ 2,\ \ldots$

The generalized Poisson regression model is a generalization of the standard Poisson regression (PR) model. When $\alpha = 0$ the probability function reduces to the PR model. Within the framework of PR model, the equality constraint is observed between the conditional mean $\mathrm{E}(Y_i \mid x_i)$ and the conditional variance $\mathrm{V}(Y_i \mid x_i)$ of the dependent variable for each observation. In practical applications, this assumption is often untrue since the variance can either be larger or smaller than the mean. If the variance is not equal to the mean, the estimates in PR model are still consistent but are inefficient, which leads to the invalidation of inference based on the estimated standard errors.

When $\alpha > 0$, the GPR model represents count data with over-dispersion and when $\alpha < 0$, the GPR model represents count data with under-dispersion. In (1), α is called the dispersion parameter and it can be estimated along with the regression coefficients in the GPR model.

The estimates of α and β in the GPR model (1) are obtained by using the method of maximum likelihood. The log-likelihood function of the GPR model and partial derivatives are given in Wulu [1999].

The initial estimate of α can be taken to be zero. Alternatively, an initial α can be obtained by equating the chi-square statistic to its degrees of freedom. This is given by

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\mathrm{V}(Y_i \mid x_i)} = n - k, \tag{3}$$

where $\mathrm{V}(Y_i \mid x_i)$ is given by (2).

3. GOODNESS OF FIT AND TEST FOR DISPERSION

The goodness-of-fit of GPR model can be based on the deviance statistic. The deviance statistic can be approximated by a chi-square distribution if the n_i's are large. We use the log-likelihood value to measure the goodness-of-fit of the regression models. The regression model with a larger log-likelihood value is better than the one with a smaller value.

The GPR model reduces to the PR model when α = 0. To assess the adequacy of the GPR model over the PR model, we test the hypothesis

$$H_0 : \alpha = 0 \ \text{ against } \ H_a : \alpha \neq 0. \tag{4}$$

This tests for the significance of the dispersion parameter. Whenever H_0 is rejected, one should use the GPR in place of the PR model for the given data. To carry out the test in (4), one can use the asymptotically normal Wald type "t" statistic defined as the ratio of the estimate of α to its standard error. An alternative test for the hypothesis in (4) is to use

the likelihood ratio test statistic, which α is approximately chi-square distributed with one degree of freedom when the null hypothesis is true.

4. APPLICATION

The data were taken from a home interview survey of 1095 U.S. households conducted in June 1978 by Cambridge Systematics, Inc. for the National Science Foundation [CSI, 1979]. The data were collected for each trip taken by all members of the sampled households during the 24 hours prior to the date of the actual interview. Terza and Wilson [1990] analyzed the CSI data, which identified five possible trips by destination such as, work, shopping, personal business, school, and social. They indicated that the frequency of total trips taken by households varies from 0 to 44. They also observed that trips to home were excluded, since this type of trip will necessarily be taken whenever at least one other type of trip is taken.

No attempt was made to deal with the issue of *trip* chaining, whereby travelers visit a number of destinations before returning home. After deleting observations containing missing values, 577 observations remain. Thus, the sample data is comprised of 577 households of whom 490 owns at least one vehicle. From the CSI data, Terza [1998] defined a count response variable and ten independent variables (or covariates).

The assumptions are as follows: (i) that unobservable factors relating to household members' taste for public transit in the trip frequency regression may be correlated with vehicle ownership; (ii) household that are positively disposed towards the use of public transportation will have higher trip frequencies because their transit mode options are less constrained such a positive attitude towards public transit may be a component of an overall favorable view of intra-metropolitan travel motivated by a desire to avail one self of various urban amenities (regardless of the mode of transportation); (iii) if this is the case, ownership of vehicle by a household and the unobservable in the trip frequency regression will be positively correlated; (iv) if one's affinity for public transit is primarily motivated by disdain for the adverse aspects of private modes of transportation (e.g., traffic congestion, parking problems, etc.) then correlation between vehicle ownership and the unobservable component of the regression model will be negative.

Terza and Wilson [1990] indicated that the frequency of total trips taken by households varies from 0 to 44. The means and variances of trips are shown (in brackets): work (1.02, 2.30); shopping (1.05, 4.04); personal business (0.55, 1.92); school (0.21, 0.51); and social (1.72, 7.43). They also observed that trips to home were excluded, since this type of trip will necessarily be taken whenever at least one other type of trip is taken.

Significant parameter estimates were discussed by Terza and Wilson [1990]. For example, they found that (i) number of household members takes a positive sign in the Poisson section of the models, and in the logit section in the case of shopping and school trips; (ii) holding the number of adults constant and increasing household members (i.e., adding children), increases the number of trips taken, and increases the odds of trips being taken for shopping or school purposes; (iii) an increase in the number of individuals in households tends to decrease the number of work and personal business trips relative to social trips; also, an increase in the number of individuals in households tends to increase the total number of trips taken; and (iv) it is conceivable that frequencies of different types of trips could vary over periods of different length.

