EXPLORATORY MULTIVARIABLE ANALYSES OF CALIFORNIA DRIVER RECORD ACCIDENT RATES

By
Michael A. Gebers

MAY 1997

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE May 1997	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE Exploratory Multivariable Analyses of California Driver Record Accident Rates	5. FUNDING NUMBERS
6. AUTHOR(S) Michael A. Gebers	

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) California Department of Motor Vehicles Research and Development Branch P.O. Box 932382 Sacramento, CA 94232-3820	8. PERFORMING ORGANIZATION REPORT NUMBER RSS-97-166

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Since 1964, the California Department of Motor Vehicles has issued several monographs on driver characteristics and accident risk factors as part of a series of analyses known as the California Driver Record Study.

This paper presents the results of a number of regression analyses of driving record variables measured over a 6-year time period (1986-91). The techniques presented consist of ordinary least squares, weighted least squares, Poisson, negative binomial, linear probability, and logistic regression models. The objective of the analyses was to compare the results obtained from several different regression techniques under consideration for use in the in-progress California Driver Record Study.

The results are informative in determining whether the various regression methods produce similar results for different sample sizes and to explore whether reliance on ordinary least squares techniques in past California Driver Record Study analyses have produced biased significance levels and parameter estimates.

The results indicate that, for these data, the use of the different regression techniques do not lead to any greater increase in individual accident prediction beyond that obtained through application of ordinary least squares regression. In addition, the methods produce almost identical results in terms of the relative importance and statistical significance of the independent variables. It therefore appears safe to employ ordinary least squares multiple regression techniques on driver accident-count distributions of the type represented by California driver records, at least when the sample sizes are large.

14. SUBJECT TERMS Motor vehicles accidents, traffic safety, accident proneness, accident rates, accident risks, accident repeater drivers, convictions, high-risk drivers, sex factor in driving.	15. NUMBER OF PAGES 31
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

NSN 7540-01-280-5500

Standard Form 298 (2-89)
Prescribed by ANSI Std. Z39-18
298-102

PREFACE

This report is issued as an internal monograph of the California Department of Motor Vehicles' Research and Development Branch rather than an official report of the State of California. The opinions, findings, and conclusions expressed in the report are those of the author and not necessarily those of the State of California.

ACKNOWLEDGMENTS

This study was conducted under the general direction of Raymond C. Peck, Research Chief, and the supervision of Robert A. Hagge, Research Manager.

The author would like to extend thanks to Alan Fung, Information Systems Division Programmer, for providing assistance in computer programming for the extraction, processing, and summarization of data and to Douglas Luong, Office Technician, for preparing the report drafts. The author would also like to thank Debbie McKenzie, Associate Governmental Program Analyst, for proofreading a latter draft of this report.

EXECUTIVE SUMMARY

Background and Objectives
- Since 1964, the California Department of Motor Vehicles has issued a number of monographs on driver characteristics and accident risk factors as part of a series of analyses known as the California Driver Record Study.

- Past California Driver Record Study analyses, and many other studies conducted by the California Department of Motor Vehicles, have utilized standard parametric techniques such as analysis of variance, analysis of covariance, and ordinary least squares multiple regression models in analyzing the relationship between a variety of independent variables and subsequent accident rates. The justification for using these techniques is based on the operation of the central limit theorem in producing approximate normality of the test statistic when sample size is extremely large.

- This paper presents the results of a number of regression analyses of driving record variables measured over a 6-year time period (1986-91). The techniques presented consist of ordinary least squares, weighted least squares, Poisson, negative binomial, linear probability, and logistic regression models. The objective of the analyses was to compare the results obtained from several different regression techniques under consideration for use in the 1996 California Driver Record Study, which is currently in progress.

- The results are informative in determining whether the various regression methods produce similar results for different sample sizes and to explore whether reliance on ordinary least squares techniques in past California Driver Record Study analyses have produced biased significance levels and parameter estimates.

Research Methods
- Data for the analyses were obtained from the driving records of a 1% random sample of licensed California drivers extracted in 1992 from the California Driver Record Study database.

