Journal of Safety Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Safety Research

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

Determinants of seat belt use: A regression analysis with FARS data corrected for self-selection

Frank Goetzke [a], Samia Islam [b]

[a] Department of Urban and Public Affairs, University of Louisville, USA
[b] College of Business and Economics, Boise State University, USA

ARTICLE INFO

Article history:
Received 20 January 2015
Received in revised form 22 April 2015
Accepted 14 July 2015
Available online xxxx

Keywords:
Seatbealt use
FARS
Peltzman effect
Safety behavior

ABSTRACT

We develop a methodology to use FARS data as an alternative to NOPUS in estimating seat belt usage. The advantages of using FARS over NOPUS are that (i) FARS is broader because it contains more variables relevant for policy analysis, (ii) FARS allows for easy multivariate regression analysis, and finally, (iii) FARS data is more cost-effective. We derive coefficient estimates for categories such as vehicle occupants' age and night time seat belt use that observational surveys like NOPUS cannot easily provide. Although this is primarily a methodological paper, we present and discuss our results in the context of public policy so that our findings are relevant for road safety practitioners, researchers, and policy makers. Our results indicate that policies should focus on passengers (as opposed to drivers), male and young vehicle occupants, and that law enforcement should focus on pick-up trucks, rural roads, and nights. We find evidence that primary seat belt laws are effective.

1. INTRODUCTION

While the effectiveness of seat belts is well established in research (Blincoe, 1994; Carpenter & Stehr, 2008; Crandall, Olson, & Sklar, 2001; Evans, 1986; Kahane, 2000; Klein & Walz, 1995; Levitt & Porter, 2001; McCartt & Northrup, 2004; Partyka, 1988; Partyka & Womble, 1989), U.S. seat belt usage rates have been relatively low compared to other developed nations (NHTSA, 2007).[1] In 1997, the federal government set targets to increase seat belt usage from 68% in 1996 to 85% by 2000, and then to 90% in 2005, both of which went unmet according to the annual National Occupant Protection Use Survey (NOPUS).[2] The National Highway Traffic Safety Administration (NHTSA) of the U.S. Department of Transportation reported that over half of passenger vehicle occupants killed in traffic accidents in 2006 died unbuckled (NHTSA, 2007).

With the rise in federal funding for highway safety initiatives and awareness programs, and more stringent primary and secondary law enforcement at the state level, the failure to meet targeted usage rates is confounding. But more importantly, it points to a need for targeted policies to incentivize usage. Before we can design such policies it is critical to determine the factors that affect vehicle occupants' decision to wear a seat belt. In the past, seat belt effectiveness studies that used NOPUS data could not address many of these factors because of the lack of certain details in the NOPUS data.

Research on seat belt usage typically utilizes one of two publicly available data sources: (a) National Occupant Protection Use Surveys (NOPUS), and (b) Fatality Analysis Reporting System (FARS). The National Center for Statistics and Analysis of the NHTSA consider NOPUS to be their most reliable data set tracking the trends in seat belt usage by motorists. However, the observational nature of NOPUS data not only subjects it to some limitations due to the probability of human error in the data collection, but also due to lack of reliable data on vehicle occupants' personal characteristics or nighttime travel behavior.

Fatality Analysis Reporting System (FARS) is the other available database for evaluating the usage rates of occupant's protection devices.[3] An advantage of using the FARS database over NOPUS is that it is more comprehensive in the reported variables, providing, for example, additional data for vehicle occupant characteristics, as well as nighttime data. However, one critical problem with FARS data is that it underestimates seat belts usage when compared to estimates obtained from observational data such as NOPUS[4] (Salzberg, Yamada, Saibel, & Moffat, 2002) due to the nature of the reporting system. To be included in FARS, a crash must result in the death of a person

E-mail addresses: f0goet01@louisville.edu (F. Goetzke), samiaislam@boisestate.edu (S. Islam).

[1] It must be noted here, however, that safety devices such as seat belts and air bags can cause injury, and even fatalities at times (Hollands, Winston, Stafford, & Shochat, 1996; Morris & Borja, 1998).

[2] NOPUS is an annual survey providing probability-based observed data on seat belt use in the United States conducted by the National Center for Statistics and Analysis of the NHTSA.

[3] FARS is a national census of motor vehicle fatalities which has been collected since 1975 and contains information on over 989,451 motor vehicle fatalities within the 50 States, the District of Columbia, and Puerto Rico as well as over 100 different coded data elements that characterize the crash, the vehicle, and the people involved.

[4] NOPUS is not the only source of probability-based observational data. States conduct their own observational surveys annually. However, several states only survey passenger cars and omit light trucks (Kahane, 2000).

http://dx.doi.org/10.1016/j.jsr.2015.07.004
0022-4375/

(occupant of a vehicle or a non-occupant) within 30 days of the crash. Since it lists only those crashes where there is at least one fatality, there is a potential issue of sample selection given that an individual's seat belt use affects his or her probability of death, which in turn influences whether the crash is included in the data because of the correlation between seat belt use and fatality. It has been shown that such sample selection leads to biased regression coefficient estimates (Angrist & Krueger, 1999; Heckman, 1979; Heckman, Ichimura, Smith, & Todd, 1996). The extent of this sample selection bias becomes even more significant when we consider that only about 0.5% of motor vehicle crashes involve a fatality, and in 90% of the incidents there is just a single death (NHTSA, 1998). Had that death not occurred, the crash would not be included in the FARS database. Therefore, empirically, the impact of sample selection can be substantial, and failing to account for it leads to estimates that systematically understate seat belt usage. Previously, Salzberg et al. (2002) investigated seat belt usage rates by comparing FARS data with observation surveys and concluded that unbelted occupants are over-represented in fatal collisions for two reasons: (a) because of a greater chance of involvement in potentially fatal collisions in the first place, and (b) because they are not afforded the protection of seat belts when a collision does occur. Their model focused on risk but they did not address the sample selection bias.

