Extensions to Poisson Regression for Hospital Admissions data

Elizabeth Stojanovski, Ian Robinson

The University of Newcastle, School of Mathematical and Physical Sciences, Callaghan 2308, NSW Australia
Elizabeth.Stojanovski@newcastle.edu.au

Abstract

Presented are models for length of hospital stay data. The Hurdle model is an extension to the Poisson model when the data structure can be considered as two separate processes as was evident with the present case study. Hospital data was considered and collected as part of a larger study over a five year period. The Hurdle model appeared the most appropriate in terms of overall goodness of fit.

Keywords: Length of stay, Poisson, Zero-inflated

1. Introduction

The number of bed days in hospital, from admission to discharge, is known as length of stay. Length of stay can be affected by many factors including hospital protocol, and has been used for many purposes including as an indicator of hospital resources. Readmission can also be an indicator of resources, as more frequent readmissions lead to greater resource consumption [1]. Modelling such data, however, does pose some statistical issues.

Literature that involved the modelling of length of stay data has focused on a single admission or time to readmission following certain admission types. A particular study [2] used logistic regression to model length of stay by dichotomising the outcome using a cut-off point of 7 days. The average number of days has been modelled using multiple linear regression [3], which has limitations in the presence of highly skewed data. In another study [4], patients were followed after surgery to their first readmission using Cox-proportional hazards models for which those without a readmission were censored. Time to readmission after undergoing a certain procedure has been assessed using Poisson regression [5], for which repeated observations were ignored for simplicity.

The primary aim of the present study is to determine the most effective way to model length of stay data. Length of stay across all readmissions over a specified time period will be considered for each patient. Of those admitted to hospital initially, not all will be readmitted while others could be readmitted more than once within the study period. Although length of stay data, a form of count data, is commonly modelled using the Poisson distribution, inferences can be biased in the presence of model overdispersion. Extensions are made here to Poisson regression to handle overdispersion and an excess number of zeros in the data, and these extensions are applied to hospital readmission data.

2. Methods

Data Adults scheduled to undergo elective surgery at a hospital in the period from January 2001 to December 2001 were recruited into this study. Hospital separations data were collected from this hospital on these participants from the time of recruitment until December 2005. Data contain all hospital separations (admission and discharge dates) recorded for the patient (linked by medical reference number). Although the procedure itself is not of direct interest here as it was elective and considered not to affect the long term health of subjects, it was used here as a means of collecting information on and following up patients.

The total length of stay for each participant will be measured across all readmissions over the study period (to December 2005) following the initial procedure. Total length of stay accounts for the number of readmissions and the length of stay for each readmission to give an overall, long term measure of hospital usage. Data on patient demographics and socio-economic status were obtained in a questionnaire given at baseline.

Generalised Linear Models The Poisson distribution is used to model count data. Given a random variable Y, the probability of observing any specific count y is given by:

$$f(y;\mu) = \frac{e^{-\mu}(\mu)^{y}}{y!}\,.$$

A unique property of the Poisson distribution is that the expected value and the variance of the random variable are equal. Deviation from this assumption results in underdispersion or overdispersion. In the case of overdispersion, standard errors will be underestimated as only the observed heterogeneity in the data will be captured while heterogeneity relating to overdispersion will be unaccounted for, likely to result in Type 1 errors.

The Negative binomial model is an extension to the Poisson which accounts for unobservable variation (overdispersion) by including a random error term (u) which is added to the conditional mean of the Poisson regression model. The probability density function (PDF) of the Negative binomial can be defined as:

$$f(y;\mu,u) = \frac{e^{-\mu u}(\mu u)^y}{y!}$$

which is essentially a Poisson model with gamma heterogeneity. This model is a mixture of Poisson and Gamma distributions and captures the observed variation from the Poisson and overdispersion from the Gamma distribution.