Using the same data set, Terza [1998] applied full information maximum likelihood, two-stage method of moments, and nonlinear weighted least squares estimation procedures to model determinants of household trip frequencies. Terza extended the count data regression models to account for endogenous switching and its two most common incarnations--endogenous treatment effects and sample selection. Terza found that both corrected and uncorrected parameter estimates indicate that vehicle ownership exerts a positive and significant influence on trip frequency; and that the corrected results reject the exogeneity of d_i (θ is significantly different from zero) and indicate that the structural effect of vehicle ownership on trip frequency will be understated if the endogeneity of d_i is ignored.

Dependent variable: Y_i is the number of trips taken by members of the ith household in the 24-hour period immediately prior to the survey interview, where i=1,2,3,....,577.

Independent variable: WORKSCHL is the percentage of total trips for work or school versus personal business or pleasure performed by individuals in the household. HHMEM is the number of individuals in the household. DISTOCBD is the distance from home to the central business district in kilometers;

AREASIZE is 1 if Standard Metropolitan Statistical Area (SMSA), where home is located, contains at least 2.5 million population. FULLTIME is the number of full time workers in household. ADULTS is the number of adults or individuals at least 16 years of age in household. DISTONOD is the distance from home to nearest transit node, in blocks. REALINC is the household income divided by median income of census tract in which household resides. WEEKEND is 1 if 24-hour survey period is either Saturday or Sunday. d_i is 1 if household owns at least one motorized vehicle, where i=1,2,...,577.

According to Terza [1998] d_i is a dummy variable in which the likelihood that the household owns a vehicle was represented by the following latent ordinal index variate: $z_i\alpha + v_i$, where v_i is the random error term and z_i includes all of the elements in the independent variables except WEEKEND and α denotes a vector of unknown parameters. The variable d_i is a binary switching variable and a heterogeneity term. The observed vehicle ownership outcomes are modeled such that d_i is 1, if and only if $z_i\alpha + v_i \geq 0$; or 0, otherwise. The mean and variance of the dependent variable were 4.5511 and 24.3554, respectively. This is an indication of over-dispersion. Therefore, the regression parameters from the standard Poisson regression model are consistently estimated, but the standard errors are biased downwards leading to the rejection of more null hypotheses. The negative binomial regression model is applicable here, since it can only be applied to count data with over-dispersion. The generalized Poisson regression is applicable here, since it can be applied to count data with over-dispersion, equi-dispersion, or under-dispersion.

The characteristics of the Poisson process are as follows: (1) the occurrences of the events are independent; (2) theoretically, an infinite number of occurrences of the event must be possible in the interval; (3) the probability of a single occurrence of the event in a given interval is proportional to the length of the interval; and (4) in any infinitesimally small portion of the interval, the probability of more than one occurrence of the event is negligible.

The two key assumptions underlying the Poisson distribution are that the rate is constant and that the counts in one interval of time or space are independent of the counts in disjoint intervals.

These assumptions are often not met. For example, suppose that insects are counted on leaves of plants, were deposited in groups, there might be clustering of the insects and the independence assumption might fail. The leaves are of different sizes and occur at various locations on different plants; the rate of infestation may well not be constant over the different locations. Furthermore, if the insects hatched from eggs that If counts occurring over time are being recorded, the underlying rate of phenomenon being studied might not be constant.

We note from the Tables 1 the estimates of dispersion parameter using GPR model is positive indicating over-dispersion. The Wald t-statistic for testing the null hypothesis of H_o: $\alpha = 0$ is significant. Thus, the dispersion parameter α is significantly different from zero. The PR model is not appropriate for this particular data set since we reject the null hypothesis H_o: $\alpha = 0$. From Table 2 the GPR model is preferred to the PR model based on all three goodness-of-fit measures: Pearson's chi-square, deviance, log-likelihood. For example, the GPR model has a smaller deviance than the PR model. The estimated log-likelihood values for GPR model is -1374.8166, whereas it is -1636.1103 for the PR model indicating better fits using GPR models.

The parameter estimates from the both models (PR, and GPR) are somewhat similar. This is expected, as estimates from these models are consistent. Tabular results indicate that not accounting for the over-dispersion, the standard errors from the PR model are under estimated. Consequently, the t-statistic for testing the significance of each regression parameter is generally upward biased for the PR model. Household family size, the number of full time workers in household, distance from home to nearest transit station, and ownership of motorized vehicle are statistically significant and directly related to household trip frequencies. The effects of the number of adults and of total household trips for work or school versus business or pleasure are negative and statistically significant. For example, the percentage of total trips for work or school versus personal business or pleasures lowers the frequency of household trips by about 34%.

Table 1: The effects of selected covariates on household trip frequencies: comparing Poisson Regression, and Generalized Poisson Regression models.

