Hacettepe Journal of Mathematics and Statistics
Volume 36 (1) (2007), 79 – 87

FACTORS INFLUENCING THE SEVERITY OF DAMAGE IN BUS ACCIDENTS IN TURKEY DURING 2002 : AN APPLICATION OF THE ORDERED PROBIT MODEL

Özge Uçar* and Hüseyin Tatlıdil*

Received 20 : 02 : 2006 : Accepted 05 : 01 : 2007

Abstract

This paper presents an application of the ordered probit model to bus accident data to reveal the most significance factors that affect the severity of vehicle damage. Data recorded by The Department of Traffic Training and Research of the General Directorate of Security Affairs of Turkey for the year 2002 was used. The dependent variable was determined as the severity of damage, and was classified according to four levels: "No damage", "Little damage", "Medium damage" and "High damage". In view of the ordinality of the dependent variable, the use of the ordered probit model was preferred.

The results suggest that, all variables indicating specific locations of the accident, having an accident at night or in bad weather conditions increase the severity of damage. On the contrary, travelling in the same direction in two-vehicle accidents, bumping into another vehicle from one side and those accidents classified as "other types of accident" lead to non-severe damage. Additionally, the severity of damage decreases for drivers graduated from high school.

Keywords: Severity of damage, Traffic crash modelling, Bus accidents, Ordered probit model.

2000 AMS Classification: 62 J 12.

*Department of Statistics, Hacettepe University, Ankara, Turkey.
E-mail: (Ö. Uçar) ozgeucar@hacettepe.edu.tr (H. Tatlıdil) tatlidil@hacettepe.edu.tr

1. Introduction

The use of ordered response models is gradually becoming widespread in a variety of research fields. Especially in surveys carried out in social sciences, a large number of ordered variables are included to the survey questionnaire. If one of these variables is the dependent variable and if the effect of all other variables on the dependent variable is investigated, ordered response models that take into consideration the ordinality of the dependent variable are chosen as appropriate analysis techniques. In the event that the dependent variable is actually ordered but we use unordered dependent variable models, parameter estimates will not be efficient. Similarly, despite the fact that the dependent variable is actually unordered but we use ordered models, parameter estimates will be biased. When we compare these two types of statistical negativeness, obtaining unbiased estimates is statistically more important than efficiency.

The aim of this study is to reveal major factors that affect the severity of vehicle damages in bus accidents in Turkey during 2002 using the ordered probit model. This study may be seen as a preliminary study for actuarial studies by insurance firms concerned with the severity of vehicle damage. According to the characteristics of drivers and vehicles, the probability of an observation belonging to each category of the dependent variable may be predicted with these models. Therefore, risk groups are constructed and firms determine insurance premiums for vehicles in advance. Additionally, in the light of the significant factors, preventive measures may be taken, which will decrease the accident rate.

Using the damage totals and age of vehicles, the severity of damages is classified according to four levels. Descriptions and coding of the categories according to the records of The Department of Traffic Training and Research of General Directorate of Security Affairs of Turkey are given as follows.

0 No-damage
1 Little-damage
2 Medium-damage
3 High-damage

Under the ordinality assumption of the dependent variable, the ordered probit model is applied to the data set.

2. Literature Review

Although there is an extensive literature including studies related to discrete choice models, examining traffic crash modelling using ordered response models have only become popular in recent years. It will be useful to review the studies related to discrete choice models in the literature.

O'Donnell and Connor (1996) assessed probabilities of four levels of injury severity as a function of driver attributes, and compared the specifications of ordered logit and probit models.

In a study made in Singapore by Quddus, Noland and Chin (2002), injuries and severity of vehicle damage were examined in motorcycle accidents. Road conditions, attributes of the drivers, environmental factors and characteristics of the motorcycles were included in the model as explanatory variables. Along with the change in the explanatory variables, the severity of injuries and vehicle damage were examined using the ordered probit model. While the first dependent variable was included in the model with three categories described as fatal injury, severe injury and non-severe injury, the second dependent variable in the model was classified into four categories described as no-damage, non-severe damage, severe damage and wreckage.

