Hacettepe Journal of Mathematics and Statistics
Volume 34 (2005), 55 – 66

APPLICATION OF THREE DISCRETE CHOICE MODELS TO MOTORCYCLE ACCIDENTS AND A COMPARISON OF THE RESULTS

Özge Uçar* and Hüseyin Tatlıdil*

Received 14 : 07 : 2004 : Accepted 04 : 03 : 2005

Abstract

In this study, the comparison of results obtained from the same data set of three alternative probability models known as the Linear Probability Model (LPM), Binary Logit Model (BLM) and Binary Probit Model (BPM) has been discussed. The use of these models is an accurate approach in the event of determining a categorical dependent variable with two levels. In order to throw light on the similarities and differences of the results, data recorded by The Department of Traffic Training and Research of the General Directorate of Security Affairs of Turkey has been examined, and also factors influencing the chance of drivers surviving or not surviving motorcycle accidents during 2002 have been determined.

Keywords: Linear probability model, Binary probit model, Binary logit model, Motorcycle accidents.

1. Introduction

For regression analysis, if the dependent variable is categorical, then the use of the Ordinary Least Square Estimation Technique (OLS) to obtain efficient parameter estimates is not acceptable since some assumptions required by OLS are violated. Since the expected value of the dependent variable (Y_i) is regarded as a probability [given in (2.1)] in all probability models, the relationship between the expected value of the dependent variable and the explanatory variables may not be linear in most cases. This relationship is expressed by an S-shaped curve. Additionally, the proposed probability value (i.e. the expected value of the dependent variable) may be smaller than zero or exceed 1 in OLS. In order to solve these problems, it has been suggested that some transformations should be applied to the proposed probability value. Therefore, with a view to establishing a linear relation between the expected value of the dependent variable and the unrestricted

*Hacettepe University, Department of Statistics, 06532 Beytepe, Ankara, Turkey.

explanatory variables, some functions - the so called link functions - have been developed. Logit models are obtained under the assumption that the link function is chosen as logit, whereas probit models are obtained when the inverse of the standard normal distribution function is chosen. In other words, the model's name is connected with the distribution assumed for the error terms. That is, while a logistic regression model is obtained under the assumption that the distribution of error terms is logistic, a probit model is obtained when the normal distribution is assumed for the error terms.

There are two main objectives of this study. The first is to reveal similarities or differences between the results obtained using LPM and other proposed alternative models in the context of an application. The second is to determine factors that increase or decrease the chance of survival of motorcyclists by examining the motorcycle accidents occurring in 2002 in Turkey using these models. The dependent variable is defined as a binary variable (Y_i), as follows.

$$Y_i = \begin{cases} 0 & \text{if the driver is dead} \\ 1 & \text{if the driver is alive.} \end{cases}$$

Here (Y_i) denotes the value of the dependent variable for observation i ($i = 1, 2, \ldots, n$). In this study, LPM, BPM and BLM have been applied to a real data set. Since all these models are probability models belonging to the family of discrete choice models, interpretations have been made based on the category coded as '1' for the dependent variable. The results have been assessed according to the signs of the parameter estimates and the odds-ratios, which can only be obtained using BLM. The methodology of models has been given as follows.

2. Methodology

The LPM was the first model to be developed under the assumption that the relationship between the explanatory variables and the dependent variable is linear, and preceded the BLM and BPM. The parameters of the LPM are estimated according to OLS, but it has been concluded that some of the assumptions of OLS are violated in LPM. For example, the changes in error variances constitute the main disadvantage of LPM due to the heteroscedasticity problem in the error variances. This heteroscedasticity can be revealed in the data set by calculating the error variance of each observation using the formulation in (2.4).

As a result of this problem, efficient parameter estimates cannot be obtained using LPM. The use of the Weighted Least Square Estimation Technique (WLS) has been proposed as a solution. If an individual data set is used in the analysis, the use of a two-step WLS is an appropriate approach. While unbiased parameter estimates are obtained in the first step, both unbiased and efficient parameter estimates can be obtained using the weighting variable, the values of which are calculated by the formulation given by (2.5) in the second step.

