Accident Analysis and Prevention 49 (2012) 512–519

Contents lists available at SciVerse ScienceDirect

Accident Analysis and Prevention

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

Factors affecting hospital admission and recovery stay duration of in-patient motor victims in Spain

Miguel Santolino*, Catalina Bolancé, Manuela Alcañiz

Department of Econometrics, Riskcenter-IREA, University of Barcelona, Avda. Diagonal, 690, 08034 Barcelona, Spain

ARTICLE INFO

Article history:
Received 10 October 2011
Received in revised form 15 March 2012
Accepted 20 March 2012

Keywords:
Body injuries
Heckit estimator
Semi-parametric estimator
Hausman test

ABSTRACT

Hospital expenses are a major cost driver of healthcare systems in Europe, with motor injuries being the leading mechanism of hospitalizations. This paper investigates the injury characteristics which explain the hospitalization of victims of traffic accidents that took place in Spain. Using a motor insurance database with 16,081 observations a generalized Tobit regression model is applied to analyse the factors that influence both the likelihood of being admitted to hospital after a motor collision and the length of hospital stay in the event of admission. The consistency of Tobit estimates relies on the normality of perturbation terms. Here a semi-parametric regression model was fitted to test the consistency of estimates, concluding that a normal distribution of errors cannot be rejected. Among other results, it was found that older men with fractures and injuries located in the head and lower torso are more likely to be hospitalized after the collision, and that they also have a longer expected length of hospital recovery stay.

1. Introduction

Healthcare expenditure is increasing rapidly in advanced societies. Over the last three decades healthcare spending in developed countries has grown on average 2.5 times faster than GDP (Hagist and Kotlikoff, 2009), and this trend is not expected to change in the near future. The reasons for this include an ageing population, advances in medical technology (their cost), and the defence strategies applied by medical practitioners to avoid malpractice liability. Although only a small proportion of the population will require hospitalization in any given year, hospitalization costs are one of the most significant annual healthcare cost drivers in healthcare systems. Peek-Asa et al. (2011) estimate that annual hospital charges exceed $1 billion in the USA, with the median hospital charge being above $25,000. These hospitalization expenditures account for nearly one third of all medical expenses for the non-institutionalized population in the USA (Machlin and Carper, 2007). A high percentage of hospitalizations are the consequence of motor collisions. In Europe road traffic-related hospitalizations are the leading mechanism of injuries requiring hospitalizations (Segui-Gomez et al., 2008).

The goal of this paper is to explore the relationship between the injury characteristics and hospital stay duration of motor victims in Spain. In the road safety literature, studies that link injury characteristics and hospital stay duration have traditionally dealt with hospital data (Guria, 1990; Forman et al., 2011; Peek-Asa et al., 2011), and their results are therefore limited to hospitalized motor victims. Given that inpatient victims could be systematically different from those who are not admitted to hospital, any conclusions drawn from these studies cannot be extrapolated to all motor collision victims. This is relevant because most victims injured in a collision do not require hospitalization. This paper deals with motor victims involved in collisions irrespective of whether hospitalization was required or not. It will therefore be possible to investigate the factors that explain hospital duration from an unconditional perspective.

The analysis is based on a Spanish motor insurance database related to motor collisions involving injury victims. The data provide details of how injuries may develop after collision. Data are censored where hospitalization duration is only observed if the victim was admitted to hospital. Regression models for censored data have been widely applied in the road safety literature (Goldstein, 1986; Anastasopoulos et al., 2008; Farah et al., 2009). Multivariate extensions are found in Anastasopoulos et al. (2012a). Here a sample selection regression model is applied to determine the factors that affect both the hospital admission of traffic accident victims and the length of hospital stay in the event of admission. The dependent variable (hospitalization duration) is quantitative but sample selection regression models have been also applied for qualitative dependent variables (Tarko and Azam, 2011). This methodology allows for the dependence of perturbation terms of both procedures. Modelling approaches that ignore dependence between hospital admission and recovery stay length would lead to biased parameter estimates, when such dependence exists (Cameron and Trivedi, 2005).

* Corresponding author. Tel.: +34 934 020 484; fax: +34 934 021 821.
E-mail address: msantolino@ub.edu (M. Santolino).

0001-4575/$ – see front matter
http://dx.doi.org/10.1016/j.aap.2012.03.025

A generalized Tobit regression is fitted in order to model the duration of hospitalization.[1] The reliability of Tobit estimators is based on distributional assumptions. Unlike least square estimates for uncensored data, Tobit estimators are not consistent if errors are not normally distributed (Mashtare and Hutson, 2011). Consequently, the hypothesis of a normal distribution of errors must be tested when the Tobit regression model is applied. To our knowledge, however, the Tobit normality hypothesis has not been previously tested in road safety applications. In this paper a semi-parametric regression model is estimated in order to test the distributional assumption of normality. The Hausman specification test is then computed so as to compare Tobit estimators—which are only consistent estimators under the null hypothesis of normality—with semi-parametric estimators, which are consistent estimators under both hypotheses (Newey, 1987, 2009).

The structure of the paper is as follows. Section 2 presents the main characteristics of parametric and semi-parametric selectivity regression models, and outlines the utility of semi-parametric methods to test distributional assumptions. In Section 3, an empirical application using a Spanish database is presented. This is followed by interpretation of the results and discussion in Section 4. Concluding remarks are summarized in Section 5.

2. Statistical analysis

Hospital admission and length of stay of in-patients may be interpreted as a sample selection issue. First, medical practitioners examine the victim and decide whether admission to hospital is required. If so, the length of the hospital stay will depend on multiple factors including the physical condition of the victim and the type and severity of injuries. This section briefly describes both parametric and semi-parametric sample selection regression models. First, we describe the most popular parametric sample selection model, namely the generalized Tobit regression model, before moving on to consider Newey's semi-parametric regression model. The Hausman test is then discussed as a way of selecting the preferred model specification.

