A Poisson Regression Model of Fatal Traffic Accidents Involving Small Passenger Sedans in Japan

Kazumitsu Nawata
Department of Advanced Social and International Studies, University of Tokyo
3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, JAPAN, e-mail: nawata@waka.c.u-tokyo.ac.jp

Abstract

This paper analyzes fatal traffic accidents of small passenger sedans in Japan. Variables that may relate to the occurrence of traffic fatal accidents are analyzed using a regression type model. Nine explanatory variables describing characteristics of drivers and driving conditions are used as explanatory variables. The maximum likelihood method based on the Poisson distribution is used for estimations for the model. Both the assumptions of the Poisson distribution and the exogeneity of explanatory variables are also tested. Finally, the problems associated with the linear probability model and sample-selection biases of the explanatory variables are considered.

1. INTRODUCTION

An analysis of factors related to the occurrence of traffic accidents is important in reducing the number of such accidents. In 1992, the UK Department of Transportation conducted a regression analysis using data aggregated by the types of cars. The analysis showed that variables, such as the sex and age of the driver and the size and speed of the car affect the occurrence of traffic accidents. Given the different traffic environments in the UK and Japan, it would seem to be useful to conduct a similar analysis for Japan. The Institute of Traffic Accident Research and Data Analysis (ITARDA) in Japan recently developed the Traffic Accident Integrated Database, making possible an integrated analysis from the stand points of the driver, vehicle, and traffic environment.

This paper analyzes fatal traffic accidents in Japan using data for Sedan-A-Class cars, whose displacement volumes are predominantly 1500 cc or less; the cars were registered at the end of 1995. Data for 28 different types of cars without missing values are used. The regression-type model is used for the first time for accident data in Japan, and nine variables describing characteristics of drivers and driving conditions are used as explanatory variables.

As some types of cars in the data set were involved in few or no fatal accidents (two types reported no fatal accidents, nine types reported 1-5 fatal accidents), they are treated as discrete data. (For details and empirical examples of the discrete data, see , Winkelmann (1997) and Cameron and Trivedi (1998).) Moreover. as there are large differences in numbers of vehicles registered according to the types of cars, the model is estimated using the maximum likelihood method based on the Poisson distribution. Both the assumptions of the Poisson distribution and the exogeneity of explanatory variables are tested. Problems associated with the linear probability model and sample selection biases of the explanatory variables are also considered.

2. MODELS

The model based on aggregated data is presented this section. Suppose that there are k different types of cars. Let y_{ij} be a dummy variable that expresses whether or not the j-th registered vehicle of the i-th type causes a fatal accident during the observation period. Specifically,

(1) $y_{ij} = \begin{cases} 0 & \text{the vehicle does not cause a fatal accident,} \\ 1 & \text{the vehicle causes a fatal accident.} \end{cases}$

Let

(2) $y_{ij}^{*} = x_{ij}'\alpha + u_{ij}$,

where x_{ij} is a vector of explanatory variables representing characteristics of the driver and driving conditions. α is a vector of unknown parameters. As with the usual analysis of binary data, a fatal accident will happen if $y_{ij}^{*} > 0$ and will not happen if $y_{ij}^{*} \le 0$. Let F be the distribution function of $-u_{ij}$. The probability of fatal accident, $p_{ij} \equiv P(y_{ij} = 1 | x_{ij})$, is given by

(3) $p_{ij} = F(x_{ij}'\alpha)$.

All the explanatory variables are dummy variables, $1 >> p_{ij}$ and $n_i >> 0$, in this study. Therefore, by the law of small numbers and the reproductivity of the Poisson distribution, the total number of fatal accidents of the i-th type of car, $Y_i = \sum_{j=1}^{n_i} y_{ij}$, is considered to follow the Poisson distribution with the expected value $\lambda_i \equiv \sum_{j=1}^{n_i} p_{ij}$

where n_i is the number of registered vehicles of the i-th type of car.

Usually, F is assumed to be either a standard normal or logistic distribution. However, data regarding individual vehicles are not available, and only the aggregated data, which give averages of variables for each type of car, are available in this study. As a result, we cannot use these assumptions directly. Here, F is assumed to be approximated by a linear probability function, and a linear probability model is used in the analysis. If F is approximated by the linear probability function, we get,

(4) $p_{ij} \approx x_{ij}'\beta$, and $\lambda_i \approx X_i'\beta = n_i x_i'\beta$,

where $X_i = \sum_{j=1}^{n_i} x_{ij}$ and $x_i = X_i / n_i$.