Even though this is an efficient way to solve this problem, it is not so easy to satisfy the other assumptions required by OLS, such as the normality of the dependent variable, the restriction on the probability associated with the categories of the dependent variable to lie in the interval $[0, 1]$, etc. Therefore models which use the Maximum Likelihood Estimation Technique (MLE), such as the BLM and the BPM, have been developed as an alternative to the LPM. Since MLE does not require an assumption about the distribution associated with the dependent variable, homogeneous error variances or a linearity assumption between the dependent variable and the explanatory variables, the use of these two models has gradually becomes widespread in a variety of research fields. However the choice between LPM, BPM and BLM is largely arbitrary, and the validity of the models depends solely on the data set used. Since the probability values associated with the categories of the dependent variable is in the range of $0 - -1$, and the linearity assumption of OLS is approximately satisfied when the values of the explanatory variables are restricted to a small range, in this case the results obtained from all three models can be expected to be similar. Otherwise, the results may not be fully consistent with one other.

As an application, a real data set described in Section 3 has been used to estimate the parameters of LPM, BLM and BPM. A comparisons of the results have been made in terms of the significance of the models and parameters, the signs of the parameters and the magnitudes of the estimated coefficients of the explanatory variables. Additionally, factors that lead to an increases or decreases in the chance of survival of motorcyclists have been determined.

In the following sub-sections the basic principles of LPM, BLM and BPM will be explained briefly.

2.1. The Linear Probability Model. If the dependent variable has two categories, it is expressed by a dummy variable coded as '0' and '1'. Linear regression applied to a binary dependent variable is called LPM. This model is the expected value of Y_i and is expressed as:

$$\begin{aligned} &Y_i = \sum b_k x_{ik} + u_i,\ i = 1, 2, \ldots, n, \\ &E(Y_i \backslash x_i) = (1)P(Y_i = 1) + (0)P(Y_i = 0) = \\ &\qquad = P(Y_i = 1) = \sum b_k x_{ik} = \Pi_i,\ i = 1, 2, \ldots, n, \end{aligned} \tag{2.1}$$

where Y_i denotes the value of the dependent variable (0 or 1) associated with observation i, the x_i are explanatory variables, the b_k parameter estimates and n the total number of observations in (2.1). The value Π_i in (2.1) is interpreted as the conditional probability of belonging to the category coded as '1' in the dependent variable according to the values of the explanatory variables of observation i.

As was mentioned in Section 2, the parameters of LPM are estimated according to OLS. Therefore, it will be useful to summarize the main assumptions of OLS, and mention proposed solutions to the problems arising when the assumptions are violated.

i. The Normal Distribution Problem

The Normal distribution is assumed for the dependent variable and error term in OLS. Since the dependent variable has only two values (0 or 1), it is not expected that the distribution of the dependent variable will be normal. Similarly, the error term can take only two values. These values are given by (2.2) and (2.3), and the error variance is given by (2.4) below.

If $Y_i = 1$ and the probability of observing the category '1' in the dependent variable is Π_i, then the error term can be expressed as,

$$\begin{aligned} u_i &= 1 - E(Y_i), \\ &= 1 - E(\alpha + \beta x_i), \\ &= 1 - \Pi_i, \end{aligned} \tag{2.2}$$

and if $Y_i = 0$ and the probability of observing the category '0' in the dependent variable is $1 - \Pi_i$, then the error term can be expressed as,

$$\begin{aligned} u_i &= 0 - E(Y_i), \\ &= 0 - E(\alpha + \beta x_i), \\ &= -\Pi_i, \end{aligned} \tag{2.3}$$

where u_i represents the error term associated with the observation i in (2.2), (2.3) and (2.4). It is clearly seen that error term associated with observation i can only take one of the two values given in (2.2) or (2.3) (i.e. $1 - \Pi_i$ or $-\Pi_i$). Therefore, the distribution of the error terms is not normal, as well.

