Accident Analysis and Prevention 41 (2009) 861–868

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

The sensitivity of estimates of regression to the mean

Mike Maher[a,*], Linda Mountain[b]

[a] Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, United Kingdom
[b] Department of Engineering, University of Liverpool, Liverpool L69 3BX, United Kingdom

ARTICLE INFO

Article history:
Received 28 January 2009
Received in revised form 20 April 2009
Accepted 22 April 2009

Keywords:
Regression-to-mean
Bayesian methods
Markov Chain Monte Carlo
WinBUGS
Predictive accident models
Traffic safety

ABSTRACT

Estimations of the effectiveness of remedial treatments in road safety analysis are frequently bedevilled by the problem of regression to the mean (RTM). The number of accidents x observed at a site in the "before" period is a "noisy" quantity: x is Poisson distributed about an (unknown) true mean m for that site, so that $x = m + e$. Sites selected for treatment tend to have a positive random error component e, which will on average be zero in the "after" period, even if no treatment is applied.

Methods for estimating RTM usually require some assumption about the underlying (prior) between-site distribution of the true means $f_0(m)$: for example, in the empirical Bayes method, a gamma distribution is assumed. The paper considers the impact of different assumptions for this distribution and, indeed, whether any distributional form needs to be assumed. Using Markov Chain Monte Carlo methods, a variety of distributional forms are assumed for $f_0(m)$ and applied to each of a number of real data sets, including that from a major study on the effectiveness of speed cameras. It is shown that, in some cases, the size of the estimated RTM effect can be quite sensitive to the choice of distribution.

1. Introduction

It is now well known that the estimation of the effectiveness of safety remedial treatments through before and after studies is often affected by the phenomenon of *regression to the mean* (RTM). Typically, sites selected for treatment are those with a higher-than-average number of accidents in the before period. Since accident frequencies are random quantities (usually assumed to follow a Poisson distribution), there is the strong likelihood that this random error component is positive in the before period, and will on average relax to zero in the after period, even if no treatment were to be applied. If no account is taken of this effect, and the treatment effect is estimated by a simple comparison of after-to-before accident frequencies, there is the danger that the treatment effectiveness will be exaggerated. Hauer (1980, 1986) was amongst the first to draw attention to this effect in the field of transport safety research, and has written extensively on the subject since that time (see, for example, his book: Hauer, 1997). A notable recent example in the debate about RTM has been the case of the UK study on the effectiveness of speed cameras (Gains et al., 2005).

Ideally, it would be preferable if the sites to receive any safety treatment could be decided in a random manner from amongst a specified set of potential sites, with those sites not selected then acting as controls. Such a strict experimental design is very rarely possible in practice, however. Alternatively, RTM could be avoided if, once the sites had been selected for treatment on the basis of their accident record in the *before* period, the implementation of treatment were delayed for some time (known as the *lag* period). The lag period can then be used to provide an unbiased estimate of the true accident rate before treatment is applied. The treatment effectiveness is then obtained by a comparison, not of after with before, but of after with lag period (see Mountain et al., 1998).

Generally, of course, the local authority is keen to implement the treatment as soon as is practicable and therefore, any lag period is typically quite short, and hence the estimate of the true before rate from the lag period would be subject to appreciable uncertainty. Nevertheless, in some cases, the lag period may be sufficiently long to make this a viable approach.

If the effect of RTM cannot be avoided by either of the two approaches above, it is necessary instead to estimate it. The most widely accepted way to achieve this is through the approach known as the empirical Bayes (EB) method. In the conventional version of the EB method, there is an assumption of a gamma distribution for the spread of the underlying true site means m (referred to as the prior distribution). This assumption is primarily employed for mathematical convenience, as it ensures that the posterior distribution is also gamma, and a simple, linear formula emerges for the posterior mean $E(m|x)$. The purpose of this paper is to examine this assumption, and to investigate the sensitivity of the estimate of the RTM effect to this assumed form of the prior distribution. Elvik (2008) has recently also looked at the "predictive validity" of the empirical Bayes method, but this concentrated on alternative ways

* Corresponding author.
E-mail addresses: m.j.maher@its.leeds.ac.uk (M. Maher), l.mountain@liverpool.ac.uk (L. Mountain).

0001-4575/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2009.04.020

of estimating the parameters in the gamma distribution, and alternative forms of the predictive accident model, and did not involve consideration of alternative forms of distribution.

The rest of the paper will be devoted to an investigation into the sensitivity of the magnitude of the RTM effect to different assumptions about the form of the prior distribution. In particular, it will be shown how Markov Chain Monte Carlo (MCMC) methods (as contained in the WinBUGS software package (Lunn et al., 2000) for example) provide a powerful and convenient way to carry out the necessary Bayesian modelling and estimation for a wide variety of forms of prior distribution. This approach is then applied to a number of real data sets of different types.