Hence we can get the estimator of β by maximizing the logarithm of the likelihood function based on the Poisson distribution given by

(5) $\log L(\beta) = \sum_{i=1}^{k} \{-n_i x_i'\beta + Y_i \log(n_i x_i'\beta) - \log(Y_i!)\}$.

Note that for the consistency of individual behavior and the aggregated model, this model is essential.

3. DATA

3.1 Fatal Accidents

The data were obtained from the Traffic Accident Integrated Database of the ITARDA. The numbers of vehicles involved in fatal accidents were 3-year totals from 1992 through 1995. The data from 28 different types of cars without missing values were selected from Sedan-A-Class cars. The total number of vehicles involved in fatal accidents was 901 among 28 types of cars. The distribution is shown in Table 3.1. No vehicles were involved in fatal accidents for two types of cars, 1-5 vehicles were reported in fatal accidents for nine types, 6-20 were reported for one type, 21-50 were reported for two types, 51-100 were reported for nine types, and 101-150 were reported for two types. The average was 32.2 and the largest number of fatal accidents was 144.

Table 3.2 shows the distribution of the numbers of vehicles registered for the 28 types of cars. Fifty-thousand or fewer vehicles were registered for 5 types, 50-100 thousand for 4 types, 50-100 for 7 types, 400 thousand to 1 million for 6 types, and 2-3 million for one type. The total number of registered vehicles was 15,920,000; the largest number of any particular type was 2,773,000, and the average number per type was 568,000.

Table 3.3 shows the fatal accident rates per 10,000 registered vehicles (= vehicles involved in fatal accidents/ registered vehicles × 10,000). The fatal accident rates per 10,000 registered vehicles were 0-0.3 for 4 types, 0.3-0.6 for 13 types, 0.6-0.9 for 7 types, 0.9-1.2 for 2 types, and 1.2-1.5 for 2 types. The highest rate was 1.34, the average was 0.57, and the standard deviation was 0.55.

3.2 Explanatory Variables

Explanatory variables for the drivers and the driving conditions were selected. The variables were chosen from the Traffic Accident Integrated Database. The variables are shown in Table 3.4. As with the fatal accident data, the variables made use of data from 3 years, 1993 through 1995. Among these variables, the rate of a single-type accident, *SINGLE*, refers to a type of an accident which is different from other variables that are determined by the driver and driving conditions and can be treated as exogenous variables. However, since *SINGLE* can be considered an indicator of "the degree of the carelessness of driving," it is included in the study. [1]

Table 3.1 Distribution of Vehicles Involved in Fatal Accidents

Number of Vehicles	Types of Cars
0	2
1-5	9
6-20	1
21-50	9
51-100	5
101-150	2
Total	28

Table 3.2 Number of Vehicles Registered

Registered Numbers (Thousands)	Types
0-50	5
50-100	4
100-500	7
500-1,000	6
1,000-2,000	5
2,000-3,000	1
Total	28

Table 3.3 Fatal Accident Rates per 10,000 Registered Vehicles

Fatal Accident Rates	Types
0-0.3	4
0.3-0.6	13
0.6-0.9	7
0.9-1.2	2
1.2-1.5	2
Total	28

Table 3.4 Definitions of Explanatory Variables

Variables	Definition
MALE	Rate of the male driver
AGE24	Rate of the driver age 24 or younger
AGE65	Rate of the driver age 65 or older
BUSINESS	Rate of the business use
LEISURE	Rate of the leisure or pleasure use
BELT	Rate of wearing a seat belt
NIGHT	Rate of night driving
SPEED	Rate of speeding over 40 km/h
SINGLE	Rate of the single-type accident

It is desirable to use data that include all registered vehicles as the values for the explanatory variables in the analysis. Unfortunately, such a database does not currently exist. Therefore, data that were collected from vehicles involved in accidents with at least one casualty are used in this

paper. These data were not obtained by random sampling, and we cannot deny the possibility of sample selection biases which may have affected the analysis. However, casualty accidents are much more frequent than fatal accidents. The vehicles involved in casualty accidents per 10 thousand vehicles were 150.3-187.5 during the period, meaning that approximately 2% of registered vehicles were investigated. Although the data are incomplete, the analysis was carried out with the assumption that these data represent explanatory variables with a good approximation.

