Predicting the Severity of Median-Related Crashes in Pennsylvania by Using Logistic Regression

Eric T. Donnell and John M. Mason, Jr.

Models of median-related crash severity were developed by using roadway inventory and crash records for Pennsylvania Interstate highways. Cross-median and median barrier crashes formed the sample of crash types considered. Data were collected to model crash severity, including cross-section, traffic volume, and environmental predictor variables. Logistic regression models were developed by using both an ordinal and a nominal response. The results indicate that modeling crash severity as an ordinal response provided appropriate results for cross-median crashes, whereas a nominal response was more appropriate for median barrier crashes. Explanatory variables such as pavement surface conditions, use of drugs or alcohol, presence of an interchange entrance ramp, horizontal alignment, crash type, and average daily traffic volumes affect crash severity. The analysis results may be used by practitioners to understand the trade-off between geometric design decisions and median-related crash severity. Approximately 0.7% median barrier crashes on the Interstate system resulted in a fatality, whereas 43% were property-damage-only crashes and about 56% were injury crashes. More than 17% of cross-median collisions were fatal, and 67% involved injury.

Each year nearly 40,000 motorists in the United States die from crashes on the highway or street network (*1*). This number of fatalities has remained unchanged for more than a decade. The highest-level design criteria are reserved for Interstate highways. Implicit in the high mobility design criteria for Interstate highways is safety; nonetheless, there are more than 5,000 fatalities per year on Interstates (*1*).

The total economic loss due to reportable traffic crashes in Pennsylvania was nearly $12 billion per year from 1994 through 1998 (*2*). In Pennsylvania, an average of 125 motorists were killed per year on the Interstate highway system during this period. Nearly half these fatalities were median-related crashes, in which motorists left the travel way to the left and entered the median. A detailed analysis of median-related crash severity will help improve safety on Pennsylvania Interstate highways. Further, crash modeling will help in the safety decision-making process.

This paper describes the methodology that was used to develop crash severity models, by using logistic regression, for cross-median and median barrier crashes in Pennsylvania. Included is a discussion about the data used for modeling, descriptive safety measures of crash severity, and a procedure used to compute the probability (odds) of a fatal, injury, or property-damage-only (PDO) crash. The significance of this work is to test an ordinal response for crash severity prediction and to determine which geometric design and environmental variables best explain severity of median-involved crashes.

E. T. Donnell, BMI-SG, P.O. Box 154, 5230 Hearthwood Lane, Alexandria, PA 16611-0154. J. M. Mason, Jr., Pennsylvania State University, 101 Hammond Building, University Park, PA 16802.

Transportation Research Record: Journal of the Transportation Research Board, No. 1897, TRB, National Research Council, Washington, D.C., 2004, pp. 55–63.

LITERATURE REVIEW

Various statistical approaches have been used to model crash severity as a function of roadway, roadside, operational, environmental, and other explanatory variables. Disaggregate models, such as logistic regression, can test a variety of factors that contribute to crash severity.

Chang and Mannering used nested logit models to study the relationship between injury severity and truck and nontruck crashes (*3*). A nested logit model divides crash severity into a hierarchy of levels to provide distinction between severity classifications. By assuming that vehicle occupancy and crash severity were generalized extreme value distributed, the severity probabilities were nested by vehicle occupancy and level of injury sustained by the most severely injured vehicle occupant. Each of the nests was estimated by using multinomial logit models. For nontruck crashes, roadway, driver, vehicle, environmental, and crash characteristics all influenced the injury severity level, regardless of occupancy levels. High speed and turning movements were found to be the most influential predictors of truck-related crash severity.

Modeling severity as a discrete outcome involves estimating the probability that a vehicular crash has a certain severity by determining the likelihood of outcomes given that a crash has occurred. Lee and Chang estimated the severity of run-off-road crashes in the state of Washington, again by using the nested logit model (*4*). Temporal, environmental, driver, roadway, and roadside characteristics were used to estimate property damage and possible injury probabilities for rural run-off-road crashes conditioned on no evident injury. The findings indicated that wet pavement surfaces resulted in possible injury, drivers younger than 25 were more likely to be involved in injury crashes, alcohol-impaired drivers were more likely to be involved in injury crashes, and crashes in the presence of a horizontal curve were more likely to involve an injury.

Other severity models have been developed by using logistic regression models (*5–8*). These models were used to investigate the injury severity of head-on highway crashes and the severity of young-driver crashes. In the case of young-driver severity, a sequential

approach was taken to compare binary response levels (i.e., no injury and possible injury; possible injury and nonincapacitating injury; nonincapacitating injury and incapacitating injury; incapacitating injury and fatality). In the binary response case, the logistic regression model takes the following form (*6*):

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \alpha + \beta' X_i \quad (1)$$

where

p_i = prob($y_i = y_1 | X_i$) is response probability, and y_1 is first ordered level of y;
α = intercept parameter;
β' = vector of coefficients to be estimated; and
X_i = vector of independent variables.

Factors that prove most influential in predicting severity in young-driver crashes included influence of alcohol or drugs, ejection in the crash, point of impact, crash location, existence of horizontal curve or vertical grades at the crash site, speed of the vehicle, and restraint device usage (*6*).

