Journal of Safety Research 40 (2009) 257–263

Contents lists available at ScienceDirect

Journal of Safety Research

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

Estimating the variance in before-after studies

Zhirui Ye [a,*], Dominique Lord [b]

[a] *Western Transportation Institute, Montana State University, P O Box 174250, Bozeman, MT 59717, USA*
[b] *Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843, USA*

ARTICLE INFO

Available online 11 June 2009

Keywords:
Before-after study
Variance estimation
Bootstrap
Resampling
Non-parametric

ABSTRACT

Problem: To simplify the computation of the variance in before-after studies, it is generally assumed that the observed crash data for each entity (or observation) are Poisson distributed. Given the characteristics of this distribution, the observed value (x_i) for each entity is implicitly made equal to its variance. However, the variance should be estimated using the conditional properties of this observed value (defined as a random variable), that is, $f(x_i|\mu_i)$, since the mean of the observed value is in fact unknown. *Method:* Parametric and non-parametric bootstrap methods were investigated to evaluate the conditional assumption using simulated and observed data. *Results:* The results of this study show that observed data should not be used as a substitute for the variance, even if the entities are assumed to be Poisson distributed. Consequently, the estimated variance for the parameters under study in traditional before-after studies is likely to be underestimated. *Conclusions:* The proposed methods offer more accurate approaches for estimating the variance in before-after studies.

1. Introduction

The before-after study is a commonly used method for measuring the safety effects of a single treatment or a combination of treatments in highway safety (Hauer, 1997). Short of a controlled and full randomized study design, this type of study is deemed superior to cross-sectional studies since many attributes linked to the converted sites where the treatment (or change) was implemented remain unchanged. Although not perfect, the before-after study approach offers a better control for estimating the effects of a treatment. In fact, as the name suggests, it implies that a change actually occurred between the "before" and "after" conditions (Hauer, 2005a). Combined with the empirical Bayes (EB) technique, the before-after study can also minimize the bias caused by the regression-to-the-mean (RTM) commonly found in crash data analyses (Persaud, Retting, Garner, & Lord, 2001; Persaud, McGee, Lyon, & Lord, 2003). Despite their large popularity, it should be mentioned that not everyone agrees about their superiority over cross-sectional studies (Tarko, Eranky, & Sinha, 1998; Noland, 2003).

Before-after studies can be grouped into three types: the simple (naïve) before-after study; the before-after study with control groups; and the before-after study using the EB technique (also using a control group). The selection of the study type is usually governed by the availability of the data, such as crashes and traffic flow, and whether the transportation safety analyst has access to entities that are part of the reference group. The selection can also be influenced by the amount of available data (or sample size).

As described by Hauer (1997), the traditional before-after study (no matter which type is used) can be accomplished using two tasks. The first task consists of predicting the expected number ($\hat{\pi}$) (in this paper, we will work with the estimated value; hence, $\hat{\pi}$ is an estimate of π) of target crashes for a specific entity (i.e., intersection, segment) or series of entities in the "after" period had the safety treatment not been implemented. The second task consists of estimating the number of target crashes ($\hat{\lambda}$) for the specific entity in the "after" period. Here, the term "after" means the time period after the implementation of a treatment; correspondingly, the term "before" refers to the time before the implementation of this treatment. In most practical cases, either $\hat{\pi}$ or $\hat{\lambda}$ can be applied to a composite series of entities where a similar treatment was implemented at each entity.

Hauer (1997) proposed a four-step process for estimating the safety effects of a treatment. The process is described as follows:

Step 1: For $j = 1,2,\ldots,n$, estimate $\lambda(j)$ and $\pi(j)$. Then, compute the summation of the estimated and predicted values, such that $\hat{\lambda} = \Sigma\lambda(j)$ and $\hat{\pi} = \Sigma\pi(j)$.

Step 2: For $j = 1,2,\ldots,n$, estimate $Var\{\hat{\lambda}(j)\}$ and $Var\{\hat{\pi}(j)\}$. For each single entity, it is assumed that observed data (e.g., annual crash counts over a long timeframe) are Poisson distributed and $\hat{\lambda}(j)$ can be approximated by the observed value in the before period. On the other hand, the calculation of $Var\{\hat{\pi}(j)\}$ will depend on the statistical methods adopted for the study (e.g., observed data in naïve studies, method of moments, regression models, EB technique). Assuming that crash data

* Corresponding author. Tel.: +1 406 994 7909.
E-mail address: jared.ye@coe.montana.edu (Z. Ye).

