Accident Analysis and Prevention 131 (2019) 327–335

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Endogenous commercial driver's traffic violations and freight truck-involved crashes on mainlines of expressway

Jungyeol Hong[a], Juneyoung Park[b], Gunwoo Lee[c,*], Dongjoo Park[a]

[a] Department of Transportation Engineering, The University of Seoul, 163 Seoulsiripdae-ro Dongdaemun-gu, Seoul, 02504, South Korea
[b] Department of Transportation & Logistics Engineering, Hanyang University, 55 Hanyangdeahak-ro, Ansan, Gyeonggi-do, 15588, South Korea
[c] Department of International Logistics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, 02504, South Korea

ARTICLE INFO

Keywords:
Freight truck-involved crash
Two-stage residual inclusion
Expressway
Traffic violation
Endogeneity

ABSTRACT

Freight truck-involved crashes result in a high mortality rate and significantly impact logistic costs; therefore, many researchers have analyzed the causes of truck-involved traffic crashes. In the existing literature, it was found that truck-involved crashes are affected by factors such as road geometry, weather, driver and vehicle characteristics, and traffic volume based on a variety of statistical methodologies; however, the endogenous impact resulting from driver traffic violation has not been considered. The goal of the study is to discover the factors influencing freight vehicle crashes and develop more accurate crash probability estimation by explaining the endogenous driver traffic violations. To achieve the purpose of this study, we applied the two-stage residual inclusion (2SRI) approach, a methodology used in the nonlinear regression analysis model for capturing the endogeneity issue. This method improves the accuracy of the model by capturing the unobserved effects of driver traffic violations. From the results, traffic violations were identified to be influenced by the driver's physical condition, as well as driver and vehicle characteristics. Furthermore, variables of driver traffic violations such as improper passing, speeding, and safe distance violation were found to be endogenous in the probability model of freight truck crashes on expressway mainlines.

1. Introduction

The share of freight transportation using roadways has increased by more than 90% in South Korea since 2011, mainly due to the small size of the national land area (99,720 km^2). Therefore, transporting domestic freight on expressways is more efficient in terms of time and cost as compared to using modes such as air, railway, or sea. The average daily volume for freight transport trucks on the expressway was 13,211veh/day in 2016 and represented 26.4% of the total average daily traffic volume. In addition, of all the 48,593 crashes that occurred on expressways during the last five years, 15,011 involved trucks and accounted for approximately one-third of all expressway traffic crashes (Korea Expressway Corporation, 2017). Freight truck-involved crashes increase not only the number of human casualties but also logistic costs. Particularly, a crash involving a freight truck transporting hazardous materials can cause disastrous socioeconomic and environmental damages to the country. Thus, concern for the safety and management of freight transport trucks and drivers on expressways is growing.

In the past, numerous studies on truck-involved crashes have placed much emphasis on roadway design and its operational characteristics (Miaou, 1994; Harkey et al., 1996; Harwood et al., 1999; Schneider et al., 2009; Poe, 2010; Lemp et al., 2011; and Dong et al., 2015). Miaou (1994) found that truck-involved crashes were associated with geometric roadway design such as horizontal curvature, vertical grade, and shoulder. Harkey et al. (1996) showed that operational characteristics of longer combination vehicles and geometric design had an impact on safety, and Harwood et al. (1999) found that larger and more massive trucks were significantly affected by lengths of horizontal curves. In a study by Schneider et al. (2009), the impact of horizontal curvature on truck crashes was also demonstrated. Poe (2010) examined the effects of facility design characteristics, operating characteristics, and corridor operating attributes on truck crashes, and Lemp et al. (2011) discovered that factors such as curved road, uphill/downhill grade, road crest, and sag affect oversized truck crashes. Dong et al. (2015) reported that passenger car average annual daily traffic (AADT), large truck AADT, segment length, degree of horizontal curvature, terrain type, land use, median type, number of lanes, lane width, right side shoulder width, lighting condition, and posted speed limits significantly affect truck-involved crashes on highways in the state of Tennessee.

However, in addition to the roadway geometry effects, the cause of

* Corresponding author.
E-mail addresses: jyhong9868@uos.ac.kr (J. Hong), juneyoung@hanyang.ac.kr (J. Park), gunwoo@cau.ac.kr (G. Lee), djpark@uos.ac.kr (D. Park).

https://doi.org/10.1016/j.aap.2019.07.026
Received 8 August 2018; Received in revised form 19 June 2019; Accepted 25 July 2019
Available online 01 August 2019
0001-4575/

truck driver-related crashes cannot be ignored in the analysis of truck-involved crashes on expressways. Raw data analysis of traffic crashes on Korean expressways (Korea Expressway Corporation, 2017) shows that 74% of the truck-involved crashes are caused by driver factors such as violation of traffic regulations, driving error, and physical condition. Some previous studies focused on the relationship between truck driver characteristics and crashes (Lyles et al., 1991; Khattak et al., 2003; Hallmark et al., 2009; Cantor et al., 2010; Zhu and Srinivasan, 2011). Campbell (1991) analyzed the association of large trucks involved in fatal crashes with driver age and found that fatal crash rates for large trucks increased with decreasing driver age. In a study by Khattak et al. (2003), the researchers concluded that risky driving behavior such as driving under the influence (DUI), speeding, or not wearing a seat belt led to increased injury severity in single-truck crashes. Hallmark et al. (2009) found that driver conditions such as fatigue, anger, distraction, or unfamiliar roadway caused more truck crashes. Also, Cantor et al. (2010) showed that driver factors such as age, weight, height, gender, employment stability, and previous violations were likely to have an effect on the occurrence of truck crashes. Zhu and Srinivasan (2011) established that driver distraction, alcohol use, and emotion were associated with higher crash severity in truck crashes. The study by Chen and Zhang (2016) discovered that crashes involving fatigued truck drivers on Jianxi and Shaanxi highways have an association with driver's attributes such as age, gender, over speeding, overloading, brake performance, roadway types, pavement condition, sharp curve and steep grade, weather, visibility, time of day, and season. Furthermore, the cause of truck driver's violations has been investigated in several previous studies (Hartley and El Hassani, 1994; Matthews, 2002; Lagarde et al., 2004; Rowden et al., 2011). Hartley and El Hassani (1994) examined the relationship between self-reported stress and traffic violations, and the study by Matthews (2002) demonstrated that driver stress was associated with higher traffic crash rates and speeding convictions. Similarly, Lagarde et al. (2004) found that life stress is likely to increase driver's traffic violations, and Rowden et al. (2011) concluded that the risk-taking trait which is a dimension of the Driver Stress Inventory had the most substantial impact on violations.

Nevertheless, studies determining the relationship between the endogeneity of drivers' traffic violations and truck-involved crashes on a specific segment type of expressway are lacking. As shown in Table 1, linear regression and logit models were used in most of the previous studies to identify the relationship between the number or probability of truck-involved crashes and multiple regressors. These modeling approaches do not account for the endogenous variable of truck driver traffic violations; therefore, in the present study, the appropriate methodology is applied to fit the model while considering of endogeneity.

Based on the reasons mentioned above, a two-stage residual inclusion (2SRI) estimation, which can identify the unobserved effect, namely, the influence of endogenous truck drivers' traffic violations on truck-involved crashes on the expressway, was applied in the present study. Terza et al. (2008) stated that the 2SRI is a comparatively simple estimation technique that addresses endogeneity bias and is appropriate in various nonlinear regression contexts. Therefore, applying the 2SRI approach in this study allowed for the development of more precise probability models for truck-involved crashes and to explain the endogenous driver traffic violation variables by using instrumental variables (IVs).

First, the effects of variables relating to the driver, vehicle, roadway geometry, roadside features, pavement, weather, season, and time on the probability of truck-involved crashes on a specific type of segment were investigated. Second, a 2SRI estimation technique was used to determine the likelihood of truck-involved crashes while capturing the endogeneity issue of truck driver traffic violations in order to predict the risk more accurately. Variables collected from raw data of freight truck-involved crashes is discussed in the next section. Then, a demonstration of the methodological approach for 2SRI and the results from the models are discussed with their significant implications. The last section presents a summary of the overall study and a discussion of the main contributions and recommendations for future studies.