Kim et al. used log-linear models to predict automobile crash and injury severity (*7*). The results suggested that alcohol or drug use and lack of seat belt use increase the odds of more severe crashes and injuries. Significant relationships were found between driver age and average daily traffic (ADT), injury severity, collision type, vehicle speed, alcohol involvement, and roadway characteristics in another study that used log-linear models (*8*). Log-linear models are useful for study of the association among categorical variables; however, logistic regression is more appropriate when a response variable is used to measure the direct effects on a set of independent variables.

Existing literature indicates that logistic regression has been frequently used to model crash severity. Explanatory variables such as geometric design elements, traffic operational measures, environmental conditions, and safety restraint use have been used to estimate severity. Past modeling efforts focused on two different modeling structures: the binary response and a nested model. Few predictive safety models have been developed to model crash severity by using an ordinal response. An ordinal response variable is considered when categories can be ranked from low to high, but the spacing between the categories is unspecified.

STUDY METHODOLOGY

This research was done to develop median-related crash severity models for Interstate highways in Pennsylvania. Models for cross-median and median barrier crashes were developed. These models use logistic regression to determine the probability of fatal, injury, or PDO crashes given various geometric, traffic operations, and environmental conditions. Such a modeling technique uses the selected independent variables to predict the probability that the dependent variable (crash severity) is of an ordinal or nominal scale (*9*, p. 157). Regression parameters are estimated by using maximum likelihood methods. The SAS LOGISTIC procedure is used to estimate the model and assess goodness of fit (*10*). Severity models were developed for the Pennsylvania Interstate highway system for sections with and without longitudinal median barriers. All crash severity models were first developed with an ordinal response variable. If the proportional odds assumption was violated, the models were reestimated with a nominal response by using the SAS CATMOD procedure (*10*).

All electronic data were provided by the Pennsylvania Department of Transportation (PennDOT). Additionally, field data were collected to supplement the electronic data. Two primary sources of electronic information were available for creating data sets that could be used to model crash severity: roadway inventory data from PennDOT's roadway management system (RMS) and crash data from PennDOT's crash reporting system (CRS). The total length of the Interstate highway system in Pennsylvania is 2,556 mi, excluding toll roads. Approximately 29% (738 mi) contain fixed longitudinal median barriers. About 12% (318 mi) of divided Interstate highways in Pennsylvania contain nontraversable median side slopes with no barrier. More than 48% (1,238 mi) of divided Interstate highways have traversable median side slopes with no barrier. The remaining sections of divided Interstate highway have independent alignments with a natural barrier separating opposing directions of travel.

Table 1 shows the roadway inventory, environmental, and traffic operational data that were available for modeling median-related crash severity. Because this research was part of an effort to evaluate median design policies in Pennsylvania, driver characteristics (age, gender, etc.) were not included in the analysis. Only geometric design and environmental variables that could be acquired from PennDOT's RMS and CRS were included in the analysis. As indicated, 13 explanatory variables were considered for estimating crash severity. Crash severity was initially considered ordered with three levels—fatal, injury, and PDO. Pennsylvania categorizes a fatal crash as one in which a person involved in the crash dies within 30 days after the crash for reasons attributed to the crash. An injury crash is one in which no one involved was killed but at least one person was injured. A PDO crash is a reportable crash in which no one was injured or killed but vehicle towing was required (*11*).

ADT, median width, and percentage of heavy vehicles in the traffic stream were the continuous variables available in the electronic data. The remaining 10 explanatory variables were categorical. With the exception of the curve, interchange ramp, and median cross-slope indicators, all categorical explanatory variables were generated from the electronic crash data. The curve, interchange, and slope data were gathered from video photolog review or field measurements.

The distribution of crash severity is shown in Table 2. Between 1994 and 1998, there were 138 cross-median collisions (CMCs) and 4,416 median barrier crashes on the Pennsylvania Interstate highway system. More than 17% of CMC crashes were fatal, and 67% involved injury. Less than 1% of median barrier crashes were fatal, and 56% involved injury. It is clear that CMC crashes are much more severe events than median barrier or all crash types. Median barrier crashes are less severe than all crash types combined.

REGRESSION THEORY

When the absolute distance between categories of a variable is unknown, yet there is a clear ordering of the categories, the variable is considered ordinal. Logistic regression analysis is concerned with models whose outcome variables are discrete. In the case of crash severity models, the ordinal response categories are fatality, injury, and PDO crashes.

An ordinal logistic regression model is derived from a measurement model in which a latent variable y' is mapped to an observed