In our 2009 paper (Islam & Goetzke, 2009), we used an identification method developed by Levitt and Porter (2001)[5] to correct for the sample selection bias in FARS data in order to obtain a credible estimate of seat belt usage in the United States. Curiously, but rather insightfully, the sample selection problem that arises because of the exclusion of a majority of non-fatal crash statistics from the FARS data set is countered by further limiting the data. We used two different specifications to correct the sample selection bias in FARS data: (a) correction based on strict independence of seat belt choice (Model 1, details of which are described in Section 3.1 below), and (b) correction based on strict dependence of seat belt choice (Model 2, described in Section 3.2 below). By applying these corrections, we showed that the FARS database can be established as an acceptable and comparable alternative to the observational NOPUS data (Islam & Goetzke, 2009). In this paper, we extend our previous work (Islam & Goetzke, 2009) on correcting the sample selection bias evident in the FARS dataset and using a multivariate regression analysis, address the following critical question: What factors affect the decision to use seat belts? This question is critical in designing effective policies. For example, are there particular socio-demographic groups that policy should focus on? Do specific times of day or regions require heightened enforcement? These specific policy-relevant questions are precisely what the corrected FARS dataset allows us to answer due to the rich cross-sectional detail that it contains on each reported event.

While in Islam and Goetzke (2009) we presented a method to correct sample selection bias in FARS data and then proceeded to conduct a univariate statistical analysis to show its applicability as an alternative to NOPUS data, in this paper we go further by presenting results from a multivariate regression analysis of different categories of vehicle occupants, times of day, regions, and vehicle types on seat belt use. We begin our analysis with model specifications similar to our 2009 paper (described above). For consistency of methodology, we use the same 2006 data that we used in our earlier paper. We produce coefficient estimates for categories such as nighttime seat belt use that NOPUS cannot provide since roadside surveys are conducted during daylight hours. We develop a methodology to use FARS data as an alternative to NOPUS. The advantages of using FARS over NOPUS are that FARS data are: (a) broader because it contains more variables relevant for policy analysis (e.g. nighttime observations), (b) better in quality (once corrected for sample selection bias) because multivariate regression analysis can be easily conducted, and (c) more cost-effective. Although this is primarily a methodological paper, we present and discuss our results in the context of public policy so that our findings become relevant for road safety practitioners, researchers, and policymakers.

The remainder of this paper is organized as follows: Section 2 describes the methodology and data used for our analysis. Section 3 describes our model specifications and variables. Section 4 presents the results and a discussion of their policy implications. Section 5 presents a concluding summary of our findings and outlines possible future extensions of this paper.

2. METHODOLOGY AND DATA

For our analysis we utilize 2006 data from the Fatality Analysis Reporting system (FARS) compiled by the National Highway Traffic Safety Administration (NHTSA).[6] We focus on the 2006 data to remain compatible with our previous analysis of NOPUS (Islam & Goetzke, 2009). We define the following categories of vehicles and passenger characteristics for easy correspondence with the NOPUS data: Occupants wearing seat belts (shoulder belts or both lap and shoulder belts) and occupants without seat belts (only lap belts or no belts); front seat passengers (seating position code 11 and 13); Passenger cars (body type codes 1–12), vans & SUVs (body type codes 14–21). 'Day' is considered to be the hours between 8 a.m. and 6 p.m., following NHTSA survey times. The categories for expressways and surface streets are derived from the reported roadway functional class. We follow the classification for urban, suburban, and rural roadways in combination with the city code.

We apply a binary logit model in our analysis to determine the likelihood of seat belt usage given various occupant, vehicle, and built environment characteristics. In general, a logistic regression model with multiple predictor variables can be characterized as follows (Hosmer & Lameshow, 2010):

$$logit(p) = log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k.$$

Each estimated coefficient is the expected change in the log-odds, $\log(p\ /\ (1-p))$, of the dependent binary choice variable, that is, using seat belts as in our case, for a unit increase in the corresponding independent variable, holding the other predictor variables constant. When a binary outcome variable is modeled using logistic regression, the interpretation of the logistic regression coefficients becomes easier after transforming of the logistic regression coefficient into odds ratios. The odds ratio is the odds of an event occurring in one group compared to the odds of it happening in another group, or a sample based estimate of that ratio.[7] An odds ratio of 1 indicates that the condition or event under study is equally likely in both groups. Using the odds ratio, it is also possible to compare the regression results of the FARS data with the NOPUS survey findings.

There are several advantages to this method that allows researchers to use FARS data instead of NOPUS in their estimations. As mentioned earlier, FARS data contains more variables relevant for policies, and, with corrections for sample section bias, allows for multivariate regression analysis to be conducted. It is also relatively inexpensive to collect and maintain.

[5] Levitt and Porter (2001) use a simple but ingenious identification strategy that allows them to directly estimate the impact of seat belts and air bags on crash survival rates, despite sample selection in the data.

[6] Source: http://www-fars.nhtsa.dot.gov/Main/index.aspx.

[7] If the probabilities of an event happening are p_1 (first group) and p_2 (second group), then the odds ratio is:
$(p_1 / (1-p_1)) / (p_2 / (1-p_2)) = p_1/q_1 / p_2 / q_2 = p_1q_2 / p_2q_1$ where $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

Please cite this article as: Goetzke, F., & Islam, S., Determinants of seat belt use: A regression analysis with FARS data corrected for self-selection, *Journal of Safety Research* (2015), http://dx.doi.org/10.1016/j.jsr.2015.07.004