Journal of Data Science **2**(2004), 287-295

On the Generalized Poisson Regression Model with an Application to Accident Data

Felix Famoye[1], John T. Wulu, Jr.[2] and Karan P. Singh[3]
[1] *Central Michigan University,* [2] *Bureau of Primary Health Care*
and [3] *UNT Health Science Center*

Abstract: In this paper a random sample of drivers aged sixty-five years or older was selected from the Alabama Department of Public Safety Records. The data in the sample has information on many variables including the number of accidents, demographic information, driving habits, and medication. The purpose of the sample was to assess the effects of demographic factors, driving habits, and medication use on elderly drivers. The generalized Poisson regression (GPR) model is considered for identifying the relationship between the number of accidents and some covariates. About 59% of drivers who rate their quality of driving as average or below are involved in automobile accidents. Drivers who take calcium channel blockers show a significantly reduced risk of about 34.5%. Based on the test for the dispersion parameter and the goodness-of-fit measure for the accident data, the GPR model performs as good as or better than the other regression models.

Key words: Dispersion, elderly drivers, goodness-of-fit measure, maximum likelihood estimation.

1. Introduction

Analyses of crash reports that are attributed to ‘driver inattention’ suggest that many types of attention failure may be involved in motor-vehicle crashes. Shinar and Scheiber (1991) estimated that 25% to 50% of motor vehicle crashes result from driver inattention. Memory and attention are mental capabilities or cognitive functions that are integral to driving. Recalling how to operate the motor vehicle, the meaning of the road signs and signals, and how to get where the driver intends to go are just part of the whole driving scenario. Attention for safe driving is critical to monitor traffic, highway, adverse weather, and vehicle conditions in every age group-elderly as well as non-elderly.

The continued use of the automobile by a high proportion of the elderly community suggests that the growing older population in the United States is almost certain to result in an increase in the number of elderly drivers. Future generations of elderly drivers are likely to be even older and drive more miles than the aged of today by virtue of their increasing numbers and their continued reliance on the car in old age (Jette and Branch, 1992). Thus, it becomes imperative for additional studies to be conducted in order to identify additional risk factors for automobile injuries among the elderly in an effort to protect the older population and the well being of the community. In recent years, Poisson type regression models have been used to model count response variable affected by one or more covariates. King (1989) and Winkelmann and Zimmermann (1994) developed the generalized event count models based on the Poisson, negative binomial, and the binomial distributions. Winkelmann and Zimmermann (1994) noted that the Poisson regression model is not appropriate when a data set exhibit over-dispersion, a condition where the variance is more than the mean.

The main objective of this study is to assess the effects of demographic factors, driving habits, and medication use on elderly drivers involved in automobile accidents by using the generalized Poisson regression (GPR) model studied by Famoye (1993). In section 2, we describe the data used in this paper. Section 3 outlines the GPR model for the number of accidents involving elderly drivers. In section 4, we review the goodness-of-fit for the GPR model. In section 5, we present the results from data analysis. In section 6, we discuss the results of data analysis.

2. Description of Accident Data

A random sample of 901 drivers who aged 65 years or older was selected from the Alabama Department of Public Safety Records for the years 1991-1996. The setting of this study was Mobile County, Alabama. Details of the study are given by McGwin *et a*l. (2000). Briefly, during a telephone interview subjects were asked if a physician, nurse, or other health care professional had told them they had certain medical conditions; and if so, whether they were taking any medications for the conditions. Subjects were also asked about their driving habits-including self-reported quality of driving, level of comfort with certain driving situations, and type of vehicle most commonly driven. Subjects were asked about the number of accidents they had during their driving from 1991 to 1995.

Accident cases or observations were excluded from the analysis, if they had missing information for questions related to any of the variables in Table 1. Thus, the final study consisted of 595 subjects, approximately 66% of 901 cases in the sample. These exclusions were necessary to set up the data matrix upon which we apply the PR and GPR models. The sample mean and sample variance of the response variable Y, the number of automobile accidents, are respectively, 0.76

and 1.33. The Poisson regression (PR) and the generalized Poisson regression (GPR) models were used to assess the effects of demographic factors, driving habits and medication use on elderly drivers involved in automobile accidents. The variables used in the regression models are presented in Table 1.

Table 1: Variable definition of automobile accidents involving elderly drivers

Dependent/response variable is NUM_ACC (Y),
the number of accidents involving elderly drivers between 1991 and 1995;
Covariates are coded as 1 if true and 0 otherwise.