The use of models developed on the basis of MLE, such as logit and probit, is proposed as a solution to this problem since it does not require the normal distribution assumption associated with the dependent variable and the error terms.

ii. The Heteroscedasticity Problem

Under OLS it is assumed that the error variances do not change from observation to observation. The error variances, which are calculated on the basis of the values of the error terms (2.2) and (2.3), are given by $V(u_i) = E(u_i^2) - [E(u_i)]^2$. Noting that

$$E(u_i) = (1 - \Pi_i)(\Pi_i) + (-\Pi_i)(1 - \Pi_i) = 0,$$

we obtain

$$\begin{aligned} V(u_i) &= E(u_i^2) \\ &= (1 - \Pi_i)^2(\Pi_i) + (-\Pi_i)^2(1 - \Pi_i) \qquad (2.4) \\ &= \Pi_i(1 - \Pi_i). \end{aligned}$$

The presence of the subscript i in (2.4) implies changes in the error variances from observation to observation. As a result, one of the main assumptions of OLS, namely that the variance be homogeneous, is violated.

Since the estimated parameters will not now be efficient, the use of WLS is proposed as a solution to this problem. Therefore, due to the fact that an individual data set has been used in the application part of the study, the methodology of a two-step WLS will be introduced in the following sub-section.

iii. The Nonlinearity Problem

The construction of OLS is based on the linearity assumption. This means that in LPM the relationship between the expected value of the dependent variable $P(Y_i = 1)$ and the explanatory variables has a linear form. However, the amount of increases or decreases in the values of the explanatory variable is not the same as the probability values. The actual relationship is defined with an S-shaped curve. In other words, the actual relationship is not linear but non-linear. Therefore, some appropriate transformations need to be applied to $P(Y_i = 1)$, such as logit and probit, that provide the linearity and also restrict the value of the probability to lie in the interval $[0, 1]$.

2.1.1. *A two-step weighted least square estimation technique in LPM.* In the first step, OLS is applied to the data set to obtain unbiased parameter estimates. Then weights are calculated for each observation using the formula in (2.5), and the weighting variable is constructed.

$$(2.5) \qquad W_i = \frac{1}{\left[\left(\sum b_k x_{ik}\right)\left(1 - \sum b_k x_{ik}\right)\right]^{1/2}} = \frac{1}{\left[\Pi_i(1 - \Pi_i)\right]^{1/2}}.$$

The expression in the denominator of (2.5) is the standard deviation of the error term u_i.

In the second step, WLS is applied to the data set taking the weighting variable into account. Consequently, both unbiased and efficient parameters are obtained. The general structure of the model estimated using WLS is given by (2.6).

$$(2.6) \qquad (W_i Y_i) = \sum (W_i b_k x_{ik}) + (W_i u_i).$$

Aldrich and Nelson [1] proved that the weighted errors are constant throughout the observations after the application of a two-step WLS.

2.2. The Binary Logit Model. The first model developed as an alternative to LPM was BLM. One of the basic problems with LPM results from the boundaries of the values assigned to the categories of the dependent variable. The fact that probabilities must lie in the unit interval $[0, 1]$ is a fundamental axiom of probability theory, but the value of $\sum b_k x_{ki}$ is unrestricted in LPM. It is suggested that the most suitable way to solve this problem is to apply a transformation to these values designed to obtain values in the unit interval which may be used as probabilities. While doing this, no restriction is needed on the explanatory variables and parameter estimates (the b_k 's). We can eliminate the upper bound, $P(Y_i = 1)$, by calculating the ratio $\frac{P(Y_i=1)}{1-P(Y_i=1)}$, due to the fact that as $P(Y_i = 1)$ approaches one, the ratio $\frac{P(Y_i=1)}{1-P(Y_i=1)}$ tends to infinity. Similarly, the most appropriate transformation to eliminate the lower boundary of zero is to take the natural logarithm of this ratio, as given by (2.7).

$$\log_e \left[\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right] = \sum b_k x_{ik}. \tag{2.7}$$

As $P(Y_i = 1)$ approaches to zero, the transformed value given in (2.7) tends towards minus infinity. Therefore, all values will lie in the interval $(-\infty, \infty)$, and as a result, the probability values will be restricted to the unit interval $[0, 1]$ (see [1], [10, 11]). This transformation applied to the probability value of LPM is called logit, and the model is called the 'logit model'. This model can be expressed in two different forms. The Logit and logistic regression expressions of the model are given by (2.8) and (2.9), respectively.

$$\log_e \left[\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right] = \sum b_k x_{ik}. \tag{2.8}$$