2.1. Generalized Tobit regression model

The Tobit regression model (Tobin, 1958) has been widely used in the social sciences, and is applied when the range of the dependent variable is censored. Censoring occurs if data show a lower threshold (left-censored) below which the observed value of the variable always takes the value of the threshold. A similar interpretation holds for data with an upper threshold (right-censored), although applications with left-censored data are more common. An extensive survey of Tobit regression models is provided by Amemiya (1984). Let y be our outcome of interest, which is the length of hospital stay of the injury victim after a collision. The outcome y is observed only if $y^* > 0$, where y^* is a latent (unobserved) variable. In our application y^* indicates the hospital admission decision. The generalized Tobit regression model is specified as follows:

$$\begin{aligned} y_i^* &= z_i'\gamma + u_i \\ y_i &= x_i'\beta + \varepsilon_i \end{aligned} \qquad (1)$$

for the ith individual, $i = 1, \ldots N$, where z_i and x_i are vectors of regressors, γ and β are vectors of parameters, and u_i and ε_i are normally distributed random errors with mean equal to zero and variance σ_1^2 and σ_2^2, respectively. Errors are correlated with covariance equal to σ_{12}. The observed variable is I which is related to y^* according to:

$$I_i = \begin{cases} 1 & \text{if} \quad y_i^* > 0 \\ 0 & \text{if} \quad y_i^* \le 0 \end{cases},$$

where I_i takes the value 1 when the ith victim is admitted to hospital and 0 otherwise. One may put $\sigma_1^2 = 1$ without any loss of generality, since only the sign of the latent variable is observed. In Amemiya's terminology (1984) this model is referred to as the type II Tobit regression model. Heckman (1976) suggested a procedure to estimate parameters in two steps. The assumption in the Heckman procedure is that ε is linearly regressed on u, where u follows a standard normal distribution. Following this procedure, also called the Heckit estimator, a probit model regression of I on z is firstly estimated. In the next step an additional regressor is included in the second equation in (1) to account for bias due to non-random sampling. β and σ_{12} may then be estimated by ordinary least squares (OLS) techniques applied to the extended regression model:

$$\begin{aligned} y_i &= x_i'\beta + \sigma_{12}\lambda_i + \eta_i \\ &= x_i'\beta + \sigma_2\rho\lambda_i + \eta_i \end{aligned} \qquad (2)$$

such that the Heckman-type lambda is $\lambda_i = \phi(z_i'\gamma)/\Phi(z_i'\gamma)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, and ρ is the error correlation, $\rho = \sigma_{12}/\sigma_2$. The error variance σ_2^2 may be estimated by the following expression:

$$\hat{\sigma}_2^2 = N_1^{-1}\sum_{i}^{N_1}[\hat{\eta}_i^2 + \hat{\sigma}_{12}^2\hat{\lambda}_i(\hat{\lambda}_i + z_i'\hat{\gamma})],$$

where $\hat{\lambda}_i = \phi(z_i'\hat{\gamma})/\Phi(z_i'\hat{\gamma})$ and N_1 is the number of victims admitted to hospital (Cameron and Trivedi, 2005).[2]

To test hypotheses, an estimator of the asymptotic covariance of coefficients is required. Note that the disturbance term in (2) is heteroscedastic due to sample selection correction. Additionally, the unknown γ is replaced by a vector of estimated parameters. OLS standard errors have to be corrected to account for these particularities. An estimator of the asymptotic covariance matrix is provided by Heckman (1979). The covariance matrix may also be estimated by means of bootstrapping techniques, as provided by most commercial statistical software.

2.2. Semi-parametric regression model

Statistical consistency of the Heckit estimator is based on the normality assumption of the perturbation term u. If u is not normally distributed then $\lambda \neq \phi(z'\gamma)/\Phi(z'\gamma)$. The model (1) is consequently misspecified and OLS techniques in (2) lead to inconsistent estimates. Newey (2009) suggests a model estimation method in which the functional form of the selection correction is now unrestricted. He showed that this estimation procedure then gives consistent estimates of β even if u is not normally distributed.

Like the Heckman procedure the Newey estimation method involves two steps. In the first step Newey (2009) suggests estimating the parameter vector γ by semi-parametric techniques as an alternative to fitting a probit regression model. Let us denote

[1] Another option to deal with time is hazard-based duration models (Washington et al., 2003; Mannering, 1993). Farah et al. (2009) analyse both model specifications to predict the risk associated with the passing behaviour. Two review articles of methodological alternatives for crash-injury severities and crash-frequency are Savolainen et al. (2011) and Lord and Mannering (2010).

[2] Parameters may also be estimated by maximum likelihood (ML), but consistency of the ML estimator depends on the hypothesis of binormality of errors. Testing binormality is not easy. Traditional binormality tests for continuous dependent variables cannot be applied to sample-selection models. The Heckit estimator has the additional advantage of being consistent in the event of jointly normal distributed errors.

the semi-parametric estimator by $\hat{\hat{\gamma}}$. Newey argues that the semi-parametric estimator provided by Klein and Spady (1993) gives the most efficient estimator, such that $\hat{\hat{\lambda}}_i = h(\hat{v}_i)$ where $\hat{v}_i = z_i'\hat{\hat{\gamma}}$. In the second step, the vector of parameters β is estimated from the following linear regression:

$$y_i = x_i'\beta + \hat{\hat{\lambda}}_i + w_i = x_i'\beta + h(\hat{v}_i) + w_i. \quad (3)$$

The unknown function $h(\cdot)$ may be approximated by power series. Let $\tau(\hat{v}_i) = \hat{\tau}_i$ be a monotonic transformation of $\hat{v}_i$ and $p^K(\hat{\tau}_i) = (p_1(\hat{\tau}_i), \ldots, p_K(\hat{\tau}_i))'$, where $p_j(\hat{\tau}_i) = \hat{\tau}_i^{j-1}$. The unknown function $h(\hat{v}_i)$ may then be approximated by means of a linear combination of the elements of $p^K(\hat{\tau}_i)$. The shape of the approximation function is based on the K value, where the effect of the sample information increases with K. Given a K value, the model (3) is then defined as:

$$y_i = x_i'\beta + \delta_1 p_1(\hat{\tau}_i) + \cdots + \delta_k p_K(\hat{\tau}_i) + w_i, \quad (4)$$

where the unknown function is approximated by $\hat{h}(\hat{v}_i) = \hat{\delta}_1 p_1(\hat{\tau}_i) + \cdots + \hat{\delta}_k p_K(\hat{\tau}_i)$, and $\hat{\delta}_1, \ldots, \hat{\delta}_k$ are the OLS estimates in (4). In this study we consider $\hat{\tau}_i = \phi(\hat{v}_i)/\Phi(\hat{v}_i)$, which is one of the three alternative monotonous transformations suggested in Newey's paper. Similarly to Heckman (1976), Newey (2009) provides an estimator of the asymptotic covariance matrix of the vector of β estimates.

