Modeling Health Care Events Using Mixed Poisson Models

A. Woehl[1]

***Abstract*— This paper reviews the application of mixed Poisson model to the health care events. The model and its parameters are estimated and applied to health care data using hospital admissions. The problem of individuals falling into the zero class is discussed and estimates are compared.**

***Keywords:* mixed Poisson models; health care data; hospital admissions; zero term problem**

I. Introduction

This paper considers the application of mixed Poisson models for the analysis of health care data. Health care data in our case is defined as hospital admissions for certain individuals in a population over a period of time. Mixed Poisson models can be used to model rare events and have been used for modeling practical applications amongst others in the field of market research [1] and accidents and sickness [2]. We refer to [3] for a discussion on the mixed Poisson processes and applications. The aim of the present paper is to statistically model hospital admissions using mixed Poisson model, whereas the process of events occurring over time is assumed to be a random process for each individual where each individual has an own intensity of event occurrence.

II. Background

A. Mixed Poisson distribution

A random variable X follows the mixed Poisson distribution (MPD) if it has the probability density function (p.d.f.)

$$p_X = P(X = x) = \int_{0-}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} dF(\lambda), x = 0,1,2.... \quad (1)$$

where $F(\lambda)$ is the cumulative distribution function (c.d.f.) of a random variable over the interval $(0, \infty)$. The outcome of $F(\lambda)$ is in our case regarded as a unknown personal λ_i for each individual i that reflect an illness proneness. The distribution $F(\lambda)$ is often called structure distribution or may be regarded as prior distribution [4]. A common structure distribution is the gamma distribution with the probability density function

$$f(\lambda) = \frac{1}{a^k \Gamma(k)} \lambda^{k-1} e^{-\lambda/a}, a > 0, k > 0, \lambda > 0 \quad (2)$$

The resulting MPD is the negative binomial distribution (NBD) with the p.d.f.

$$p_X = \frac{\Gamma(k+x)}{x!\Gamma(k)} (\frac{1}{1+a})^k (\frac{a}{1+a})^x, x = 0,1,2..., k > 0, a > 0 \quad (3)$$

where a is the scale parameter and k the shape parameter of the NBD. The NBD can be re-parameterized by (m, k), where $m=ak$ denotes the mean of the distribution. The maximum likelihood and all natural moment based estimators for (m, k) are asymptotically uncorrelated for an independent and identically distributed (i.i.d.) NBD sample.

In the case of hospital admissions (1) has the following interpretation: the number of hospital admissions over the analysis period for each individual i follows the Poisson distribution with an unknown λ_i (mean of the Poisson distribution) and if this has the c.d.f. $F(\lambda)$ this means the number of events for a random individual follow the MPD for this fixed time interval.

B. Zero term problem

As described above the NBD is used for the description of data following a MPD when the structure distribution is a gamma distribution. However, observed data can be truncated, meaning that individuals falling into the zero category cannot be entirely observed. This is a typical problem in healthcare data.

To overcome this problem truncated distributions can be used. To remove the zero probability from the NBD and receive the truncated NBD (TNBD) the distribution (3) has to be divided by $P(X=0)$. Defining $p=1/(1+a)$ it follows that:

$$p_X = \frac{\Gamma(k+x)}{x!\Gamma(k)} (\frac{p^k}{1-p^k})(1-p)^x, x = 1,2,3..., k > 0, p > 0 \quad (4)$$

is the p.d.f. of the TNBD.

III. Data

The health care data is taken from the Cardiff and the Vale of Glamorgan area in Wales, United Kingdom. The dataset containing inpatient, outpatient, biochemistry and mortality data has undergone record linkage to identify those records belonging to the same individual. Patients considered for

[1] Cardiff Research Consortium, The Medicentre, Heath Park, Cardiff, CF14 4UJ, Email: anette.woehl@crc-limited.co.uk.

The author would like to thank Christopher Morgan from CRC ltd for providing the health care data.

analysis were those with a first inpatient admission coded with a diagnosis of diabetes as either a primary or secondary cause recorded on discharge diagnosis. Only new cases from April 1995 were included and all identified cases were resident within Cardiff and the Vale of Glamorgan. Data was collected over 11 years until 2005. During this analysis period individuals (patients) enter the dataset at any time during the analysis period depending on the first event, i.e. hospital admission, occurring at random time. The total number of individuals in the dataset was 15,277.