2. Data description

In the event of a traffic crash on the expressway, crash investigators from the Korea Expressway Corporation (KEC) are dispatched to the site to collect crash-related information such as roadway geometry, weather, day of week and time, detailed information of vehicles and drivers involved in the crash, cause of the crash, and damage resulting from the crash. The raw data of these individual crashes are accumulated in the highway crash database system and are used for crash risk assessment and safety management. This study used data collected by the KEC, which comprises all crashes that occurred on all expressways in South Korea from 2012 to 2016.

Individual crash data includes information on the types of vehicle (passenger car, bus, truck, and tractor-trailer) and their respective sizes (small, medium, large). Trucks defined as small-sized weigh less than 2.5 tons, mid-sized trucks are those that weigh more than 2.5 tons and less than 10 tons, and large-sized trucks weigh more than 10 tons. Considering the effect of truck size is critical in safety analysis since it has a complex relationship with crash type and fatalities (Vallette et al., 1981; Jones et al., 1983; Castillo-Manzano et al., 2016).

We preprocessed the raw crash data and built a truck-specific crash dataset by extracting truck-involved crashes from the raw data. First, some pieces of information were missing during the investigation process since this study used crash data collected by the highway crash investigators. In this case, we omitted the observation. Thus, the number of truck-involved crashes observed during a five-year period across 17 provinces categorized into expressway segments in South Korea was 15,011. We set up 88 explanatory variables from the crash data in binary, count, and continuous forms. These consisted of variables, including annual average daily traffic (AADT) and the ratio of truck traffic volume. Further, the driver's traffic violation variables were divided into improper passing, speeding, unsafe distance, and other traffic violations. The driver's attributes include age and physical conditions such as fatigue, illness, drowsy, and normal; each attribute is set as a dummy variable. The roadway geometry considered in terms of horizontal and vertical alignments, median and shoulder types, and weather condition, month of year, day of week, and time of day were all transformed to dummy variables. Table 2 shows the descriptive statistics of all variables used in this study.

3. Methodology

The explanatory variable is endogenous when the error term and an independent variable are correlated. Statistically, if at least one explanatory variable is endogenous, the Ordinary Least Squared (OLS) model which can be written in matrix notation as $y = \beta X + u$ produces a biased estimate since the $cov(x, u) \neq 0$. This clearly defines an endogeneity issue. The three main causes of endogeneity are simultaneous causality, correlated omitted variable bias, and measurement error in the independent variable, x. Therefore, Instrumental Variables (IV) methods are used to capture potential bias from unmeasured confounders by specifying an IV that correlates with the independent variable but not the error term (Basu et al., 2017).

The 2SLS (Two-Stage Least Squares) approach which uses IV's has been mainly used in previous research; however, given the binary outcome variable in various analyses, the 2SLS is deemed not suitable since it is estimated by minimizing the sum of squared residuals in a way similar to that of the linear regression. The 2SRI (Two-Stage Residual Inclusion) approach, which is another IV method was introduced as a proper method for a model that has a dependent variable of a binary or nonlinear form (Blundell and Smith, 1989, 1993; Newey, 1987; Rivers and Vuong, 1988; Smith and Blundell, 1986; Terza et al.,

Table 1
Previous studies for truck-involved crashes.

Author(s) and Years	Methodology	Key factors
Anderson and Dong, 2017	Mixed Logit Model	time of day, weather, vehicle characteristics, drivers' demographics
Dong et al., 2017	Bayesian Negative Binomial Model	weather, vehicle characteristics, drivers' demographics
Elshamly et al., 2017	Logit Model	drivers' demographics
Kim et al., 2017	Logit Model, Ordered Logit Model	roadway geometry, time of day, weather, drivers' demographics
Teoh et al., 2017	Matched Case Control Design	time of day, vehicle characteristics
Uddin and Huynh, 2017	Mixed Logit Model	roadway geometry, time of day, weather, drivers' demographics
Yuan et al., 2017	Logit Model	time of day, weather, drivers' demographics
Bin Islam and Hernandez, 2016	Random Parameters Tobit Model	roadway geometry, weather, vehicle characteristics, drivers' demographics
Chen and Zhang, 2016	Stepwise logistic regression model	driver's gender, age, driving experience, and over speeding, vehicle's commercial status, overloading and brake performance, roadway types, slippery pavement, existence of sharp curve, long steep grade, time of day, season, weather and visibility
Dong et al., 2015	Multinomial Logit Model	roadway geometry, weather, drivers' demographics
Islam, 2015	Random Parameters Probit, Mixed Logit Model	roadway geometry, time of day, weather, vehicle characteristics, drivers' demographics
Pahukula et al., 2015	Random Parameters Logit Model	roadway geometry, time of day, weather, drivers' demographics
Chen and Xie, 2014	Logit Model	time of day
Klassen et al., 2014	Mixed Logit Model	roadway geometry, time of day, weather, vehicle characteristics, drivers' demographics
Islam and Hernandez, 2013	Mixed Logit Model	roadway geometry, time of day, weather, drivers' demographics
Sharma and Landge, 2013	ZINB Model	roadway geometry
Kotikalapudi, 2012	Cross Classification Analysis, Logit Model	time of day, weather
Chen and Chen, 2011	Mixed Logit Model	roadway geometry, time of day, weather, vehicle characteristics, drivers' demographics
Islam and Hernandez, 2011	Random Parameters Tobit Model	roadway geometry, time of day, weather, vehicle characteristics
Bezwada, 2010	Multinomial Logit Model	roadway geometry, weather, drivers' demographics
Cantor et al., 2010	Poisson Model	drivers' demographics
Park and Jovanis, 2010	Logit Model	time of day
Hanowski et al., 2009	Logit Model	time of day
Rescot et al., 2009	Multiple Linear Regression Model	weather
Iragavarapu and Lord, 2007	Negative Binomial Model	roadway geometry, weather, vehicle characteristics, drivers' demographics
Chang and Chen, 2005	Data Mining	roadway geometry
Khorashadi et al., 2005	Multinomial Logit Model	weather, vehicle characteristics, drivers' demographics
Golob and Regan, 2004	Logit Model	time of day, weather
Rodriguez et al., 2003	Zero-Inflated Poisson Model	work condition and wage, drivers' demographics

2008; Basu et al., 2017; Terza, 2017). Based on studies conducted by Terza et al. (2008, Terza, 2017), the basic framework for nonlinear modeling is described in Eq. (1) as follows:

$$\mathrm{E}[\mathrm{y}x_e,\ x_0,\ x_u] = M(x_e\beta_e + x_o\beta_o + x_u\beta_u) \tag{1}$$

$$x_e = [x_{e1},\ x_{e2},\ x_{e3}\cdots,x_{eS}],\ x_0 = [x_{o1},\ x_{o2},\ x_{o3}\cdots,x_{oK}] = x_u = [x_{u1},\ x_{u2},\ x_{u3}\cdots,x_{uS}]$$

where M(·) is a nonlinear function. x_e represents a $1 \times \mathrm{S}$ vector of endogenous independent variables, and x_o denotes a $1 \times \mathrm{K}$ vector of exogenous independent variables. x_u is a $1 \times \mathrm{S}$ vector of unobservable latent confounder variables that affect the dependent variable y. For this study, we define the dependent variable y as the probability of truck-involved crash occurrence and x_e as a vector of independent endogenous traffic violation variables namely improper passing, overspeeding, unsafe distance, and other violations. The independent endogenous variables listed as elements of the vector x_e have an impact on the truck-involved crash (y), while controlling for the variables contained in the vector of exogenous variables (x_o). The exogenous independent variables contained in the vector x_o include weather condition, driver and vehicle attributes, geometry features, month of year, and time of day. x_u is correlated with the probability of truck-involved crash occurrence as well as endogenous traffic violation variables. β_e, β_o, and β_u are vectors of regression parameters. Therefore, the base model is defined as in Eq. (2) below:

$$\mathrm{y} = M(x_e\beta_e + x_o\beta_o + x_u\beta_u) + e \tag{2}$$

where e denotes the random error and can be explained by $\mathrm{y} - M(x_e\beta_e + x_o\beta_o + x_u\beta_u)$ when $\mathrm{E}[ex_e,\ x_0,\ x_u] = 0$.