More zeros than would be expected under these probability distribution functions can also result in overdispersion [6]. A zero-inflation component allows for this overdispersion with data assumed to come from two distributions. These are modelled as two latent classes: observations comprising only zeros while the other latent class models values that are not zero with the second class dependent on the probability of excess zeros from the first. Structural zeros from a binary model are mixed with the non-negative integer outcomes (which includes sample zeros) from a count distribution. The zero-inflated Poisson distribution accounts for the observable differences in the data as well as the over dispersion attributed to the excess zeros with the PDF defined as:

$$f(y;\mu) = \begin{cases} p + (1-p)e^{-\mu}, & ,y = 0 \\ (1-p)\dfrac{e^{-\mu}(\mu)^y}{y!} & ,y > 0 \end{cases}$$

where p is the logistic probability of the observation being greater than zero. The zero-inflated Negative binomial is defined as:

$$f(y;\mu,r) = \begin{cases} p + (1-p)\dfrac{1}{(1+r\mu)^{1/r}} & ,y = 0 \\ (1-p)\dfrac{\Gamma(y+1/r)}{\Gamma(y+1)\Gamma(1/r)}\dfrac{(r\mu)^y}{(1+r\mu)^{y+1/r}} & ,y > 0 \end{cases}$$

where p is as defined for the zero-inflated Poisson. Again, the probability of an excess zero is based on the weight of both the structural and the sample zeros from the Negative binomial distribution. As with the Negative binomial, the zero-inflated Negative binomial captures overdispersion and observable variation.

Hurdle models, also referred to as two part models [7], are a further extension to zero-inflated models which can be applied in situations where the data structure is such that there are two *separate* processes. The first generates zero counts (binary model) and the second generates positive non-zero counts. This separation allows for positive counts to be assessed based on the threshold or Hurdle being passed. For the present study, a threshold value of zero is considered most meaningful. The Hurdle model for the Negative binomial is given by:

$$f(y;\mu,r) = \begin{cases} p & ,y = 0 \\ \dfrac{(1-p)\dfrac{\Gamma(y+1/r)}{\Gamma(y+1)\Gamma(1/r)}\left(\dfrac{1}{1+\mu r}\right)^{1/r}\left(\dfrac{\mu r}{1+\mu r}\right)^y}{1-\left(\dfrac{1}{1+\mu r}\right)^{1/r}} & ,y > 0 \end{cases}$$

where p is the probability associated with exceeding the threshold. The second part of the model is conditional on the threshold probability. As this Hurdle model includes the zero-truncated Negative binomial model for the second component, this accounts for the observed and unobserved variation in the data in the same manner as does the standard Negative binomial model.

3. Results

There were 1601 adult participants who were eligible to be included in the present study. These comprised patients with valid questionnaire data and medical record numbers who were not readmitted for dialysis, renal failure, myeloplastic syndrome or refectory anaemia at any time during data collection. As readmission is of interest, hospital separations data regarding the initial procedure undertaken were removed from the dataset.

Participants who were not readmitted to the hospital during the study period were assigned a value of zero for total length of stay. For those with at least one readmission, total length of stay is the sum of all separate lengths of stay for the participant.

The number of readmissions ranged from 0 to 19 (Median=0, IQR=2). The histogram in Figure 1 shows the distribution of number of readmissions to be highly right skewed (Skewness=2.6). A histogram of the distribution for total length of stay is shown in Figure 2. Total length of stay has a peak at zero (Median=0) and is highly skewed to the right (Skewness=5.38, Mean=5.87). Total length of stay ranged from 0 to 192 days (IQR=6.0). As expected, the variance is large (226.95) and exceeds the mean, violating the equidispersion assumption (Z= 37892, p<.001).

The negative binomial regression model which accounts for unobserved variability via a random effect is consequently fitted to model length of stay. The value of the dispersion parameter still suggests a large amount of unexplained variability. The likelihood ratio test (p<0.001) shows the Negative binomial as an improvement in terms of explaining overdispersion relative to the Poisson although a large amount of variation remains unexplained, potentially due to the large number of zeros. Fit indices AIC and BIC

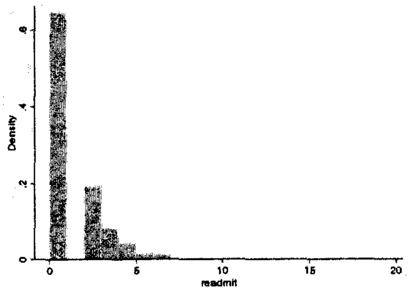

Figure 1: Histogram of number of readmissions

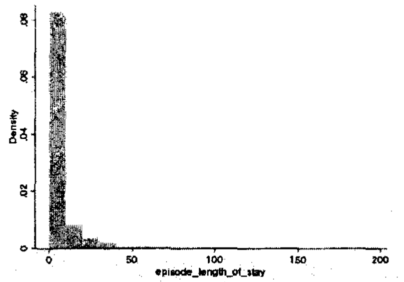

Figure 2: Histogram of total length of stay

respectively suggest a marked improvement in fit relative to the Poisson (6488, 6515 compared with 29855, 20876 respectively) suggesting that the addition of a random error term captures much of the heterogeneity that the Poisson model fails to.