When applying the mixed Poisson model we assume a Gamma-Poisson process. The occurrence of a hospital admission for a random individual follows a Poisson process with the mean of λ_i over a time interval $t=1, 2...T$ where X_{it} represents the number of hospital admissions for individual i up until time t. The distribution of hospital admissions for an individual is given by (1). For fixed i the random variable $(X_{i,t}|\Lambda=\lambda_i)$ therefore follows a Poisson distribution with the mean $t\lambda_i$. For a fixed time t the number of hospital admissions across all individuals follows the NBD with parameters (mt, k), where $m=ak$.

IV. Estimators

Theoretically the Maximum Likelihood Estimation (MLE) is the preferred method to estimate parameters. However, moment-based estimators can be as efficient as the MLE [5]. To estimate the gamma-Poisson parameters (m, k) in this paper the Method of Moments (MOM) and the Zero Term Method (ZTM) are used. For both methods, as well as for MLE, the estimator for m is simply given by the sample mean:

$$\hat{m} = \overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

Equating the population mean and variance to the corresponding sample value we obtain the MOM estimator for k:

$$\hat{k}_{MOM} = \frac{\overline{x}^2}{s^2 - \overline{x}} \text{ where } s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1} \tag{6}$$

The ZTM estimator for k is defined as the optimal solution z in the following equation:

$$\hat{p}_0 = (1+\frac{\overline{x}}{z})^{-z} \tag{7}$$

The MOM estimators for the TNBD are obtained as follows. Using the fact that $p=k/(m+k)$, where p is the probability of success in one Bernoulli trial, and solving the first two moment equations, we obtain:

$$\hat{k}_{ZMOM} = \frac{\hat{p}(s^2 + \overline{x}^2) - \overline{x}}{\overline{x}(1-\hat{p})} \tag{8}$$

and

$$\overline{x}\hat{p}(1-\hat{p}^k) = \hat{k}(1-\hat{p})$$

Solving these equation we obtain the parameter estimates for (m, k). The parameters describing the TNBD have the same definition as for NBD, but m does no longer equal the mean of the distribution. The mean of the TNBD is equals the average number of hospital admissions per individual who has non-zero number of events (i.e. being admitted to hospital). The

$$w = E(X|X \geq 1),\ w > 1 \tag{9}$$

The probability that one individual has at least one event is defined as

$$b = 1 - p_0 \text{ with b=} \overline{x} \text{ /w.} \tag{10}$$

V. Results

To test how applicable the MPD is for modeling hospital admissions the parameters (m, k) were estimated using the methods as mentioned above. Due to the nature of the data the overall panel size is unknown, since the number of individuals (in our case diabetics) cannot be determined definitely. Thus we estimated firstly the number of zeros and thus the total panel size with the cumulative data up to 2005. Due to the setup of the data there were no zeros in the data for this analysis time. The total number of zeros for the dataset was estimated to be 22,181 using the TNBD. Using this estimation and correcting the zero term for all other years the parameters (m, k) were estimated for the MPD. Results are shown in Table I. The NBD is the one dimensional marginal distribution of the MPD whose parameters k remains constant in time and whose m increases linearly with time. Estimated parameters in Table I show that k increases in time. This is evidence for not adequate fit of the model.

The first estimation does not allow for changing population and prevalence figures with time. The second estimation takes rising population and prevalence figures into consideration. This led to a new estimate for the zero terms for the analysis periods prior to 2005. The results are shown in Table II. The parameter k stabilized within the second estimation.

Table I: Estimated parameters *(m, k)* for hospital admissions

Year	m	k_{MOM}	k_{ZTM}
1995	0.199	0.096	0.106
1996	0.418	0.125	0.131
1997	0.648	0.143	0.150
1998	0.868	0.154	0.165
1999	1.094	0.156	0.179
2000	1.323	0.167	0.189
2001	1.542	0.179	0.198
2002	1.731	0.192	0.207
2003	1.944	0.203	0.214
2004	2.182	0.215	0.216
2005	2.216	0.217	0.217

Table II: Estimated parameters *(m, k)* for hospital admissions corrected for population and prevalence changes

Year	m	k_{MOM}	k_{ZTM}
1995	0.332	0.172	0.197
1996	0.656	0.211	0.233
1997	0.960	0.227	0.254
1998	1.215	0.230	0.265
1999	1.453	0.219	0.269
2000	1.669	0.221	0.266
2001	1.851	0.223	0.260
2002	1.980	0.226	0.254
2003	2.123	0.226	0.244
2004	2.277	0.226	0.231
2005	2.216	0.217	0.217

The empirical distributions and estimated NBD for the adjusted zero terms due to population and prevalence changes over time are shown in Fig. I-IV in the appendix.