In order to test the different injury levels in single-vehicle crashes, two-vehicle crashes and all types of crashes as a function of driver attributes, vehicle characteristics and type of collision, Kockelman and Kweon (2002) used the ordered probit model in their study in the USA. Descriptions of the levels coded as $0, 1, 2, 3$ are given as no-injury, non-severe injury, severe injury and fatal injury. Under the assumption of the ordinality of the dependent variable, they used the ordered probit model.

In order to test the effects of driver attributes, vehicle factors and type of collision on injuries and rollover of vast vehicles between the years 1996-1998 in North Carolina; Khattak, Schneider and Targa (2002) used the ordered and binary probit model, respectively.

3. Methodology

In the event that the dependent variable has more than two categories, there are various estimation techniques under different assumptions, two of which are the ordered and multinomial logit and probit models. When the categories of the dependent variable are ordered and a basic assumption called the 'Parallel Slopes Assumption' is satisfied, the use of ordered response models is preferred. If the categories of the dependent variable are unordered and the IIA (Independence of Irrelevant Alternatives) assumption is satisfied, the choice of multinomial models helps one to obtain more accurate results. Since we used the ordered probit model, we only discus the methodology of these models and the readers may refer to Powers and Xie (2000), Maddala (1983), Long (1997) and Borooah (2002) for more information about multinomial models.

Ordered response models are a natural extension of binary response models, which are based on the latent variable approach. The latent regression model is given as follows,

$$Y^* = \sum_{k=1}^{K} \hat{b}_k X_k + \varepsilon, \tag{3.1}$$

where Y^* denotes the latent variable that indicates the underlying tendency of the observations, the x_k 's are explanatory variables and the $\hat{b}_k$'s are the estimated linear regression model parameters that are related to the x_k 's.

It is assumed that the error term follows a definite symmetric distribution with zero mean, such as the normal or logistic distribution. If it follows a normal distribution, the model is called the ordered probit model, and when the logistic distribution is assumed for the distribution of the error term, the model is called the ordered logit model.

The relationship between observed levels of the dependent variable and the underlying tendency is given as follows,

$$\begin{aligned} &Y_i = 1,\ Y_i^* \le \mu_1 (= 0) \\ &Y_i = 2,\ \mu_1 < Y_i^* < \mu_2 \\ &Y_i = 3,\ \mu_2 < Y_i^* < \mu_3 \\ &\quad \vdots \\ &Y_i = J,\ \mu_{J-1} < Y_i^*, \end{aligned} \tag{3.2}$$

where the μ_j 's represent threshold parameters that separate the adjacent categories of dependent variables with J ordered categories. Using the latent variable approach, the probability of belonging to the category j can in general be expressed as follows:

$$(3.3)\qquad \begin{aligned} P(Y=j) &= F\left[\mu j \sum_{k=1}^{K} \hat{b}_k\, x_k\right] - F\left[\mu_{j-1} - \sum_{k=1}^{K} \hat{b}_k\, x_k\right] \\ P(Y \leq j) &= F\left[\mu_j - \sum_{k=1}^{K} \hat{b}_k\, x_k\right] \end{aligned}$$

The total number of threshold parameters to be estimated is $(J-2)$. When the model includes a constant term, one of the threshold parameters cannot be estimated by the maximum likelihood estimation technique in LIMDEP. As a solution, Greene (2000) suggests that the first threshold parameter should be set to zero ($\mu_1 = 0$) . In order to obtain a positive probability value in Equation (3.3), the threshold parameters must satisfy the relations given by Equation (3.4). Since the first threshold parameter is zero, all estimated threshold parameters will be positive.

$$(3.4)\qquad \mu_1 = 0 < \mu_2 < \cdots < \mu_{J-1}.$$

3.1. The parallel slopes assumption. This assumption indicates that the coefficient for an explanatory variable, which affects the likelihood of it being in the ordered categories of an observation, must be the same along all categories of the dependent variable. That is, the coefficient of an explanatory variable changes along the categories in multinomial models whereas only estimating one parameter vector for each category will be adequate in ordered models.