Among the explanatory variables, the values of *BELT* were obtained from all vehicles involved in casualty accidents and the values for the other variables were obtained from vehicles reporting at least one casualty. The expected signs are positive for *NIGHT, SPEED,* and *SINGLE* and negative for *BELT*; one-sided tests are employed for these variables. The means and standard deviations of the explanatory variables among 28 types of cars are shown in Table 3.5.

Table 3.5 Means and Standard Deviations of the Explanatory Variables

Variable	Mean	Standard Deviation
MALE	50.9%	12.4%
AGE24	29.1%	10.9%
AGE65	3.1%	2.0%
BUSINESS	9.6%	3.6%
LEISURE	8.5%	3.1%
BELT	86.0%	2.6%
NIGHT	33.5%	3.3%
SPEED	71.6%	14.6%
SINGLE	5.7%	1.7%

4. RESULTS OF THE ESTIMATION

4.1 Poisson Regression Models

The results of the estimation for the Poisson regression models given in Section 2 are presented here. The estimation was performed using TSP 4.4 and the following steps. First, the model, which contains all nine explanatory variables, was estimated. Next, by eliminating the variable with the smallest absolute t-value, models without less important explanatory variables were estimated. It is necessary to calculate the totals (= means × registered numbers) of explanatory variables for each type of car and to make them be X_i. Since the number of registered vehicles was large, X_i was calculated using 10,000 vehicles as be a unit. As a result, the rates of fatal accidents were per 10,000 vehicles. The results of estimation are given in Table 4.1.

For the model with all explanatory variables, *SINGLE* is significant at the 1% level and *AGE24* and *LEISURE* are significant at the 5% level. Although the absolute t-values are small and not significant at the 5% level, *NIGHT* is the expected sign, and *BELT* and *SPEED* are the opposite signs. For all models with the unimportant variables eliminated, *SINGLE* has the expected sign and is significant at the 1% level, and it is considered an important variable affecting the causalities of fatal accidents. Based on the *AIC* (Akaike Information Criterion), the model with 6 explanatory variables was selected. The selected model contains *MALE, AGE24, BUSINESS, LEISURE, NIGHT,* and *SINGLE* as explanatory variables. For this model, *MALE, AGE24,* and *LEISURE* are significant at the 5% level. The t-value of *SINGLE* is 4.679, and its *p-value* (one-sided) is 6.416×10^{-5}, which means that *SINGLE* is significant at any conventional level. These results indicate that factors such as the driver being male or age 24 or younger reduce the probability of fatal accidents; also, factors such as the car being driven for leisure or pleasure or involving a single-type accident increase the probability of fatal accidents.

Table 4.1 Results of Estimation of the Poisson Regression Models (t-values in parentheses)

Numbers of Explanatory Variables	9	8	7	6	5
Constant	0.047	-0.034	-0.020	-0.068	0.004
	(0.312)	(0.249)	(-0.167)	(-1.563)	(0.249)
MALE	-1.409	-1.468	-1.417	-1.441	-1.025
	(-1.869)	(-1.962)	(-2.202)	(-2.073)	(-2.111)
AGE24	-2.406	-2.410	-2.392	-2.413	-1.576
	(-2.303)	(-2.336)	(-2.462)	(-2.476)	(-2.429)
AGE65	1.019	1.324			
	(0.315)	(0.446)			
BUSINESS	3.482	3.499	3.455	3.358	2.568
	(1.797)	(1.915)	(1.901)	(1.858)	(1.722)
LEISURE	8.271	8.158	7.975	7.861	6.829
	(2.246)	(2.352)	(2.449)	(2.569)	(2.301)
BELT	-0.427	-0.617	-0.550		
	(-0.267)	(-0.459)	(-0.412)		
NIGHT	3.949	3.904	3.380	3.382	
	(1.353)	(1.363)	(1.578)	(1.601)	
SPEED	-0.128				
	(-0.196)				
SINGLE	9.896	9.668	9.888	10.595	12.450
	(2.803)	(3.606)	(3.817)	(4.679)	(5.660)
Log L	-83.366	-83.391	-83.489	-83.537	-84.694
AIC	184.731	182.782	180.978	179.074	179.387