Let $\hat{\beta}_H$ be the two-step Heckman estimator of the model (2) and $\hat{\beta}_N$ be the two-step semi-parametric estimator of the regression equation (4). The hypothesis of normality in (1) may be tested by means of the well-known Hausman test. Under the null hypothesis both $\hat{\beta}_H$ and $\hat{\beta}_N$ are consistent estimates but $\hat{\beta}_H$ is more efficient. However, only $\hat{\beta}_N$ is consistent under the alternative hypothesis.[3]

3. Spanish database

The data consist of a random sample of 16,081 non-fatal victims involved in traffic collisions in Spain. The database was provided by a Spanish motor insurance company. All of the victims suffered at least 1 day of temporary disability. Information included in the database was recorded by the insurer during the processing of claims in order to track them until settlement.

3.1. Insurance data

Crash databases are usually compiled from police reports. Third-liability insurance data has particularities that merit discussion. Insurance databases provide information related to the injury consequences of collisions, but limited information on the circumstances of the collision. According to Spanish motor law, drivers who are at fault in the collision are not entitled to compensation. Consequently, the insurance company did not record injury information of at-fault drivers, and this was not included in the database. Passengers in the at-fault vehicle are entitled to compensation and their injury information was recorded.

The non-recording of injury information of at-fault drivers has implications for research. First, differences between not-at-fault and at-fault drivers regarding the expected severity cannot be investigated. Huang et al. (2008) showed that drivers sustain more severe injuries and/or automobile damages when are the offending party. However, Abdel-Aty (2003) found that at-fault drivers experience less severe injuries. Linked with the offender's behaviour, several studies have investigated the relationship between driver severity and aggressive driving behaviour (Paleti et al., 2010; Nevarez et al., 2009). A second characteristic is that certain types of crashes may be infra reported (for instance, crashes involving a single vehicle). Ulfarsson and Mannering (2004) showed that factors affecting single and multiple vehicle crashes may be substantially different. Underreporting of less severe crashes in police reported data is known in road safety literature (Hauer and Hakkert, 1989; Elvik and Mysen, 1999) and statistical inference constraints when underreporting have been extensively investigated (Ye and Lord, 2011; Yamamoto et al., 2008). Finally, a lower number of crashes with multiple injured victims is observed. In our case, only 10% of the crashes involved multiple injured victims, and of those that caused victims requiring hospitalization only 2.6% involved more than one hospitalized victim.

3.2. Description of variables

Our goal here is to model the length of hospital stay of victims involved in motor collisions. The duration of hospital stay is observed for all the sample victims. Victims must be fully recovered or with stable injuries, and therefore discharged from hospital, before being compensated by the insurer. Sample casualties were compensated for personal damages by the insurer in the year 2007, although the motor collision may have occurred before that year. Specifically, 28% of collisions took place in 2007, 57% in 2006, 13% in 2005, and 2% in 2004 or before.

The distribution (both unconditional and conditional) of the number of days in hospital for our data is presented in Fig. 1. The mean length of hospital stay for sample victims is 2.10 days, with a standard deviation of 13.85 and median value of 0. More than 87% of sample victims were not admitted to hospital after the collision. Specifically, 1999 sample victims had to be admitted to hospital after the collision. For those victims who were admitted to hospital, the conditional distribution of the length of stay is strongly asymmetric (Fig. 1). The mean stay in hospital for inpatients is 16.87 days, with a standard deviation of 35.97 and median value of 7 days.

The dependent variable of the model regression specification is the number of days in hospital, on a natural logarithmic scale. The indicator variable that records whether the victim was admitted to hospital after the collision and the logarithm of the number of days in hospital are regressed on the same set of explanatory variables. The same explanatory variables are included in the two processes because the same level of information is available for hospitalized and non-hospitalized victims. However, the Tobit model specification (1) is not restricted to it. Different explanatory variables may be considered to explain the hospital admission and duration (for instance, whether hospital discharge information includes additional victim information). The description of variables and main statistics are shown in Table 1. Explanatory variables are classified into three groups: general factors, factors related to the region of the body that was injured and factors describing the nature of the injury.

General factors include information of the at-fault driver, such as his/her age. A number of authors have shown a link between driver age and the frequency and severity of motor crashes, for example: associated with different crash types (Richardson et al., 1996), at multiple locations (Abdel-Aty, 2003), at a two-state of safety (Malyshkina and Mannering, 2009), under different road-surface conditions (Morgan and Mannering, 2011) or for motorcyclists (Savolainen and Mannering, 2007). The driver age effect has been traditionally associated with the fact that young drivers are prone to more risky driving behaviour, whereas older drivers are more likely to have slower reaction times (Awadzi et al., 2008; Huang et al., 2008; Abdel-Aty et al., 1998; Neyens and Boyle, 2008).

Other general factors cover attributes of the victim, such as gender and age, and information related to the type of victim.

[3] The Hausman statistic is computed as $T = (\hat{\beta}_H - \hat{\beta}_N)'[\mathrm{Var}(\hat{\beta}_N) - \mathrm{Var}(\hat{\beta}_H)]^{-1}(\hat{\beta}_H - \hat{\beta}_N)$, which is χ^2 distributed under the null hypothesis with p degrees of freedom, where p is the number of parameters in the vector β without the constant term.

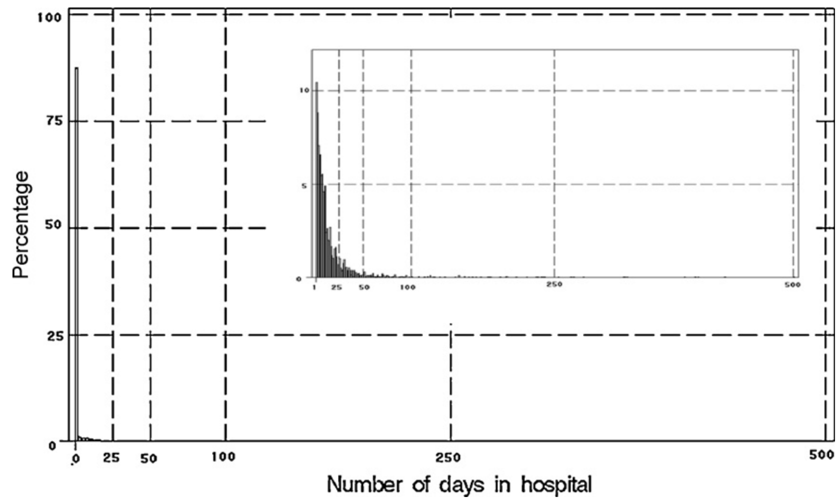

Fig. 1. Histogram for the number of days in hospital. Note: Histogram of the hospital recovery time required by motor victims (in days). Inset: histogram of the hospital recovery time (in days) for those victims who were hospitalized.