TABLE 1 Interstate Crash Severity Data

Variable Name	Variable Description	Variable Type	Range of Values
Severity (Response)	Cross-median collision severity	Categorical	1 = Fatality 2 = Injury 3 = Property-damage only
Illumination (Predictor)	Indicator of day vs. nighttime driving conditions	Categorical	1 = Daylight 2 = Dark, but lighted 3 = Dark
Weather (Predictor)	Indicator of ambient weather conditions	Categorical	1 = No adverse conditions 2 = Rain 3 = Snow/icy
Surface (Predictor)	Indicator of roadway surface condition	Categorical	1 = Dry 2 = Wet/icy
Drugs (Predictor)	Indicator of drug use by driver	Categorical	1 = No drugs 2 = Driver using drugs or alcohol
ADT (Predictor)	Average Daily Traffic volume at crash site (vehicles/day)	Continuous	Range: 5,035 to 80,597
MedWid (Predictor)	Median Width at crash site (feet)	Continuous	Range: 6 to 131
Curve (Predictor)	Indicator of horizontal curve direction at crash site	Categorical	1 = No curve 2 = Curve to the right 3 = Curve to the left
Interchange (Predictor)	Indicator of interchange entrance ramp within 1,500 feet of crash site	Categorical	1 = No ramp 2 = Ramp within 1,500 feet of crash site
Slope (Predictor)	Indicator of median cross-slope at crash site	Categorical	1 = Flatter than 6:1 2 = Steeper than 6:1
Speed (Predictor)	Speed limit (miles per hour)	Continuous	Range: 40 to 65
Type (Predictor)	Indicator of collision type	Categorical	1 = Rear-end 2 = Head-on 3 = Angle 4 = Sideswipe 5 = Single Vehicle 6 = Other
Cause (Predictor)	Indicator of crash cause	Categorical	1 = Following too closely 2 = Driving too fast 3 = Driver maneuver error 4 = Driver condition 5 = Poor/wet pavement 6 = Other
Trucks (Predictor)	Proportion of trucks in traffic stream	Continuous	Range: 6 to 43

1 ft = 0.3048 m.

TABLE 2 Cross-Median and Median Barrier Crash Severity

Severity Level	Frequency (%) of Crashes	
	Number of Crashes	Percent
Cross-Median Collisions		
Fatal	24	17.4
Injury	93	67.4
Property damage only	21	15.2
Total	138	100.0
Median Barrier Collisions		
Fatal	31	0.7
Injury	2,471	56.0
Property damage only	1,914	43.3
Total	4,416	100.0
All Crash Types Combined		
Fatal	412	1.3
Injury	15,827	51.7
Property damage only	14,415	47.0
Total	30,654	100.0

variable *y*. These variables are related according to the following equation (*12*):

$$y_i = m \quad \text{if } \tau_{m-1} \le y_i' < \tau_m \quad \text{for } m = 1 \text{ to } J \tag{2}$$

The τ are cutpoints on the measurement scale that are used to distinguish the ordinal categories, as shown in Figure 1. Category 1 (e.g., fatality) is defined by the open-ended interval on the lower end of the measurement scale; Category 3, or *J* (e.g., PDO), is defined as the portion of the scale above cutpoint τ_2. Category 2 (e.g., injury) is the portion in Figure 1 between the cutpoints.

The regression equation used for an ordinal response is $y_i' = x_i\beta + \epsilon_i$. The error term has a logistic distribution with mean zero and a variance of $\pi^2/3$. By using the SAS LOGISTIC procedure, the cutpoints are estimated, and the software assumes that $\beta_0 = 0$ (*10*). Interpretation of ordinal response variables can be performed according to odds ratios. In this analysis, the proportional odds model is used to interpret odds ratios for cumulative probabilities. The cumulative probability that the outcome is less than or equal to *m* is (*13*)

$$\Pr(y \le m|x) = \sum_{j=1}^{m} \Pr(y = j|x) \qquad \text{for } m = 1, \ldots, J-1 \tag{3}$$

The odds that an outcome is *m* or less versus greater than *m* given a set of explanatory variables *x* are (*12*)

$$\Omega_m(x) = \frac{\Pr(y \le m|x)}{1 - \Pr(y \le m|x)} = \frac{\Pr(y \le m|x)}{\Pr(y > m|x)} = \exp(\tau_m - x\beta) \tag{4}$$

For this paper, the odds of being in a fatal crash (e.g., $m \le 1$) could be compared to injury or PDO severity crashes. The set of explanatory variables in these severity models may be either continuous or categorical. The general form of the ordinal logistic regression model is (*13*)

$$L_m(x) = \alpha_m + \beta' x, m = 1, \ldots, J-1 \tag{5}$$

Since the proportional odds model assumes that the odds ratio for all values of *m* are the same, a parallel regression assumption must be tested. For instance, does a change in the weather have the same effect on the odds of a fatal crash versus an injury or PDO crash? A score test is used to test a set of $J-1$ binary logits. The constraint that the β_m are equal for the $J-1$ regressions is tested. The score test evaluates how the log-likelihood changes if the constraint is removed.

Fisher scoring algorithms are used to fit ordinal logistic regression models. Model fit assessment tools include the Akaike information criterion and –2 log-likelihood. The Fisher scoring algorithm and –2 log-likelihood are used to test the global null hypothesis that all the parameters associated with covariates are zero. Maximum likelihood was used for parameter estimation in the model.

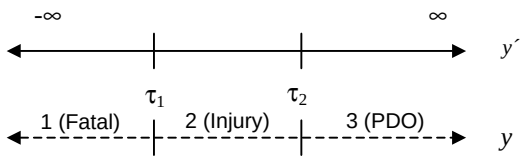

FIGURE 1 Ordinal measurement scale (*11*).

Interpreting the proportional odds model for three response variable categories implies that the odds ratios for a crash that is fatal versus an injury or a PDO crash and for a fatal crash or an injury crash versus a PDO crash are the same. Positive values for the intercept imply that the predicted probability of a fatality, as well as the cumulative probability of a fatal or an injury crash, is higher for higher values of the explanatory variables. When the intercept is negative, the predicted probability of a fatality, as well as the cumulative probability of a fatal or injury crash, is lower for higher values of the explanatory variables.