0022-4375/$ – see front matter. © 2009 National Safety Council and Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsr.2009.04.004

in the before and after periods are mutually independent, then $Var\{\hat{\lambda}\} = \Sigma Var\{\hat{\lambda}(j)\}$ and $Var\{\hat{\pi}\} = \Sigma Var\{\hat{\pi}(j)\}$.

Step 3: Estimate the parameters δ and θ, where $\hat{\delta} = \hat{\pi}$ - $\hat{\lambda}$ (again, referring to estimated values) is defined as the reduction (or increase) in the number of target crashes between the predicted and estimated values, and $\hat{\theta} = \hat{\lambda}/\hat{\pi}$ is the ratio between these two values. The term θ has also been referred to in the literature as the index of effectiveness (Persaud et al., 2001). Hauer (1997) suggests that when less than 500 crashes are used in the before-after study, θ should be corrected to remove the bias caused by the small sample size using the following adjustment factor $1/[1 + Var\{\hat{\pi}\}/\hat{\pi}^2]$.

Step 4: Estimate the variances $Var\{\hat{\delta}\}$ and $Var\{\hat{\theta}\}$. These two variances are calculated using the following equations (note: $Var\{\hat{\theta}\}$ is also adjusted for the small sample size) below:

$$Var\left\{\hat{\delta}\right\} = Var\left\{\hat{\lambda}\right\} + Var\left\{\hat{\pi}\right\} \tag{1}$$

$$Var\left\{\hat{\theta}\right\} = \frac{\hat{\theta}^{\,2}\left[\left(Var\left\{\hat{\lambda}\right\}/\hat{\lambda}^2\right) + \left(Var\left\{\hat{\pi}\right\}/\hat{\pi}^2\right)\right]}{\left[1 + \left(Var\left\{\hat{\pi}\right\}/\hat{\pi}^2\right)\right]^2} \tag{2}$$

The four-step process provides a simple way for conducting before-after studies. One important assumption with this process is related to the computation of the variance $Var(\hat{\lambda})$ (or $Var\{\hat{\pi}\}$). As described above, observed crash data are assumed to be Poisson distributed for each entity and the observed data are directly used in the analysis. However, as noted by Hauer (1997), the variance $Var(\hat{\lambda})$ is in fact unknown. The properties of the Poisson distribution are in essence used to simplify the computation of the variance. In this case, the observed crash counts, used here as random variables, are used as a substitute for estimating the variance for each entity (i.e., the observation is assumed to be equal to its mean).[1] Given the fact the mean of the observed value is unknown, the variance should be estimated using the conditional properties of the observed value (i.e., $f(x_i|\mu_i)$) (e.g., Cook & Wei, 2001; Diggle, Liang, & Zeger, 2002). Consequently, there is a need to evaluate how these conditional properties affect the estimation of the variance in the context of a before-after study.

The objectives of this paper are to evaluate whether or not the assumption that crash data should be used as a direct substitute to the variance is valid, even when one assumes the data are Poisson distributed for each entity, and if not, to examine how this may affect the estimation of the variance for calculating the inferences associated with the parameters used to estimate the safety effects in a before-after study. To accomplish the objectives of this study, parametric and non-parametric bootstrap resampling methods are investigated to evaluate this assumption. The bootstrap method is first applied to data simulated using a Poisson distribution and a Negative Binomial (or Poisson-gamma) distribution, and a mixture of these two distributions to evaluate its applicability in this research. Then, both methods are applied to two datasets of observed before and after crash data taken from the literature. The proposed methods are used to estimate $\hat{\lambda}$, $\hat{\pi}$, $Var\{\hat{\lambda}\}$ and $Var\{\hat{\pi}\}$ and the output is compared with the traditional before-after method to compute these values.

The rest of this paper is divided into six sections. The first section presents the parametric method for estimating the variance of conditional random variables. The second section presents the characteristics of the bootstrap method used in this study. The third section covers the evaluation of the bootstrap method using simulated data. The fourth section presents the application of the methods to observe before and after data taken from the literature. The fifth section describes important discussion points associated with before-after studies and offers avenues for further work. The last section summarizes the key findings of this study.