$P(Y_i = 1)$ is obtained from (2.8) by a straightforward calculation, as follows. Firstly (2.8) gives

$$\frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \exp\left(\sum b_k x_{ik}\right),$$

and solving this linear equation for $P(Y_i = 1)$ gives

$$P(Y_i = 1) = \frac{\exp\left(\sum b_k x_{ik}\right)}{1 + \exp\left(\sum b_k x_{ki}\right)}. \tag{2.9}$$

2.3. The Binary Probit Model. The second alternative model to LPM was suggested by Bliss [2], and is called BPM. It assumes that the distribution of the error terms is normal (see [1]). The model was first developed by Finney [3]. The general formulation of BPM is as follows:

$$\Phi^{-1}(\mu) = \sum_{k=1}^{K} b_k x_k, \tag{2.10}$$

where Φ denotes the standard normal cumulative distribution function, μ is the mean of the dependent variable, the x_k are explanatory variables and b_k the unknown parameters.

The standard normal distribution function is given as follows.

$$\Phi(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)du. \tag{2.11}$$

$$P(Y=1) = 1 - L\Big(-\sum_{k=1}^{K} b_k x_k\Big) = L\Big(\sum_{k=1}^{K} b_k x_k\Big) = \Phi\Big(\sum_{k=1}^{K} b_k x_k\Big). \tag{2.12}$$

In (2.12), L represents the general expression for the cumulative distribution function assumed for the error terms. If in particular the standard normal distribution function is chosen for L, the expression for the probability in BPM is given by the right-hand side of (2.12) (see [9]).

In this section the general methodology of LPM, BLM and BPM has been introduced. In the following section, a numerical example will be given associated with these models.

3. A Numerical Example

In this section, 1814 motorcycle accidents recorded in Turkey by The Department of Traffic Training and Research of the General Directorate of Security Affairs of Turkey in 2002 have been examined.

All explanatory variables included in the model are believed to have an effect on the dependent variable and 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 of the accident, the number of vehicles involved and the type of collision. The remaining two classes are related to characteristic of the road, including day-weather conditions and surface type, and the age of the vehicles, respectively.

Since all variables are coded in LIMDEP 7.0 using a technique known as *dummy variable coding*, the mean values of all variables must lie in the interval $[0,1]$ since LIMDEP 7.0 does not provide any parameter estimates unless this restriction is satisfied (see [4]). This has necessitated a small change to the continuous variable that represents the age of the vehicle. All values of this variable have been divided by the value 100 since, as explained in [5], this provides a mean value in the interval $[0,1]$.

In the study, the parameters of LPM and BLM have been estimated using SPSS 9.0 software, whereas BPM has been constructed using LIMDEP 7.0. The OLS results of LPM, which constitute the first step of WLS, are given in Table 1. The general structure of LPM (for OLS and WLS estimators) has been given by (2.1). The LPM model is expressed as follows in terms of the OLS results.

$$\begin{aligned} &P(Y_i = 1) = \\ &\quad 0.999 - 0.156\text{Superhighway} - 0.0762\text{ProvinceRoad} + \cdots + 0.0701\text{VehicleAge}. \end{aligned} \tag{3.1}$$

As mentioned before, these parameter estimates are unbiased but inefficient. In order to verify that the error variances associated with each accident differ from one another, the probabilities (Π_i) of belonging to the category '1' in the dependent variable for two accidents are calculated from (3.1) under the related accident characteristics. The error variances related to these two accidents are obtained using these probabilities by the formula given in (2.4). The first accident is characterized as follows:

Characteristics of the First Accident:

Avenue or Street
One-way traffic
Clear Weather
Urban area
Single Vehicle
Road surface, other
Age Group (18-25)
Age of Vehicle (0)
Crash Type, other
Daytime
Primary School

Table 1. OLS Results (The First Step of WLS)