- For each subject, information was collected on driver (a) age; (b) gender; (c) presence of a physical or mental condition code on record; (d) presence of license restrictions on record; (e) number of total citations occurring during 1986-88; and (f) number of total accidents occurring during 1986-88.

- Ordinary least squares, weighted least squares, Poisson, negative binomial, linear probability, and logistic regression were used to identify which combination of variables in the pool provided the most accurate equation for predicting the accident criterion measure.

- Analyses are presented for two types of models: (1) those using frequency data, where the dependent (criterion) variable represents the actual number of accident involvements, from 0 to K accidents, and (2) those using categorical data, where the accident criterion measure is a binary variable (equal to 0 if no accidents and 1 if one or more accidents).

Results
- The results of the analyses are consistent with those of prior traffic safety research, with all of the models indicating that increased accident involvement was associated with the following:
  – Increased prior citation frequency
  – Increased prior accident frequency
  – Possessing a commercial driver license
  – Being young
  – Being male
  – Having a medical condition on record
  – Having a driver license restriction on record

- The use of different regression techniques do not lead to any greater increase in individual accident prediction beyond that obtained through application of ordinary least squares regression.

- Any generalization about driving performance from the present analyses is limited by the absence of exposure data (e.g., miles driven) and territorial data (e.g., driver record information by ZIP Code and U.S. census variables).

Recommendations
- The results indicate that, for these data, the use of the different regression techniques do not lead to any greater increase in individual accident prediction beyond that obtained through application of ordinary least squares regression. In addition, the methods produce almost identical results in terms of the relative importance and statistical significance of the independent variables.

- It therefore appears safe to employ ordinary least squares multiple regression techniques on driver accident-count distributions of the type represented by California driver records, at least when the sample sizes are large.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF CONTENTS (continued)

LIST OF TABLES (continued)

INTRODUCTION

This paper presents the results of a number of regression analyses of driving record variables measured over a 6-year time period (1986-91). The objective of the analyses was to compare the results obtained from several different regression techniques under consideration for use in the 1996 California Driver Record Study, which is currently in progress. Because this latter effort will both include and extend the present dataset and analyses, no detailed interpretation of the results will be presented here. Rather, the following presents only the multiple regression equations and highlights the major findings.

Past California driver record studies, and many other studies of the California DMV, have utilized standard parametric techniques such as analysis of variance, analysis of covariance and ordinary least squares (OLS) multiple regression models in analyzing the relationship between a variety of independent variables and subsequent accident rates. The justification for using OLS-based parametric methods on non normally-distributed accident count variables is provided in several previous California DMV publications, such as Peck and Kuan (1983), DeYoung (1995) and Gebers, DeYoung, and Peck (1997). In general, these justifications are based on asymptotic arguments–i.e., the operation of the central limit theorem in producing approximate normality of the test statistic when N is extremely large. The results of a number of Monte Carlo studies have also been cited as an additional defense, including the effects of violating the homoscedasticity assumption when heteroscedasticity is not extreme.

Recent years have witnessed the development and increased availability of techniques which are less reliant on asymptotic arguments and more anchored in formal mathematical derivation. Among these techniques are Poisson and negative binomial regression, weighted least squares regression, logistic regression, and probit regression. It is therefore becoming more common to see unqualified indictments against the use of OLS-based techniques on highly non-normal data such as accident frequencies. A number of authors (Maddala, 1991; Kleinbaum et al., 1988) point out that highly skewed Poisson-like variables produce heteroscedastistic residuals, thereby introducing "inconsistency" into the parameter estimates produced by OLS techniques. Draper and Smith (1966) note that use of OLS-multiple regression in the presence of heteroscedasticity and non-normality results in regression models which, though not biased, do not satisfy the property of minimum variance. Other authors point out that OLS models can produce parameter estimates (Y's) that reside outside the permissible range of observed values (e.g., negative values or values above 1.0 for binary dependent variables).