3.2. The ordered probit model. The equations of the probabilities associated with the J ordered categories are given as follows,

$$(3.5)\qquad \begin{aligned} P(Y=1) &= \Phi\left[-\sum_{k=1}^{K} \hat{b}_k\, x_k\right] \\ P(Y=2) &= \Phi\left[\mu_2 - \sum_{k=1}^{K} \hat{b}_k\, x_k\right] - \Phi\left[-\sum_{k=1}^{K} \hat{b}_k\, x_k\right] \\ &\vdots \\ P(Y=J) &= 1 - \Phi\left[\mu_{J-1} - \sum_{k=1}^{K} \hat{b}_k\, x_k\right], \end{aligned}$$

where Φ represents the cumulative standard normal distribution function.

The probability of an observation belonging to, or being smaller than, category j is given by the following equation in the ordered probit model.

$$(3.6)\qquad P(Y \leq j) = P(Y^* \leq \mu_j) = \Phi\left[\mu_j - \sum_{k=1}^{K} \hat{b}_k\, x_k\right].$$

The basic methodology of ordered response models is described above. To obtain more information about the estimation techniques, the equations associated with marginal effects of models and interpretations, see Borooah (2002), Maddala (1983), Long (1997), Powers and Xie (2000).

4. The data set

Data used in this study came from the records for 2002 of The Department of Traffic Training and Research of the General Directorate of Security Affairs of Turkey. The total number of bus accidents that occurred in the year 2002 was 3467. The descriptive statistics suggest that while 78.8% of these accidents occurred in settlement locations, 65.7% occurred on avenues or streets. That the majority of bus accidents occurred in the daytime and on dry surfaces, which imply clear weather, does not comply with our expectations and since almost all of the bus drivers are male, the variable indicating the sex of drivers was removed from the study.

Explanatory variables may be classified into four categories. The first category indicates the attributes of the drivers including education and age. The second indicates accident characteristics including the location, the number of vehicles in the accident and the type of collision. The remaining two categories are the characteristic of the road, including day-weather conditions, the road surface, and the age of the vehicles, respectively.

Table 1. Explanatory Variables

Variable	Mean	S.D.	Description	
x_0	1.00	0.000	Constant	
x_1	0.018	0.13	= 1 if accident occurs on a superhighway (SW)	= 0 otherwise
x_2	0.31	0.46	= 1 if accident occurs on a provincial road (PR)	= 0 otherwise
x_3	0.019	0.14	= 1 if accident occurs on a road other than, SW or PR	= 0 otherwise
x_4	0.29	0.45	= 1 if accident involves two vehicles (same direction)	= 0 otherwise
x_5	0.17	0.38	= 1 if accident involves two vehicles (adjacent directions)	= 0 otherwise
x_6	0.11	0.31	= 1 if accident involves two vehicle (opposite directions)	= 0 otherwise
x_7	0.11	0.31	= 1 if accident involves more than two vehicles	= 0 otherwise
x_8	0.20	0.40	= 1 if collision is bumping rear-end (RE);	= 0 otherwise
x_9	0.32	0.47	= 1 if collision is bumping from one side (OSV)	= 0 otherwise
x_{10}	0.016	0.12	= 1 if collision is bumping into stationary vehicle (SV)	= 0 otherwise
x_{11}	0.017	0.13	= 1 if collision is bumping into stationary material (SM)	= 0 otherwise
x_{12}	0.019	0.14	= 1 if collision is a rollover (R)	= 0 otherwise
x_{13}	0.31	0.46	= 1 if collision is other than RE, OSV, SV, SM or R	= 0 otherwise
x_{14}	0.97	0.17	= 1 if the road surface is asphalt;	= 0 otherwise
x_{15}	0.012	0.11	= 1 if the road surface is parquet	= 0 otherwise
x_{16}	0.004	0.06	= 1 if the road surface is gravel	= 0 otherwise
x_{17}	0.24	0.43	= 1 if the accident occurs at night	= 0 otherwise
x_{18}	0.04	0.19	= 1 if the accident occurs at dawn	= 0 otherwise
x_{19}	0.015	0.12	= 1 if the weather is foggy (FW)	= 0 otherwise
x_{20}	0.09	0.29	= 1 if the weather is rainy (RW)	= 0 otherwise
x_{21}	0.028	0.16	= 1 if the weather is snowy (SW)	= 0 otherwise
x_{22}	0.15	0.35	= 1 if the weather is other than FW, RW or SW	= 0 otherwise
x_{23}	0.33	0.47	= 1 if the age of the driver is in group (26-35)	= 0 otherwise
x_{24}	0.38	0.48	= 1 if the age of the driver is in group (36-45)	= 0 otherwise
x_{25}	0.19	0.39	= 1 if the age of driver is in group (46-55)	= 0 otherwise
x_{26}	0.03	0.16	= 1 if the driver is 56 years old or older	= 0 otherwise
x_{27}	0.68	0.47	= 1 if the driver graduated from primary school	= 0 otherwise
x_{28}	0.17	0.38	= 1 if the driver graduated from secondary school;	= 0 otherwise
x_{29}	0.14	0.34	= 1 if the driver graduated from high school;	= 0 otherwise
x_{30}	0.08	0.07	age of vehicle (years) divided by 100	