4.2 Regression Models for the Single-Type Accident Rate

The analysis in the previous section shows that *SINGLE* is considered to be an important variable affecting the occurrence of fatal accidents. However, unlike other explanatory variables, *SINGLE* represents the result of an accident and is not determined by the driver or the driving conditions. In this section *SINGLE* is analyzed using the regression models, where *SINGLE* is regressed in relation to the other explanatory variables. The weighted least squares (WLS) method, weighted by the registered numbers, is used for the estimation. As before, the model with all variables is estimated first. Then, models are estimated by eliminating variables with small absolute t-values. The expected signs are positive for *AGE24, NIGHT,* and *SPEED*, and negative for *BELT*, and one-sided tests are employed for these variables.

The results of the estimation are given in Table 4.2. *BELT* is negative and significant at the 1% level for all models, and a strong relationship

with *SINGLE* is suggested. *SPEED* and *NIGHT* are positive for all models, and they are significant at the 5% level for models with 6 or fewer explanatory variables. (*NIGHT* is significant at the 1% level for the model with 4 explanatory variables.) Therefore, one may conclude that factors such as not wearing a seat belt, driving over 40 km/hour, and night driving affect fatal accidents through *SINGLE*, which represents the "degree of violence and carelessness" of driving.

Table 4.2 Results of the Estimation of the Models for *SINGLE* (t-values in parentheses)

Number of Explanatory Variables	8	7	6	5	4
Constant	0.2368	0.2368	0.2315	0.2219	0.2126
	(2.307)	(2.368)	(2.439)	(2.412)	(2.394)
MALE	-0.0008				
	(-0.023)				
AGE24	0.0156	0.0142	0.1979	0.0128	
	(0.254)	(0.379)	(0.711)	(0.520)	
AGE65	0.2180	0.2194	0.2273	0.2367	0.2085
	(1.173)	(1.278)	(1.384)	(1.471)	(1.399)
BUSINESS	0.0522	0.0536	0.0579		
	(0.428)	(0.512)	(0.576)		
LEISURE	0.0369	0.0401			
	(0.163)	(0.224)			
BELT	-0.3501	-0.3506	-0.3490	-0.3305	-0.3261
	(-3.351)	(-3.529)	(-3.605)	(-3.674)	(-3.700)
NIGHT	0.2245	0.2261	0.2340	0.2290	0.2695
	(1.481)	(1.718)	(1.890)	(1.883)	(2.936)
SPEED	0.0345	0.0348	0.0383	0.0419	0.0373
	(1.122)	-1.2730	(1.749)	(2.026)	(2.030)
R^2	0.8390	0.8332	0.8253	0.8221	0.8200

5. TESTS OF THE POISSON DISTRIBUTION AND EXOGENEITY OF SINGLE

5.1 The Test of the Poisson Distribution

In this paper, the analysis is carried out with the Poisson distribution. However, the assumptions of the Poisson distribution may not be satisfied, and the mean and variance may not be equal. Several testing methods have been proposed. When the mean and variance are not equal, generalized Poisson methods based on the binomial and negative binomial distributions and Poisson empirical Bayes methods are used in the analysis. (For details, see Collings and Margolin(1985), Lee (1986), Cameron and Trivendi (1990), Consul and Famoye (1992), and Christiansen and Morris (1997).) In this paper, the method suggested by Cameron and Trivendi (1990) is used to test the Poisson distribution.

Let $\mu_i = EY_i, \sigma_i^2 = V(Y_i)$. For the Poisson distribution, $\sigma_i^2 = \mu_i$. Following Collings and Margolin (1985), let the null and alternative hypotheses be:

(6) $H_0 : \sigma_i^2 = \mu_i, \quad H_1 : \sigma_i^2 = \mu_i + \alpha\mu_i^2, \quad \alpha \neq 0.$

In this case the test statistic is given by:

(7) $T = [\sum(\frac{\hat{\mu}_i^2}{2})]^{-1/2}[\sum\frac{1}{2}\{(Y_i - \hat{\mu}_i)^2 - Y_i\}]$,

where $\hat{\mu}_i$ is the estimator of μ_i.