Variable	Poisson Regression (PR)			Generalized Poisson Regression (GPR)		
	Estimate	SE	t-value	Estimate	SE	t-value
Constant	-0.4159	0.1268	-3.2778*	-0.5014	0.1612	-3.1110*
WORKSCHL	-0.4457	0.0695	-6.4141*	-0.3362	0.1386	-2.4259*
HHMEM	0.2192	0.0141	15.5978*	0.2391	0.0316	7.5574*
DISTOCBD	-0.0018	0.0015	-1.2194	-0.0011	0.0025	-0.4476
AREASIZE	-0.0284	0.0442	-0.6425	-0.0366	0.0802	-0.4568
FULLTIME	0.3073	0.0283	10.8547*	0.3818	0.0583	6.5527*
ADULTS	-0.1929	0.0281	-6.8564*	-0.2007	0.0587	3.4218*
DISTONOD	0.0054	0.0012	4.4067*	0.0063	0.0025	2.5377*
REALINC	0.0269	0.0062	4.3543*	0.0245	0.0130	1.8799
WEEKEND	-0.0635	0.0489	-1.3002	0.0093	0.0900	0.1029
d_i	1.4058	0.1232	11.4152*	1.3040	0.1496	8.7153*
α				0.1649	0.0138	11.9433*

SE denotes asymptotic standard error; *significant at 0.05 level

Table 2: Good of fit test measure.

Goodness-of-fit test measures	PR	GPR
Pearson's Chi-square	1841.0340	1867.5256
Deviance	1729.1169	655.7078
Log-Likelihood	- 1636.1103	- 1374.8166

5. DISCUSSION

In summary, we found that (i) holding the number of adults constant and increasing household members (e.g., adding children), increases the frequency of trips being taken; (ii) an increase in the number of trips for work or school taken by household members tends to decrease the number of trips for personal business or pleasure; (iii) an increase in the number of individuals in households tends to increase the total number of household trips taken; (iv) an increase in the number of full-time individuals in the households tends to increase the total number of household trips taken; (v) household income is positively related to the number of household trips taken; (vi) the closer household members are to transit stations the more trips are taken; and (vii) the number of household trips taken during weekdays are inversely proportional to the number of trips taken during the weekends.

In this paper, we described nonlinear regression techniques appropriate for the analysis of household trip frequencies data. The GPR and BNR models confirms results from Terza and Wilson [1990] and Terza [1998]. It has been shown that GRP model gives better fits than standard Poisson regression model when estimating determinants of household trip frequencies.

6. ACKNOWLEDGEMENT

The authors thank Dr. J. V. Terza for providing the Household trip frequency data. They also appreciate the suggestions made by Drs J. V. Terza and Felix Famoye in improving the manuscript.

7. REFERENCES

Breslow N., Tests of hypotheses in over-dispersed Poisson regression and other quasi-likelihood models, *Journal of the American Statistical Association,* 85:565-571, 1990.

Consul, P.C., Generalized Poisson distribution: Properties and applications, Dekker, Inc., New York, 1989.

CSI, National Transportation Study Survey Data File Documentation Manual, Cambridge, MA: Cambridge Systematics, Inc., 1979.

Cox D.R., Some remarks on over-dispersion, *Biometrika,* 70:269-27, 1983.

Efron B., Double exponential families and their use in generalized linear regression, *Journal of the American Statistical Association,* 81:709-721, 1986.

Lawless, J.F., Negative binomial and mixed Poisson regression, *Canadian Journal of Statistics,* 15:209-225, 1987.

Lu, W.S., and C.N. Morris, Estimation in the generalized linear empirical Bayes model using the extended quasi-likelihood, *Communication in Statistics, Part A-Theory and Methods,* 23:661-688, 1994.

Manton, K.G., M.A. Woodbury, and E.A. Stallard, Variance components approach to categorical data models with heterogeneous cell population: analysis of spatial gradients in lung cancer mortality rates in North Carolina counties, *Biometrics,* 37:259-269, 1981.

Nelder, J.A., and R.W. Wedderburn, Generalized Linear Models, *Journal of the Royal Statistical Society Series A,* 135:370-384, 1972.

Singh, K.P., and F. Famoye, Analysis of rates using a generalized Poisson regression model, *Biometrical Journal,* 25, 917-923, 1993.

Stein, G.Z., and J.M. Juritz, Linear models with an inverse Gaussian-Poisson error distribution. *Communication in Statistics-Theory and Methods,* 557-571, 1988.

Terza, J.V. and N. Wilson, Analyzing frequency of several types of events: A mixed Poisson approach, *Review of Economics and Statistics,* 72, 108-115, 1990.

Terza, J.V., Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects, *Journal of Econometrics,* 84, 129-154, 1998.

Wulu, J.T., Generalized Poisson regression models with applications, Ph.D. dissertation, University of Alabama at Birmingham, USA, 1999.