There are many variables that might affect the severity of vehicle damage, but after obtaining the correlation matrix which revealed any strong dependencies among variables, we decided to remove one of the correlated variables from the study to eliminate a dependency that could lead to incorrect statistical results. Therefore, the independency of the explanatory variables, that is the main assumption of the discrete choice models, was satisfied too. Another reason of these changes result from the fact that the LIMDEP software by Greene (1995) does not give parameter estimates if there is a significant correlation between the explanatory variables.

Table 2. Parameter Estimates

Reference Category	Variable	$\hat{\beta}$	Std Error (S.E)	b/S.E	P-values
Avenue or Street	x_1	**1.2958**	**0.1427**	**9.083**	**0.00**
	x_2	**0.9951**	**0.0488**	**20.39**	**0.00**
	x_3	**0.3661**	**0.1448**	**2.528**	**0.01**
Single Vehicle	x_4	**-0.3012**	**0.1196**	**-2.520**	**0.01**
	x_5	-0.1894	0.1401	-1.352	0.18
	x_6	-0.1689	0.1417	-1.192	0.23
	x_7	-0.1081	0.1361	-0.794	0.43
Head-on Crash	x_8	0.1763	0.1024	-1.723	0.09
	x_9	**-0.1715**	**0.0873**	**-1.965**	**0.04**
	x_{10}	-0.2597	0.1818	-1.428	0.15
	x_{11}	0.2464	0.1898	1.298	0.19
	x_{12}	0.1594	0.1896	0.841	0.40
	x_{13}	**-1.6085**	**0.1459**	**-11.02**	**0.00**
Concrete	x_{14}	0.0656	0.1843	0.356	0.72
	x_{15}	-0.3008	0.3056	-0.985	0.32
	x_{16}	-0.0906	0.4189	-0.216	0.83
Daytime	x_{17}	**0.2755**	**0.0485**	**5.685**	**0.00**
	x_{18}	0.0451	0.1247	0.361	0.72
Clear Weather	x_{19}	**0.3339**	**0.1667**	**2.003**	**0.04**
	x_{20}	**0.3547**	**0.0679**	**5.223**	**0.00**
	x_{21}	**0.6413**	**0.1033**	**6.209**	**0.00**
	x_{22}	0.0017	0.0604	-0.027	0.98
Age Group (18-25)	x_{23}	0.1101	0.0939	1.172	0.24
	x_{24}	0.1061	0.0939	1.129	0.26
	x_{25}	0.1333	0.0987	1.350	0.18
	x_{26}	0.0364	0.1517	0.240	0.81
Higher Edu.	x_{27}	-0.3012	0.1821	-1.655	0.09
	x_{28}	-0.2355	0.1874	-1.257	0.21
	x_{29}	**-0.3757**	**0.1884**	**-1.994**	**0.04**
	x_{30}	**-1.7215**	**0.3121**	**-5.516**	**0.00**
Threshold (μ_1)		**2.33346**	**0.0458**	**50.935**	**0.00**
Threshold (μ_2)		**3.35174**	**0.05671**	**59.108**	**0.00**
Constant		**1.7273**	**0.3069**	**5.629**	**0.00**
-2LLR		**1690.06**			**0.00**