2. The principles of the empirical Bayes method

In the EB method, it is assumed that x, the number of accidents at a site in the before period, is Poisson distributed about an unknown mean m. In the most general version of the problem, an estimate for m is available from a *predictive accident model*, in which the predicted number of accidents, denoted by m_0, is a function of site characteristics (flows, design, etc.), typically obtained by regression analysis. It is conventionally assumed that the true site mean m is gamma distributed about m_0. In the Bayesian framework, the prior distribution $f_0(m)$ is this gamma distribution, and the likelihood $L(x|m)$ is Poisson. It then follows that the posterior distribution $f_1(m|x)$ is also gamma distributed, with a mean value m_1 given by

$$m_1 = E(m|x) = \alpha m_0 + (1-\alpha)x \qquad \text{where} \quad \alpha = \frac{1}{1+(m_0/k)} \tag{1}$$

where k is the shape parameter of the gamma distribution, which measures the precision of the predictive accident model. The posterior estimate of the true before mean is therefore given, not by the observed before frequency x, but by this weighted average of the prediction m_0 and the observed value x.

As an example, consider the data in the first two rows of Table 1, showing the observed numbers of accidents at the 9603 sites in the North Lanarkshire region of Scotland over a 3-year period. It is important to note that, in North Lanarkshire, the sites are of a wide variety of types: junctions and links, long links and short links, urban sites and rural sites. Sites with at least five accidents are referred to as "cluster sites" and are normally earmarked for remedial treatment. A naïve estimate of the effectiveness of the treatment would be given by a comparison of the total number of accidents at the treated sites in the after period with the corresponding number in the before period. But this would be likely to exaggerate the treatment effectiveness, as the before total will be higher than the true mean value, due to the regression to mean effect. Put another way, the addition of the Poisson noise at each site means that the distribution of the observed numbers of accidents is broader than the distribution of the underlying true mean values. Those sites with a high observed number of accidents will almost certainly have a true mean value that is less than the observed. In order to obtain a reliable, unbiased estimate of treatment effectiveness, we need to "shrink" the observed values towards the overall mean.

Now in this particular version of the problem, there is no predictive accident model as such: the data from all the sites in the region form a "reference group", and the "prediction" m_0 is the same for all sites. Then, in (1), the weight α is given by the fraction of the total variance of the x's that is attributable to the Poisson noise:

$$\alpha = \frac{1}{1+(m_0/k)} = \frac{m_0}{m_0+\sigma_0^2} = \frac{m_0}{\sigma^2} \tag{2}$$

where σ^2 is the variance of the x's and σ_0^2 is the estimated variance of the unknown true means. In the data in Table 1, the mean number of observed accidents per site is 0.327, and $\sigma^2 = 0.584$. Assuming that the observed accidents at any site are Poisson distributed, the variance of the underlying distribution of true mean values is, by the method of moments, $\sigma_0^2 = 0.584 - 0.327 = 0.257$.

Here, we have $m_0 = 0.327$ and the shape $k = (m_0/\sigma_0)^2 = 0.415$, so that $\alpha = 0.559$. Therefore the EB posterior mean estimates m_1 or $E(m|x)$ are given in Table 1, from where the shrinkage towards the mean can be seen. So, for those 13 sites with an observed number of accidents of $x = 6$, we should compare the after accident number with 2.83 per site rather than with 6: that is, a regression to mean effect of $(2.83-6)/6$ or −53%.

However, this RTM estimate is on the basis of the assumption of a gamma distribution for the true site means. The gamma is assumed in the EB method because it is the conjugate prior to the Poisson likelihood, and hence the formula for the posterior mean in (1) is particularly simple to derive, and is a linear combination of the prior mean and the observed value. The objective in this paper is to examine the effect of deviating from this assumption of a gamma, as there seems no obvious reason to support that choice rather than any other.

Hauer (1997) points out that, if the posterior mean is a linear combination of prior mean and observed value, then use of the weight α given by (2) gives the minimum-variance estimator of m_1 – *whatever the form of the prior distribution*. However, for forms of prior distribution other than gamma, the posterior mean will *not* generally be a linear combination of prior mean and observed value. This is evident in the plots of m_1 versus x shown in Wright et al. (1988), for a variety of forms of prior distribution, and will also be confirmed later in this paper.

3. A distribution-free method

First, however, we note that Hauer (1980) pointed out the rather remarkable result that, for this version of the problem, it is possible to obtain the expected value of the posterior mean without making *any* assumption about the form of the prior distribution.

$$\begin{aligned} E(m|x) &= \frac{\int_m m(\exp(-m)m^x/x!)f_0(m)\,\mathrm{d}m}{\int_m(\exp(-m)m^x/x!)f_0(m)\,\mathrm{d}m} \\ &= \frac{(x+1)\int_m(\exp(-m)m^{x+1}/(x+1)!)f_0(m)\,\mathrm{d}m}{\int_m(\exp(-m)m^x/x!)f_0(m)\,\mathrm{d}m} \\ &= \frac{(x+1)P(x+1)}{P(x)} \end{aligned} \tag{3}$$

In passing, we observe that in the case where the prior is gamma distributed, so that the $P(x)$ are from a negative binomial distribution, (3) reduces to the expression in (1).

Table 1
North Lanarkshire data: number of sites $N(x)$ at which x accidents occurred.

x	0	1	2	3	4	5	6	7	8	9	11	13
$N(x)$	7411	1645	341	117	38	26	13	7	2	1	1	1
$E(m\|x)$	0.18	0.62	1.06	1.50	1.95	2.39	2.83	3.27	3.71	4.15	5.03	5.91