MODELING APPROACH AND RESULTS

A bivariate analysis of each variable was performed to examine the effect of each explanatory variable (Table 1) on CMC crash severity. Point estimates and odds ratios were reviewed to identify undesirable logistic regression variables (*14*). Point estimates for one of the odds ratios equal to zero or infinity may yield undesirable logistic regression results if included in a multivariate model. Table 3 shows the results of the bivariate analysis, including the variable name, likelihood ratio chi-squared test with $k-1$ degrees of freedom (*k* is number of levels of independent variable), and *p*-values for each level independent variable. The individual odds ratios are also shown. Review of each individual explanatory variable and its corresponding effect on crash severity reveals that the cause, drugs, curve, weather, and slope predictors have the greatest influence.

To develop a final ordinal logistic regression model of CMC crash severity, a stepwise procedure was used with all explanatory variables considered. Interaction terms were also included in the procedure—the interaction terms included the drugs predictor as well as the weather predictor. For example, drugs–weather, drugs–cause, drugs–curve, and drugs–slope were analyzed in the stepwise procedure. The PROC LOGISTIC statement was used with a significance level of 0.10 to retain variables in the model. The final ordinal logistic regression results are shown in Table 4.

Interpretation of the final CMC crash severity model in Table 4 indicates that all the predictor variables are statistically significant (*p*-value less than 0.05). The score test for the proportional odds assumption has a *p*-value of 0.0651 (5 degrees of freedom), which indicates that the proportional odds model adequately fits the data because the hypothesis that the regression lines for cumulative logits are parallel is not rejected. The likelihood ratio test *p*-value of 0.0008 (5 degrees of freedom) indicates that the null hypothesis is rejected, and the conclusion is that the predictor variables given in the model affect the severity of CMC crashes, or the model with independent variables is statistically better than the model with only the intercept.

The odds ratio is used to quantify the effect of significant independent variables on the dependent variable. The odds ratio is simply exp(parameter estimate) and can be used to explain the relative effects of a unit change in the variable on the severity of a crash. The relative effect of a driver not under the influence of drugs versus a driver under the influence of drugs is $\exp(0.6552) = 1.926$. This indicates that the odds of a crash severity of 3 (PDO) versus crash severity of 1 or 2 (fatality or injury) are 1.926 times higher for drivers who are not using drugs than for drivers who are using drugs. Alternatively stated, the odds of severity of 1 (fatality) versus severity of 2 or 3 (injury or PDO) increase by 93 when the driver is under the

TABLE 3 Bivariate Logistic Regression Analysis Results for CMC Crashes

Variable	Degrees of Freedom	Likelihood Ratio Chi-square (p-value)	Odds Ratio Estimates		
			Effect	Point Estimate	95% Wald Confidence Limits
Illumination	2	0.1469 (0.9292)	Illumination 1 vs. 3	0.935	(0.407, 2.150)
			Illumination 2 vs. 3	1.171	(0.307, 4.471)
Weather	2	1.8912 (0.3884)	Weather 1 vs. 3	0.652	(0.245, 1.676)
			Weather 2 vs. 3	1.110	(0.393, 3.131)
Surface	1	1.0664 (0.3018)	Surface 1 vs. 2	0.691	(0.342, 1.397)
Cause	5	7.0573 (0.2164)	Cause 1 vs. 6	2.404	(0.090, 64.072)
			Cause 2 vs. 6	0.859	(0.042, 17.524)
			Cause 3 vs. 6	1.246	(0.061, 25.405)
			Cause 4 vs. 6	0.446	(0.019, 10.310)
			Cause 5 vs. 6	3.236	(0.127, 82.663)
Drugs	1	4.9659 (0.0259)	Drugs 1 vs. 2	3.924	(1.295, 11.891)
ADT	1	0.0078 (0.9296)	Coefficent = 7.27E-7	SE = 8.216E-6	
Speed	1	0.0004 (0.9849)	Coefficent = -0.000665	SE = 0.0350	
MedWid	1	0.4798 (0.4885)	Coefficent = -0.00583	SE = 0.00828	
Curve	2	2.6735 (0.2627)	Curve 1 vs. 3	2.149	(0.865, 5.339)
			Curve 2 vs. 3	1.344	(0.592, 3.052)
Interchange	1	0.0692 (0.7925)	Interchange 1 vs. 2	0.908	(0.440, 1.872)
Slope	1	1.4873 (0.2226)	Slope 1 vs. 2	0.514	(0.179, 1.470)
Type	3	1.2004 (0.7529)	Type 1 vs. 4	1.373	(0.437, 4.313)
			Type 2 vs. 4	0.731	(0.319, 1.678)
			Type 3 vs. 4	1.059	(0.361, 3.112)
Trucks	1	0.1658 (0.6839)	Coefficent = 0.00738	SE = 0.0182	

influence of drugs or alcohol. Similarly, the odds of crash severity 3 (PDO) versus crash severity 1 or 2 (fatality or injury) are 3.933 times higher for tangent sections of roadway versus roadway sections that are curved to the left. Or, the odds of severity 1 (fatality) versus severity 2 or 3 (injury or PDO) increase by 393 when people drive on sections of roadway curved to the left versus those that are tangent. The odds of a PDO crash versus an injury or fatal crash on sections of roadway curved to the right are about one-third (33.6%) as high as those occurring on sections curved to the left. This indicates that the odds of a cross-median fatal crash versus an injury or PDO crash increase by nearly 300% when people drive on sections curved to the right as opposed to those curved to the left.