These concerns have also been raised in connection with studies of both driver and roadway accident rates. In a recent article on the pitfalls of using R^2 as a measure to evaluate goodness-of-fit of accident prediction models, Miaou, Lu, and Lum (1996) list a number of disadvantages of OLS-based methods, as noted below:

> Because accident prediction models are nonnormal and functional forms are typically nonlinear, this study showed through simulated examples that R^2 is not an appropriate measure to make the decisions and comparisons described.

Furthermore, three properties were identified as desirable for any alternative measure to appropriately evaluate the models: (a) it should be bounded between 0 and 1–a value of 0 if no covariate (other than the intercept) is included in the model and a value of 1 if all the necessary covariates are included, (b) it should increase proportionally as equally important, independent covariates are selected and added to the model one at a time regardless of their order of selection, and (c) it should be invariant with respect to the mean (i.e., the value of the measure should not change by simply increasing or decreasing the value of the intercept term in the model). (p. 13)

The above attitude is also reflected in a monograph by Davis (1990), which was commissioned by the National Highway Transportation Safety Administration (NHTSA) in connection with the final report by Stock et al. (1983) on the Dekalb driver training project. Davis (1990) was highly critical of the authors' use of analysis of variance and multiple regression on grounds similar to Miaou et al., as evidenced by the following:

The use of ANOVA techniques when the underlying assumptions are *moderately* violated may give results which are quite similar to those obtained from more appropriate methods of statistical analysis. However, based on the extremely skewed, non-normal distributions of the dependent variables in this study, the use of ANOVA methods is inappropriate. (p. 11)

With respect to the statistical methods used by DeWolf and Smith (1988), the multiple regression approach is subject to the same criticisms as were mentioned for the Stock et al. (1983) and Smith and Blatt (1987) reports, in that the distributions of the dependent variables (number of accidents, number of violations) do not satisfy the general linear model assumptions. (p. 22)

It is instructive to note that none of the above authors acknowledge the possible role of very large N in mitigating the effects of non-normality in the parent distribution.

It is therefore informative to examine whether various methods produce similar results at different sample sizes and to explore whether reliance on OLS techniques in past California driver record studies have produced significance levels and parameter estimates that are materially biased.

METHODOLOGY

Subjects

Data for the analyses were obtained from the driving records for a 1% random sample of licensed California drivers (n = 152,931) extracted in 1992 from the California Driver Record Study database. Detailed information on this database is provided by Peck, McBride, and Coppin (1971), Peck and Kuan (1983), and Peck and Gebers (1992).

To be eligible for selection into the sample, drivers had to meet the following criteria: (a) have a valid driver license at the beginning of the study period; (b) be alive as of

the May 1992 extraction date; and (c) possess a driver license that had not been expired for over 6 months as of the extraction date.

For each subject, information was collected on (a) age; (b) gender; (c) presence of a physical or mental (P&M) condition code on record; (d) presence of license restrictions on record; (e) number of total citations occurring during 1986-88; and (f) number of total accidents occurring during 1986-88. The following displays descriptive statistics for the biographical and driver-record variables:

Variable	n = 152,931
Total accidents (1989-91)	
$\overline{X}$	0.1517
SD	0.4138
Variance	0.1713
Total accidents (1986-88)	
$\overline{X}$	0.1706
SD	0.4380
Variance	0.1918
Total citations (1986-88)	
$\overline{X}$	0.6414
SD	1.1964
Variance	1.4313
Age	
$\overline{X}$	42.67
SD	15.33
Variance	234.96
% Class 1/A or 2/B license	3.3
% one or more P&M conditions	1.4
% one or more restrictions	34.0
% male	52.4

Analysis

Multiple regression analysis was used to identify which combination of variables in the pool provided the most accurate equation for predicting the criterion measure, total accident frequency during 1989-91. To be included in the analyses, drivers had to be licensed for the entire 1986-91 period.

In the following sections, results are presented for two types of regression models: (1) those using frequency data, where the dependent (criterion) variable represents the actual number of accident involvements, from 0 to K accidents, and (2) those using categorical data, where the accident criterion measure is a binary variable (equal to 0 if no accidents and 1 if one or more accidents).