APPENDIX

Figure I: Histogram and estimated NBD for population and prevalence change adjustment of the zero term: year: 1995

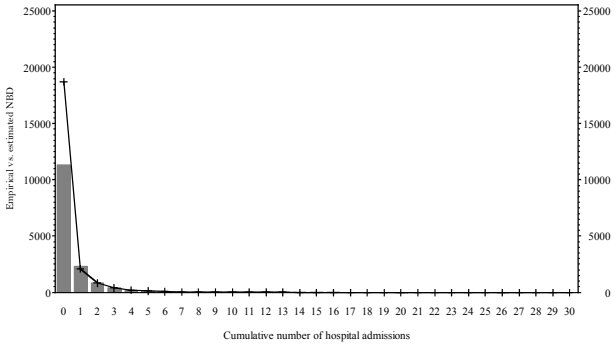


Figure II: Histogram and estimated NBD for population and prevalence change adjustment of the zero term: year: 1998

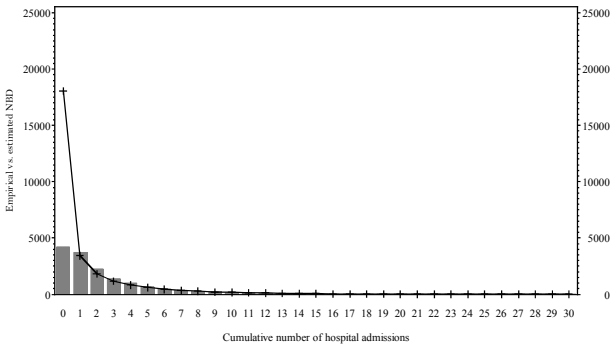


Figure III: Histogram and estimated NBD for population and prevalence change adjustment of the zero term: year: 2002

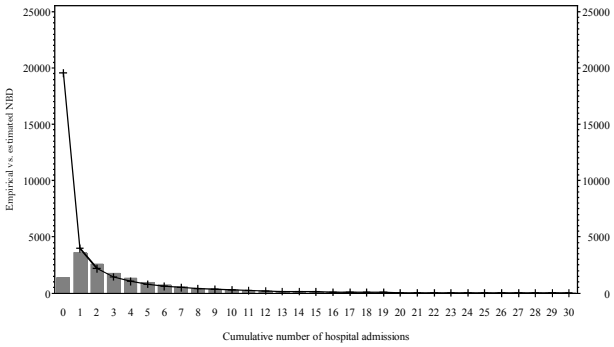
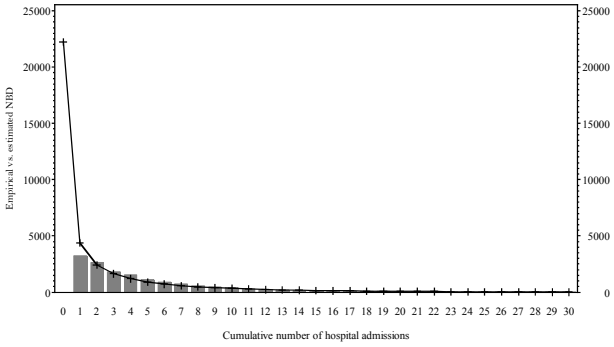


Figure IV: Histogram and estimated NBD for population and prevalence change adjustment of the zero term; year 2005



REFERENCES

[1] A.S.C Ehrenberg. *Repeat Buying: Facts, Theory and Applications.* London: Charles Griffin & Company Ltd.; New York: Oxford University Press, 1988.

[2] O. Lundberg. *On Random Processes and Their Application to sickness and Accident Statistics.* Uppsala: Almquist and Wiksells, 1964.

[3] V. Savani and A. A. Zhigljavsky. Modelling recurrent events using mixed Poisson models. *Present volume.*

[4] J. Grandell. *Mixed Poisson Processes,* volume 77 of *Monographs on Statistics and Applied Probability.* London: Chapman & Hall, 1997.

[5] V. Savani and A. A. Zhigljavsky. Efficient Estimation of Parameters of the Negative Binomial Distribution. *Communications in Statistics – Theory and Methods*, 35: 767-783, 2006.