On the basis of the results of the ordinal logistic regression analysis for CMC crash severity, the following regression equations can be written:

$$\log\left[\frac{p_1}{(1-p_1)}\right] = -2.2212 + 0.6552X_1 + 1.3694X_2 - 1.0591X_3 - 1.1884X_4 + 1.3088X_5 \quad (6)$$

$$\log\left[\frac{(p_1+p_2)}{p_3}\right] = 1.4074 + 0.6552X_1 + 1.3694X_2 - 1.0591X_3 - 1.1884X_4 + 1.3088X_5 \quad (7)$$

TABLE 4 Final Ordinal Logistic Regression Model for CMC Crashes

Analysis of Effects		Effect	Degrees of Freedom (df)	Wald Chi-square	P-value
		Drugs	1	4.2612	0.0390
		Curve	2	11.5855	0.0030
		Drugs*Curve	2	11.2707	0.0036
Parameter	**Estimated Coefficient**	**Odds Ratio**	**Estimated Standard Error**	**Wald Statistic**	**P-value**
Intercept 3	-2.2212	--	0.3618	37.6874	<0.0001
Intercept 2	1.4074	--	0.3531	15.8867	<0.0001
Drugs 1	0.6552	1.926	0.3174	4.2612	0.0390
Curve 1	1.3694	3.933	0.4227	10.4974	0.0012
Curve 2	-1.0591	0.336	0.4574	5.3617	0.0206
Drugs*Curve 1 1	-1.1884	0.305	0.4166	8.1362	0.0043
Drugs*Curve 1 2	1.3088	3.702	0.4617	8.0348	0.0046
Likelihood Ratio Test: χ^2 = 21.1107 (5 d.f.); p-value = 0.0008 Score Test for Proportional Odds Assumption: χ^2 = 10.3810 (5 d.f.); p-value = 0.0651 Akaike Information Criterion (AIC) = 229.331 -2 Log L = 215.331					

where

p_1 = P(Y = fatal);
p_2 = P(Y = injury);
p_3 = P(Y = PDO);
X_1 = drug or alcohol use indicator (1 if not using, 0 otherwise);
X_2 = horizontal alignment indicator (1 if tangent, 0 otherwise);
X_3 = horizontal alignment indicator (1 if curve to right, 0 otherwise);
X_4 = interaction between drug use and horizontal alignment indicator (1 if no drug use and tangent section, 0 otherwise); and,
X_5 = interaction between drug use and horizontal alignment indicator (1 if no drug use and curved section to the right).

The predicted probabilities can then be computed as follows:

$$p_{\text{fatal}} = \frac{e^{\text{Equation 6}}}{1 + e^{\text{Equation 6}}} \quad (8)$$

$$p_{\text{fatal}} + p_{\text{injury}} = \frac{e^{\text{Equation 7}}}{1 + e^{\text{Equation 7}}} \quad (9)$$

$$p_{\text{PDO}} = 1 - (p_{\text{fatal}} + p_{\text{injury}}) \quad (10)$$

On the basis of the possible values of the indicator variables and the predicted probability equations, the probability of a fatal crash is between 0.098 and 0.243; the probability of injury crash is between 0.681 and 0.705; the probability of a PDO crash ranges between 0.052 and 0.221. The observed probabilities are 0.174, 0.674, and 0.152 for fatal, injury, and PDO crashes, respectively.

To build the median barrier crash severity model, a procedure identical to the CMC crash severity estimate was used. The final model results are shown in Table 5.

The categorical ADT variable used in the analysis shown in Table 5 contains seven levels, indicating directional traffic volumes: Level 1 is volumes between 0 and 5,000; 2 is volumes between 5,001 and 10,000; 3 is volumes between 10,001 and 15,000; 4 is volumes between 15,001 and 20,000; 5 is volumes between 20,001 and 25,000; 6 is volumes between 25,001 and 30,000; and 7 is volumes greater than 30,000. ADT was initially modeled as a continuous variable; however, the results indicated that ADT was not statistically significant (p-value > 0.90). Because it was hypothesized that ADT would have an effect on crash severity, it was coded as a categorical variable to better examine its effect on crash severity. The categorical levels were chosen on the basis of existing PennDOT median barrier warrant criteria (*2*), wherein ADT is considered a decision-making variable in evaluation of the need for median barrier on divided highways. The ADT categories shown are simply the directional ADT used in the existing criteria.

The variables pavement surface, various crash cause, drugs, various ADT categories, interchange entrance ramp, and interaction between drugs and interchange entrance ramp are statistically significant. However, the score test for the proportional odds assumption is violated ($p < 0.0001$). The likelihood ratio test indicates that the explanatory variables given in the model impact crash severity.