If x_e and x_u are correlated, the issue of endogeneity occurs. Hence, capturing the bias caused by the endogeneity issue through the inclusion of IV's is necessary. In order to validate the association between x_e and x_u, and to capture endogeneity by using IV's, Terza et al. (2008, Terza, 2017) defined the nonlinear auxiliary Eq. (3) as follows:

$$x_e = \gamma(w;\alpha) + x_u \tag{3}$$

where $\gamma(w;\alpha)$ is the conditional mean of x_e given w= $[x_o\, w^+]$ and α is a vector of parameters, $(\mathrm{K}\times S^+) \times 1$. $w^+ = [w_1^+,\ \cdots,w_{S+}^+]$. According to Terza et al. (2008, Terza, 2017) $w^+$is defined as a $1 \times S^+$ vector of IV's and must meet three conditions as follows. First, the elements of w^+cannot be correlated with the vector of unobservable latent confounder variables, x_u. Second, the elements of w^+ must be strongly correlated with the vector of endogenous independent variables, x_e. Lastly, they cannot have a direct impact on the dependent variable y, and cannot be correlated with the error term e.

In the first stage of 2SRI, the appropriate nonlinear least square estimates of the vector α ($\hat{\alpha}$) are computed using Eq. (3), and predictors of x_e are calculated using Eq. (4) below.

$$\hat{x_e} = \gamma(w;\hat{\alpha}) \tag{4}$$

Where $\hat{\alpha}$ indicates the consistent estimate of α in the first stage of 2SRI. The nonlinear least square estimator for the first stage is the maximizer of $\sum_{s=1}^{n}\ln\{g(x_{es}w_s;\alpha)\}$. x_{es} and w_s are the values of x_e and w for the s *th* observation in the sample.

Eq. (5) below describes the residuals in the first stage of the model and is defined as follows:

$$\hat{x_u} = x_e - \gamma(w;\hat{\alpha}) \tag{5}$$

The four residuals $\hat{x_u}$ of the traffic violation models developed in the first stage were applied in the second stage model. In the second stage of the 2SRI, the model is developed by including the residuals ($\hat{x_u}$), original endogenous regressor variables (x_e), and exogenous independent variables (x_o) as shown in Eq. (6) below:

$$\mathrm{y} = M(x_e\beta_e + x_o\beta_o + \hat{x_u}\beta_u) + e^{2SRI} \tag{6}$$

where e^{2SRI} is the regression error term. Each β can be obtained using

Table 2
Descriptive Statistics.

Variable descriptions	Mean	Std.
Annual average daily traffic (vehicles/a day)	61115	45903
Ratio of truck traffic volume (truck volume/vehicle volume)	0.288	0.077
Crash location: mainline (1 = yes, 0 = otherwise)	0.602	0.489
Improper passing (1 = yes, 0 = otherwise): Violating the overtaking method prescribed by the Road Traffic Act	0.011	0.105
Speeding (1 = yes, 0 = otherwise): Exceeding the speed limit specified by the Road Traffic Act	0.025	0.157
Unsafe distance (1 = yes, 0 = otherwise): Violating the safety distance rule defined by the Road Traffic Act	0.191	0.393
Other traffic violation (1 = yes, 0 = otherwise): Wrong-way driving, violation of HOV lane usage rule, and violation of regulations on Road Traffic Act other than improper passing, speeding, and unsafe distance.	0.089	0.285
Drowsy driving (1 = yes, 0 = otherwise)	0.165	0.371
Negligent driving (1 = yes, 0 = otherwise)	0.251	0.433
Driver is under normal condition before crash (1 = yes, 0 = otherwise)	0.900	0.300
Driving under influence (1 = yes, 0 = otherwise)	0.008	0.087
Driver was tired (1 = yes, 0 = otherwise)	0.078	0.268
Drivers fault unknown (1 = yes, 0 = otherwise)	0.008	0.09
Driver illness (1 = yes, 0 = otherwise)	0.001	0.028
Obstructions on the road (1 = yes, 0 = otherwise)	0.044	0.205
Falling loads (1 = yes, 0 = otherwise)	0.036	0.185
Brake failure (1 = yes, 0 = otherwise)	0.020	0.139
Vehicle breakdown (1 = yes, 0 = otherwise)	0.020	0.141
Tire burst (1 = yes, 0 = otherwise)	0.060	0.238
Weather: snowy (1 = yes, 0 = otherwise)	0.031	0.172
Weather: sunny (1 = yes, 0 = otherwise)	0.636	0.481
Weather: rainy (1 = yes, 0 = otherwise)	0.188	0.391
Weather: foggy (1 = yes, 0 = otherwise)	0.005	0.073
Weather: cloudy (1 = yes, 0 = otherwise)	0.140	0.347
Horizontal alignment: no curvature (1 = yes, 0 = otherwise)	0.796	0.403
Horizontal left curvature > 1000 m (1 = yes, 0 = otherwise)	0.095	0.293
Horizontal right curvature > 1000 m (1 = yes, 0 = otherwise)	0.105	0.306
Horizontal left curvature > 500 m, < 1000 m (1 = yes, 0 = otherwise)	0.001	0.035
Horizontal right curvature > 500 m, < 1000 m (1 = yes, 0 = otherwise)	0.001	0.031
Horizontal left curvature < 500 m (1 = yes, 0 = otherwise)	0.001	0.025
Horizontal right curvature < 500 m (1 = yes, 0 = otherwise)	0.001	0.03
Vertical alignment: no curvature (1 = yes, 0 = otherwise)	0.684	0.465
Vertical grade uphill < 1% (1 = yes, 0 = otherwise)	0.043	0.202
Vertical grade downhill < 1% and > 3% (1 = yes, 0 = otherwise)	0.087	0.282
Vertical grade uphill > 3% (1 = yes, 0 = otherwise)	0.030	0.172
Vertical grade uphill < 1% and > 3% (1 = yes, 0 = otherwise)	0.077	0.267
Vertical grade downhill < 1% (1 = yes, 0 = otherwise)	0.045	0.207
Vertical grade downhill > 3% (1 = yes, 0 = otherwise)	0.033	0.18
Median type: guardrail (1 = yes, 0 = otherwise)	0.088	0.283
Median type: 127 cm fixed barrier (1 = yes, 0 = otherwise)	0.334	0.472
Median type: 81 cm fixed barrier (1 = yes, 0 = otherwise)	0.175	0.38
Median type: others (1 = yes, 0 = otherwise)	0.067	0.25
Median type: grass median strip (1 = yes, 0 = otherwise)	0.021	0.145
Median type: moving (1 = yes, 0 = otherwise)	0.011	0.102
No median (1 = yes, 0 = otherwise)	0.289	0.453
Shoulder type: guardrail (1 = yes, 0 = otherwise)	0.385	0.487
Shoulder type: cable (1 = yes, 0 = otherwise)	0.001	0.02
Shoulder type: pipe (1 = yes, 0 = otherwise)	0.001	0.031
Shoulder type: fence (1 = yes, 0 = otherwise)	0.007	0.084
Shoulder type: rock (1 = yes, 0 = otherwise)	0.018	0.134
Shoulder type: concrete (1 = yes, 0 = otherwise)	0.089	0.285
Trailer (1 = yes, 0 = otherwise)	0.147	0.354
Special freight truck (1 = yes, 0 = otherwise)	0.020	0.139
Truck size: large (1 = yes, 0 = otherwise)	0.199	0.399
Truck size: mid (1 = yes, 0 = otherwise)	0.315	0.465
Truck size: small (1 = yes, 0 = otherwise)	0.311	0.463
Driver gender: male (1 = yes, 0 = otherwise)	0.102	0.303
Drivers' age: under 20 years old (1 = yes, 0 = otherwise)	0.057	0.232
Drivers' age: > = 20 and < 30 years old (1 = yes, 0 = otherwise)	0.005	0.071
Drivers' age: > = 30 and < 40 years old (1 = yes, 0 = otherwise)	0.013	0.111
Drivers' age: > = 40 and < 50 years old (1 = yes, 0 = otherwise)	0.021	0.145
Drivers' age: > = 50 and < 60 years old (1 = yes, 0 = otherwise)	0.017	0.13
Drivers' age: > = 60 years old (1 = yes, 0 = otherwise)	0.006	0.077
Crash date: Monday (1 = yes, 0 = otherwise)	0.163	0.37
Crash date: Tuesday (1 = yes, 0 = otherwise)	0.179	0.384
Crash date: Wednesday (1 = yes, 0 = otherwise)	0.159	0.365
Crash date: Thursday (1 = yes, 0 = otherwise)	0.159	0.366
Crash date: Friday (1 = yes, 0 = otherwise)	0.168	0.374
Crash date: Saturday (1 = yes, 0 = otherwise)	0.115	0.32
Crash date: Sunday (1 = yes, 0 = otherwise)	0.056	0.23
January (1 = yes, 0 = otherwise)	0.072	0.258
February (1 = yes, 0 = otherwise)	0.064	0.245
March (1 = yes, 0 = otherwise)	0.075	0.263
April (1 = yes, 0 = otherwise)	0.084	0.278
May (1 = yes, 0 = otherwise)	0.088	0.283
June (1 = yes, 0 = otherwise)	0.09	0.286
July (1 = yes, 0 = otherwise)	0.102	0.303
August (1 = yes, 0 = otherwise)	0.093	0.29
September (1 = yes, 0 = otherwise)	0.089	0.285
October (1 = yes, 0 = otherwise)	0.088	0.283
November (1 = yes, 0 = otherwise)	0.081	0.273
December (1 = yes, 0 = otherwise)	0.074	0.262
Crash occurrence time: 1AM - 4AM (1 = yes, 0 = otherwise)	0.027	0.162
Crash occurrence time: 4AM - 8AM (1 = yes, 0 = otherwise)	0.026	0.161
Crash occurrence time: 8AM - 12 P M (1 = yes, 0 = otherwise)	0.025	0.157
Crash occurrence time: 12 P M - 4 P M (1 = yes, 0 = otherwise)	0.026	0.159
Crash occurrence time: 4 P M - 8 P M (1 = yes, 0 = otherwise)	0.033	0.178
Crash occurrence time: 8 P M - 0AM (1 = yes, 0 = otherwise)	0.043	0.203