Variable	Description	Percentage of 1's
GENDER	subjects who are male	49.2
EDUC	subjects who attain Tech/College educational level	34.3
BLACK	subjects who are Black or African-American	21.0
DRIVAVE	subjects who rate their quality of driving as average or below	15.1
EVERYDAY	subjects who drive everyday per week	51.6
HWAY	subjects who are comfortable driving on highway/freeway	53.4
WALK	subjects who need help or have difficulty walking at least 1/4 mile	11.4
OBJECTS	subjects with no difficulty noticing objects off to the side while walking	89.1
WORK	subjects who work full time or part-time	20.2
CA_BLO	subjects who take CA-CHANNEL BLOCKER as a medication	9.7
VASODIL	subjects who take VASODILATOR as a medication	3.7
GLAUCMED	subjects who take GLAUCMED as a medication	4.5

3. The Generalized Poisson Regression Model

Suppose Y_i is a count response variable that follows a generalized Poisson distribution. To model accident data, we define $Y_i, (i = 1, 2, \ldots, 595)$ as the number of automobile accidents involving elderly drivers. The probability function of Y_i is given by

$$f_i(y_i, \mu_i, \alpha) = \left(\frac{\mu_i}{1+\alpha\mu_i}\right) \frac{(1+\alpha y_i)^{y_i-1}}{y_i!} \exp\left[\frac{-\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right], \tag{3.1}$$

$y_i = 0, 1, 2, \ldots$, and $\mu_i = \mu_i(x_i) = \exp(x_i\beta)$, where x_i is a $(k-1)$ dimensional vector of covariates including demographic factors, driving habits and medication use, and β is a k-dimensional vector of regression parameters. For details on the generalized Poisson regression model, the reader is referred to Famoye (1993). The mean and variance of Y_i are, respectively, given by

$$E(Y_i|x+i) = \mu_i \tag{3.2}$$

and

$$V(Y_i|x_i) = \mu_i(1+\alpha\mu_i)^2 \tag{3.3}$$

The generalized Poisson regression model (3.1) is a generalization of the standard Poisson regression (PR) model. When $\alpha = 0$ the probability function in (3.1) reduces to the PR model. Within the framework of PR model, the equality constraint is observed between the conditional mean $E(Y_i|x_i)$ and the conditional variance $V(Y_i|x_i)$ of the dependent variable for each observation. In practical applications and in "real" situations, this assumption is questionable 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. If $\alpha < 0$, (3.1) gets truncated and it may not sum to 1, Famoye (1993). However, if $\alpha > 0$, (3.1) will always sum to 1 and this is the case in the application presented in section 5 [see Appendix for the proof]. In (3.1), α is called the dispersion parameter and it can be estimated along with the regression coefficients in the GPR model. Using the method of maximum likelihood the estimates of α and β in the GPR model (3.1) are given by Famoye (1993).

4. Goodness-of-fit and Test for Dispersion

The goodness-of-fit of GPR model can be based on the deviance statistic that is defined by Famoye (1993). The deviance statistic can be approximated by a chi-square distribution when μ_i's are large. For the accident data, this is not the case as our dependent variable has a mean of 0.76. 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 log-likelihood value.

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

$$H_0 : \alpha = 0 \quad \text{against} \quad H_a : \alpha \neq 0 \tag{4.1}$$

The test of H_0 in (4.1) is for the significance of the dispersion parameter. Whenever H_0 is rejected, it is recommended to use the GPR model in place of the PR model. To carry out the test in (4.1), one may 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 null hypothesis in (4.1) is to use the likelihood ratio test statistic, which is approximately chi-square distributed with one degree of freedom when the null hypothesis is true.

5. Results

About 59% of drivers who rate their quality of driving as average or below are involved in automobile accidents. Nearly 59% of African Americans are involved in automobile crashes. Drivers who take calcium channel blockers show a significantly reduced risk of about 34.5%. Fifty six percent of males are involved in automobile accidents. The parameter estimates and their standard errors using the PR and the GPR models are given in Table 2.

In comparing the sample mean 0.76 of the response variable to its sample variance 1.33, the data suggests a case of over-dispersion. The estimated dispersion parameter from the GPR model is positive, which is an indication of over-dispersion. The asymptotic "t"-statistic for testing the null hypothesis in (4.1) is approximately 2.68 as given in Table 2. Thus, the dispersion parameter α is significantly different from zero (5% level). The Poisson regression model is not appropriate for this data since we reject the null hypothesis given in (4.1). The log-likelihood values for the PR and GPR models are -673.3 and -667.0, respectively, which also indicate that modeling over-dispersed data using the GPR model is more appropriate than the PR model.