Since the proportional odds assumption was violated for the median barrier crash data, the response variable (crash severity) was

TABLE 5 Final Ordinal Logistic Regression Model for Median Barrier Crashes

Analysis of Effects		**Effect**	**Degrees of Freedom (df)**	**Wald Chi-square**	**P-value**
		Surface	1	28.1599	<.0001
		Cause	5	10.2195	0.0716
		Drugs	1	16.9268	<.0001
		ADT	6	23.1512	<.0001
		Ramp	1	4.0916	0.0431
		Drugs*Ramp	1	9.1637	0.0025
Parameter	**Estimated Coefficient**	**Odds Ratio**	**Estimated Standard Error**	**Wald Statistic**	**P-value**
Intercept 3	-0.4690	--	0.0770	37.0592	<0.0001
Intercept 2	4.8344	--	0.1917	635.9604	<0.0001
Surface 1	-0.1947	0.678	0.0367	28.1599	<0.0001
Cause 1	-0.3875	0.601	0.1343	8.3227	0.0039
Cause 2	0.0289	0.912	0.0752	0.1474	0.7011
Cause 3	0.0036	0.889	0.0641	0.0032	0.9551
Cause 4	0.0333	0.916	0.1013	0.1078	0.7427
Cause 5	0.2006	1.083	0.1039	3.7296	0.0535
Drugs 1	0.2771	1.319	0.0674	16.6298	<0.0001
ADT 1	0.4134	1.983	0.1195	11.9731	0.0005
ADT 2	0.1046	1.456	0.0629	2.7646	0.0964
ADT 3	-0.0848	1.205	0.0794	1.1399	0.2857
ADT 4	-0.0571	1.239	0.0754	0.5741	0.4486
ADT 5	-0.0104	1.298	0.0742	0.0198	0.8880
ADT 6	-0.0943	1.194	0.0826	1.3034	0.2536
Ramp 1	-0.1219	0.885	0.0603	4.0916	0.0431
Drugs*Ramp 1 1	0.1817	--	0.0600	9.1637	0.0025
Likelihood Ratio Test: χ^2 = 144.6431 (15 d.f.); p-value = <0.0001 Score Test for Proportional Odds Assumption: χ^2 = 54.1356 (15 d.f.); p-value = <0.0001 Akaike Information Criterion (AIC) = 6269.387 -2 Log L = 6235.387					

modeled as an unordered variable (nominal) and the logistic regression analysis redone. A multinomial logit model can be thought of as one in which a series of binary logits are estimated for all of the possible outcome categories. In the case of the crash severity models, the multinomial logit model compares fatal to injury, fatal to PDO, and injury to PDO outcomes. Consideration of the independence from irrelevant alternatives assumption of the multinomial logit model may be violated for certain severity classification schemes. However, the severity classifications (fatal, injury, and PDO) used in this research were considered distinctly different—ones that could be weighed independently in the eyes of the decision maker, such as a police officer investigating the crash.

If y is the response variable with J nominal outcomes, then the assumption of the multinomial logit model is that the categories 1 through J are not ordered. Also, let $\Pr(y = m|x)$ be the probability of observing outcome m given the set of independent variables x. The model for y is constructed as follows:

- Assume that $\Pr(y = m|x)$ is a linear combination $x\beta_{m.}$ The vector $\beta_m = (\beta_{0m} \ldots \beta_{km} \ldots \beta_{Km})$ contains the intercept β_{0m} and coefficients β_{km} for the effect of x_k on outcome m. This is an opposing view from the ordinal response model because the parameter estimates are assumed different for each outcome.
- To ensure nonnegativity for the probabilities, the exponential of $x\beta_m$ is taken.
- For the probabilities to sum to 1, the following normalization is needed:

$$\Pr(y_i = m|x_i) = \frac{\exp(x_i\beta_m)}{\sum_{j=1}^{J}\exp(x_i\beta_j)} \tag{11}$$

To identify the set of parameters that generates the probabilities, a constraint must be imposed. It is common to impose the constraint that one of the parameter estimates equals 0 (i.e., $\beta_1 = 0$). Imposing such a constraint allows the model to be written as follows:

$$\Pr(y_i = 1|x_i) = \frac{1}{1 + \sum_{j=2}^{J}\exp(x_i\beta_j)} \tag{12}$$

$$\Pr(y_i = m|x_i) = \frac{\exp(x_i\beta_m)}{1 + \sum_{j=2}^{J}\exp(x_i\beta_j)} \qquad \text{for } m > 1 \tag{13}$$

Maximum likelihood estimation is used to determine parameter estimates. To develop the final nominal logistic regression model of median barrier crash severity, a stepwise procedure was used. The analysis included the predictor variables shown in Table 1. Interaction terms were also included in the procedure—the interaction terms included the drugs predictor. For example, drugs–surface and drugs–ramp were analyzed in the stepwise procedure. The PROC CATMOD (*10*) statement was used with a significance level of 0.1 to retain variables in the model. The results for the stepwise procedure are shown in Table 6.

From the results presented in Table 6, a high p-value (0.5989, 48 degrees of freedom) for the likelihood ratio test indicates that the model is a good fit. This indicates that the response variable (crash severity) is more appropriately modeled as a nominal variable rather than as an ordinal variable. Independent variables with low p-values (<0.10) are statistically significant.