the appropriate nonlinear least square in Eq (2), and $\hat{x_u}$ is applied instead of x_u. The nonlinear least square estimator for the second stage is the maximizer of $\sum_{s=1}^{n} \ln\{f(y_s x_{es},\ x_{os},\ \hat{x_{us}};\beta)\}$. y_s is the observed values of y, and x_{os} represents the observed x_o values for the s *th* observation in the sample. $\hat{x_{us}}$ denotes the residual from the first stage model for the s *th* observation in the sample data.

4. Results

The purpose of our study is to analyze the probabilities of truck crash occurrence at a mainline with endogenous driver's traffic violation variables. Thus, the dependent variable for the final model is of a binary and non-linear form. Moreover, we test the hypothesis that the variables of driver's traffic violations are correlated with the error term. With the model assumptions, we employed the 2SRI approach to estimate the probability of truck-involved crashes occurring on a mainline. The statistical tool, Stata 15, was used for the analysis.

The first stage of the model addressed the relationships between the probability of a truck driver's specific traffic violation and driver/vehicle attributes. The second component of the model predicts the probability of truck-involved crash occurrence on a mainline with a variety of exogenous regressors such as roadway geometry, roadside features, weather, time of day, season, and traffic conditions, as well as residuals obtained from the first stage of the model and endogenous variables of truck driver traffic violations. In the first-stage model of the 2SRI analysis, the probabilities of truck driver traffic violations were estimated in the form of a logit model. The probability of improper passing violation was significantly associated with truck driver condition, type of truck, and driver age. Table 3 shows that the probability of improper passing increases when drivers are between 20 and 40 years of age but decreases when driving trailers and special trucks (hazardous material transporting truck, special freight truck, and tow truck).

The pattern of truck-involved crashes based on driver speeding violation shows different results when compared with the variable for improper passing. The probability of speeding increases with small and mid-size trucks but decreases with large-size trucks. Supposedly, quickly increasing speed is difficult for larger trucks; therefore, drivers of larger trucks rarely violate the speed limit. In addition, we can

Table 3
First-stage logit models for the probability of driver's traffic violations.

Variables	Models			
	Improper Passing	Speeding	Unsafe distance	Other violations
Driver's condition: normal	2.967 (2.95)	2.263 (12.87)	3.242 (4.56)	–
Driver's condition: DUI	–	–	–	2.838 (14.17)
Driver's condition: tiresome	–	–	–	−2.538 (-7.12)
Driver's condition: drowsy	–	–	–	1.547 (7.56)
Driver's condition: illness	–	–	–	4.794 (4.55)
Trailer	−2.413 (-5.19)	–	–	0.425 (5.59)
Special freight truck	−2.011 (-2.91)	1.281 (3.40)	–	0.565 (3.17)
Truck size: large	−1.974 (-4.74)	−0.385 (-4.54)	–	–
Truck size: mid	−1.921 (-4.8)	0.400 (5.88)	–	–
Truck size: small	−1.656 (-4.2)	0.759 (11.51)	–	–
Male driver	–	−0.576 (-3.06)	–	−0.614 (-4.14)
Driver age: under 20 years old	0.770 (2.84)	−0.705 (-3.65)	2.086 (15.45)	
Driver age: > = 20 & < 30 years old	2.000 (3.79)	–	1.141 (1.91)	
Driver age: > = 30 & < 40 years old	2.085 (6.33)	−1.178 (-2.97)	2.097 (8.01)	
Driver age: > = 40 & < 50 years old	1.274 (3.60)	−0.713 (-2.50)	2.134 (10.54)	
Driver age: > = 50 & < 60 years old	0.855 (1.84)	−0.957 (-2.92)	2.073 (9.21)	
Driver age: > = 60 years old	–	–	1.715 (3.96)	
Driver age: > = 40 & < 60 years old	–	–	–	−0.595 (-2.09)
Constant	−5.712 (-5.33)	−3.840 (-20.98)	−7.320 (-10.29)	−2.355 (-65.69)
Log-likelihood	−829.31	−6,519.54	−1,507.1	−4,048.22

Note: () represents t-statistics.

consider the effect of legal restrictions on the results. The South Korean government has been obliged to install Digital Tachographs (DTG) in commercial trucks to monitor drivers' behavior. However, according to the freight statistics (Ministry of Land Infrastructure and Transport, 2018), only 5.1% of all small trucks are commercial trucks, while that of mid-size and large-size trucks is 45% and 81%, respectively. Since small non-commercial trucks are not speed monitored by the DTG, it can be assumed that they contribute heavily to the increase in the likelihood of truck crashes due to speed violations. It is also the case that large trucks which comprise more than 80% of commercial vehicles have a negative association with the probability of crashing due to speed violations.

Driver age had an impact on crash probability due to safe distance violation, and all ages showed a positive coefficient. The drivers between 30 and 60 years of age were more likely to violate the safe distance regulation compared to other ages. From the results of "other violation" models, we found that driver conditions such as fatigue, illness, or drowsiness had a considerable influence on crash probability caused by other traffic violations. In addition, the results showed that crash probability increased for truck drivers in poor health.

We computed residuals ($\hat{x_u}$) of the probability of a truck driver's specific traffic violation from the four logit models in the first stage. Then we developed the second stage model for confirming the relationship between the probability of the truck-involved crashes on expressway mainlines and traffic conditions (AADT, portion of truck), vehicle error, weather, pavement conditions, roadway geometry, roadside features, day, month, and time of day. The residuals and original variables of driver traffic violation probability are also included in the second stage model. Table 4 shows the results of the model.

The logarithm of AADT is considered significant with a negative coefficient denoting that an increment of AADT reduces the probability of truck crashes on expressway mainlines. This finding suggests that high traffic volumes on expressways lead to increased truck driver's concern regarding safety. However, a higher proportion of truck traffic volume is likely to increase the probability of truck crashes, indicating that more interaction among trucks results in higher crash probability.