In both PR and GPR models, seven independent variables (drivave, everyday, hway, walk, ca_blo, objects, and work) are significant at 5% level. The variable, gender, is significant under the PR model at 10% level but this is not the case under the GPR model. The parameter estimates from both models are very similar; however, the standard errors from the PR model are under estimated. The standard errors from the GPR model are more appropriate in this case since the model accounts for the over-dispersion exhibited by the data. At 5% level, the effect of elderly working drivers is statistically significant and is positively associated with the number of automobile accidents. This implies that elderly

Table 2: Determinants of elderly automobile accidents

	Poisson		GPR	
Variable	Estimate±se	t-value	Estimate±se	t-value
Intercept	$-0.5924 \pm .1849$	-3.20^*	$-0.6309 \pm .1996$	-3.16^*
Black	$0.1856 \pm .1138$	1.63	$0.2015 \pm .1226$	1.64
Ca_blo	$-0.4644 \pm .2010$	-2.31^*	$-0.4686 \pm .2098$	-2.23^*
Drivave	$0.2725 \pm .1245$	2.19^*	$0.2908 \pm .1348$	2.16^*
Everyday	$0.2250 \pm .0998$	2.25^*	$0.2167 \pm .1068$	2.03^*
Gender	$0.1735 \pm .0997$	1.74	$0.1689 \pm .1063$	1.59
Glaucmed	$-0.2288 \pm .2469$	-0.93	$-0.1883 \pm .2626$	-0.72
Walk	$0.6461 \pm .1232$	5.24^*	$0.5965 \pm .1359$	4.39^*
Vasodil	$-0.5904 \pm .3603$	-1.64	$-0.6075 \pm .3726$	-1.63
Hway	$0.4338 \pm .1404$	3.09^*	$0.4289 \pm .1487$	2.88^*
Objects	$-0.4582 \pm .1310$	-3.50^*	$-0.3977 \pm .1443$	-2.76^*
Work	$0.2828 \pm .1108$	2.55^*	$0.2450 \pm .1206$	2.03^*
Educ	$-0.1453 \pm .1048$	-1.39	$-0.1275 \pm .1119$	-1.14
α			$0.0794 \pm .0296$	2.68^*
Log-likelihood	-673.3		-667.0	

* means significant at 0.05 level, se = standard error

drivers with part time or full time work are involved in more automobile crashes than the others. Elderly drivers who rate their quality of driving as average or below significantly contributed to number of automobile accidents. Elders who need help or have difficulty walking at least 1/4 mile were involved in more accidents than the other group. Elders who drove everyday were involved in more accidents than those who did not. Elderly drivers who take calcium channel blockers show a significantly reduced risk of automobile accidents.

6. Discussion

With the growing population of older adults, the number of persons aged 65 years or older driving continues to increase. In 1985, there were 15.5 million American drivers (9.8% of all drivers) aged 65 years or older (Reuben *et al.*, 1998). With driving being so closely associated with independence and personal autonomy, it is not likely that this estimate of elderly drivers will significantly decrease in subsequent years. Jette and Branch (1992), in a ten-year longitudinal study, reported that the elderly continue to rely on the automobile as a primary mode of transportation into their eighth and ninth decades of life. Additionally, the study revealed that more than three-quarters of all people rely on the automobile as their primary means of travel and that this pattern of reliance changed little during the subsequent decades of their lives.

When a data set has too many zeros, Lambert (1992) suggested the use of zero-inflated Poisson regression (ZIP) model. In the accident data, the observed percentages of 0, 1, and 2 are, respectively, 47.2%, 36.6% and 12.1%. van den Broek (1995) proposed a score test for zero inflation in a Poisson regression model. The score statistic has an asymptotic chi-square distribution with 1 degree of freedom under the null hypothesis of no zero inflation. For this data, the score statistic is computed to be 0.67, which is not significant. Based on this result, it does not appear that there is zero inflation in the data. Therefore, we did not consider the use of zero-inflated PR model for the data. Also, the data is over-dispersed which indicates that the PR model is not appropriate either.

To model over-dispersion, the GPR model discussed in section 3 and the negative binomial regression (NBR) model are among the suitable models. We applied the NBR model to the data and found the results to be similar to that of GPR model. Thus, we decided to exclude the parameter estimates of the NBR model to save space in the paper. If we know before hand that the data is over-dispersed, either the NBR model or the GPR model can be used. However, if the type of dispersion is unknown, the choice should be GPR model since it is more flexible.

