Accident Analysis and Prevention 49 (2012) 360–365

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Bias caused by self-reporting distraction and its impact on crash estimates

Paul F. Hanley*, Nikhil Sikka

University of Iowa, Public Policy Center, 227 South Quadrangle, Iowa City, IA 52242-1192, United States

ARTICLE INFO

Article history:
Received 28 August 2011
Received in revised form 7 February 2012
Accepted 12 February 2012

Keywords:
Driver distractions
Self-reporting bias
Injury severity
Two-stage correction

ABSTRACT

Over the last decade, driver distractions, such as cell phone use and texting, have become a significant contributor to roadway crashes. Some states now have legislation that severely restricts or bans driver activities deemed distracting. However, many policies and engineered countermeasures are based on self-reported crash data. This raises the issue of potential bias and when not controlled for in analysis supporting policy decisions, can lead to poor allocation of public resources. This study explores the impact of self-reporting driver distraction on the likelihood estimates of the injury severity category of vehicle crashes. Using a two-step correction technique, the presence of bias is tested, when present corrected, and its impact is interpreted. The findings show that self-reporting bias is present in the national database, a database often used to help evaluate policy and engineering options, self-reporting bias understates the true effect of driver distraction on injury severity, and it is not uniform across injury categories. As a result, the forecast of potential savings of countermeasure policies or in-vehicle devices will be distorted leading to inefficient allocation of public resources.

1. Introduction

Over the last decade, driver distraction, such as cell phone use and texting has become a significant contributor to roadway crashes. In US, states have acknowledged this by considering legislation that severely restricts or bans driver activities deemed distracting. Six states have passed laws against the use of hand-held cell phones while driving and 25 states and Washington, DC ban text messaging for all drivers (Governors Highway Safety Association, 2010). Further restrictions are in place for novice and school bus drivers. However, even with the bans in place, distracted driving remains a concern. As per 2008 National Highway Traffic Safety Administration (NHTSA) estimates, driver distractions were responsible for almost 6000 fatalities in roadway crashes—16% of all fatal crashes—and more than half a million were injured (Ascone et al., 2009). Also, the share of distracted drivers in fatal crashes has increased from 8% in 2004 to 11% in 2008.

In response to an increasing need to quantify the impact of driver distraction, the US Department of Transportation now includes within the National Automotive Sampling System General Estimates System (GES) information about driver distraction. The GES is produced from a multi-stage stratified sample selected from over 7 million annual police-reports in the US that contain crash information, which for the most part is self-reported, especially driver distraction. This raises the potential self-reporting bias—a bias caused by individuals desire to avoid providing self-incriminating information about their driving behavior—that confound the observed differences between groups thus preventing the realization of the true distinctiveness of the group or individuals. When self-reporting bias is not controlled in analysis that supports policy decisions, bias will lead to poor allocation of public resources (Larzelere et al., 2004). In this paper, the term self-reporting is synonymous with self-selection as used in previous literature.

Based on earlier studies (Amoros et al., 2006; Bos et al., 2009; Elvik and Mysen, 1999; Holdridge et al., 2005; Malik et al., 2003), which found high rates of under-reporting and non-reporting of data by police, one would conclude that self-reported distraction, as contained in the GES crash database, also has a high likelihood of being under-reported. Several researchers have acknowledged this potential (Bunn et al., 2005; Lam, 2002; McEvoy et al., 2006; Neyens and Boyle, 2008; Stutts et al., 2001; Young and Lenné, 2010), though do not explicitly correct for bias. Therefore, a goal of this study is to test for self-reporting bias in distraction reporting, correct for the bias (if present) and assess its impact on crash injury severity levels.

To correct for potential bias caused by the self-reported distraction on injury severity, a two-stage model is estimated. In the first stage, the probability of a driver being distracted is estimated and the results are used to correct for self-reporting bias in the second stage. The magnitude of the bias is then estimated and discussed.

2. Background

Controlling for selection bias has been a part of the safety literature and researchers have considered sample selection on the

* Corresponding author. Tel.: +1 319 335 6800; fax: +1 319 335 6801.
E-mail address: paul-hanley@uiowa.edu (P.F. Hanley).

0001-4575/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2012.02.008

Table 1
Variables used in stage 1 regression.

Variable	Categories	
Distraction	Distracted	
	Not distracted	
Critical event	Vehicle failure	Entering intersection
	Poor road conditions	Vehicle decelerating
	Loss control	Other vehicle stopped
	Lane departure	Traveling in same direction with lower speed
	Roadway departure	Traveling in same direction with higher speed
	Turning at intersection	Other critical events (=ref.)
Movement prior to critical event	Going straight	Turning right or left
	Starting in traffic lane	Changing lanes
	Stopped in traffic lane	Other events (=ref.)
Violation	Alcohol/drugs related	Other violations
	Speeding	No violation charged (=ref.)
Occupants	Total occupants	
Driver's age	Up to 19	45–64
	20–29	65 and greater
	30–44 (=ref.)	