Weather conditions were also closely associated with truck crashes. The probability of truck crashes tends to be higher in snowy and rainy weather conditions; crash probability was higher in snowy weather than rainy weather. Potholes in pavement had a considerable influence on truck crashes and are likely to increase the probability of truck crashes. Potholes cause a significant strain on truck suspension, potentially resulting in serious property damage.

Type of roadway median variable showed to be ineffective in terms of safety in truck crashes. All variables related to median presence were significant at the 99% confidence level with a positive coefficient and indicate that the existence of medians on roadways resulted in an increased probability of truck crashes. Truck crash probability varied depending on median type. For instance, the fixed barrier type showed a higher probability of truck crashes.

The pipe type shoulder reduced the probability of truck crashes, whereas the installation type shoulder such as cable, fence, or guardrail was likely to increase truck crashes. This finding shows that shoulder features are related to a location on the expressway. Barriers on roadsides are installed in high-risk areas such as bridges and curves; therefore, large-size vehicles such as trucks have a higher possibility to incur traffic crashes.

Based on the results from the second-stage logit model shown in Table 4, the probability of truck crashes was higher on Thursday compared to other days. The summer months of July and August had more truck crashes, while truck crashes were more likely to occur in February than in other months. The time of day between 5:00 am and 8:00 am showed less probability of truck crashes. Unlike the results of our analysis, several previous studies insisted that driving very early in the morning results in higher probability of commercial vehicle crashes (Williamson and Boufous, 2007; Woods and Grandin, 2008). However, South Korea has a relatively small land area, and transporting time is short and mainly during the daytime. Therefore, truck traffic volume on the expressway is mostly high between 9 a.m. and 6 pm. As such, the probability of observing a freight truck crash early in the morning is relatively low. According to KEC's toll collecting system statistics, it can be seen that truck-involved crashes mostly occur from 1 pm to 4 pm rather than at night or early in the morning.

The four types of truck driver traffic violations were significant at the 99% confidence level. Improper passing and safe distance violation had a positive coefficient, indicating that the violations were likely to increase the probability of truck crashes, and the specific violations were crucial risk factors on the expressway.

For this study, we assumed that the driver traffic violation variables in the model for predicting the probability of truck-involved crashes on mainlines were endogenous. In this case, the variables for the traffic violations by each driver were endogenous regressors, and driver and

Table 4
Second-stage logit model of the truck crash probability at mainline on expressway.

Variables	Dependent Variable: truck crash probability at mainline			
	2SRI (a)		Naïve logit (b)	
	Coef.	z	Coef.	z
Logarithm of AADT	−0.206	−5.08	−0.205	−5.44
Percent of truck traffic volume	2.736	6.95	2.814	7.18
Obstructions on the road	1.451	8.14	1.549	8.80
Falling loads	−1.028	−6.29	−1.028	−6.35
Brake failure	−0.959	−4.08	−0.907	−4.21
Vehicle breakdown	0.916	3.03	0.956	2.91
Tire burst	1.230	8.72	1.145	8.33
Weather: snowy	1.283	6.58	1.266	7.14
Weather: cloudy	−0.163	−1.89	−0.180	−2.16
Weather: rainy	0.390	5.99	0.368	5.76
Pavement condition: normal	0.276	4.48	0.331	5.45
Pavement condition: pothole	3.491	3.01	3.441	3.08
Median type: guardrail	5.002	5.03	5.081	5.03
Median type: 127 cm fixed barrier	8.229	8.28	8.310	8.22
Median type: 81 cm fixed barrier	8.187	8.22	8.264	8.16
Median type: others	3.729	3.75	3.781	3.74
Median type: grass median strip	5.963	5.96	6.050	5.95
Median type: moving	4.003	3.96	4.032	3.92
No median	3.111	3.12	3.163	3.13
Shoulder type: guardrail	0.508	5.04	0.491	4.87
Shoulder type: cable	1.941	2.04	1.748	1.88
Shoulder type: pipe	−2.213	−2.27	−2.157	−2.37
Shoulder type: fence	1.621	2.61	1.594	3.55
Shoulder type: rock	2.194	7.11	2.207	6.96
Shoulder type: concrete	0.665	4.55	0.702	5.09
No shoulder	0.590	5.30	0.598	5.46
Thursday	0.157	2.03	0.157	2.00
February	−0.289	−2.24	−0.290	−2.38
July	0.264	2.66	0.253	2.62
August	0.367	3.81	0.358	3.72
Time: 5am to 8am	−0.178	−2.21	−0.180	−2.31
Improper passing	7.891	2.29	1.252	3.79
Speeding	−1.160	−3.91	−0.588	−7.62
Unsafe distance	8.365	6.06	1.400	7.71
Other traffic violation	−3.445	−7.62	−0.671	−6.48
Residual: improper passing	6.761	1.96	–	–
Residual: unsafe distance	7.429	5.25	–	–
Residual: speeding	−0.653	−2.07	–	–
Residual: other traffic violation	−2.971	−6.32	–	–
Constant	−4.172	−3.89	−4.410	−4.03
Log-likelihood	−4,066.01		−4,138.90	
AIC	8,212.02		8,349.79	
BIC	8,514.76		8,622.26	
Pseudo R^2	0.577		0.569	
Mean Square Error (MSE)	0.082		0.085	
Root Mean Square Error (RMSE)	0.286		0.291	
Mean Absolute Error (MAE)	0.167		0.170	
Mean Absolute Percentage Error (MAPE)	0.138		0.141	

vehicle attributes were set as IVs. Table 4 shows that the four residual variables obtained from the first stage logit model were all significant at the 95% confidence level in the second stage model. More specifically, the residuals are significantly different from zero, so we can confirm that there is an endogeneity bias in the OLS estimates, and IVs are necessary to correct the issue. The results verify our assumption that the four types of traffic violations were endogenous; thus, we conclude that the 2SRI approach captured the endogeneity bias in the model. The log-likelihood was better in the 2SRI model comparing to naïve logit approach, and the AIC and BIC values in Table 4(a) and (b) were found to be 8,212.02 and 8,514.76, respectively for 2SRI and 8,349.79 and 8,622.26, respectively for the naïve logit model. These results show that the 2SRI model exhibited a better fit than the naïve logit model excluding IVs. Pseudo R^2 was slightly higher in the 2SRI model compared to that of the naïve logit model.

We computed accuracy measures such as the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) as shown in Table 4 in order to compare the accuracy of the 2SRI predictions to those produced by the Naïve logit model. For the 2SRI model, the estimated accuracy metrics were found to be 0.082 for MSE, 0.286 for RMSE, 0.167 for MAE, and 0.138 for MAPE, whereas for the naïve logit model the metrics were 0.085 for MSE, 0.291 for RMSE, 0.170 for MAE, and 0.141 for MAPE. It shows that the estimates for the 2SRI model were a little bit lower than that of the naïve logit model, which means the 2SRI model fit became slightly better when the endogeneity bias was captured.

A prediction can be classified as positive if $p_{obs} \geq cut\ off$ and negative otherwise. The cutoff value was set as 0.5. We treat the classification as correct if it is positive and actual truck crash occurrence is equal to 1 or if it is negative and truck crash occurrence is equal to 0. According to the classification statistics, the overall rate of correct classification for the 2SRI model is estimated to be 89.1% while the naïve logit model showed 88.8%.

Margins are statistics estimated from predictions of a model at fixed values of some covariates and averaging or integrating over the remaining covariate (Williams, 2012). First, we estimated the predicted probabilities evaluated at the mean of the covariate. The 2SRI model indicated that the probability of truck crash occurrence at a mainline equal to 1 is 74% given that all predictors are set to their mean values. On the other hand, the naive logit model showed 73%.