Given the data was shown to be highly overdispersed, a zero-inflated Negative binomial model was fitted to the data. This model is compared to the standard model using the Vuong test (*Z*-value=8.57, p<.001) suggesting the model to explain more heterogeneity than the Negative binomial, although the likelihood ratio test (χ^2= 54.45, p<.001) indicates that the model still fails to capture most of the data heterogeneity. In terms of goodness of fit, the values of AIC and BIC do slightly improve (6309, 6346 versus 6488, 6515 respectively).

Hurdle models can be interpreted as a two art model with the first model a logistic model predicting the probability of readmission. The second model separates from the first model in that only those who have a readmission are assessed. The second model calls on the zero-truncated Negative binomial. The model fitted was total length of stay as the outcome with age and gender as potential predictors for demonstration purposes. Table 1 shows the results of this model. Age appeared a statistically significant predictor of total length of stay with older patients more likely to have greater length of stays than younger patients (p<0.001). Substantial improvement in the fit of the Hurdle model suggests it as the most appropriate model for these data (AIC=3558, BIC=3604).

Table 1: Parameter estimates from Hurdle regression model

Parameter	Estimate	Robust s standard error	z-value	p-value	95% confidence Interval
Logistic					
Intercept	-1.27	0.20	-6.34	<.001	-1.66, 0.88
Age	0.01	0.003	3.35	<.001	0.00, 0.02
Gender (male)	0.03	0.11	0.24	0.81	-0.19, 0.25
Zero-truncated Negative binomial					
Intercept	1.62	0.19	8.67	<.001	1.25, 1.98
Age	0.02	0.01	5.95	<.001	0.01, 0.03
Gender (male)	-0.03	0.12	-0.25	0.81	-0.27, 0.21

4. Conclusions

The superiority of Hurdle and zero-inflated Negative binomial models over the standard Poisson and Negative binomial regression models for modelling total length of stay was demonstrated. In particular, the unexplained heterogeneity and/or excess number of zeros can be characterised by the Negative binomial regression models. Hurdle models were found to be the most appropriate extension.

Length of stay was assessed using individual level data as opposed to aggregated data enabling a patient's activities in terms of readmissions to be tracked.

5. Discussion

All hospital separations were recorded at only one hospital and therefore exclude admissions to other hospitals. Hospital outpatient usage was also not considered. Furthermore the analysis did not account for reason of primary admission. Length of stay can potentially be affected by diagnosis and/or procedure undertaken. Another limitation of this study is that it failed to account for mortality of participants over the course of the study.

References

[1] R. Dales, G. Dionne, J. Leech, M. Lunau, I. Schweitzer "Preoperative Prediction of pulmonary complications following thoracic surgery", *Chest,* 194, 155—159, 1993.

[2] H. Lazar, C. Fitzgerald, S. Gross, T. Heeren, G. Aldea, R. Shemin. "Determinants of Length of Stay after Bypass Graft Surgery", *American Heart Association,* 92, 20-24, 1995.

[3] O. Sadr-Azodi, R. Bellocco, K. Eriksson, J. Adami. "The impact of tobacco use and body mass index on the length of stay in hospital and the risk of post-operative complications among patients undergoing total hip replacement", *Journal of Bone and Joint Surgery,* 88, 1316-1320, 2006.

[4] J. Steuer, P. Blomqvist, F. Granath, B. Rydh, A. Ekbom, U. de Faire, E. Stahle. "Hospital Readmission After Coronary Artery Bypass Grafting: Are Women Doing Worse?", *Annuals of Thoracic Surgery,* 73, 1380-1386, 2002.

[5] J. Garcia-Aymerich, E. Farrero, M. Félez, J. Izquierdo, M. Marrades, J. Antó. "Risk factors of readmission to hospital for a COPD exacerbation: a prospective study", *Thorax,* 58, 100-105, 2003.

[6] J. Elhai, P. Calhoun, J. Ford, "Statitsical procedures for analysing mental health services data", *Psychiatry Research,* 160, 129-136, 2007.

[7] Y. Asada, G. Kephart. "Equity in health services use and intensity of use in Canada", *BMC Health Services Research,* 7, 41-53, 2007.