As in the study of O'Donnell and Connor (1996), a small change was made to the continuous variable (age of the vehicle), where all values were divided by 100. As explained by Greene (2000), the reason for such a change is the necessity to have the mean in the interval (0-1). Since all variables are coded using dummy variable coding in LIMDEP, the mean value of all variables must be in the interval (0-1) and LIMDEP does not give parameter estimations unless this restriction is satisfied. After removing irrelevant variables from the study, the remaining variables are shown in Table 1.

5. Model Results

The results are assessed according to the signs of the coefficients and marginal effects of the explanatory variables on the probabilities.

Parameter estimates are given in Table 2. Table 3 consists of the marginal effects of the explanatory variables on the probabilities.

Before interpreting the maximum likelihood estimates of the unknown parameters (β) and the thresholds, the validity of the model and the ordinality of the dependent variable must be verified. The –2LLR (-2 Logarithmic Likelihood Ratio) statistics in Table 2 indicates the validity of the model. All threshold parameters are statistically significant as seen in Table 2, and this result confirms our prior assumption about the ordinality of the dependent variable.

According to the reference categories of each variable group, different interpretations are provided here. First, the results are interpreted according to the signs of the coefficients.

5.1. Signs of the coefficients. Positive and significant coefficients indicate more severe damages whereas negative and significant coefficients indicate less severe damages. We conclude that having an accident at night (x_{17}) on a superhighway (SW-x_1), provincial road (PR-x_2), roads other than SW and PR (x_3) such as village and forest roads in all type of weather conditions such as foggy (FW-x_{19}), rainy (RW-x_{20}) and snowy (SNW-x_{21}) are all associated with more severe damage.

When we compare all factors leading to an increase in the probability of more severe damage, having an accident on a superhighway (1.2958) leads to an increased rate of severe damage. In comparison with accidents occurring on provincial roads (0.9951), the risk of being exposed to more severe damage to the vehicle is about 1.3 (1.2958/0.9951) times higher for accidents occurring on superhighways (1.2958). This ratio rises to 3.5 (1.2958/0.3661) for accidents occurring on roads other than SW and PR (0.3661).

Having an accident in snowy weather (0.6413), at night (0.2755) causes more severe crashes and more severe damage. Foggy (0.3339), rainy (0.3547) and snowy (0.6413) weather conditions have increasing effects on the severity of vehicle damage. The risk of being exposed to a high degree of damage to vehicles in accidents occurring in snowy weather is about 1.92 (0.6413/0.3339) times higher than in foggy weather, and about 1.81 (0.6413/0.3547) times higher than in rainy weather.

All other significant variables have a negative coefficient and therefore a decreasing effect on the severity of damage.

In comparison with accidents involving bumping one side of the vehicle (-0.1715), the chance of being exposed to less severe damage of the vehicles is about 9.38 (-1.6085/-0.1715) times higher with other types of accident (-1.6085).

A one-year increase in the age of the vehicle causes a decrease in the probability of severe damages.

All other variables are not statistically significant.

Interpretations according to the signs of the coefficients are only a guide which reveal the direction of the changes in the probabilities. In order to determine the magnitudes of these changes associated with a unit change in the value of an explanatory variable; marginal effects of explanatory variables on probabilities must be interpreted.

5.2. Marginal effects on probability. Marginal effects of all statistically significant explanatory variables on each probability associated with the four categories of severity of vehicle damages are shown in Table 3.