The marginal effects show the change in probability when either the independent dummy variable changes from 0 to 1 or the continuous variable increases by one unit. The marginal effects for all selected variables in both the 2SRI and naïve logit truck-involved crash models were interpreted as follows. From the results of the 2SRI model, we found that the change in probability for one instant change in the logarithm of AADT is -1.7%, and that of the variable for the percent of truck traffic volume is 22.9%. Similar to this result, the probabilities of truck crash occurrence decreased by 1.8% and 24%, respectively when the logarithm of AADT and percent of truck traffic volume increased by one unit in the naïve logit model. When "cable type" of the shoulder variable changes from 0 to 1, the probability of truck-involved crash occurrence increased by 16.4% in the 2SRI model whereas it is raised by 14.9% in the naïve logit model. The variables for the various types of medians show more marginal effects. The change in probability for one instant change in the median of guardrail type is 41.9%, in the median with a 127 cm fixed barrier is 68.9%, and in the median with an 81 cm fixed barrier is 68.6%. The coefficients of marginal effect for the guardrail dummy variables in the naïve logit model were found to be similar to the coefficients obtained compared to that of the 2SRI model. However, the driver's traffic violation variables such as improper passing, violating the unsafe distance, over-speeding, and other traffic violations differed mainly in the marginal effect of the two models significantly. The marginal effects of driver's traffic violation variables in both models were found to be significant at 95% confidence level. The probability of a truck-involved crash in the 2SRI model showed that the likelihood of improper passing increases by 66.1%, and violating the unsafe distance also increases by 70.1%.

On the other hand, changing the variables for speeding and "other traffic violations" from 0 to 1 led to a decrease in the probability of truck-involved crash occurrence by 9.7%, 28.8%, respectively. For the naïve logit model, the change in probability when the improper passing

Table 5
Marginal effects for the 2SRI and naïve logit models.

Variable	2SRI		Naïve logit	
	Coef.	z	Coef.	z
Logarithm of AADT	−0.017	−5.400	−0.018	−5.45
Percent of truck traffic volume	0.229	6.920	0.240	7.20
Obstructions on the road	0.122	8.240	0.132	8.88
Falling loads	−0.086	−6.210	−0.088	−6.35
Brake failure	−0.080	−4.430	−0.077	−4.22
Vehicle breakdown	0.077	2.710	0.082	2.91
Tire burst	0.103	8.860	0.098	8.39
Weather: snowy	0.107	7.230	0.108	7.18
Weather: cloudy	−0.014	−1.940	−0.015	−2.16
Weather: rainy	0.033	6.040	0.031	5.77
Pavement condition: normal	0.023	4.480	0.028	5.46
Pavement condition: pothole	0.292	3.130	0.293	3.08
Median: guardrail	0.419	4.950	0.433	5.03
Median: 127 cm fixed barrier	0.689	8.150	0.708	8.23
Median: 81 cm fixed barrier	0.686	8.100	0.705	8.17
Median: others	0.312	3.690	0.322	3.74
Median: grass median strip	0.499	5.870	0.516	5.95
Median: moving	0.335	3.890	0.344	3.92
No median	0.261	3.080	0.270	3.13
Shoulder: guardrail	0.043	4.990	0.042	4.87
Shoulder: cable	0.163	2.100	0.149	1.88
Shoulder: pipe	−0.185	−2.480	−0.184	−2.37
Shoulder: fence	0.136	3.610	0.136	3.56
Shoulder: rock	0.184	6.870	0.188	6.98
No shoulder	0.049	5.350	0.051	5.47
Shoulder: concrete	0.056	4.800	0.060	5.10
Thursday	0.013	1.990	0.013	2.00
February	−0.024	−2.350	−0.025	−2.38
July	0.022	2.730	0.022	2.62
August	0.031	3.770	0.031	3.72
Time: 5am to 8am	−0.015	−2.270	−0.015	−2.32
Improper passing	0.661	2.030	0.107	3.79
Speeding	−0.097	−4.070	−0.050	−7.64
Unsafe distance	0.701	5.870	0.119	7.77
Other traffic violation	−0.288	−8.290	−0.057	−6.49
Residual: improper passing	0.566	1.64	–	–
Residual: unsafe distance	0.622	5.11	–	–
Residual: speeding	−0.055	−2.18	–	–
Residual: other traffic violation	−0.249	−6.91	–	–

dummy variable is altered from 0 to 1, the probability of truck-involved crash occurrence is increased by 10.7%, and when the variable for violating the unsafe distance dummy is changed from 0 to 1, the probability of truck-involved crash occurrence by is also increased by 11.9%. The marginal effects for the variables relating to over-speeding and "other violation" were found to be -0.05 and -0.057, respectively.

The marginal effects of the four residual variables were significant at approximately 95% confidence level. Especially, variables such as improper passing and unsafe distance from the first stage represented the instantons change given that the unit was big. The change in probability for one instant change in the improper passing variable is 56.6%, and the change in probability when the unsafe distance dummy variable changes from 0 to 1 increases the probability of truck-involved crash occurrence by 62.2%. Table 5 shows the marginal effects of truck crash occurrence using both 2SRI and naïve logit models.

5. Conclusions

Freight truck-involved crashes occur due to a variety of reasons such as unsafe driver behavior, vehicle defects, and dangerous roadway conditions. However, many researchers have only focused on the effects of roadway geometry and vehicle characteristics on the probability of truck-involved crashes, because data collection and analysis of human factors are very complicated. In this study, the endogenous relationship between truck driver traffic violations and crash occurrences were determined. From the results, this study shows significant effects of the traffic condition, pavement condition, weather, roadway geometry, roadside features, time of the day, and season on the probability of truck-involved crashes. There are several methodologies used to capture endogeneity problems; however, most researchers mainly used linear regression models. The 2SRI approach addresses endogeneity issues using nonlinear regression analysis with binary data. Therefore, the 2SRI approach was applied in the study, and a prediction model was developed for the probability of truck crashes on expressway mainlines. In the 2SRI process, the results from the first-stage logit model showed that specific types of driver traffic violations were influenced by the driver's physical condition as well as driver and vehicle characteristics.

The key results in the first-stage model are as follows:

- Driving either a trailer or a special freight truck such as a hazardous material transporting tank lorry is likely to decrease the probability of the driver engaging in an improper passing violation. Drivers of age less than 60 years have an increased likelihood of violating improper passing violation.
- Truck sizes are associated with a driver's speeding violation. A medium or small sized truck driver is likely to commit more speed violations, and drivers between the ages of 30 and 60 years tend to have a decreased probability of committing a speed violation.
- All ages show a positive coefficient for violating the unsafe distance regulation, but drivers of ages between 40 and 50 years have a higher probability of committing the unsafe distance violation than other age groups.
- The results show that the probability of committing "other violations" is affected by a driver's condition. Driving under the influence, drowsy driving, and driver's illness increase the likelihood of committing "other violations".

Several researchers have conducted studies aimed at analyzing the relationship between a driver's age and traffic violations. Jonah (1990) claimed that the age group between 20 and 24 years was more likely to show risky behavior than the age group between 16 and 19 years. The study by Matthews and Moran (1986) indicated that young drivers have more confidence in driving compared to old drivers, which might lead them to commit traffic violations. Yagil (1998) also insisted that younger drivers were less willing to comply with traffic regulations compared to older drivers. However, as analyzed in this study, there is a lack of detailed analysis for discovering different characteristics between the age groups of truck drivers depending on the types of traffic violations. The results from our study demonstrate that truck-involved crashes due to each type of traffic violation are different by each age group. In terms of the drowsy/fatigue driving and driving under the influence (DUI), previous studies discovered that the drivers under those conditions were likely to engage in a traffic violation such as wrong-way driving (Knipling and Wang, 1994; Pack et al., 1995; Baratian-Ghorghi et al., 2014; Zhou et al., 2015). Zhou et al. (2015) found that approximately 58% of all wrong way driving cases were DUI related. These previous studies support our results that the driver's condition, such as drowsy driving, driver's illness, and DUI increases the likelihood of truck-involved crashes.

As the next step, each residual obtained from the four first-stage models was incorporated into the original variables of driver traffic violations and other factors in the second-stage model. Consequently, the residuals and original variables of driver traffic violation were significant; therefore, we found that the variables were endogenous in the model for estimating the probability of truck crashes on expressway mainlines. The other results from the second-stage logit model are as follows:

- A high ratio of truck traffic volume increases the truck-involved crash probability at a mainline on the expressway.
- Snowy and rainy weather conditions are likely to increase the probability of having truck-involved crashes at a mainline on the

expressway.

