Accident Analysis and Prevention 84 (2015) 112–127

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Pooling data from fatality analysis reporting system (FARS) and generalized estimates system (GES) to explore the continuum of injury severity spectrum

Shamsunnahar Yasmin[a], Naveen Eluru[a,*], Abdul R. Pinjari[b]

[a] Department of Civil, Environmental and Construction Engineering, University of Central Florida, United States
[b] Department of Civil and Environmental Engineering, University of South Florida, United States

ARTICLE INFO

Article history:
Received 19 March 2015
Received in revised form 19 June 2015
Accepted 11 August 2015
Available online 3 September 2015

Keywords:
Fatality
Fatality Analysis Reporting System (FARS)
Generalized Estimates System (GES)
Data Pooling
Generalized ordered logit model

ABSTRACT

Fatality Analysis Reporting System (FARS) and Generalized Estimates System (GES) data are most commonly used datasets to examine motor vehicle occupant injury severity in the United States (US). The FARS dataset focuses exclusively on fatal crashes, but provides detailed information on the continuum of fatality (a spectrum ranging from a death occurring within thirty days of the crash up to instantaneous death). While such data is beneficial for understanding fatal crashes, it inherently excludes crashes without fatalities. Hence, the exogenous factors identified as critical in contributing (or reducing) to fatality in the FARS data might possibly offer different effects on non-fatal crash severity levels when a truly random sample of crashes is considered. The GES data fills this gap by compiling data on a sample of roadway crashes involving all possible severity consequences providing a more representative sample of traffic crashes in the US. FARS data provides a continuous timeline of the fatal occurrences from the time to crash – as opposed to considering all fatalities to be the same. This allows an analysis of the survival time of victims before their death. The GES, on the other hand, does not offer such detailed information except identifying who died in the crash. The challenge in obtaining representative estimates for the crash population is the lack of readily available "appropriate" data that contains information available in both GES and FARS datasets. One way to address this issue is to replace the fatal crashes in the GES data with fatal crashes from FARS data thus augmenting the GES data sample with a very refined categorization of fatal crashes. The sample thus formed, *if statistically valid*, will provide us with a reasonable representation of the crash population.

This paper focuses on developing a framework for pooling of data from FARS and GES data. The validation of the pooled sample against the original GES sample (unpooled sample) is carried out through two methods: (1) univariate sample comparison and (2) econometric model parameter estimate comparison. The validation exercise indicates that parameter estimates obtained using the pooled data model closely resemble the parameter estimates obtained using the unpooled data. After we confirm that the differences in model estimates obtained using the pooled and unpooled data are within an acceptable margin, we also simultaneously examine the whole spectrum of injury severity on an eleven point ordinal severity scale – no injury, minor injury, severe injury, incapacitating injury, and 7 refined categories of fatalities ranging from fatality after 30 days to instant death – using a nationally representative pooled dataset. The model estimates are augmented by conducting elasticity analysis to illustrate the applicability of the proposed framework.

1. Introduction

Traffic crashes result in physical and emotional trauma as well as huge financial losses for the individuals involved, their families and the society at large. Across the world, these crashes account for 18 deaths and 1136 disability-adjusted life years (DALY) lost per 100,000 individuals annually (WHO, 2013a,b). Researchers and

* Corresponding author.
E-mail addresses: shamsunnahar.yasmin@mail.mcgill.ca (S. Yasmin), naveen.eluru@ucf.edu (N. Eluru), apinjari@eng.usf.edu (A.R. Pinjari).

http://dx.doi.org/10.1016/j.aap.2015.08.009
0001-4575/

practitioners are constantly seeking remedial measures to reduce the burden of these unfortunate events. Toward this end, literature in transportation safety has evolved along two major streams: the first stream of research is focused on identifying attributes that result in traffic crashes and propose means to reduce the occurrence of traffic crashes (see Lord and Mannering (2010) for a review of these studies); the second stream of work examines crash events and identifies factors that impact the crash outcome and suggests countermeasures to reduce crash related consequences (injuries and fatalities) (see Savolainen et al. (2011) and Yasmin and Eluru (2013) for a review). The current research study contributes to the second stream of literature with a specific focus on driver injury severity analysis.

A number of studies have explored the impact of various factors on vehicle occupant injury severity at disaggregate level (see Bédard et al., 2002; Fredette et al., 2008; and Yasmin and Eluru, 2013 for a detailed review). These studies can broadly be categorized as: (a) studies that focus exclusively on crashes involving only fatalities (employing a sample of crashes involving fatalities) and (b) studies that examine crashes that involve all levels of injury severity – ranging from no injury to fatality (employing a random sample of traffic crashes that compile different levels of injury severity). In the United States (US), the former category of studies predominantly use the Fatality Analysis Reporting System (FARS) database (see Evans and Frick, 1988; Preusser et al., 1998; Zador et al., 2000; Gates et al., 2013) while the latter group of studies primarily employ the General Estimates System (GES) database (see Kockelman and Kweon, 2002; Eluru and Bhat, 2007; Yasmin and Eluru, 2013).