Table 3. Marginal Effects of Explanatory Variables on Probabilities

Indicator Categories	$P(Y = 0)$	$P(Y = 1)$	$P(Y = 2)$	$P(Y = 3)$
Superhighway	**-0.2591**	-0.0053	0.2160	**0.0484**
Provincial road	-0.1990	-0.0041	0.1659	0.0372
Other roads	-0.0732	-0.0015	0.0610	0.0137
Two vehicle (same direction)	0.0602	0.0012	-0.0502	-0.0113
Bumping from one side	0.0343	0.0007	-0.0286	-0.0064
Other crash type	0.3216	0.0066	-0.2681	-0.0601
Night	-0.0551	-0.0011	0.0459	0.0103
Foggy	-0.0668	-0.0014	0.0557	0.0125
Rainy	-0.0709	-0.0014	0.0591	0.0133
Snowy	-0.1282	-0.0026	0.1069	0.0240
High-education	0.0751	0.0015	-0.0626	-0.0140
Age of the vehicle	**0.3442**	0.0070	-0.2869	**-0.0643**

We conclude that in moving from an accident occurring on an avenue or street to an accident occurring on a superhighway, the probability of being exposed to no damage to the vehicle decreases 0.2591 unit and the probability of being exposed to high damage increases 0.0484 unit.

Similarly, a one-year increase in the age of the vehicle causes a 0.3442 unit increase in the probability of being exposed to no damage to the vehicle and a 0.0643 unit decrease in the probability of being exposed to high damage to the vehicle.

Other values in Table 3 may be interpreted in the same manner. When interpreting the results, the most important point to be kept in mind is that the changes in probabilities occur along with changes from the reference categories to the indicator categories.

6. Conclusion

In this study we have determined significant factors that affect the severity of vehicle damage in bus accidents in Turkey during 2002. Assuming that the dependent variable that indicates the severity of vehicle damage is ordered with four categories, the ordered probit model has been applied to the data set.

When a general assessment of the results is made according to the signs of the coefficients, accidents occurring on a superhighway lead to the most severe damage whereas older vehicles cause the least hazardous accidents.

Other significant factors that lead to an increase in the severity of damage are accidents occurring on a provincial road, ‘ other roads such as forest or village roads, all types of bad weather condition, especially snowy weather, and having an accident at night.

Factors such as two vehicle accidents (same direction), drivers graduated from high school, vehicles bumping from one side, or accidents coming under the label ‘other type of crashes’ such as bumping into a pedestrian or an animal, and the age of vehicle, have a decreasing effect on the severity of vehicle damage.

As a continuation of this study, the probabilities of belonging to the categories of the dependent variable may be calculated for each vehicle by eliminating insignificant explanatory variables. Therefore, according to the characteristics of the vehicles, the group to which each vehicle belongs may be determined. Additionally, as was discussed above, insurance firms may forecast the premiums of members in advance by using all this information.

References

[1] Borooah, V. K. *Logit and Probit (Ordered and Multinomial Models)* (Sage University Papers, London, 07-138, 2002).

[2] Greene, W. H. *LIMDEP Version 7.0 User's Manual* (Bellport: Econometric Software Inc, 1995).

[3] Greene, W. H. *Econometric Analysis* (New York University, Prince Hall, Upper Saddle River, New Jersey 07458, ISBN: 0-13-013297 7, 2000).

[4] Khattak, A. J., Schneider, R. J. and Targa, F. *Risk factors in large truck rollovers and injury severity: Analysis of single-vehicle collisions*, Transportation Research Board 82nd Annual Meeting, Washington, D.C., 2002.

[5] Kockelman, K. M. and Kweon, Y. J. *Driver injury severity: An application of ordered probit models*, Accident Analysis and Prevention **34**, 313–321, 2002.

[6] Long, J. S. *Regression Models for Categorical and Limited Dependent Variables* (Sage Publications, Advance Qualitative Techniques in the Social Sciences Series; No:7, Thousand Oaks, London, 1997).

[7] Maddala, G. S. *Limited-Dependent and Qualitative Variables in Econometrics*, Econometric Society Monographs **3** (Cambridge University Press, 1983).

[8] O'Donnell, C. J. and Connor, D. H. *Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice*, Accident Analysis and Prevention, **28**(6), 739–753, 1996.

[9] Powers, D. A. and Xie, Y. *Statistical Methods for Categorical Data Analysis* (Academic Press, 2000).

[10] Quddus, M. A., Noland, R. B. and Chin, H. C. *An analysis of motorcycle injury and vehicle damage severity using ordered probit models*, Journal of Safety Research **33**, 445–462, 2002.