Male-female differences in the determinants of injury severities are known (Ulfarsson and Mannering, 2004; Islam and Mannering, 2006). However, the effect of gender on severity is ambiguous. Many crash severity studies have found that females involved in motor collisions suffer more serious injuries than do males (Hill and Boyle, 2006; Evans, 2001; Kockelman and Kweon, 2002), whereas other authors suggest that males are involved in more serious crashes (Tay and Rifaat, 2007; Valent et al., 2002). The extant literature indicates that the victim's age positively influences the frequency and severity of injuries (O'Donnell and Connor, 1996; Delen et al., 2006; Boucher and Santolino, 2010), although some studies have found a non-linear relationship between age and severity (Ulfarsson and Mannering, 2004; Newgard, 2008; Kockelman and Kweon, 2002). Regarding the type of victim, a

Table 1
Description of variables and statistics.

Variable	Label	Description	N	Mean	SD	Min.	Max.
Dependent variables							
y_1	Hospital admission	1 if victim is admitted to hospital; 0 otherwise	16,081	0.124	0.329	0	1
y_2	Hospital stay	Length of hospital stay (number of days on natural logarithmic scale)	1999	2.014	1.208	0	6.215
Regressors							
General factors							
x_1	At-fault driver age	Age of the at-fault driver (divided by 100)	16,081	0.412	0.144	0.14	0.990
x_2	Victim age	Age of the victim (divided by 100)	16,081	0.383	0.167	0	0.980
x_3	Victim age squared	Victim age squared (divided by 10,000)	16,081	0.175	0.150	0	0.960
x_4	Gender	1 if the injured victim is male; 0 otherwise	16,081	0.455	0.498	0	1
x_5	Driver	1 if the injured victim was the driver; 0 otherwise	16,081	0.510	0.500	0	1
x_6	Passenger	1 if the injured victim was the passenger; 0 otherwise	16,081	0.373	0.483	0	1
x_7	Pedestrian/cyclist	1 if the injured was a non-motorized road user; 0 otherwise	16,081	0.117	0.321	0	1
Factors related to the injured body region							
x_8	Head	1 if injury located in head; 0 otherwise	16,081	0.129	0.336	0	1
x_9	Neck	1 if injury located in neck; 0 otherwise	16,081	0.699	0.458	0	1
x_{10}	Upper torso	1 if injury located in upper torso (thorax/dorsal); 0 otherwise	16,081	0.250	0.433	0	1
x_{11}	Lower torso	1 if injury located in lower torso (abdomen/lumbar); 0 otherwise	16,081	0.193	0.395	0	1
x_{12}	Upper extremities	1 if injury located in upper extremities; 0 otherwise	16,081	0.257	0.437	0	1
x_{13}	Lower extremities	1 if injury located in lower extremities; 0 otherwise	16,081	0.238	0.426	0	1
x_{14}	Multiple regions	1 if multiple body regions; 0 otherwise	16,081	0.063	0.243	0	1
Factors related to the nature of the injury							
x_{15}	Wound	1 if open wound. 0 otherwise	16,081	0.077	0.267	0	1
x_{16}	Fracture	1 if fracture; 0 otherwise	16,081	0.174	0.379	0	1
x_{17}	Contusion	1 if contusion; 0 otherwise	16,081	0.542	0.498	0	1
x_{18}	Sprain/strain	1 if sprain/strain; 0 otherwise	16,081	0.738	0.440	0	1
x_{19}	Abrasion	1 if abrasion/burn; 0 otherwise	16,081	0.023	0.151	0	1
x_{20}	Internal injury	1 if internal injury (nerves, blood vessels, etc.); 0 otherwise	16,081	0.087	0.282	0	1

Table 2
Heckman two-step estimates, standard errors (*p*-values in parenthesis) and marginal effects of the generalized Tobit regression.

Variable	Label	Hospital admission				Length of hospital stay			
		Coeff.	Estim.	Std. error	Marginal effect ([a])	Coeff.	Estim.	Std. error	Marginal effect ([a])
	Intercept	γ_0	−1.039	0.092 (<0.001)	–	β_1	0.774	0.346 (0.032)	–
x_1	At-fault driver age	–	–	–	–	–	–	–	–
x_2	Victim age	γ_2	−1.260	0.395 (0.001)	−0.153	–	–	–	–
x_3	Victim age squared	γ_3	1.232	0.437 (0.005)	0.149	–	–	–	–
x_4	Gender	γ_4	0.190	0.034 (<0.001)	0.023	–	–	–	–
x_6	Passenger ([b])	γ_6	0.150	0.038 (<0.001)	0.019	β_6	0.157	0.062 (0.012)	0.086
x_7	Pedestrian/cyclist	γ_7	0.227	0.049 (<0.001)	0.031	β_7	0.192	0.069 (0.006)	0.087
x_8	Head	γ_8	0.477	0.046 (<0.001)	0.076	β_8	0.396	0.098 (<0.001)	0.176
x_9	Neck	γ_9	−0.586	0.060 (<0.001)	−0.085	β_9	−0.400	0.132 (0.013)	−0.128
x_{10}	Upper torso	γ_{10}	0.174	0.042 (<0.001)	0.023	β_{10}	0.237	0.067 (<0.001)	0.156
x_{11}	Lower torso	γ_{11}	0.326	0.046 (<0.001)	0.046	β_{11}	0.401	0.086 (<0.001)	0.249
x_{12}	Upper extremities	–	–	–	–	β_{12}	−0.171	0.056 (0.002)	−0.171
x_{13}	Lower extremities	γ_{13}	0.187	0.048 (<0.001)	0.025	β_{13}	0.375	0.064 (<0.001)	0.287
x_{14}	Multiple regions	–	–	–	–	β_{14}	0.255	0.092 (0.006)	0.255
x_{15}	Wound	γ_{15}	0.112	0.050 (0.027)	0.014	–	–	–	–
x_{16}	Fracture	γ_{16}	0.896	0.037 (<0.001)	0.166	β_{16}	0.723	0.149 (<0.001)	0.315
x_{17}	Contusion	γ_{17}	−0.484	0.040 (<0.001)	−0.061	β_{17}	−0.672	0.098 (<0.001)	−0.445
x_{18}	Sprain/strain	γ_{19}	−0.231	0.058 (<0.001)	−0.031	β_{18}	−0.409	0.092 (<0.001)	−0.301
x_{19}	Abrasion	γ_{20}	−0.496	0.100 (<0.001)	−0.041	–	–	–	–
x_{20}	Internal injury	γ_{21}	0.652	0.041 (<0.001)	0.115	β_{20}	0.584	0.108 (<0.001)	0.285
						σ_{12}	0.552	0.210 (0.009)	
						σ_2	1.183		
						rho	0.466		

