ELSEVIER

Accident Analysis and Prevention 37 (2005) 267–273

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Confidence and prediction intervals for generalised linear accident models

G.R. Wood*

Department of Statistics, Macquarie University, Sydney 2109, NSW, Australia

Abstract

Generalised linear models, with "log" link and either Poisson or negative binomial errors, are commonly used for relating accident rates to explanatory variables. This paper adds to the toolkit for such models. It describes how confidence intervals (for example, for the true accident rate at given flows) and prediction intervals (for example, for the number of accidents at a new site with given flows) can be produced using spreadsheet technology.

Keywords: Generalised linear model; Negative binomial; Poisson

1. Introduction

Generalised linear models have gathered recognition in recent years (Maycock and Hall, 1984; Hauer et al., 1988; Maher and Summersgill, 1996) as useful tools for relating the number of accidents, of a specified type, to explanatory variables such as vehicle flows. For the single flow model, the true mean number of accidents μ is modelled as $\beta_0 x^{\beta_1}$, where x denotes the flow. The distribution of the observed number of accidents, for a given flow, is assumed to be either Poisson, or more generally, negative binomially distributed about this mean value. The negative binomial distribution occurs naturally when we allow for variation of safety M between sites, with a given flow, to be modelled by a gamma distribution, and then variation of the number of accidents Y within a site, with safety M, to be modelled by a Poisson distribution with mean M. A detailed description of these models has been given in the companion paper (Wood, 2002), where methods for assessing goodness of fit were described.

Once goodness of fit is established for a model, it is of interest to provide confidence intervals (for model parameters) and prediction intervals (for dependent variables); this is routinely carried out when working with linear models. Such intervals provide information about the extent of variation in these quantities. In this context, the intervals of interest, for a given flow, are:

(i) A confidence interval for μ, the true accident rate.
(ii) (a) For a Poisson model, a prediction interval for y, the accident rate at a new site.
(b) For a negative binomial model, a prediction interval for m, the safety of a new site, and a prediction interval for y, the accident rate at a new site.

The purpose of this paper is to provide formulae, in Section 2, which enable construction of these intervals, and to illustrate their use with real accident data in Section 3. The required calculations can be carried out on a spreadsheet. Exposition is generally in terms of models with a single flow; models with more than one explanatory variable are handled in an extended, but similar, fashion. Notation and terminology used in this paper are as in Wood (2002).

Standard texts, for example, McCullagh and Nelder (1989), discuss confidence intervals for generalised linear model parameters; the author, however, has not found the approach discussed here in the literature, other than in Maher and Summersgill (1996). Here, we clarify, amplify and extend that work.

Specifically, approximate confidence and prediction intervals appropriate for a given flow are developed. We caution

* Fax: +61 2 9850 7669.
E-mail address: gwood@efs.mq.edu.au.

0001-4575/$ – see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2004.10.005

that the confidence level necessarily decreases if we wish to make statements about many flow values. For this, so-called simultaneous (and necessarily wider) confidence bands are needed. The development of simultaneous confidence bands is a topic of current research; the work of Sun et al. (2000) produces such confidence bands for the mean in a generalised linear model.

This paper can be read in two ways. A reader interested in the practical construction of confidence and prediction intervals should skim Section 2, then work carefully through the examples of Section 3, referring to Section 2 and Appendix A for formulae as needed (Table 4 provides an overall summary). For the reader interested in the underlying theory, careful reading of Section 2 and Appendix A is recommended.

2. Confidence and prediction intervals

A confidence interval for the true mean, for both the Poisson and negative binomial models, is developed in Section 2.1. In Section 2.2, a prediction interval for a predicted number of accidents at a new site is derived for the Poisson model, while in Section 2.3, prediction intervals for safety and predicted number of accidents at a new site are produced for negative binomial models.