3

RESULTS

Frequency Data: Ordinary Least Squares, Weighted Least Squares, Poisson, and Negative Binomial Regression Models

Table 1 summarizes the results from the nonconcurrent 6-year (1986-88; 1989-91) ordinary least squares and weighted least squares multiple regression analyses. All seven candidate variables were statistically significant predictors of accident involvement at the .10 level of probability, and, therefore, all were included in both regression equations. The directions (positive or negative) of the regression coefficients indicate that increased accident involvement is associated with:

- Increased prior citation frequency
- Increased prior accident frequency
- Having a commercial driver license (which is mostly held by high-mileage professional drivers)
- Being young
- Being male
- Having one or more P&M conditions on record
- Having one or more driver license restrictions on record

Table 1

Summary of Nonconcurrent 6-Year (1986-88; 1989-91) Multiple Regression
Equation for Predicting Total Accidents Using Ordinary Least Squares
and Weighted Least Squares Regression Models (n = 152,931)

Predictor variable	Ordinary least squares				Weighted least squares			
	Regression coefficient	Standard error	F	p	Regression coefficient	Standard error	F	p
Constant	0.211	0.005	1741.49	.000	0.207	0.005	1875.48	.000
Prior total citations	0.029	0.009	965.10	.000	0.030	0.001	857.88	.000
Prior total accidents	0.060	0.002	595.97	.000	0.059	0.003	489.40	.000
License class	0.108	0.006	331.47	.000	0.108	0.007	265.73	.000
Age	-0.001	0.000	247.50	.000	-0.001	0.000	271.93	.000
Gender	-0.028	0.002	170.83	.000	-0.029	0.002	188.94	.000
P&M indicator	0.060	0.009	44.83	.000	0.061	0.010	40.38	.000
Restriction status	0.008	0.002	11.06	.000	0.007	0.002	9.45	.002

F for the equation = 546.81
p = .000
R^2 = .024

F for the equation = 503.33
p = .000
R^2 = .023

Note. Only independent variables that contributed significantly (p < .10) to the prediction of the criterion measure were included in the models. The criterion measure, total accidents during 1989-91, had a mean of 0.152 and standard deviation of 0.414.

The ordinary least squares regression equation takes the following form:

$$Y' = A + B_1 X_1 + B_2 X_2 + \ldots B_K X_K$$

where Y' is the predicted value (either a positive or negative continuous value) on the dependent variable, A is the Y-intercept (the value of Y when all the X values are zero), the Xs represent the various independent variables (of which there are K), and the Bs are the regression coefficients assigned to each of the independent variables. Linear regression models are usually additive models (from which one can estimate increments in absolute risk) rather than multiplicative models (from which one can estimate relative risks). Parameters can be included for estimating nonlinear and interactive (nonadditive) relationships, but the model is still linear and additive in the context of the fitted parameter vector.

The reader should note that the use of multiple regression involves meeting the following assumptions: (1) Independence–the Y observations are statistically independent of one another, (2) Linearity–the value of Y' is a linear function of $X_1, X_2, \ldots X_K$, (3) Homoscedasticity–the variance of Y' is the same for any fixed combination of $X_1, X_2, \ldots X_K$, (4) Normality–the errors of prediction are normally distributed at all levels of Y', (5) Measurement infallibility–the variates are free of measurement error, and (6) Additivity–the effect terms (coefficients) of the parameters can be combined in an additive fashion to estimate Y. Failure to meet the above assumptions are potential threats to the accuracy of the parameter estimates.

As stated above, a fundamental assumption underlying unweighted least squares linear regression analysis is that all random errors have the same variance at different levels of the explanatory variable. The homogeneity of residual error assumption is invariably violated with accident data because of the direct proportional relationship between the means and variances of the arrays, thereby introducing heteroscedasticity into the distribution of the residuals.