N = 16,081; McFadden R^2: 0.344 (binary model); McKelvey and Zavoina R^2: 0.339 (two-stage model).
[a] Marginal effects for the probability of hospital admission and the expected length of stay conditional on being admitted are computed at the means of the independent variables. In case of binary variables, marginal effects are computed for discrete changes from 0 to 1. Marginal effects on the length of hospital stay are not displayed for variables not included in the second stage.
[b] Driver is the base category.

distinction is made between driver, passenger and non-motorized road user. The seating position of the victim has likewise been related to the severity of injuries (O'Donnell and Connor, 1996; Smith and Cummings, 2004; Newgard et al., 2005). Yamamotoa and Shankar (2004) applied a bivariate ordered-response probit to account for correlation between driver and passenger injury severities. The crash severity behaviour of non-motorized road users has also been extensively investigated (Chong et al., 2010; Tarko and Azam, 2011; Eluru et al., 2008; Lee and Abdel-Aty, 2005).

Injury factors provide a description of injuries resulting from the accident. The injury information recorded is based on medical examinations carried out by the insurance company during the period in which victims are recovering from their injuries. Injuries resulting from a collision are stated by judicial decision, or agreed upon between parties (insurer and victim), based on both medical reports provided by parties and forensic examinations. The medical examinations have to be made in accordance with a legislative medical scale in force since 1995. The medical scale describes 475 injuries and provides severity point scores for each injury. The motor financial compensation for permanent injuries is then assessed in function of the severity scores (for more details, see Santolino, 2010; Boucher and Santolino, 2010). The average number of injuries in the sample was 1.55 injuries per victim involved in a motor collision.

Injuries described in the legislative scale are classified according to their nature and the body location in order to reduce the number of injuries on the legal scale into a manageable number of diagnostic categories, inspired by the diagnostic matrix of the International Classification of Diseases (ICD)-9-CM codes developed by Barell et al. (2002). Specifically, a two-dimensional array is used to describe each injury. The first element indicates the location of the injury, while the second records its type. Seven factors are related to the region of the body that was injured and six factors to the nature of the injuries. In comparison with Barell's matrix classification, the number of factors related to the location of injuries was reduced to a simpler classification in order to make identification of the injury location easier for non-medical practitioners (for instance, police officers at the crash scene). As regards the factors related to the injury's nature, the *Internal injury* category records information from the *Internal*, *Blood vessels* and *Nerves* categories of the Barell matrix, due to the low frequency of the last two categories (less than 0.5%). Low frequency was also the reason for excluding the *Amputations* and *Crush* categories from the regression model. All the injury factors are included in the regression as binary variables which take the value 1 if the characteristic is observed and zero otherwise. Victims may suffer more than injury and, therefore, factors are not mutually exclusive.

4. Results

Prior to analysing coefficient estimates it is necessary to test whether the Heckman two-step estimates of β are consistent and, therefore, reliable. As previously indicated, the normality hypothesis testing involves a semi-parametric estimation of the model (1). The K value must be selected to approximate the unknown function $h(\hat{v}_i)$ in the semi-parametric method. The analysis is performed for K = 4, K = 5 and K = 6. In all cases the null hypothesis cannot be rejected at the 5% significance level (see Annex A.1 for details). We therefore conclude that the Heckit estimator of the model (1) is consistent. Additionally, error dependence between injured victims from the same crash was investigated.[4]

The two-step Heckman estimates of the generalized Tobit regression model are shown in Table 2. In order to achieve a parsimonious model, explanatory variables with associated coefficients

[4] The unit of analysis in the model is individual victim. The dependence of observations from the same crash was assessed assuming alternative correlation structure types for the error specification in the modeling of the hospitalization probability. In all the cases the regression produced a poorer fit to the data than in the case of correlation independence, probably due to the low number of crashes involving multiple victims. As a result, it was assumed independence between observations.

not significantly different from zero at the 5% significant level were withdrawn from the regression model. Note that the covariance coefficient σ_{12} is significantly different from zero at the 1% significance level. This shows that the endogeneity assumption is realistic for these data, which means that the decision to admit a victim to hospital is not independent of the length of the stay in hospital once the victim is admitted.[5]

In the road safety literature the driver's age is traditionally associated with the severity of collisions. According to our results, however, the age of the at-fault driver did not have explanatory power as regards the probability of being admitted to hospital and the subsequent length of hospitalization. Therefore, results do not suggest that either the aggressive behaviour or physical limitations associated with different driver age groups affects crash severity. This is consistent with the results of Tay (2006), who found that increasing the number of licenses issued to ageing drivers did not significantly increase the number of fatal crashes. Here we also tested a quadratic relationship, including as a regressor a variable recording the squared age of the driver. However, the coefficients were not significantly different from zero.

Four of the six coefficients of the variables related to victim attributes were statistically significant. The factor associated with the victim's gender showed significant positive coefficients in both equations, especially in the first regression. This means that males injured in a collision were more likely to be hospitalized after the crash than were females, and their average recovery stay in hospital was also longer. Therefore, males seem to suffer more serious injuries than do females, a finding that is consistent with the results of Tay and Rifaat (2007) and Valent et al. (2002). Similarly, Abdel-Aty (2003) showed that males have a higher probability of a severe injury whether the crash occurs. As regards age, there was a quadratic relationship between the victim's age and the probability of being hospitalized. Specifically, young and older victims were more likely to be admitted to hospital than were middle-aged victims, with the inflexion point being around the age of 50 years. From this age on, the probability of hospitalization increased with increasing ages. A previous study by Newgard (2008) similarly found that young and old victims were more likely to be seriously injured, and that the risk for serious injury rose more steeply after the age of 50. Ulfarsson and Mannering (2004) showed that the influence of age on the severity is similar for both genders. The positive relationship between age and severity is traditionally associated with the greater fragility and pre-existing medical conditions of elder victims (Awadzi et al., 2008; Savolainen and Mannering, 2007; Li et al., 2003), whereby physical decline would be intensified from the age of fifty onwards.