The parameter estimates provided in Table 6 show two intercepts, one for each generalized logistic regression model; two regression

TABLE 6 Final Nominal Logistic Regression Results for Median Barrier Crashes

		Source	Degrees of Freedom	Chi-square	Pr > Chi-square
Analysis of Variance		Intercept	2	240.44	<0.0001
		Surface	2	48.33	<0.0001
		Drugs	2	30.64	<0.0001
		Ramp	2	4.01	0.1347
		ADT	6	21.57	0.0014
		Drugs*Ramp	2	7.18	0.0276
		Likelihood Ratio	48	44.94	0.5989
Parameter	**Function Number**	**Estimate**	**Standard Error**	**Chi-square**	**Pr > Chi-square**
Intercept	1	-3.8315	0.3000	163.13	<0.0001
Intercept	2	0.4357	0.0625	48.58	<0.0001
Surface 1	1	0.6948	0.2733	6.46	0.0110
Surface 1	2	0.2120	0.0320	43.90	<0.0001
Drugs 1	1	-1.0691	0.2156	24.59	<0.0001
Drugs 1	2	-0.2038	0.0607	11.28	0.0008
Ramp 1	1	0.3743	0.2161	3.00	0.0833
Ramp 1	2	0.0798	0.0606	1.74	0.1876
ADT 1	1	-0.4272	0.3559	1.44	0.2300
ADT 1	2	-0.2239	0.0525	18.20	<0.0001
ADT 2	1	0.0659	0.2995	0.05	0.8258
ADT 2	2	0.0989	0.0526	0.00	0.9850
ADT 3	1	-0.0286	0.2999	0.01	0.9240
ADT 3	2	-0.00552	0.0530	0.01	0.9170
Drugs*Ramp 1 1	1	-0.3256	0.2127	2.34	0.1258
Drugs*Ramp 1 1	2	-0.1477	0.0602	6.01	0.0142

slopes for the surface, drugs, ramp, and drugs–ramp indicator variables; and six regression slopes for the ADT indicator variable. The ADT indicator was modeled as a categorical variable with four categories, each indicating a directional traffic volume: Category 1 is traffic volumes between 0 and 10,000 vehicles per day; 2 is traffic volumes between 10,001 and 20,000 vehicles per day; 3 is traffic volumes between 20,001 and 30,000 vehicles per day; and 4 is traffic volumes greater than 30,000 vehicles per day. These categories are slightly different from those used for the ordinal logistic regression model (see Table 5). This was done because the ordinal model indicated that only lower ADT categories significantly affected crash severity.

The parameter estimates not shown in Table 6 are opposite in sign for the categorical variables with two levels. The fourth category of the ADT variable has a parameter estimate equal to the sum of the three categories shown with a negative sign convention.

Logit equations can be written by using the estimates from Table 6, as follows:

$$\log\left[\frac{P(Y = \text{fatal})}{P(Y = \text{PDO})}\right] = -3.8315 + 0.6948X_1 - 1.0691X_2 + 0.3743X_3 - 0.4272X_4 - 0.3256X_5 \quad (14)$$

$$\log\left[\frac{P(Y = \text{injury})}{P(Y = \text{PDO})}\right] = 0.4357 + 0.2120X_1 - 0.2038X_2 + 0.0798X_3 - 0.2239X_4 - 0.1477X_5 \quad (15)$$

where

X_1 = pavement surface indicator (1 if dry, 0 otherwise);
X_2 = drug or alcohol use indicator (1 if no drugs or alcohol, 0 otherwise);
X_3 = interchange entrance ramp indicator (1 if no ramp influence, 0 otherwise);
X_4 = ADT indicator (1 if fewer than 10,000 vehicles per day, 0 otherwise); and
X_5 = interaction effect for drugs and ramp (1 if no ramps or drugs, 0 otherwise).

Interpretation of the model results is straightforward. The 16 parameters included in Table 6 are the intercepts and slope coefficients for the two equations predicting the log odds of fatal versus PDO and injury versus PDO crash severities. The effects of pavement surface on fatal versus PDO crashes are statistically significant, as are drug effects. The regression parameters for these two variables are 0.695 and –1.069, respectively. The relative effect of a driver not under the influence of drugs versus a driver under the influence of drugs is exp(–1.069) = 0.343. This indicates that the odds of crash severity being a fatal crash versus a PDO crash are 66% lower if the driver is not using drugs or alcohol. The relative effect of dry pavement surface versus wet or icy pavement surface is 2.0. This indicates that the odds of a fatal crash severity versus PDO crash severity are 100% higher for driving on a dry pavement surface. When one compares injury crash severity to PDO crash severity, the surface, drugs, and ADT indicators are all statistically significant. These coefficients are 0.212, –0.204, and –0.224, respectively. This indicates that the relative effects of an injury versus PDO crash severity are 23% higher, 19% lower, and 20% lower, respectively. In summary, the odds of an injury crash versus PDO crash are 23% higher on dry pavement surfaces than on wet or icy pavement surfaces. Not using drugs or alcohol reduces the likelihood of an injury crash versus a PDO crash. And lower traffic volume levels decrease the likelihood of an injury crash versus a PDO crash.

The predicted probabilities (as based on the results shown in Table 6) can easily be computed on the basis of the following equations:

$$p_{\text{fatal}} = \frac{e^{\text{Equation(14)}}}{[1 + e^{\text{Equation(14)}} + e^{\text{Equation(15)}}]} \quad (16)$$

$$p_{\text{injury}} = \frac{e^{\text{Equation(15)}}}{[1 + e^{\text{Equation(14)}} + e^{\text{Equation(15)}}]} \quad (17)$$

$$p_{\text{PDO}} = 1 - (p_{\text{fatal}} + p_{\text{injury}}) \quad (18)$$

On the basis of the values of the indicator variables and the predicted probability equations, the probability of a fatal median barrier crash is between 0.005 and 0.008; the probability of an injury crash ranges between 0.535 and 0.602; the probability of a PDO crash ranges between 0.390 and 0.460. The observed crash severity distribution was 0.007, 0.560, and 0.433 for fatal, injury, and PDO crashes, respectively.