Under the null hypothesis, T asymptotically follows the standard normal distribution. The value of T is 0.545 for the model with 6 explanatory variables, which minimizes *AIC*. The null hypothesis is not rejected at the 5% level, and the difference from the Poisson distribution is not admitted.

5.2 Test of the Exogeneity of *SINGLE*

As mentioned, although *SINGLE* is a variable representing the "degree of violence and carelessness" of driving, it is the result of an accident and does not directly measure the "degree of violence and carelessness". As a result, *SINGLE* might be correlated with the error term of the equation, and it might not function as an exogenous variable. Therefore, the exogeneity of *SINGLE* is tested using Hausman's (1978) principle (for details, see Grogger (1990)) for the model with 6 explanatory variables. The null hypothesis is that *SINGLE* is an exogenous variable. The testing procedure is as follows.

i) (4) gives a linear regression model given by

(8) $Y_i \approx X_i'\beta + \varepsilon_i, \quad E\varepsilon_i = 0.$

Considering *SINGLE* as an endogenous variable, estimate (8) by the instrumental variable method. Let $\hat{\gamma}_{IV}$ and $\hat{V}(\hat{\gamma}_{IV})$ be the instrumental variable estimator of the coefficient of *SINGLE* and its estimated variance. (Note that $\hat{V}(\hat{\gamma}_{IV})$ must be obtained considering heteroskedasticity of the error term.) $\hat{\gamma}_{IV}$ is consistent under both the null and alternative.

ii) Calculate the test statistic h given by

(9) $h = (\hat{\gamma}_{IV} - \hat{\gamma}_{ML})^2 / [\hat{V}(\hat{\gamma}_{IV}) - \hat{V}(\hat{\gamma}_{ML})]$

where $\hat{\gamma}_{ML}$ and $\hat{V}(\hat{\gamma}_{ML})$ are maximum likelihood estimators of the coefficient of *SINGLE* and its estimated variances. h follows the χ^2 distribution with one degree of freedom under the null.

The values of estimates are $\hat{\gamma}_{IV}$ = 12.062, $\hat{V}(\hat{\gamma}_{IV})$ =11.664, $\hat{\gamma}_{ML}$ =10.595, and $\hat{V}(\hat{\gamma}_{ML})$ = 4.679. The value of h is 0.308 and the null is not rejected at the 5% level. Therefore, it is not necessary to treat *SINGLE* as an endogenous variable.

6. PROBLEMS WITH THE ANALYSIS

In this paper, the analysis was carried out using the Traffic Accident Integration Database. However, because of restrictions in the data there are two potential problems that may affect the analysis. One is the assumption of the linear probability model. The other problem is that the means of explanatory variables are obtained only from vehicles involved in casualty accidents. These two problems are considered theoretically in this section.

6.1 Linear Probability Model

The linear probability model is used in the analysis. However, this model is thought to be problematic. Here, the linear model is compared

with the probit model, where the error term u_{ij} follows the standard normal distribution, and the appropriateness of the model is considered. Let the distribution of u_{ij} be the standard normal distribution and the distributions of the explanatory variables be:

(10) $x_{ij}'\alpha \sim N(\mu_i, \sigma_i^2)$, $\mu_i = E(x_{ij})$, and $\sigma_i^2 = V(x_{ij}'\alpha)$.

This assumption makes it possible to treat the problem analytically. Although the actual explanatory variables are dummy variables, the analysis gives important information regarding the appropriateness of the model.

Under these assumptions, the probability of the occurrence of fatal accidents in the i-th type of cars, P_i, is given by,

$$\begin{aligned} P_i &= E_x E_u(Y_{ij} \mid x_{ij}) = E_x\{P(Y_{ij}=1 \mid x_{ij})\} \\ &= E_x\{\Phi(x_{ij}'\alpha)\} = \int_{-\infty}^{\infty} \Phi(x)\frac{1}{\sigma_i}\phi\{(x-\mu_i'\alpha)/\sigma_i\}dx \\ &= \Phi(\mu_i'\alpha/\sqrt{1+\sigma_i^2}), \end{aligned} \tag{11}$$

where ϕ and Φ are the density and distribution functions of the standard normal distribution, and E_u and E_x express the expected values with respect to u_{ij} and x_{ij}. Suppose that the variances are the same for all types of cars so that $\sigma_i^2 = \sigma^2$ for $i = 1,2,..,k$.