The weighted-least squares method of analysis is a modification of standard regression analysis procedures and is used when a regression model is to be fit to data for which the assumptions of variance homogeneity and/or independence, stated above, do not hold. The least squares residual sum of squares is:

$$\Sigma(Y_i - B_0 - B_1X_i - \ldots B_mX_m)^2.$$

The weighted least squares residual sum of squares is:

$$\Sigma w_i(Y_i - B_0 - B_1X_i - \ldots B_mX_m)^2.$$

where w_i is the nonnegative weight assigned to an individual observation. Observations with small weights contribute less to the sum of squares and thus provide less influence to the estimation of parameters, and vice versa for observations with larger weights. Therefore, it is logical to assign small weights to observations whose large error of prediction make them more unreliable, and likewise to assign larger weights to observations with smaller error of prediction. It can, in fact, be shown that best linear unbiased estimates are obtained if the weights are inversely proportional to the individual errors (Kleinbaum, Kupper, & Muller, 1988). For the current analysis, it is assumed that the mean and standard deviation of the

accident criterion are proportional to one another. Although this assumption is not perfectly met, the amount of variance overdispersion is small.

In the present analysis, an ordinary least squares regression was run, and predicted scores and residuals were computed for all observations. The sample was subdivided into quartiles on the basis of the distribution of the predicted scores. The standard deviation of the residuals was calculated for each quartile. The individual weights used in the follow-up weighted least squares regression were defined as the reciprocals of the standard deviations.

A comparison of the results from the unweighted and weighted regression analyses in Table 1 shows the effect of weighting. The regression coefficients obtained by the two methods are remarkably similar, and all significance levels (p values) except one are identical through three digits. That one exception (restriction status) differed slightly on the third digit (p = .000 versus .002).

As displayed in Table 1, R^2 actually declined slightly, from .024 using ordinary least squares to .023 using weighted least squares. Since OLS has the property of maximizing the R^2, the quantity computed from the weighted least squares may be smaller than the OLS R^2. This can lead to an interpretive quandary when the discrepancy is larger than exists here–namely preferring models with the highest R^2, but acknowledging that, on theoretical grounds, a model with lower R^2 is preferred.

As noted earlier, attempts to model accident frequencies using least squares regression techniques have been criticized in prior research (e.g., Boyer, Dionne, & Vanasse, 1990; Grogger, 1990; Davis, 1990). Linear models often assume a normal distribution of data and allow for the prediction of negative values.

The above concern has led to the development and advocacy of the Poisson regression model. The Poisson distribution lends itself to the modeling of either count or means data by virtue of its discrete, nonnegative integer distribution. In the case of traffic accidents, the Poisson distribution gives

$$\Pr\ (Y = K) = (e^{-\lambda})\frac{\lambda^K}{K!}$$

where Pr (Y = K) is the probability that the number of accidents, Y, will equal K, e = 2.718 (base of the natural logarithm), and λ is the expected number of accidents. Given a vector of variables, λ for an individual driver can be estimated by the equation,

$$\ln \lambda_i' = X_i B$$

where X is a vector of variables (e.g., age, gender, prior citations) and B is a vector of estimable coefficients.

Poisson models are generally multiplicative. Poisson regression models are not restricted to all of the assumptions noted above for multiple linear regression and are

specifically applicable to discrete count data where the probability of a given event (e.g., accidents) is relatively infrequent and can be approximated by a Poisson probability function.

The Poisson distribution, however, suffers from a potentially important limitation, namely that the dependent variable's mean and variance are constrained to be equal. Data overdispersion (in which the variance is greater than the mean) or underdispersion (in which the variance is less than the mean) violates this constraint and leads to biased estimates of the significance of the regression coefficients. If overdispersion is present, the negative binomial regression model is employed as an alternative.

The negative binomial model is an extension of the Poisson regression model which allows the variance of the process to differ from the mean. The negative binomial model is

$$\ln \lambda_i = B\,X_i + \mathcal{E},$$

where $\exp(\mathcal{E})$ has a gamma distribution with mean 0 and variance γ.