Reference Category	Variable Name	Unstandardized Coeff.		Standardized Coeff.	Sig.
		B	Std. Error	Beta	
Avenue / Street	Superhighway	**-0.1560**	**0.055**	**-0.068**	**0.005**
	Province road	**-0.0762**	**0.016**	**-0.134**	**0.000**
	Other road	-0.0303	0.042	-0.017	0.466
Single Vehicle	Two vehicles (same direction)	-0.0037	0.027	-0.008	0.892
	Two vehicles (opp. direction)	0.0038	0.030	0.007	0.900
	Two vehicles (adj. direction)	0.0147	0.031	0.024	0.631
	> two vehicles	-0.0364	0.038	-0.027	0.341
Head-on Crash	Rear-end	0.0270	0.024	0.036	0.260
	One-side	0.0269	0.017	0.063	0.115
	Stationary obj.	-0.0028	0.039	-0.002	0.942
	Rollover	**0.0840**	**0.035**	**0.097**	**0.018**
	Other crash type	0.0583	0.031	0.128	0.059
Concrete	Asphalt	-0.0351	0.034	-0.049	0.306
	Other surface	-0.0232	0.038	-0.029	0.542
Daytime	Night	**-0.0336**	**0.011**	**-0.070**	**0.002**
	Dawn	-0.0343	0.027	-0.029	0.198
Clear Weather	Fog or rain	0.0070	0.026	0.006	0.786
	Other weather conditions	0.0139	0.016	0.019	0.394
Age Group (18-25)	26-35	0.0075	0.011	0.017	0.506
	36-45	-0.0113	0.015	-0.019	0.441
	46-55	-0.0054	0.019	-0.007	0.777
	56+	**-0.0689**	**0.024**	**-0.069**	**0.004**
Higher Edu.	Primary school	**-0.0462**	**0.020**	**-0.109**	**0.018**
	Secondary school	-0.0141	0.022	-0.024	0.530
	High school	-0.0261	0.022	-0.050	0.226
Urban area	Rural area	**-0.0988**	**0.022**	**-0.133**	**0.000**
One-way traffic	Two-way traffic	**0.0262**	**0.010**	**0.062**	**0.008**
	Vehicle Age	0.0701	0.064	0.025	0.272
Constant		**0.999**	**0.051**	-	**0.000**

Under these characteristics, the probability of the driver having survived the first accident is obtained as follows.

$$\begin{aligned}P(Y_1=1) &= \Pi_1 = 0.99-0.156(0)-0.0762(0)-0.0303(0)-0.0037(0)+0.0038(0)\\&+0.0147(0)-0.0364(0)+0.0270(0)+0.0269(0)-0.0028(0)+0.0840(0)+0.0583(0)\\&-0.0351(0)-0.0232(1)-0.0336(0)-0.0343(0)+0.0070(0)+0.0139(0)+0.0075(0)\\&-0.0113(0)-0.0054(0)-0.0689(0)-0.0462(1)-0.0141(0)-0.0261(0)-0.0988(0)\\&+0.0262(0)+0.0701(0)\approx 0.98\end{aligned}$$

The error variance of the first accident is calculated as follows.

$$V(u_1) = \Pi_1(1 - \Pi_1) = (0.98)(1 - 0.98) = 0.0196.$$

Characteristics of the Second Accident:

Province Road — Two-Vehicle (same direction) — Rear-End
One-way traffic — Asphalt — Night Time
Weather conditions, other — Age group (18-25) — Primary School
Urban area — Age of vehicle (0.02)

Under these characteristics, the probability of the driver having survived the second accident is obtained as follows.

$$\begin{aligned} P(Y_1 = 1) &= \Pi_1 = 0.99 - 0.156(0) - 0.0762(1) - 0.0303(0) - 0.0037(1) + 0.0038(0) \\ &+ 0.0147(0) - 0.0364(0) + 0.0270(1) + 0.0269(0) - 0.0028(0) + 0.0840(0) + 0.0583(0) \\ &- 0.0351(1) - 0.0232(0) - 0.0336(1) - 0.0343(0) + 0.0070(0) + 0.0139(1) + 0.0075(0) \\ &- 0.0113(0) - 0.0054(0) - 0.0689(0) - 0.0462(1) - 0.0141(0) - 0.0261(0) - 0.0988(0) \\ &+ 0.0262(0) + 0.0701(0) \approx 0.85 \end{aligned}$$

The error variance of the second accident is calculated as follows.

$$V(u_1) = \Pi_1(1 - \Pi_1) = (0.85)(1 - 0.85) = 0.1275.$$

The difference in the error variances between the two accidents is obvious. Therefore, in order to obtain both unbiased and efficient parameter estimates, a two-step WLS is used in LPM. Using the formulation given by (2.5) and the parameter estimates, weights for each individual are obtained. Then, the weighted dependent variable (W_iY_i) regresses on the weighted explanatory variables ($\sum w_i b_k x_{ik}$), and the two-step WLS results for LPM given in Table 2 are obtained. One level of all categorical explanatory variables (generally the first or the last level is chosen) must be determined as a reference category. Therefore, results are interpreted according to the reference categories. All results associated with three alternative models are given in Table 2.