The type of victim also determines the probability of hospitalization and the length of stay. Compared with drivers, we found that passengers and non-motorized road users were more likely to be admitted and also to have longer recovery periods in hospital. Other studies also suggest that passengers sustain more serious injuries. O'Donnell and Connor (1996) showed that the risk of injury increases when the victim is seated in passenger position. Similarly, Hutchinson (1986) showed that passengers were more seriously injured than drivers in non-overturning accidents, while Hill and Boyle (2006) found that front passengers were 1.1 times more likely to be seriously injured than were drivers. Regarding non-motorized road users, Pucher and Dijkstra (2003) estimated that pedestrians were 23 times more likely to be killed than were car occupants in the USA, with cyclists being 12 times more likely to be killed than car occupants. Here we found that both pedestrians and cyclists were more likely to be hospitalized after a collision and their recovery stay in hospital is longer. This is an expected result due to the lack of protective elements for mitigating the impact suffered by non-motorized road users.

In relation to the injury location factors, they almost all showed significant coefficients. Head injuries, followed by those located in the lower torso and the lower extremities, were the injuries which had the strongest positive effect on the probability of being admitted to hospital and on the recovery duration of in-patient motor victims. Previous studies that link severity and injury location support these results. For example, Norin et al. (1997) showed that head injuries were predominant at higher levels of disability, accounting for 40% of injuries. They also found that while abdomen injuries were more common at higher disability levels, leg injuries were stable across all disability levels. In the same context Fildes et al. (1994) reported that one third of severe crashes involve lower limb injuries to front seat occupants in frontal crashes.

Two injury location factors (*Neck* and *Upper extremities*) showed negative coefficients. Thus, neck injuries were negatively related with the probability of being hospitalized and with recovery stay length. Note that injuries located in the neck are the most frequent motor injuries (Table 1) and include mild injuries such as whiplash. The *Upper extremities* factor only had one significant coefficient. Indeed, injuries located in the upper extremities had no explanatory power as regards the probability of being hospitalized. In the event of admission, however, the expected length of hospital stay was lower for victims with upper extremity injuries. Norin et al. (1997) found that injuries to the neck and arms account for almost 40% of injuries at lower levels of disability, although they represent fewer than 5% at higher disability levels.

Finally, hospitalization and recovery duration are also explained by the nature of the injuries, as indicated by the significance of the associated coefficients. Fractures and internal injuries increased both the likelihood of being admitted to hospital and the expected recovery period as an in-patient. By contrast, minor injuries such as contusions, sprains and strains were negatively related to the likelihood of hospitalization and the length of hospital stay. Wounds and abrasions did have an influence on the probability of being admitted to hospital, but not on the length of stay. In fact, while wound injuries increased the probability of hospitalization, this decreased when the victim suffered abrasions as a result of the collision. One explanation for the negative effect of abrasions on the hospitalization likelihood would be that most of these injuries are minor abrasions for which first aid only is required. Similar findings have been reported by other authors. Peek-Asa et al. (2011) showed that in the USA motor-injured teenagers with fractures, internal injuries or intracranial injuries were associated with both higher hospital charges and longer lengths of hospital stay. Thygerson et al. (2011) found that fracture and wound injuries accounted for only a fifth of visits to emergency departments but almost two thirds of hospitalizations of injured drivers. By contrast, bruises and abrasions were much more frequent causes of emergency department visits than of hospitalizations. To conclude, fractures and internal injuries,

[5] Heterogeneity between units of analysis was also analyzed. Unobserved heterogeneity between the units of analysis may be considered by means of a random-parameters Tobit regression model, as made by Anastasopoulos et al. (2012b) to account for unobserved heterogeneity between road segments. To consider the possibility of unobserved heterogeneity, we estimated a random parameter model. However, a likelihood ratio test comparing fixed- and random-parameter models indicated that the two approaches where not statistically different and thus, in this case, random parameters were not warranted.

Table A1
Results of the Hausman test.

K value	Hausman statistic value	*p*-Value
4	9.596	0.962
5	8.098	0.986
6	7.662	0.990

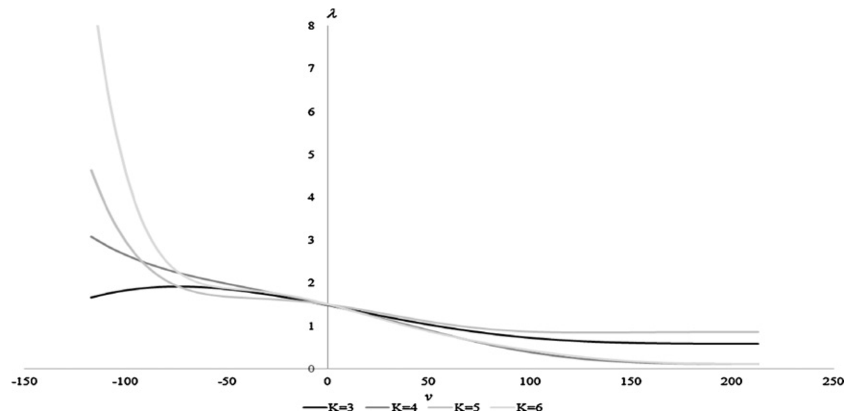

Fig. A1. Power series estimation of $\hat{\hat{\lambda}}_i = h(\hat{v}_i)$ for $K=3$, $K=4$, $K=5$ and $K=6$. Note that the shape of the $h(\hat{v}_i)$ function is more flexible for larger K values. This is because the influence of the sample information increases with K. Additionally, the resulting approximation is a monotonic decreasing function when $K=4$, $K=5$ and $K=6$. Similarly, a monotonic decreasing function is obtained if a Probit model is applied to estimate $h(\hat{v}_i)$. However, the resulting approximation is not a monotonic decreasing function when the function is approximated by a second-order power series ($K=3$). Consequently, the analysis is performed for $K=4$, $K=5$ and $K=6$.

followed by wounds, seem to be the most serious types of injuries in terms of hospital admission and/or length of hospital stay.