2.1. Confidence interval for μ

The generalised linear model we have described uses a "log" link function; the logarithm of μ is linear in the model parameters β_0' and β_1, since $\eta = \log\mu = \log\beta_0 + \beta_1 \log x = \beta_0' + \beta_1 \log x$, for the single flow model. Standard generalised linear model theory gives that asymptotically the estimates b_0' and b_1, of β_0' and β_1, respectively, have a bivariate normal distribution (Dobson, 1990), in particular

$$\begin{bmatrix} b_0' \\ b_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0' \\ \beta_1 \end{bmatrix}, I^{-1} \right),$$

so they are unbiased, with covariance matrix the inverse of the information matrix I. It follows that $\hat{\eta} = b_0' + b_1 \log x$ has asymptotically a normal distribution and since $\hat{\eta} = \log\hat{\mu}$, where $\hat{\mu} = \mathrm{e}^{b_0'} x^{b_1}$, $\hat{\mu}$ has an approximately lognormal distribution.

This enables us to write down an approximate 95% confidence interval for η, when the flow is x, as

$$b_0' + b_1 \log x \pm 1.96\sqrt{\mathrm{Var}(b_0' + b_1 \log x)},$$

whence a 95% confidence interval for $\mu = \mathrm{e}^{\eta}$ is given by

$$\left[\mathrm{e}^{b_0' + b_1 \log x - 1.96\sqrt{\mathrm{Var}(b_0' + b_1 \log x)}}, \; \mathrm{e}^{b_0' + b_1 \log x + 1.96\sqrt{\mathrm{Var}(b_0' + b_1 \log x)}} \right].$$

The lower boundary is closer to the estimate $\hat{\mu}$ of μ than is the higher boundary, reflecting the right skewed lognormal distribution of the estimate $\hat{\mu}$. Here,

$$\begin{aligned} &\mathrm{Var}(b_0' + b_1 \log x) \\ &\quad = \mathrm{Var}(b_0') + 2 \log x \,\mathrm{Cov}(b_0', b_1) + (\log x)^2 \,\mathrm{Var}(b_1) \\ &\quad = I_{11}^{-1} + 2 \log x \, I_{12}^{-1} + (\log x)^2 \, I_{22}^{-1}. \end{aligned}$$

Illustrative real examples are given in Section 3. Note that $\hat{\eta} = (1, \log x)(b_0', b_1)^{\mathrm{T}}$ (where "T" denotes transpose) so in practice $\mathrm{Var}(\hat{\eta})$ is most easily calculated as

$$\mathrm{Var}(\hat{\eta}) = (1, \log x) I^{-1} (1, \log x)^{\mathrm{T}}.$$

There are two ways to find the components of I^{-1}. If the model is fitted using a statistical package then options are generally available which output the covariance matrix I^{-1} of the parameters. On the other hand, if using the first principles method described in Wood (2002, (A.3)), then the required covariance matrix is $(X^{\mathrm{T}} W X)^{-1}$, where X is the design matrix and W a diagonal matrix.

A final remark in this subsection: the lognormal distribution of $\hat{\mu}$ discussed can be approximated by a normal distribution, or

$$\hat{\mu} \sim N(\mu_0 = \mu, \sigma_0^2 = \mu^2 \,\mathrm{Var}(\hat{\eta})),$$

as in Maher and Summersgill (1996, Eq. (14)). This approximate sampling distribution for $\hat{\mu}$ is fundamental in the sequel.

2.2. Poisson model

We consider the case of the Poisson model and an interval for a predicted number of accidents, y. Under the model, given a true mean accident rate of μ, the conditional distribution of accidents Y is Poisson with mean μ. A confidence interval for the number of accidents Y, however, must now accommodate the approximately normal variation in $\hat{\mu}$, our estimator of μ, as $N(\mu_0, \sigma_0^2)$. Table 1 summarises the variables involved.

The marginal distribution of Y is thus a mixture of Poisson distributions, on the mean, by a normal distribution. It can be shown that the distribution of Y, supported by $\{0, 1, 2, \ldots\}$ has mean μ_0 and variance $\sigma_0^2 + \mu_0$. (The key to this calculation is the observation that a central moment of a mixture is the mixture of the central moments of the distributions being mixed.) Our intuition does tell us that this variance should depend on that of $\hat{\mu}$, namely σ_0^2, and also should increase

Table 1
The two levels of variation, first in μ, then in Y given μ, to be considered when forming a prediction interval for y in the Poisson model

Variable description	Variable notation	Distribution
Accident rate, given true rate μ	$Y\|\mu$	Poisson(μ)
Estimator of true mean accident rate	$\hat{\mu}$	$N(\mu_0, \sigma_0^2)$