- Depending on the types of median and shoulder, it has different effects on the probability of truck-involved crashes.
- February and time of day between 5 a.m. to 8 a.m. tend to decrease the probability of having a truck-involved crash.

Previous research supports our conclusion that as truck traffic increases, the probability of truck-involved crash occurrence also increases. Chang and Chen (2005) discovered that high truck ADTs increase the risk of truck crashes, and Dong et al. (2015) claimed that truck crashes are significantly positively related to truck ADT. Moreover, our findings prove that weather conditions such as snow and rain make the surface of the roadway slippery, and heavy trucks with a high center of weight are vulnerable to these roadway surface conditions as presented in the results of previous studies (Rodriguez et al., 2006; Rescot et al., 2009; Kotikalapudi, 2012; Kim et al., 2017; Dong et al., 2015*)*. Considering the types of median and shoulder, our results are in line with the study by Dong et al. (2015) which found that a specific shoulder and median type affected the truck-involved crash risk. The relationship between the factors such as the month of the year, time of day, and truck-involved crash risk were different from the results provided by previous studies. This result is due to the unique characteristics of South Korea's weather, geographical features, and travel patterns of freight trucks.

Finally, the computed marginal effects show that engaging in an improper passing and driving without adhering to the safe distance guideline increases the probability of truck-involved crashes by 66.1% and by 70.1%, respectively; the change in probability for a unit change in speeding is −9.7%, and in other traffic violation incidents is -28.8%.

To the best of our knowledge, this is the first study in which the 2SRI has been applied to investigate truck-related crashes with unobservable variables that are potentially correlated with the driver's traffic violations. This study showed that the 2SRI approach is relatively simple, but a plausible method for capturing endogeneity bias. Therefore, risk factors in a broader scope should be considered using this method, and a strategy for safe roadway management of freight trucks should be designed in future studies.

Further, other methods such as the copula approach which sets no restrictions on the probability distribution and captures the potential correlations due to unobserved factors could be considered to jointly analyze the relationship between the probability of driver's traffic violation and truck crash occurrence in future studies. Recently, Zou et al. (2019) studied wild-life vehicle crash analysis using a copula-based approach, and Laman et al. (2018) developed a joint model of traffic incident duration with a copula approach due to its advantages. Another study by Wang et al. (2019) applied a copula-based multivariate ordered probit model to estimate the four intersection crash consequence factors simultaneously. The 2SRI approach, which was applied for this study focuses on accounting for the endogeneity issue. Therefore, we suggest comparing two joint methods in terms of roadway safety in future studies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank Dr. Venky Shankar for his helpful advice on the econometric analysis in this paper. This research was a part of the project titled, "Smart Port IoT convergence and operation technology development" (No. 20190399), funded by the Ministry of Oceans and Fisheries, Republic of Korea.

References

Anderson, J.C., Dong, S., 2017. Heavy-vehicle driver injury severity analysis by time of week: a mixed logit approach using HSIS crash data. Inst. Transp. Eng. ITE J. 87 (9), 41.

Baratian-Ghorghi, F., Zhou, H., Shaw, J., 2014. Overview of wrong-way driving fatal crashes in the United States. ITE J. 84 (8), 41–47.

Basu, A., Coe, N., Chapman, C.G., 2017. Comparing 2SLS vs 2SRI for Binary Outcomes and Binary Exposures (No. w23840). National Bureau of Economic Research.

Bezwada, N.N.K., 2010. Characteristics and contributory Causes Associated With Fatal Large Truck Crashes. Kansas State University, USA.

Bin Islam, M., Hernandez, S., 2016. Fatality rates for crashes involving heavy vehicles on highways: a random parameter tobit regression approach. J. Transp. Saf. Secur. 8 (3), 247–265.

Blundell, R.W., Smith, R.J., 1989. Estimation in a class of simultaneous equation limited dependent variable models. ReV. Econ. Stud. 56 (1), 37–57.

Blundell, R.W., Smith, R.J., 1993. 5 Simultaneous Microeconometric Models with Censored or Qualitative Dependent Variables.

Cantor, D.E., Corsi, T.M., Grimm, C.M., Özpolat, K., 2010. A driver focused truck crash prediction model. Transp. Res. Part E: Logist. Transp. Rev. 46 (5), 683–692.

Castillo-Manzano, J.I., Castro-Nuño, M., Fageda, X., 2016. Exploring the relationship between truck load capacity and traffic accidents in the European Union. Transp. Res. Part E: Logist. Transp. Rev. 88, 94–109.

Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyse freeway crash frequency. J. Saf. Res. 36, 365–375.

Chen, C., Xie, Y., 2014. Modeling the safety impacts of driving hours and rest breaks on truck drivers considering time-dependent covariates. J. Saf. Res. 51, 57–63.

Chen, C., Zhang, J., 2016. Exploring background risk factors for fatigue crashes involving truck drivers on regional roadway networks: a case control study in Jiangxi and Shaanxi, China. SpringerPlus 5 (1), 582.

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single-and multi-vehicle crashes on rural highways. Crash Anal. Prev. 43 (5), 1677–1688.

Dong, C., Richards, S.H., Huang, B., Jiang, X., 2015. Identifying the factors contributing to the severity of truck-involved crashes. Int. J. Inj. Control Saf. Promotion 22 (2), 116–126.

Dong, C., Dong, Q., Huang, B., Hu, W., Nambisan, S.S., 2017. Estimating factors contributing to frequency and severity of large truck–involved crashes. J. Transp. Eng. Part A: Syst. 143 (8), 04017032.

Elshamly, A.F., El-Hakim, R.A., Afify, H.A., 2017. Factors affecting crashes risks among truck drivers in Egypt. In MATEC Web of Conferences 124, 04009.

Golob, T.F., Regan, A.C., 2004. Truck-involved crashes and traffic levels on urban freeways. Paper Presented at the 83rd Annual Meeting of Transportation Research Board.

Hallmark, S.L., Hsu, Y.Y., Maze, T.H., McDonald, T.J., Fitzsimmons, E.J., 2009. Investigating Factors Contributing to Large Truck Lane Departure Crashes Using the Federal Motor Carrier Safety Administration's Large Truck Crash Causation Study (LTCCS) Database.

Hanowski, R.J., Hickman, J.S., Olson, R.L., Ocanegra, J., 2009. Evaluating the 2003 revised hours-of-service regulations for truck drivers: the impact of time-on-task on critical incident risk. Crash Anal. Prev. 41 (2), 268–275.

Harkey, D., Council, F., Zegeer, C., 1996. Operational characteristics of longer combination vehicles and related geometric design issues. Transp. Res. Rec.: J. Transp. Res. Board 1523, 22–28.

Hartley, L.R., El Hassani, J., 1994. Stress, violations and crashes. Appl. Ergon. 25 (4), 221–230.

Harwood, D.W., Mason, J.M., Brydia, R.E., 1999. Design policies for sight distance at stop-controlled intersections based on gap acceptance. Transp. Res. Part A: Policy Pract. 33 (3–4), 199–216.

Iragavarapu, V., Lord, D., 2007. Examining Potential Factors Affecting the Safety Performance and Design of Exclusive Truck Facilities (Unpublished Master'S Thesis). Texas A&M University, Texas.

Islam, M.B., Hernandez, S., 2011. An Empirical Analysis of Fatality Rates for Large Truck Involved Crashes on Interstate Highways. 3rd International Road Safety and Simulation. pp. 14–16.

Islam, M., Hernandez, S., 2013. Large truck–involved crashes: exploratory injury severity analysis. J. Transp. Eng. 139 (6), 596–604.

Islam, M., 2015. Multi-vehicle crashes involving large trucks: a random parameter discrete outcome modelling approach. J. Transp. Res. Forum 54 (1).

Jonah, B.A., 1990. Age differences in risky driving. Health Educ. Res. 5 (2), 139–149.

Jones, I., Stein, H., Zador, P., 1983. Influence of Truck Size and Weight on Highway Crashes. Insurance Institute for Highway Safety.