5. Conclusion

Motor traffic accidents are the leading mechanism of hospitalizations in Europe. Most of the road safety literature analysing the factors which influence the length of hospital stay is based on hospital data and, therefore, the results are limited to motor victims who were admitted to hospital. Here a generalized Tobit regression model was applied to analyse jointly the factors that influence both the likelihood of being admitted to hospital after a motor collision and the length of hospital stay in the event of admission. These two processes were shown to be statistically dependent and, therefore, biased estimates would be obtained if they were modelled separately. Although the reliability of Tobit estimates depends on distributional hypotheses, these are seldom tested (to the best of our knowledge, they have never been tested in road safety applications). Here the consistency of parameter estimates was tested by means of semi-parametric techniques, which did not reject the hypothesis of normality of errors.

The age, gender and type of victim, as well as the location and nature of injuries, were found to be factors that influence the likelihood of being admitted to hospital and/or the length of hospital stay required to recover from injuries sustained in a motor collision. These findings are of particular interest for road safety policy makers, since the analysis of factors that explain hospitalization can be used to target road safety policies so as to reduce the hospitalization rate and length of hospital stay of motor victims. For instance, fractures and injuries located in the head are associated with higher hospitalization rates and longer recovery periods in hospital. A priority for policy makers should therefore be to identify and reduce the types of collisions associated with these injuries. The present analysis can also help policy planners to tackle the motor injury problem not only from the viewpoint of the severity of disabilities caused by collisions but also in terms of the financial consequences borne by society, since hospital charges may be measured financially. Finally, understanding the relationship between hospital admissions and length of hospital stay for motor victims is also of great relevance to medical practitioners, since collisions are a major cause of hospitalizations.

The above discussion shows that the analysis of hospitalization factors is of interest for multiple research areas in road safety. From the methodological point of view, the modelling of hospital duration presents fascinating challenges. A potential line of study for the future is to extend the econometrics to hazard-based duration modelling in order to capture the duration effects in survival analysis. The application of mixture hazard-based models to analyse jointly cases with positive probability of occurrence of an event and cases where it is impossible that the event occurs is an interesting approach that is worth exploring.

Acknowledgements

The authors are grateful to the two anonymous referees for the general and specific comments. We also wish to acknowledge the support of the Spanish Ministry of Education and Science and FEDER Grant ECO2008-01223.

Annex.

A.1. Normality test results

The normality hypothesis testing involves a semi-parametric estimation of the model (1), which leads to consistent estimates in both of the hypotheses. However, a K value must be selected to approximate the unknown function $h(\hat{v}_i)$ in the semi-parametric method. Fig. A1 shows the semi-parametric estimation of $h(\hat{v}_i)$ for $K=3$, $K=4$, $K=5$ and $K=6$.

Table A1

References

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34, 597–603.
Abdel-Aty, M., Chen, C., Schott, J., 1998. An assessment of the effect of driver age on traffic accident involvement using log-linear models. Accident Analysis and Prevention 30, 851–861.
Amemiya, T., 1984. Tobit models: a survey. Journal of Econometrics 24, 3–61.
Anastasopoulos, P., Tarko, A., Mannering, F., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis and Prevention 40, 768–775.
Anastasopoulos, P., Mannering, F., Shankar, V., Haddock, J., 2012a. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45, 110–119.
Anastasopoulos, P., Shankar, V., Haddock, J., Mannering, F., 2012b. A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis and Prevention 45, 628–633.
Awadzi, K., Classen, S., Hall, A., Duncan, R.P., Garvan, C.W., 2008. Predictors of injury among younger and older adults in fatal motor vehicle crashes. Accident Analysis and Prevention 40, 1804–1810.
Barell, V., Aharonson-Daniel, L., Fingerhut, L.A., Mackenzie, E.J., Ziv, A., Boyko, V., Abargel, A., Avitzour, M., Heruti, R., 2002. An introduction to the Barell body region by nature of injury diagnosis matrix. Injury Prevention 8, 91–96.

Boucher, J.-P., Santolino, M., 2010. Discrete distributions when modeling the disability severity score of motor victims. Accident Analysis and Prevention 42 (6), 2041–2049.

Cameron, A., Trivedi, P., 2005. Microeconometrics. Cambridge University Press, Cambridge.

Chong, S., Poulos, R., Olivier, J., Watson, W., Grzebieta, R., 2010. Relative injury severity among vulnerable non-motorised road users: comparative analysis of injury arising from bicycle–motor vehicle and bicycle–pedestrian collisions. Accident Analysis and Prevention 42, 290–296.

Delen, D., Sharda, R., Bessonov, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accident Analysis and Prevention 38, 434–444.

Eluru, N., Bhat, C., Hensher, D., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis and Prevention 40, 1033–1054.

Evans, L., 2001. Female compared with male fatality risk from similar physical impacts. The Journal of Trauma: Injury, Infection, and Critical Care 50, 281–288.

Elvik, R., Mysen, A.B., 1999. Incomplete accident reporting: Meta-analysis of studies made in 13 Countries. Transportation Research Record 1665, 133–140.

Farah, H., Bekhor, S., Polus, A., 2009. Risk evaluation by modeling of passing behavior on two-lane rural highways. Accident Analysis and Prevention, 887–894.

Fildes, B., Lane, J., Lenard, J., Vulcan, A.P, 1994. Passenger cars and occupant injury: side impact crashes. Federal Office of Road Safety, Report CR 134.

Forman, J., Lopez-Valdes, F., Pollack, K., Heredero-Ordoyo, R., Molinero, A., Mansilla, A., Fildes, B., Segui-Gomez, M., 2011. Injuries among powered two-wheeler users in eight European countries: a descriptive analysis of hospital discharge data. Accident Analysis and Prevention, http://dx.doi.org/10.1016/j.aap.2011.02.020.

Goldstein, J.P., 1986. The effect of motorcycle helmet use on the probability of fatality and the severity of head and neck injuries. Evaluation Review 10, 355–375.

Guria, J., 1990. Length of hospitalization—an indicator of social costs of disabilities from traffic injuries. Accident Analysis and Prevention 22 (4), 379–389.

Hagist, C., Kotlikoff, L.J., 2009. Who's going broke? Comparing healthcare costs in ten OECD countries. Hacienda Pública Española/Revista de Economía Pública 188, 55–72.

Hauer, E., Hakkert, S., 1989. The extent and some implications of incomplete crash reporting. Transportation Research Record 1185, 1–10.

Heckman, J.J., 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Annals of Economic and Social Measurement 5, 475–492.