effectiveness of seat-belt restraints (Dee, 1997; Evans, 1996; Levitt and Porter, 1999). Similar to the GES dataset, the datasets used in those studies relied on self-reported data and the results clearly show the importance of correcting for self-reporting bias in making valid conclusions. Evans (1996) empirically discusses sample selection problem in determining the relationship between crash severity and seat-belt usage. He used the combined National Accident Sampling System (NASS) data from year 1982 to year 1991 to show that unbelted-drivers in high severity crashes are over-represented and concluded that if sample selection is not corrected for, seat-belt effectiveness estimates are biased upwards to a great deal. Levitt and Porter (1999) measured the seat belt and air bag effectiveness on crash survival. The authors dealt with the sample selection problem by only including the crashes in which someone in a different vehicle died. Another similar study was conducted by Dee (1997) who used the Center for Disease Control and Prevention's (CDC) annual Behavioral Risk Factor Surveillance System (BRFSS) telephone surveys to examine the sharp increase in the seat belt usage in the late 1980s and early 1990s and its impact on crash fatalities. He concluded that previous evaluations overstated the impact of seat-belt laws on its usage by about 60%. Also, the results showed non-homogeneous effects of seat-belt policies on its use—high risk drivers were significantly less sensitive to the enactment of seat belt laws and their enforcement status. However, missing in the referenced studies on injury severity is the testing and control for self-reporting bias when examining the relationship between injury severity and drivers distraction.

3. Empirical approach

To test and correct for self-reporting bias, the two-stage regression methodology as originally suggested by Heckman (1979) and modified by Lee (1983) is used in this paper. The first stage, referred to as the selection process, estimates the logit values of driver distraction using multinomial logistic regression using vehicle crash and driver characteristics data contained in the GES database. The outcome variable is binary (1 = driver reported as distracted and 0 = driver reported as not distracted or non-response) and the explanatory variables as presented in Table 1 and further explained in the next section. In the first stage logit analysis, the direct effects of the explanatory variables (given by Xi) on the self-reported distraction variable are estimated, as well as the unmeasured characteristics of distraction that are captured by the residuals. The purpose of this stage is to estimates the influence of unknown factors, which remain in the distraction variable after eliminating the direct effects of the known factors.

Table 2
Variables used in stage 2 regression.

Variable	Category	
Injury severity	No injury	Non-incapacitating
	Possible injury	Incapacitating or fatal injury
Distraction	Distracted	Not distracted
λ, bias control factor	Estimates from Stage 1	
Sex (male)	Female	Male
Restraint system in place	Yes	No
Adverse weather conditions	Yes	No
Traffic controls present	Yes	No
Alcohol/drugs involved	Yes	No
Driver's age	Up to 19	45–64
	20–29	65 and greater
	30–44 (=ref.)	
Vehicle's age	Years	

The second stage, referred to as the outcome equation, uses the transformed logit values from stage 1—as suggested by Lee (1983). The multinomial logit model is shown in Eq. (1) where the dependent variable, C, contains five crash injury severity categories (i.e. no injury, possible injury, non-incapacitating injury, and fatal or incapacitating injury). The Y_j are the explanatory variables, β_j are the associated parameters in Table 2 and further explained in the next section. D is an indicator for distraction ($D = 1$ when driver was reported as distracted, $D = 0$ otherwise). The δ parameter captures the net impact of distraction on injury severity. The model includes the transformed logit values, referred to as a bias control factor, λ. As a separate predictor variable with its associated parameter θ. ε is assumed to be independently and identically distributed according to logit error terms.

$$C = \beta_j Y_j + \delta D + \theta\lambda + \varepsilon_2^* \tag{1}$$

The conditional probability function of D given first stage explanatory variables X_i is shown in Eq. (2).

$$p_i = Pr\left(D = 1 \,\middle|\, X_i\right) = E\left(D \,\middle|\, X_i\right) \tag{2}$$

The estimated p_i values are transformed using Eq. (3). The transformed logit values, given by q_i, are then used to calculate the selection bias control factor λ for each observation using Eqs. (4) and (5). The interpretation of λ is through θ; if the estimate of θ is statistically significant, bias due to the self-reporting of distraction and the sign of the coefficient is used to state if the bias either under or over estimates the effects of distraction on injury severity.

$$q_i = \Phi^{-1}\frac{\exp(xi'\beta)}{1+\exp(xi'\beta)} = \Phi^{-1}[P_{\text{logit}}] \tag{3}$$

and

$$\lambda_i = \frac{(1/sqrt(2*\pi))*(\exp(-q*q*0.5))}{\Phi_i(q)} \quad \text{for} \quad D = 1 \tag{4}$$

$$\lambda_i = \frac{-((1/sqrt(2*\pi))*(\exp(-q*q*0.5)))}{1-\Phi_i(q)} \quad \text{for} \quad D = 0 \tag{5}$$

where $\Phi(\ )$ denotes the cumulative density function of standard normal distribution, $\Phi^{-1}(\ )$ is the inverse standard normal probability density function, and D is the state of distraction.