Table 2 presents the results from the Poisson and negative binomial regression analyses. As stated above, overdispersion is a phenomenon that sometimes occurs in data that are arguably inappropriately modeled with a Poisson distribution. If the estimate of dispersion (variance divided by the mean) is greater than 1, then the data

Table 2

Summary of Nonconcurrent 6-Year (1986-88; 1989-91)
Multiple Regression Equation for Predicting Total Accidents Using Poisson
and Negative Binomial Models (n = 152,931)

Predictor variable	Poisson regression				Negative binomial regression			
	Regression coefficient	Standard error	χ^2	p	Regression coefficient	Standard error	χ^2	p
Constant	-1.35	0.032	1842.69	.000	-1.36	0.024	3095.26	.000
Prior total citations	0.112	0.004	725.35	.000	0.114	0.003	1195.01	.000
Prior total accidents	0.274	0.012	545.11	.000	0.275	0.009	897.50	.000
License class	0.458	0.027	282.07	.000	0.458	0.021	462.94	.000
Age	-0.009	0.001	365.23	.000	-0.009	0.000	612.68	.000
Gender	-0.218	0.014	240.36	.000	-0.217	0.011	402.02	.000
P&M indicator	0.279	0.045	37.88	.000	0.284	0.035	64.59	.000
Restriction status	0.049	0.015	11.20	.000	0.049	0.011	18.60	.000
	Deviance = 95,760 Pseudo-R^2 = .03577				Deviance = 89,006 Pseudo-R^2 = .03626			

Note. Only independent variables that contributed significantly ($p < .10$) to prediction of the criterion measure were included in the models. The criterion measure, total accidents during 1989-91, had a mean of 0.152 and standard deviation of 0.414.

may be overdispersed. On the other hand, if the dispersion estimate is less than 1, then the data may be underdispersed. If the value is within the typically-acceptable 0.8 to 1.2 range, the model can be considered to be correctly specified (SAS Institute Inc., 1993; Hilbe, 1994). As displayed in Table 2, the dispersion statistic associated with the Poisson model is 1.08, which indicates that overdispersion may not be a problem with these data. (For a discussion on tests for detecting overdispersion in Poisson regression models, the interested reader is referred to Dean and Lawless [1989].)

As can be seen, the results for the Poisson and negative binomial models are quite similar. Since the Poisson model is a particular case of the negative binomial model, the difference in the deviance goodness-of-fit statistics for the two models can be compared to decide if there is any gain in "model fit" from using a negative binomial regression (the better fitting model having the lower deviance score). The difference of 6,754 between the deviance statistics is significant ($p < .000$), indicating that the negative binomial model performs significantly better than does the Poisson model. However, each of the two models explains about the same amount of variance–the pseudo-R^2 is .0358 for the Poisson model and .0363 for the negative binomial model.

It should be noted that the results from the Poisson and negative binomial regressions parallel those from the linear models presented in Table 1, since the directions (i.e., signs) and p values of the regression coefficients are essentially identical. In fact, the p values for the variables in Table 2 are identical to those for the OLS model through three digits. (These pseudo-R^2 statistics are usually not comparable to the "true" R^2 produced by ordinary least squares and weighted least squares regression.)

Elasticities of independent variables were estimated from the Poisson parameters to determine the impact of these variables on accident frequency. Elasticities can be roughly defined as the percentage change in the number of accidents resulting from a 1% change in the independent variable. Elasticities for each individual observation were computed, and then an average elasticity was estimated for the sample. Because the elasticities of binary variables are not meaningful, Table 3 presents elasticities only for the continuous predictor variables.

Table 3

Accident Frequency Elasticity Estimates

Independent variable	Elasticity (%)
Prior total citations	0.072
Prior total accidents	0.047
Age	-0.434

Note. Elasticity is defined as the percentage change in the average number of accidents that would be expected to result from a 1.000% change in the independent variable.