Heckman, J.J., 1979. Sample selection as a specification error. Econometrica 47, 153–161.

Hill, J., Boyle, L., 2006. Assessing the relative risk of severe injury in automotive crashes for older female occupants. Accident Analysis and Prevention 38, 148–154.

Huang, H., Chin, H.-C., Haque, M.-M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40, 45–54.

Hutchinson, T., 1986. Statistical modelling of injury severity, with special reference to driver and front seat passenger in single-vehicle crashes. Accident Analysis and Prevention 18, 157–167.

Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single vehicle accident injuries: some additional evidence. Journal of Safety Research 37, 267–276.

Klein, R., Spady, R., 1993. An efficient semiparametric estimator for binary response models. Econometrica 61, 387–421.

Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis and Prevention 34, 313–321.

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis and Prevention 37, 775–786.

Li, G., Braver, E., Chenb, L.-H., 2003. Fragility versus excessive crash involvement as determinants of high death rates per vehicle-mile of travel among older drivers. Accident Analysis and Prevention 35, 227–235.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291–305.

Machlin, S.R., Carper, K., 2007. Expenses for Inpatient Hospital Stays 2004. Statistical Brief #164. Agency for Healthcare Research and Quality, Rockville, MD.

Malyshkina, N., Mannering, F., 2009. Markov switching multinomial logit model: an application to accident-injury severities. Accident Analysis and Prevention 41, 829–838.

Mannering, F., 1993. Male/female driver characteristics and accident risk: some new evidence. Accident Analysis and Prevention 25, 77–84.

Mashtare Jr., T., Hutson, A., 2011. Utilizing the flexibility of the epsilon-skew-normal distribution for Tobit regression problems. Communications in Statistics: Theory and Methods 40, 408–423.

Morgan, A., Mannering, F., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. Accident Analysis and Prevention 43, 1852–1863.

Nevarez, A., Abdel-Aty, M.A., Wang, X., Santos, J., 2009. Large scale injury severity analysis for arterial roads: modeling scheme and contributing factors. Presented at the 88th Transportation Research Board Annual Meeting, Washington, DC.

Newey, W.K., 1987. Specification tests for distributional assumptions in the Tobit model. Journal of Econometrics 34, 124–145.

Newey, W.K., 2009. Two-step series estimation of sample selection models. The Econometrics Journal 12, S217–S229.

Newgard, C., 2008. Defining the "older" crash victim: the relationship between age and serious injury in motor vehicle crashes. Accident Analysis and Prevention 40, 1498–1505.

Newgard, C., Lewis, R., Kraus, J., McConnell, J., 2005. Seat position and the risk of serious thoracoabdominal injury in lateral motor vehicle crashes. Accident Analysis and Prevention 37, 668–674.

Neyens, D., Boyle, L., 2008. The influence of driver distraction on the severity of injuries sustained by teenage drivers and their passengers. Accident Analysis and Prevention 2008, 254–259.

Norin, H., Kraff, M., Korner, J., Nygren, A., Tingvall, C., 1997. Injury severity assessment for car occupants in frontal impacts, using disability scaling. Journal of Clinical Epidemiology 50, 95–103.

O'Donnell, C.J., Connor, D.H., 1996. Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice. Accident Analysis and Prevention 28, 739–753.

Paleti, R., Eluru, N., Bhat, C., 2010. Examining the influence of aggressive driving behaviour on driver injury severity. Accident Analysis and Prevention 42, 1839–1854.

Peek-Asa, C., Yanga, J., Ramireza, M., Hamanna, C., Cheng, G., 2011. Factors affecting hospital charges and length of stay from teenage motor vehicle crash-related hospitalizations among United States teenagers, 2002–2007. Accident Analysis and Prevention 43, 595–600.

Pucher, J., Dijkstra, L., 2003. Promoting safe walking and cycling to improve public health: lessons from The Netherlands and Germany. American Journal of Public Health 93, 1509–1516.

Richardson, J., Kim, K., Li, L., Nitz, L., 1996. Patterns of motor vehicle crash involvement by driver age and sex in Hawaii. Journal of Safety Research 27, 117–125.

Santolino, M., 2010. Determinants of the decision to appeal against motor bodily injury judgements made by Spanish trial courts. International Review of Law & Economics 30, 37–45.

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39, 955–963.

Savolainen, P., Mannering, F., Lord, D., Quddus, M., 2011. The statistical analysis of crash-injury severities: a review and assessment of methodological alternatives. Accident Analysis and Prevention 43, 1666–1676.

Segui-Gomez, M., Martinez-Gonzalez, M.A., de Irala, J., Ewert, U., 2008. Injury-related Hospitalizations in Europe 2004. European Center for Injury Prevention. Universidad de Navarra.

Smith, K., Cummings, P., 2004. Passenger seating position and the risk of passenger death or injury in traffic crashes. Accident Analysis and Prevention 36, 257–260.

Tarko, A., Azam, M.S., 2011. Pedestrian injury analysis with consideration of the selectivity bias in linked police-hospital data. Accident Analysis and Prevention 43, 1689–1695.

Tay, R., 2006. Ageing drivers: storm in a teacup? Accident Analysis and Prevention 38, 112–121.

Tay, R., Rifaat, S., 2007. Factors contributing to the severity of intersection crashes. Journal of Advanced Transportation 41, 245–265.

Thygerson, S., Merrill, R., Cook, L., Thomas, A., 2011. Comparison of factors influencing emergency department visits and hospitalization among drivers in work and nonwork-related motor vehicle crashes in Utah, 1999–2005. Accident Analysis and Prevention 43, 209–213.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24–36.

Ulfarsson, G., Mannering, F., 2004. Differences in male and female injury severities in sport-utility vehicle, pickup and passenger car accidents. Accident Analysis and Prevention 36, 135–147.

Valent, F., Schiava, F., Savonitto, C., Gallo, T., Brusaferro, S., Barbone, F., 2002. Risk factors for fatal road traffic accidents in Udine, Italy. Accident Analysis and Prevention 34, 71–84.

Washington, S., Karlaftis, M., Mannering, F., 2003. Statistical and Econometric Methods fro Transportation Data Analysis. Chapman & Hall/CRC.

Yamamoto, T., Hashiji, J., Shankar, V., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis and Prevention 40, 1320–1329.

Yamamotoa, T., Shankar, V., 2004. Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. Accident Analysis and Prevention 36, 869–876.

Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models. Transportation Research Record 2241, 51–58.