CONCLUSION

Both CMC and median barrier crash severities were initially modeled by using ordinal logistic regression. In each case, the response (crash severity) contained three levels—fatality, injury, or PDO. In the case of CMC crashes, the use of drugs and the presence of a curvilinear alignment increased the odds that a crash is fatal when compared to an injury or PDO crash. The model developed for CMC crash severity is statistically significant, and the assumption of parallel regression lines (ordinal response) was appropriate.

For Interstate median barrier crashes, a wet roadway surface, the use of drugs or alcohol, the presence of an interchange entrance ramp, the crash type, and the ADT volume all change the odds of a fatal crash. In all the models developed, a drug or alcohol interaction term was statistically significant. The reported results indicate that ordinal logistic regression provides an adequate fit to the median barrier crash severity data; however, the proportional odds assumption of parallel regression lines was violated when median barrier crash data for Pennsylvania Interstates were used. This could be because of the small number of fatal crashes that occurred on the system between 1994 and 1998. Only 0.7% of median barrier crashes on the Interstate system resulted in a fatality, whereas 43% were PDO crashes, and about 56% were injury crashes. As such, the median barrier crash severity models were reestimated by using nominal logistic regression. The results suggest that an unordered response is appropriate. For the median barrier crash models, the roadway surface condition, use of drugs or alcohol, presence of an interchange entrance ramp, traffic volumes, and the interaction between drug use and the presence of an interchange ramp all affect crash severity.

The results of the crash severity modeling could be used by practitioners to determine the probability of fatal, injury, and PDO as based on a set of geometric and environmental explanatory variables. The relative effects of the geometric design variables, such as interchange entrance ramp presence, horizontal curvature, and traffic volumes, show the impact that design decisions have on

crash severity. The modeling output suggests that curved alignments increase crash severity when compared to tangent alignment sections. Further, the presence of interchange entrance ramps increases the likelihood of fatal and injury crashes compared to PDO crashes. Use of the logistic regression models, combined with crash frequency models, would provide very useful decision-making tools for median-involved crashes and permit their dual use to assess median design and safety policies.

ACKNOWLEDGMENTS

The authors recognize Martin T. Pietrucha, Konstadinos G. Goulias, and William L. Harkness of Pennsylvania State University, who provided significant guidance on this research effort. The assistance of the Bureau of Highway Safety and Traffic Engineering of the Pennsylvania Department of Transportation is acknowledged for providing the data needed to complete the study.

REFERENCES

1. *Fatal Crash Reporting System.* NHTSA, Washington, D.C., 2001.
2. *Publication 13M, Design Manual.* Pennsylvania Department of Transportation, Harrisburg, 1998.
3. Chang, L. Y., and F. Mannering. Analysis of Injury Severity and Vehicle Occupancy in Truck- and Non-Truck-Involved Crashes. *Crash Analysis and Prevention,* Vol. 31, No. 1, 1999, pp. 579–592.
4. Lee, J., and F. Mannering. Impact of Roadside Features on the Frequency and Severity of Run-off-Roadway Crashes: An Empirical Analysis. *Crash Analysis and Prevention,* Vol. 34, No. 6, 2002, pp. 149–161.
5. Mercier, C. R., M. C. Shelley, II, J. B. Rimkus, and J. M. Mercier. Age and Gender as Predictors of Injury Severity in Head-On Highway Vehicular Crashes. In *Transportation Research Record 1581,* TRB, National Research Council, Washington, D.C., 1997, pp. 37–46.
6. Dissanayake, S., and J. Lu. Analysis of Severity of Young Driver Crashes: Sequential Binary Logistic Regression Modeling. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1784,* TRB, National Research Council, Washington, D.C., 2002, pp. 108–114.
7. Kim, K., L. Nitz, J. Richardson, and L. Li. Personal and Behavioral Predictors of Automobile Crash and Injury Severity. *Crash Analysis and Prevention,* Vol. 27, No. 4, 1995, pp. 469–481.
8. Abdel-Aty, M. A., C. L. Chen, and J. R. Schott. An Assessment of Driver Age on Traffic Crash Involvement Using Log-Linear Models. *Crash Analysis and Prevention,* Vol. 30, No. 6, 1998, pp. 851–861.
9. Bauer, K. M., and D. W. Harwood. *Statistical Models of At-Grade Intersection Crashes.* FHWA-RD-96-125. FHWA, U.S. Department of Transportation, 1996.
10. *SAS/STAT User's Guide, Version 8.* SAS Publishing, Cary, N.C., 1999.
11. *Pennsylvania Crash Facts and Statistics*. Bureau of Highway Safety and Traffic Engineering, Pennsylvania Department of Transportation, Harrisburg, 2003.
12. Long, J. S. Regression Models for Categorical and Limited Dependent Variables. In *Advanced Quantitative Techniques in the Social Sciences,* Vol. 7, Sage Publications, Thousand Oaks, Calif., 1997.
13. Agresti, A. *Categorical Data Analysis.* John Wiley and Sons, New York, 1990.
14. Hosmer, D. W., and S. Lemeshow. *Applied Logistic Regression.* John Wiley and Sons, New York, 1989.

Publication of this paper sponsored by Safety Data, Analysis and Evaluation Committee.