2. Parametric Method

Since the mean of an entity is unknown, the analysis of the random variable must be carried out using the conditional properties of this variable with respect to the mean (i.e., $f(x|\eta)$), where $\eta = \{\eta(1), \eta(2),\ldots, \eta(j)\}$ (Cook & Wei, 2001; Agresti, 2002; Bolstad, 2004). Furthermore, the conditional property entails that the values can be approximated using any suitable distribution.

Researchers who have conducted before-after studies (non-randomized trials) using the same dataset in the before and after periods have analyzed the data using the conditional properties described above (note: in many cases, the mean of the observation is modeled as a random-effect variable). Examples of such observational studies where the mean of the Poisson distribution was modeled as a random-effect variable can be found in medicine (Cook & Wei, 2001), epidemiology (Laird & Ware, 1982; Diggle et al., 2002), and animal science (Schaik, Shoukri, Martin, Schukken, Nielen, 1999). Very recently, researchers in highway safety have also started using this conditional property for before-after studies (Persaud, Lan, Lyon, & Bhim, 2009; Park, Park, & Lomax, 2009). Depending upon the assumptions, the random-effect variable has been modeled using different marginal distributions. For example, Diggle et al. (2002) have proposed the Gaussian distribution for modeling the mean of the Poisson model. Because of the properties associated with the Poisson-gamma distribution (i.e., the closed form of the conjugate distribution), other researchers have proposed to model the mean using the gamma distribution (Cook & Wei, 2001; Diggle et al., 2002; Persaud et al., 2009; Park et al., 2009). As discussed by Lord, Washington, and Ivan (2005), it is important to point out that the Poisson-gamma distribution (as well as other mixed-Poisson distributions, such as the Poisson-lognormal) is used to approximate the true characteristics of the motor-vehicle crash process. It should be pointed out that by allowing the mean to follow a given distribution, the variance estimated will not be underestimated when a regression model is used in a before-after study.

Getting back to the primary objective of this analysis, if the means of the Poisson distributions for entities 1,2,…,j are assumed to follow a gamma distribution ($\eta = \{\eta(1),\ldots\eta(j)\} \sim Gamma(\phi,\ \mu/\phi)$), it can be shown that the marginal distribution becomes the conjugate Poisson-gamma distribution, where ϕ is defined as the inverse dispersion parameter of the Poisson-gamma distribution. The mean and variance of η are μ and μ^2/ϕ, respectively.

Using the theorem proposed by Casella and Berger (1990), referred to as Conditional Variance Identity (CVI), it is possible to estimate the variance for a series of observed values, when each value is conditional upon the mean. This theorem states that "for any two random variables X and Y, $Var\{Y\} = E[Var(Y|X)] + Var[E(Y|X)]$, provided that the expectations exist." For the curious reader, Agresti (2002) provides a very good discussion about the application of the CVI properties to Poisson random variables. His discussion in fact supports this work. This author states that the CVI needs to be used to estimate the variance of random variables because μ varies (i.e., unknown) due to unmeasured factors.

[1] To examine whether this assumption in before-after studies is reasonable, one can look into a single entity. If the actual expected number of crash counts ($\eta(j)$) for the jth entity in either the before or the after period, then it can be shown that $Var\{\hat{\eta}|\hat{\eta}(j)\} = \Sigma Var\{\hat{\eta}(j)\} = \Sigma\eta(j)$. In practice, $\eta(j)$ is approximated using observed data $x(j)$ (a random variable) and the true mean is therefore not known with certainty. By using the probability mass function (PMF) of the Poisson distribution, it is straightforward to compute the probability that $Var\{\eta(j)\} = \eta(j) = x(j)$, assuming that $x(j)$ represents the crash count over one year time period (or other very short time periods). It is obvious that the probability for the observed count (X) to equal the mean decreases as the mean $\eta(j)$ increases. For example, the probability is about 40 percent when $\eta(j) = 1$, while it decreases to 10 percent when $\eta(j) = 15$. This entails that $Var\{\hat{\eta}(j)\} = x(j)$ may not be reasonable, since the count has a large probability not being equal to the "true" mean of an entity (if known). Thus, it is safe to assume that $x(j)$ cannot be a good approximation of $\eta(j)$.