Khattak, A.J., Schneider, R.J., Targa, F., 2003. Risk factors in large truck rollovers and injury severity: analysis of single-vehicle collisions. In 82nd Annual Meeting of the Transportation Research Board.

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005. Differences in rural and urban driver-injury severities in crashes involving large-trucks: an exploratory analysis. Crash Anal. Prev. 37 (5), 910–921.

Kim, J., Doustmohammadi, M., Anderson, M., 2017. Analysis of not at-fault truck crashes in Alabama. Int. J. Traff. Transp. Eng. 6 (2), 28–35.

Klassen, J., El-Basyouny, K., Islam, M.T., 2014. Analyzing the severity of bicycle-motor vehicle collision using spatial mixed logit models: a city of edmonton case study. Saf. Sci. 62, 295–304.

Knipling, R.R., Wang, J.S., 1994. Crashes and Fatalities Related to Driver drowsiness/fatigue. National Highway Traffic Safety Administration., Washington, DC.

Korea Expressway Corporation, 2017. Open Data Portal, Traffic Crash Statistics. https://www.data.go.kr/dataset/3038489/fileData.do.

Kotikalapudi, S., 2012. Characteristics and Contributory Causes Related to Large Truck Crashes (Phase-II)-All Crashes (Doctoral Dissertation). Retrieved from Kansas State University. (MATC-KSU: 352 Final Report 25-1121-0001-352).

Lagarde, E., Chastang, J.F., Gueguen, A., Coeuret-Pellicer, M., Chiron, M., Lafont, S., 2004. Emotional stress and traffic crashes: the impact of separation and divorce. Epidemiology 762–766.

Laman, H., Yasmin, S., Eluru, N., 2018. Joint modeling of traffic incident duration components (reporting, response, and clearance time): a copula-based approach. Transp. Res. Rec. 2672 (30), 76–89.

Lemp, J.D., Kockelman, K.M., Unnikrishnan, A., 2011. Analysis of large truck crash severity using heteroskedastic ordered probit models. Crash Anal. Prev. 43 (1), 370–380.

Lyles, R.W., Campbell, K.L., Blower, D.F., Stamatiadis, P., 1991. Differential Truck Crash Rates for Michigan (No. HS-041 372).

Matthews, M.L., Moran, A.R., 1986. Age differences in male drivers' perception of accident risk: the role of perceived driving ability. Accid. Anal. Prev. 18 (4), 299–313.

Matthews, G., 2002. Towards a transactional ergonomics for driver stress and fatigue. Theoretical Issues Ergon. Sci. 3 (2), 195–211.

Miaou, S.P., 1994. The relationship between truck crashes and geometric design of road sections: poisson versus negative binomial regressions. Crash Anal. Prev. 26 (4), 471–482.

Ministry of Land Infrastructure and Transport, 2018. National Freight Statistics. http://molit.go.kr/english/intro.do.

Newey, W.K., 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. J. Econom. 36 (3), 231–250.

Pack, A.I., Pack, A.M., Rodgman, E., Cucchiara, A., Dinges, D.F., Schwab, C.W., 1995. Characteristics of crashes attributed to the driver having fallen asleep. Accid. Anal. Prev. 27 (6), 769–775.

Pahukula, J., Hernandez, S., Unnikrishnan, A., 2015. A time of day analysis of crashes involving large trucks in urban areas. Accid. Anal. Prev. 75, 155–163.

Park, S.W., Jovanis, P., 2010. Hours of service and truck crash risk: findings from three national us carriers during 2004. J. Transp. Res. Board 2194, 3–10.

Poe, C.M., 2010. Geometric Design and Operational Factors that Impact Truck Use of Toll Roads (No. FHWA/TX-11/0-5377-1). Texas Transportation Institute, Texas A & M University System.

Rescot, R., Jasrotia, R., Hovey, K., Li, Y., Schrock, S.D., 2009. Predicting and mitigating wind-induced truck crashes on I-70 in Kansas. Proceedings of the 2009 Mid-Continent Transportation Research Symposium, Iowa. Iowa Department of Transportation.

Rivers, D., Vuong, Q.H., 1988. Limited information estimators and exogeneity tests for simultaneous probit models. J. Econom. 39 (3), 347–366.

Rodriguez, D., Rocha, M., Khattak, A., Belzer, M., 2003. Effects of truck driver wages and working conditions on highway safety: case study. Transp. Res. Rec.: J. Transp. Res. Board 1833, 95–102.

Rodriguez, D.A., Targa, F., Belzer, M.H., 2006. Pay incentives and truck driver safety: a case study. ILR Rev. 59 (2), 205–225.

Rowden, P., Matthews, G., Watson, B., Biggs, H., 2011. The relative impact of work-related stress, life stress and driving environment stress on driving outcomes. Crash Anal. Prev. 43 (4), 1332–1340.

Schneider IV, W., Zimmerman, K., Van Boxel, D., Vavilikolanu, S., 2009. Bayesian analysis of the effect of horizontal curvature on truck crashes using training and validation data sets. Transp. Res. Rec.: J. Transp. Res. Board 2096, 41–46.

Sharma, A.K., Landge, V.S., 2013. Zero inflated negative binomial for modelling heavy vehicle crash rate on Indian rural highway. Int. J. Adv. Eng. Technol. 5 (2), 292–301.

Smith, R.J., Blundell, R.W., 1986. An exogeneity test for a simultaneous equation Tobit model with an application to labor supply. Econometrica: J. Econom. Soc. 679–685.

Teoh, E.R., Carter, D.L., Smith, S., McCartt, A.T., 2017. Crash risk factors for interstate large trucks in North Carolina. J. Saf. Res. 62, 13–21.

Terza, J.V., Basu, A., Rathouz, P.J., 2008. Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. J. Health Econ. 27 (3), 531–543.

Terza, J.V., 2017. Two-stage residual inclusion estimation: a practitioners guide to Stata implementation. Stata J. 17 (4), 916–938.

Uddin, M., Huynh, N., 2017. Truck-involved crashes injury severity analysis for different lighting conditions on rural and urban roadways. Crash Anal. Prev. 108, 44–55.

Vallette, G.R., McGee, H.W., Sanders, J.H., Enger, D.J., 1981. The effect of truck size and weight on accident experience and traffic operations, volume 3. Accident Experience of Large Trucks. US Department of Transportation, Federal Highway Administration, Washington, DC Report No. FHWA/RD-80/137.

Wang, K., Bhowmik, T., Yasmin, S., Zhao, S., Eluru, N., Jackson, E., 2019. Multivariate copula temporal modeling of intersection crash consequence metrics: A joint estimation of injury severity, crash type, vehicle damage and driver error. Accid. Anal. Prev. 125, 188–197.

Williams, R., 2012. Using the margins command to estimate and interpret adjusted predictions and marginal effects. Stata J. 12, 308–331.

Williamson, A., Boufous, S., 2007. A data-matching study of the role of fatigue in work-related crashes. Transp. Res. part F: Traffic Psychol. Behave. 10 (3), 242–253.

Woods, J., Grandin, T., 2008. Fatigue: a major cause of commercial livestock truck accidents. Veterinaria Italiana 44 (1), 259–262.

Yagil, D., 1998. Gender and age-related differences in attitudes toward traffic laws and traffic violations. Transp. Res. Part F: Traffic Psychol. Behav. 1 (2), 123–135.

Yuan, Q., Lu, M., Theofilatos, A., Li, Y.B., 2017. Investigation on occupant injury severity in rear-end crashes involving trucks as the front vehicle in Beijing area, China. Chin. J. Traumatol. 20 (1), 20–26.

Zou, Y., Zhong, X., Tang, J., Ye, X., Wu, L., Ijaz, M., Wang, Y., 2019. A copula-based approach for accommodating the underreporting effect in wildlife–vehicle crash analysis. Sustainability 11 (2), 418.

Zhou, H., Zhao, J., Pour-Rouholamin, M., Tobias, P.A., 2015. Statistical characteristics of wrong-way driving crashes on Illinois freeways. Traffic Inj. Prev. 16 (8), 760–767.

Zhu, X., Srinivasan, S., 2011. A comprehensive analysis of factors influencing the injury severity of large-truck crashes. Crash Anal. Prev. 43 (1), 49–57.