In summary, the estimated dispersion parameter from the data is positive and it is significantly different from zero. Based on the goodness-of-fit measure for the accident data, the GPR model seems to perform better than the PR model in identifying demographic factors, driving habits and medication use associated with the number of accidents involving elderly drivers. Additional studies should be conducted in order to identify additional risk factors for automobile accidents involving the elderly to improve traffic safety.

Acknowledgments

This work was done in collaboration with the co-authors while Professor Famoye, Central Michigan University, was on his sabbatical leave at the Department of Biostatistics, School of Public Health, University of North Texas at Fort Worth. The authors thank Dr. G. McGwin of the University of Alabama at Birmingham for providing the accident data used in this paper. We also thank the Referee and the Editor for valuable suggestions.

Appendix: Model (3.1) sums to 1

The Lagrange expansion [see Whittaker and Watson (1927, p.133)] of $f(t) = e^{\theta(t-1)}$ under the transformation $u = t/g(t) = te^{-\alpha\theta(t-1)}$ is given by

$$\begin{aligned} f(t) &= f(0) + \sum_{y=1}^{\infty} \frac{u^y}{y!} \left(\frac{\partial}{\partial t}\right)^{y-1} \left[g^y(t) f'(t)\right]_{t=0} \\ &= e^{-\theta} + \theta \sum_{y=1}^{\infty} \frac{u^y}{y!} \left(\frac{\partial}{\partial t}\right)^{y-1} \left[\exp[\theta(1+\alpha y)(t-1)]\right]_{t=0} \\ &= e^{-\theta} + \sum_{y=1}^{\infty} u^y \frac{\theta^y}{y!} (1+\alpha y)^{y-1} \left[\exp[-\theta(1+\alpha y)]\right] \end{aligned}$$

$$f(1) = 1 = \sum_{y=0}^{\infty} \frac{\theta^y (1+\alpha y)^{y-1}}{y!} \exp[-\theta(1+\alpha y)]$$

From Famoye (1993), $\theta = \mu/(1+\alpha\mu)$ and so by using this value of θ in the last summation, we get

$$1 = \sum_{y=0}^{\infty} \left(\frac{\mu}{1+\alpha\mu}\right)^y \frac{(1+\alpha y)^{y-1}}{y!} \exp\left[\frac{-\mu(1+\alpha y)}{1+\alpha\mu}\right] \qquad \text{(A.1)}$$

The terms in the above summation are given by (3.1). If $\alpha < 0$, the right hand side of (A.1) gets truncated and it may not sum to 1. However, if $\alpha > 0$, the right hand side of (A.1) will always sum to 1. Thus, the probabilities in (3.1) will sum to 1 when $\alpha > 0$.

References

Famoye, F. (1993). Restricted generalized Poisson regression model. *Communications in Statistics – Theory and Methods* **22**, 1335-1354.

Jette, A. M. and Branch, L. G. (1992). A ten-year follow-up of driving patterns among the community-dwelling elderly. *Human Factors* **34**, 25-31.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science* **33**, 762-784.

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

McGwin, G., Sims, R. V., Pully, L. and Roseman, J. M. (2000). Relationship among chronic medical conditions, medications and automobile crashes in the elderly: A population-based case control study. *American Journal of Epidemiology* **152**, 424-431.

Reuben, D. B., Silliman, R. A. and Traines, M. (1998). The aging driver-medicine, policy and ethics. *Journal of the American Geriatrics Society* **36**, 1135-1142.

Shinar, D. and Scheiber, F. (1991) Requirements for safety and mobility in older drivers. *Human Factors* **33**, 507-519.

van den Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics* **51**, 738-743.

Whittaker, E. T. and Watson, G. N. (1927). *A Course of Modern Analysis.* Cambridge University Press, Cambridge.

Winkelmann, R. and Zimmermann, K. F. (1994). Count data models for demographic data. *Mathematical Population Studies* **4**, 205-221.

Received April 22, 2003; accepted June 12, 2003.

Felix Famoye
Department of Mathematics
Central Michigan University
Mt. Pleasant, Michigan 48859, USA
felix.famoye@cmich.edu

John T. Wulu, Jr.
DHHS, Hlth. Res. and Svs. Adm.
Bureau of Primary Health Care
Bethesda, MD 20814, USA
jwulu@hrsa.gov

Karan P. Singh
School of Public Health
UNT Health Science Center
Forth Worth, TX 76107, USA
ksingh@hsc.unt.edu