The model formation of LPM is expressed in terms of the probabilities given in (3.2):

$$\begin{aligned} &P(Y_i = 1) = \\ &\quad 1.0170 - 0.1470\text{Superhighway} - 0.070\text{ProvinceRoad} + \cdots + 0.0586\text{VehicleAge} \end{aligned} \tag{3.2}$$

The model formation of BLM is expressed in terms of the probabilities given in (3.3):

$$P(Y_i = 1) = \frac{\exp(4.9674 - 1.8120\text{Superhighway} - 1.3945\text{ProvinceRoad} + \cdots + 1.2680\text{VehicleAge})}{1 + \exp(4.9674 - 1.8120\text{Superhighway} - 1.3945\text{ProvinceRoad} + \cdots + 1.2680\text{VehicleAge})} \tag{3.3}$$

The model formation of BPM is expressed in terms of the probabilities given in (3.4):

$$\begin{aligned} &P(Y_i = 1) = \\ &\quad \Phi(2.8979 - 0.9005\text{Superhighway} - 1.6722\text{ProvinceRoad} + \cdots + 0.8686\text{VehicleAge}) \end{aligned} \tag{3.4}$$

Before interpreting the results, the validity and goodness-of-fit of the models must be tested. While the validity of LPM is tested using the Regression Sum of the Squares (RSS), it is tested using the Likelihood Ratio approach [-2LLR (-2 Logaritmic Likelihood Ratio)] in BPM and BLM. Additionally, the goodness-of-fit of the models can be assessed by means of the Correct Classification Rate (CCR) (see [6]).

The results are given in Table 3.

Table 2. Parameter Estimates

Since the parameter estimates obtained from OLS are inefficient, interpretations using these parameters do not reflect reality and interpretations should therefore be made according to the parameter estimates shown below.

Reference Category	Variable Name	LPM	Prob.	BLM	Prob.	Exp (B)	BPM	Prob.
Avenue / Street	Superhighway	**-0.1470**	**0.034**	**-1.8120**	**0.018**	**0.1633**	**-0.9005**	**0.0362**
	Province road	**-0.070**	**0.000**	**-1.3945**	**0.000**	**0.2480**	**-0.6722**	**0.0000**
	Other road	-0.0158	0.584	-0.5715	0.502	0.5647	-0.2298	0.6132
Single Vehicle	Two vehicles (same direction)	**-0.032**	**0.016**	-0.1280	0.877	0.8799	-0.1529	0.6779
	Two vehicles (opp. direction)	-0.0247	0.116	0.0880	0.922	1.0920	-0.0656	0.8715
	Two vehicles (adj. direction)	-0.0175	0.268	0.3997	0.671	1.4914	0.1232	0.7735
	> two vehicles	-0.0527	0.080	-0.5167	0.590	0.5965	-0.3578	0.4294
Head-on Crash	Rear-end	0.0187	0.313	0.7258	0.168	2.0665	0.3506	0.1870
	One-side	0.0182	0.173	0.6202	0.116	1.8592	0.2813	0.1551
	Stationary obj.	-0.0347	0.267	0.1667	0.863	1.1814	-0.0108	0.9811
	Rollover	0.0217	0.275	1.9671	0.090	7.1496	0.8420	0.1058
	Other crash type	0.0054	0.767	1.6410	0.063	5.1606	0.6639	0.0982
Concrete	Asphalt	-0.0199	0.192	-1.0165	0.375	0.3618	-0.6061	0.3246
	Other surface	-0.0130	0.440	0.0416	0.978	1.0425	-0.2282	0.7527
Daytime	Night	**-0.0224**	**0.003**	**-0.7204**	**0.007**	**0.4866**	**-0.4094**	**0.0017**
	Dawn	-0.0247	0.205	-1.0167	0.079	0.3618	-0.5189	0.0699
Clear Weather	Fog or rain	-0.0006	0.972	0.2997	0.622	1.3494	0.1134	0.7066
	Other weather conditions	0.0094	0.245	0.5830	0.272	1.7913	0.3430	0.1895
Age Group (18-25)	26-35	0.0007	0.905	0.2207	0.509	1.2469	0.1024	0.5207
	36-45	-0.0094	0.294	-0.2290	0.539	0.7953	-0.1511	0.4041
	46-55	-0.0007	0.950	-0.1403	0.764	0.8691	-0.0719	0.7573
	56+	**-0.0561**	**0.019**	-0.8535	0.067	0.4259	-0.4539	0.0601
Higher Edu.	Primary school	**-0.0261**	**0.005**	-1.4650	0.058	0.2311	**-0.7889**	**0.0338**
	Secondary school	-0.0105	0.297	-0.4570	0.610	0.6332	-0.3273	0.4357
	High school	-0.0175	0.073	-0.8908	0.282	0.4103	-0.5017	0.2078
Urban area	Rural area	**-0.0999**	**0.000**	**-1.1148**	**0.001**	**0.3280**	**-0.6294**	**0.0006**
One-way traffic	Two-way traffic	**0.0176**	**0.002**	**0.7259**	**0.009**	**2.0666**	**0.3979**	**0.0034**
	Vehicle Age	0.0586	0.095	1.2680	0.460	3.5537	0.8686	0.3178
Constant		**1.0170**	**0.000**	**4.9674**	**0.002**	-	**2.8979**	**0.0003**