Table 3 provides some interesting insights. For example, a 1.000% increase in the number of prior total citations is associated with a 0.072% increase in subsequent accident frequency. Similarly, a 1.000% increase in prior total accidents is associated with a 0.047% increase in subsequent accident frequency. This suggests that, at least for these variables, accident likelihood may be more sensitive to prior citations than to prior accidents.

To gain some understanding of the relative importance of the binary variables included in the Poisson regression model, a computation can be performed to provide an idea of the relative effect of these variables on mean accident frequency (λ_{ij}). This is accomplished using the coefficients in Table 2. For example, λ_{ij} can be said to increase 58.1% ($e^{0.458}/e^0$) if a driver holds a commercial driver license.

Table 4 presents the percentage change in λ_{ij} associated with each binary independent variable. These percentage changes are somewhat analogous to the above elasticity coefficients except they represent increases in relative risk rather than additive increments per unit of change.

Table 4

Percentage Change in Mean Accident Frequency (λ_{ij})
Due to Binary Independent Variables

Independent variable	% change in λ_{ij}
License class	58.1
Restriction status	5.0
P&M status	32.2
Gender	-19.6

Categorical Data: Linear Probability and Logistic Regression Models
Models used to estimate the probability of accidents from individual driver characteristics usually involve categorical data where the dependent variable is binary (0 for no accidents and 1 for one or more accidents) (Boyer, Dionne, & Vanasse, 1990). This section presents the results from the standard ordinary least squares linear probability model and the logistic regression model. The predictor and criterion variables used in these models are the same as those used in the frequency-data models presented above.

Table 5 presents results from the linear probability and logistic regression analyses.

Table 5

Summary of Nonconcurrent 6-Year (1986-88; 1989-91) Multiple Regression
Equation for Predicting Total Accidents Using Linear Probability
and Logistic Regression Models (n = 152,931)

Predictor variable	Linear probability regression				Logistic regression			
	Regression coefficient	Standard error	F	p	Regression coefficient	Standard error	Wald χ^2	p
Constant	0.185	0.004	1975.43	.000	-1.336	0.037	1325.36	.000
Prior total citations	0.022	0.001	819.51	.000	0.139	0.006	605.09	.000
Prior total accidents	0.041	0.002	409.41	.000	0.287	0.015	356.65	.000
License class	0.073	0.005	225.64	.000	0.481	0.035	187.47	.000
Age	-0.001	0.000	259.11	.000	-0.010	0.001	301.54	.000
Gender	-0.022	0.002	153.79	.000	-0.212	0.016	174.63	.000
P&M indicator	0.037	0.007	25.67	.000	0.269	0.058	21.94	.000
Restriction status	0.005	0.002	8.28	.004	0.047	0.017	7.74	.005
	F for the equation = 445.57 p = .000 R^2 = .020				-2 Log L for intercept only = 120045.03 -2 Log L for intercept and covariates = 117348.76 Chi-square for covariates = 3056.82, df = 7, p = .000			

Note. Only independent variables that contributed significantly (p < .10) to prediction of the criterion measure were included in the models. The criterion measure, total accidents during 1989-91, had a mean of 0.152 and standard deviation of 0.414.

The standard ordinary least squares linear probability model is defined as

$$Y'_i = A + B_1 X_1 + B_2 X_2 + \dots B_K X_K$$

where the Xs represent the independent variables, and the Bs are the regression coefficients assigned to the independent variables. The dependent variable Y' is dichotomous: $Y'_i = 1$ if the i-th individual had one or more accidents during 1989-91, $Y'_i = 0$ otherwise. The expected value of Y can be interpreted as the probability that Y = 1 or more.

The results from the linear probability model indicate that the significant predictor variables explain part of the differences in accident probabilities between drivers. For example, each additional total citation during 1986-88 increases the probability of being involved in an accident in 1989-91 by 2.2 percentage points; an additional accident during 1986-88 increases the probability of being involved in a subsequent accident by 4.1%. Being female reduces the probability of accident involvement by 2.2 percentage points.