The FARS database is a census (not a sample) of all fatal crashes in the US; i.e., crashes that lead to at least one fatality within thirty consecutive days from the time of crash. The GES database, on the other hand, comprises a sample of road crashes across the US involving at least one motor vehicle traveling on a roadway and resulting in property damage, injury or death to the road users. The two datasets employed in the safety literature have their own advantages and limitations. The FARS focuses exclusively on fatal crashes. Therefore, one cannot reliably use this data to analyze the factors that increase or decrease the probability of fatality (because the data does not include crashes that do not lead to fatalities). The GES fills this gap by compiling data on a sample of roadway crashes involving all possible severity consequences (no injury, possible injury, non-incapacitating injury, incapacitating injury and fatality) providing a more representative sample of traffic crashes in the US. One of the advantages of FARS, however, is that the collected information includes the date and time of occurrence of the fatalities resulting within a 30-day time period from the crash. This detailed information provides us a continuous timeline of the fatal occurrences from the time to crash (instead of considering all fatalities to be the same). This allows for an analysis of the survival time of victims before their death. The GES, on the other hand, does not offer such detailed information except identifying who died in the crash.

Examining the impact of various exogenous factors on all levels of injury severity as well as on the survival time of fatalities can potentially play a critical role in field triage – screening process to determine the more severe cases. Preclinical trauma care is one of the most important factors affecting the outcome of motor vehicle crash (MVC) victims (Chalya et al., 2012; Palanca et al., 2003). In prehospital setting, along with the anatomic and physiological conditions of MVC victims, different mechanism-of-injuries (vehicle intrusion, occupant ejection, vehicle telemetry and death in same passenger compartment) are also considered by emergency medical service (EMS) personnel as conditions for trauma triage of victims (Sasser et al., 2012; Isenberg et al., 2011). In fact, it is evident from previous studies (Stewart, 1990) that prolonging survival beyond the first hour can potentially help avoid fatality with proper preclinical care. Hence, a refined specification of fatality might allow us to identify potential survivors that might benefit by providing pre- and post-hospital treatment.

In an effort to identifying exogenous factors that help in prolonging survival time, using detailed information available in FARS data, Yasmin et al. (2015) examined fatal crashes from a new perspective. The authors recognize that fatality is an aggregation of a continuous spectrum ranging from dying instantly to dying within thirty days of crash (as reported in the FARS data). Keeping all else same, a fatal crash that results in an immediate fatality is clearly much more severe than another crash that leads to fatality after several days. Therefore, it is useful to explicitly recognize the different levels of severity among fatal crashes. Such refined definition of fatal crashes, as opposed to lumping all fatal crashes into a single category, allows one to differentiate fatal crashes based on the survival time and to derive insights on factors that can prolong survival time. A disadvantage of the study by Yasmin et al. (2015) is that, as discussed before, the FARS dataset focuses exclusively on fatal crashes. While using the FARS data is very helpful for understanding the differences across different fatal crashes, it inherently excludes crashes with other possible, non-fatal injury severity outcomes. This makes it difficult to generalize the findings to the overall crash population. Besides, while analyzing the survival time of only fatal crash victims (using FARS data) helps in deriving the influence of various exogenous factors on survival time conditional upon the occurrence of a fatality, it doesn't allow the analyst to derive the influence of those factors in increasing the chances of survival. This is because the FARS data doesn't provide a representative sample of non-fatal crashes.

One way to address this issue is combining information from both the FARS and GES datasets into a single, disaggregate crash-level database.[1] This will bring together the strengths of both datasets – the representativeness of crashes with all injury severity outcomes from the GES data and the detailed information on fatal crashes from the FARS data. The challenge, however, lies in combining the two datasets in a statistically appropriate way. Since FARS is a census of all fatal traffic crashes in the US, all fatal crashes in the GES sample for a year should be available in the FARS data for that year. Now, if one could identify these crashes directly, it would be easy to augment the fatal crash records in GES with the detailed information from FARS. However, there is no mechanism to easily link crashes across these two databases because the datasets do not have a common identifier. Hence, an alternative, statistically valid method needs to be used for fusing information from both the datasets.

The approach is a proof of concept investigation of data pooling from two datasets while ensuring statistical validity. While, there could be various other alternative datasets for such investigation, given the extensive use of GES and FARS datasets in safety literature, they serve as good candidates for the research exercise. In this context, this paper is geared toward addressing the challenge of pooling data from GES and FARS databases. While several approaches exist in the literature to fuse information from different data sources without a common identifier (Konduri et al., 2011; Sivakumar and Polak, 2013), a simple approach is to replace fatal crashes from the GES sample by a random sample from the FARS census of crashes. We conduct statistical tests to assess if this approach suffices for

[1] To be sure, the reader would note that there have been compilation of GES and FARS datasets to obtain the Annual Traffic Safety Facts (see NHTSA, 2010). However, in these efforts, there is no attempt to pool disaggregate level data from the two sources. The report provides trends separately for FARS and GES datasets. Further, in our research, we examine the effect of exogenous variables on severity in pooled and unpooled data.