Table 3. Validity Tests of the Models

LPM		BLM		BPM	
RSS	27.015	-2LLR	154.678	-2LLR	159.6179
Significance	0.000	Significance	0.000	Significance	0.000

The fact that the significance values of each model (0.00 for each model) are less than the 0.05 critical point indicates the validity of the models. The CCR of BLM and BPM are 95.09% and 95.26%, respectively. These ratios imply a high degree of quality of the models in assigning the observations into the correct group of the dependent variable.

When the model results in Table 3 are carefully examined, it is concluded that the factors having an effect on the dependent variable do not considerably differ in LPM, BLM and BPM. This conclusion indicates that each alternative model may be used for the same objective by examining the structure of the data set. Since LPM is a well-known simple linear regression model that is applied to binary dependent variable, the application of this model and its interpretation are easier than for BLM and BPM. Therefore, researchers with a limited knowledge of probability models commonly prefer LPM to BLM or BPM. However, BLM especially has some advantages in terms of the richness of interpretation, such as the odds-ratio interpretation, compared to LPM and BPM.

The odds-ratios are given under the heading 'Exp (B)' in Table 2, and significant factors were emphasized by being given in bold.

In the following sub-section, interpretations in terms of the signs of the coefficients and the odds-ratios will be assessed.

3.1. Interpretations of the Coefficients and the Odds-Ratios. The signs of the parameter estimates indicate the direction of change in the probability of belonging to the category coded as '1' in the dependent variable. All interpretations are assessed on the basis of reference categories of the categorical explanatory variables. While positive parameter estimates imply an increase in the probability of belonging to the category coded '1', negative parameter estimates indicate a decrease. Only accidents occurring on a two-way road lead to an increase in the probability of the driver surviving an accident. All other coefficients are negative and lead to decreases in this probability (see [7] and [8]).

Compared to an avenue or street, having an accident on a super highway or province road leads to a decreases in the probability of a positive response coded as '1'. This is a result that is valid for each alternative model. Super highways lead to a greater decrease in the probability of the driver surviving than do province roads. Signs of the parameters give information about the direction of the changes in the probability, whereas in BLM, odds-ratios give information about the quantity of this change relative to the reference category. By examining odds-ratios associated with super highways and province roads, it is concluded that while having an accident on an avenue or street is about 6 $\left(\frac{1}{0.1663}\right)$ times safer than for a super highway, a province road is about 4 $\left(\frac{1}{0.2480}\right)$ times safer for drivers.