Since $\Phi(\mu_i'\alpha/\sqrt{1+\sigma^2}) = \Phi(\mu_i'\alpha^*)$ where $\alpha^* = \alpha/\sqrt{1+\sigma^2}$, it is possible to estimate the model only from the information provided by μ_i in this case. (All coefficients of the model become $1/\sqrt{1+\sigma^2}$ of the original coefficients.) Moreover, the probability of a fatal accident is about 0.3～1.0 per 10,000 vehicles, and the linear approximation of $\Phi(z)$ is very good in this range. The errors caused by the linear approximation are considered to be small. Therefore, although the estimated probabilities may not be reliable, the analysis in which the variables related to fatal accidents are determined can be carried out using the linear probability model.

Even if the variances are not a constant value, the same analysis can be carried out when the fluctuation is not very large compared to that of $\mu_i'\alpha$. Suppose that σ_i^2 is a function of $\mu_i'\alpha$. (In this data, the values of P_i are small and negative.) When σ_i^2 is a decreasing function of $|\mu_i'\alpha|$, P_i becomes an increasing function of $\mu_i'\alpha$. If σ_i^2 is an increasing function of $|\mu_i'\alpha|$, P_i becomes an increasing function of $\mu_i'\alpha$ unless the change is extreme.

For example, if $\sigma_i^2 = \gamma_2 \cdot |\mu_i'\alpha|$, $\gamma_2 > 0$, then $\mu_i'\alpha/\sqrt{1+\sigma_i^2} = \mu_i'\alpha/\sqrt{1+|\gamma_2 \cdot \mu_i'\alpha|}$. This means that P_i is an increasing function of $\mu_i'\alpha$. It is necessary that σ_i^2 increase by the order of $(\mu_i'\alpha)^2$ or higher so that P_i becomes a non-increasing function of $|\mu_i'\alpha|$.

If P_i is an increasing function of $\mu_i'\alpha$, the analysis can be carried out by the linear approximation of P_i as in the constant variance case. Moreover, even if other functional forms of σ_i^2 are considered, it is necessary for the variances to change dramatically so that the coefficients take opposite signs. In this study, only Sedan-A-Class cars are analyzed. It is therefore unlikely that there would be much change in the variance and covariance structures. Hence we may conclude that it is reasonable to use the linear probability model.

6.2 Effects of Sample Selection Biases on the Explanatory Variables

In this study, the explanatory variables were obtained only from vehicles involved in casualty accidents. As a result, we cannot deny the possibility that sample selection biases may have affected the analysis. In this section, the effects of sample selection biases are analyzed, assuming that the explanatory variables follow a multivariate normal distribution. Although the actual explanatory variables are dummy variables, the analysis provides important information regarding the effects of sample selection biases.

Let η_{ij} be a dummy variable such that $\eta_{ij} = 1$ if the j-th vehicle of the i-th type of car causes a casualty accident and $\eta_{ij} = 0$ otherwise. Suppose that

$$\eta_{ij}^* = \gamma_0 + \gamma_1 x_{1j} + x_{2j}'\gamma_2 + v_{ij}, \tag{12}$$

$$\eta_{ij} = \begin{cases} 1 & \text{if } \eta_{ij}^* > 0 \\ 0 & \text{if } \eta_{ij}^* \le 0, \end{cases}$$

where v_{ij} is a random variable that is independent of x_{ij} and follows the standard normal distribution, x_{i1j} is the first explanatory variable, x_{i2j} is a vector of other explanatory variables such that $x_{ij}' = (1, x_{i1j}, x_{i2j}')$, and γ_0, γ_1 and γ_2 are unknown parameters. Let $\mu_{i1} = Ex_{i1j}$ and $\mu_{i2} = Ex_{i2j}$. Then,

$$\eta_{ij}^* = \gamma_0^* + \gamma_1 x_{i1j}^* + x_{i2j}^{*\prime}\gamma_2 + v_{ij} = \gamma_0^* + w_{ij}, \tag{13}$$

where $\gamma_0^* = \gamma_0 + \gamma_1\mu_{i1} + \mu_{2j}'\gamma_2$, $x_{i1j}^* = x_{i1j} - \mu_{i1}$, $x_{i2j}^* = x_{i2j} - \mu_{i2}$, and $w_{ij} = \gamma_1 x_{i1j}^* + x_{i2j}^{*\prime}\gamma_2 + v_{ij}$.

Let $x_{ij}^{*\prime} = (x_{i1j}, x_{i2j}')$, the variance-covariance matrix for the i-th type of car be Ω_i, and $\gamma^{*\prime} = (\gamma_1, \gamma_2')$. The probability of a casualty accident is given by

$$P(\eta_{ij}=1) = E_x E_v(\eta_{ij}) = \Phi(\gamma_0^*/\sigma_{w,i}), \tag{14}$$

where $\sigma_{w,i}^2 = V(w_{ij}) = \gamma^{*\prime}\Omega_i\gamma^* + 1$.

Since x_{ij}^* follows the multivariate normal distribution, $x_{i2j}'\gamma_2$ is rewritten as

$$x_{i2j}'\gamma_2 = a_i x_{i1j} + \zeta_{ij}, \tag{15}$$

where ζ_{ij} is a random variable independent of x_{i1j} and follows a normal distribution with a mean of zero. Substituting (15) into (13), we get

$$\eta_{ij}^* = \gamma_0^* + \xi_{ij} + \upsilon_{ij}, \tag{16}$$

where $\xi_{ij} = (\gamma_1 + a_i)x_{i1j}^*$ and $\upsilon_{ij} = v_{ij} + \zeta_{ij}$.

Now, let

$$(17)\ \xi_{ij}^{*} = \begin{cases} \xi_{ij} & \text{if } \eta_{ij} = 1 \\ 0 & \text{if } \eta_{ij} = 0. \end{cases}$$

Then, we get,

$$E(\xi_{ij}^{*} \mid \upsilon_{ij}) = \int_{-(\gamma_0^{*}+\upsilon_{ij})}^{\infty} \frac{1}{\sqrt{2\pi}} \frac{\xi}{\sigma_{i,\xi}} \phi(\frac{\xi}{\sigma_{i,\xi}}) d\xi = \sigma_{i,\xi} \phi(-\frac{\gamma_0^{*}+\upsilon_{ij}}{\sigma_{i,\xi}}),$$

where $\sigma_{i,\xi}^2 = V(\xi_{ij})$. Let $\sigma_{i,\upsilon}^2 = V(\upsilon_{ij})$. Since $\sigma_{i,w}^2 = \sigma_{i,\xi}^2 + \sigma_{i,\upsilon}^2$,

$$E\xi_{ij}^{*} = \int_{-\infty}^{\infty} \sigma_{i,\xi} \phi(-\frac{\gamma_0^{*}+\upsilon}{\sigma_{i,\xi}}) \frac{1}{\sigma_{i,\upsilon}} \phi(\frac{\upsilon}{\sigma_{i,\upsilon}}) d\upsilon = \frac{\sigma_{i,\xi}^2}{\sigma_{i,w}} \phi(\frac{\gamma_0^{*}}{\sigma_{i,w}^2}).$$

Since $E(x_{i1j}^{*} \mid \eta_{ij} = 1) = (1/\gamma_1^{*}) E\xi_{ij}^{*} / P(\eta_{ij} = 1)$ and $V(\xi_{ij}) = \gamma_{1i}^{*2} V(x_{i1j})$,

$$(18)\quad E(x_{i1j}^{*} \mid \eta_{ij} = 1) = \gamma_{1i}^{*} \frac{\sigma_{i,1}^2}{\sigma_{i,w}} \lambda(\frac{\gamma_0^{*}}{\sigma_{i,w}}), \text{ and}$$

$$E(x_{i1j} \mid \eta_{ij} = 1) = \mu_{i1} + \gamma_{1i}^{*} \frac{\sigma_{i,1}^2}{\sigma_{i,w}} \lambda(\frac{\gamma_0^{*}}{\sigma_{i,w}}),$$

where $\gamma_{1i}^{*} = \gamma_1 + a_i$, $\sigma_{i,1}^2 = V(x_{i1j})$,and $\lambda(z) = \phi(z)/\Phi(z)$. Therefore, x_{i1j} has a sample selection bias given by

$$(19)\ b_{1i} = \gamma_{1i}^{*} \frac{\sigma_{i,1}^2}{\sigma_{i,w}} \lambda(\frac{\gamma_0^{*}}{\sigma_{i,w}}).$$