Similarly, two vehicle crashes (in the same direction) decrease the chance of a drivers survival compared to single vehicle crashes in LPM. However, this factor is not significant in BLM and BPM at a 5% significance level. It is also concluded that having an accident at night in a rural area decreases the probability of a driver's survival in LPM, BLM and BPM. Odds-ratios obtained from BLM indicate that while having an accident at night is about 2 $\left(\frac{1}{0.4866}\right))$ times more risky than during the daytime, being in an urban area is about 3 $\left(\frac{1}{0.328}\right)$ times safer than being in a rural area.

Other factors that decrease the chance of a driver's survival are being in the 56+ age group, and only having a primary school education. The 56+ age group is a significant factor only in LPM, whereas having only a primary school education is significant in both LPM and BPM. Compared to having a high school education, drivers with a primary school education run about a 4 $\left(\frac{1}{0.2311}\right)$ times greater risk.

The last factor that is significant in LPM, BLM and BPM is associated with the traffic flow on the road. Since the coefficient of 'two-way' is positive, this factor increases the chance of survival of the drivers. From odds-ratio interpretations of BLM, it is concluded that accidents occurring on a two-way road are about 2 times safer than a one-way road for drivers. Other factors have no effect on the dependent variable.

The fundamental property of probability models is that it is possible to estimate the probability values of belonging to the various categories of the dependent variable for all observations based on the explanatory variable values. These probabilities may be obtained using LPM, BLM and BPM by means of model formulations given by (3.2), (3.3) and (3.4), respectively. If the restrictions related to the data set and the assumptions of OLS are satisfied, probabilities and other interpretations obtained using LPM will be close to the results from BLM and BPM.

4. Conclusion

LPM, BLM and BPM are well-known probability models developed as alternatives to each other. Since the choice of models is arbitrary, the most appropriate model that reflects the correct structure of the data set must be selected. LPM was developed first, and due to the simplicity of interpretations, it was used until BLM and BPM were proposed. Actually, the use of LPM may not be an appropriate approach in most cases, due to some assumption violations of OLS in connected with the data set. Therefore, WLS is applied to the data set with a view to eliminating some of these violations. Hence, the results from LPM approach the results of BPM and BLM after the use of WLS. As a result, if it is possible to obtain similar results using LPM, BLM and BPM, the use of LPM is an acceptable approach in view of the the simplicity of application and interpretation. However, if a researcher wants to interpret the results in detail, BLM must be preferred due to the presence of odds-ratio interpretations.

In our study, we examined motorcycle accidents occurring in 2002 in Turkey, both to reveal significant factors that have an effect on the dependent variable and to examine whether there are considerable differences in model results obtained from the three alternative models. Firstly, we concluded that results obtained using LPM, BLM and BPM were almost the same in terms of the signs of the coefficients for the same data set. Secondly, the general profile of drivers who have a chance of survival after the accident was determined by reference to the significant factors that are simultaneously significant in LPM, BLM and BPM. As a result, it was determined that drivers who prefer to travel on a two-directed road, in an urban area, during the daytime and on an avenue or street have a better chance of survival in motorcycle accidents. One essential point to be emphasized is that all these findings are acceptable only for the data set used in this study, due to the fact that the model choice is arbitrary. When another data set is used, considerably different results may be obtained using LPM, BLM and BPM. Therefore, studies should be continued on the basis of appropriate selection criterion for these models.

References

[1] Aldrich, J. H. and Nelson, F. D. *Linear probability, logit and probit models*, (Sage Publications, London, 07-045, 1984).

[2] Bliss, C. I. *The method of probits*, Science **79**, 409–410, 1934.

[3] Finney, D. J. *Probit analysis*, (Cambridge: Cambridge University Press, 1971).

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

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

[6] Hosmer, D. W. and Lemeshow, S. *Applied logistic regression*, (John Wiley and Sons, New York, 1989).

[7] 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).

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

[9] Liao, T. F. *Interpreting probability models (Logit, probit and other generalized linear models)*, (Sage Publications, Thousand Oaks, London, 07-101, 1994).

[10] Menard, S. *Applied logistic regression analysis*, 2. Edition, (Sage University Papers, Thousand Oaks, London, 07- 106, 2002).

[11] Powers, D. A. and Xie, Y. *Statistical methods for categorical data analysis*, (Academic Press, 2000).