When the sample selection biases are the same for all types of cars, the analysis is not affected. The sample selection biases become a serious problem if they cause a change in the orders of expected values. Namely,

$$(20)\quad E(x_{i1j} \mid \eta_{ij} = 1) < E(x_{i1j} \mid \eta_{ij} = 1) \text{ despite } \mu_{i1} > \mu_{i1}.$$

The necessary conditions of (20) are: i) there are variables strongly related to casualty accidents, and ii) the variance-covariance structures of the explanatory variables are very different from each other according to the types of cars. It is necessary for us to assume extreme variance-covariance structures for the explanatory variables to be (20). Therefore, we may conclude that fatal accidents can be analyzed properly by the models and methods employed in this paper unless the variance-covariance structures are very special.

7. CONCLUSIONS

In this paper, fatal traffic accidents for Sedan-A-Class cars, whose displacement volumes are predominantly 1500 cc or less, were analyzed. For the analysis, the Traffic Accident Integrated Database, developed by the Institute of Traffic Accident Research and Data Analysis was used. The data for 28 types of cars without missing values were used. This database did not exist until recently in Japan, and this paper represents the first attempt to analyze fatal accidents by a regression-type model. Since the number of fatal accidents was small for some types of cars and the numbers of registered vehicles for each type of car were different, the Poisson regression model was employed and estimated by the maximum likelihood method.

Nine variables, including the sex and age of the driver, the purpose of driving, seatbelt usage, night driving, speed, and single-type accidents were considered in the analysis. The results of the estimation show that male drivers and drivers age 24 or younger reduce the probability of fatal accidents; however, leisure or pleasure use and single-type accidents increase the probability of fatal accidents. Although they are not significant at the 5% level, the coefficients for business use and night driving are positive. In the analysis of single-type accidents, seatbelt usage was shown to have a negative effect and night driving and driving over 40 km/hour were shown to have positive effects.

Note

1) A single-type accident is an accident that does not involve other vehicles, pedestrians, or trains. It is caused by just the accident vehicle and it could be prevented by careful driving in many cases.

Acknowledgments

The author would like to thank Michael McAleer, Hajime Wago, and Katsumi Matsuura for their helpful comments.

References

[1] Cameron, A. C., and P. K., Trivendi, 1990, "Regression-based Tests for Overdispersion in the Poisson Model," *Journal of Econometrics*, 46, 347-364.

[2] Cameron, A.C., and P.K. Trivedi, 1998, *Regression Analysis of Count Data*, Econometric Society Monographs No. 30, Cambridge University Press, New York.

[3] Christiansen, C. L., and C. N. Morris, 1997, "Hierarchical Poisson Regression modeling," *Journal of the American Statistical Association*, 92, 618-632.

[4] Collings, B. J., and Margolin B. H., 1985, "Testing Goodness of Fit for Poisson Assumption when Observations are not Identically Distributed," *Journal of the American Statistical Association*, 80, 411-418.

[5] Consul, P.C. and F. Famoye, 1992, Generalized Poisson Regression Model," *Communications in Statistics: Theory and Method*, 21,89-109.

[6] Department of Transportation, United Kingdom, 1992, *Cars: Make and Models: The Risk of Driver Injury and Car Accident Rates in Great Britain.*

[7] Grogger, J., 1990, "A Simple Test for Exogeneity in Probit, Logit, and Poisson Regression Models," *Economics Letters*, 33, 329-332.

[8] Gourieroux, C., A. Montfort, and A. Trognon, 1984, "Pseudo Maximum likelihood Methods: Applications to Poisson Models," *Econometrica*, 52, 701-720.

[9] Hausman, J., 1978, "Specification Tests in Econometrics," *Econometrica*, 46, 1251-1271.

[10] Lee, L., "Specification Test for Poisson Regression Models," 1986, *International Economic Review*, 27, 689-706.

[11] Winkelmann, R., 1997, *Econometric Analysis of Count Data* (Second Edition), Springer, New York.