The parameter estimates from the linear probability model are proportionally similar to the ordinary least squares estimates presented in Table 1. The shrinkage in R^2 from .024 in the ordinary least squares model to .020 in the linear probability model is attributable to the loss of information resulting from using a binary rather than continuous accident criterion measure.

The linear regression model has several potential limitations when the dependent variable is binary (Maddala, 1991). First, there is a serious heteroscedasticity problem, meaning that the estimates of the prediction errors are not normally distributed. Secondly, the predicted value could possibly be outside the range of 0 to 1 for certain values of the predictor variables. This is particularly troublesome if the expected value is interpreted as a probability. For this application, a non-linear model such as logistic regression is more appropriate. However, the non-linearity in the expected values only emerges as p approaches 0 or 1, and the two models tend to yield very similar results for p in the range of .20-.80.

Because logistic regression models are nonlinear, the equations used to describe the outcomes are more complex than those for OLS multiple regression models. The outcome variable, Y'_i is the probability of having one outcome or another (0 vs. 1 or more accidents) based on a nonlinear function of the best linear combination of predictors. For two possible outcomes:

$$Y'_i = \frac{e^U}{1 + e^U}$$

where Y'_i is the estimated probability that the ith case (i = 1, 2, … n) is in one of the categories and U is the linear regression equation:

$$U = A + B_1 X_1 + B_2 X_2 + … + B_K X_K$$

with constant A, coefficients B_j, and predictor, X_j for K predictors (j = 1, 2, … K).

This linear regression equation creates the logit, or log of the odds ratio:

$$\ln \left(\frac{Y'}{1 - Y'} \right) = A + \Sigma B_j X_{ij}$$

That is, the linear regression equation is the natural log of the probability of having one outcome (accident-free) divided by the probability of having the other outcome (accident-involved). The procedure for estimating coefficients is maximum likelihood, and the goal is to find the best linear combination of predictors to maximize the likelihood of obtaining the observed outcome frequencies (Hosmer and Lemeshow, 1989).

Again, the signs (positive or negative) and p values of the logistic regression coefficients in Table 5 are similar to those of the prior analyses.

Table 6 presents the odds ratios obtained from the logistic regression equation. The odds ratio as applied here refers to the relative odds of being accident-involved, as a function of a predicted driver-record category. For example, the odds ratios in Table 6 indicate that:

11

- Drivers with two prior citations are 1.32 times as likely to have a subsequent accident than are drivers with no prior citations.

- Drivers with two prior accidents are 1.78 times as likely to have a subsequent accident than are drivers with no prior accidents.

Table 6

Odds Ratios for Prediction of Total Accident Involvement from Logistic Regression Analysis of 6-Year Nonconcurrent Data (1986-88; 1989-91) (n = 152,931)

Predictor variable	Odds-ratio
Prior total citations (1986-88)	
2	1.32
4	1.75
6	2.31
Prior total accidents (1986-88)	
2	1.78
4	3.15
Age (years)	
5	0.95
10	0.91
15	0.86
License class	1.62
Driver license restriction	1.05
Physical and mental condition	1.31
Gender	1.24

- The risk of a subsequent accident is 1.62 times higher for commercial drivers than it is for non-commercial drivers.

- The risk of a subsequent accident is 1.24 times higher for men drivers than it is for women drivers.

So that the reader may get an idea about the order of magnitude of the accident estimates generated from the Poisson, OLS, and weighted least squares models, values of predicted accident rates are displayed in Table 7 for selected values of the predictor variables. While the values are arbitrary, they are within the range of values for the predictor variables. All models produce very similar estimates of accident risk for the portrayed driver groups. The Poisson and negative binomial regression models do yield higher estimated risk levels for drivers with extremely elevated counts of total citations and total accidents. However, such "deviant" records are rare. Approximately 2% of the sample had five or more citations during the prior 3-year (1986-88) period. As is shown later, these differences in expected values do not alter the comparative accuracy of the equations in predicting whether a given driver is accident-free or accident-involved.

Table 7

Predicted Frequency of Accidents from Multiple Regression Equations
at Various Values of the Predictor Variables

Independent variable combination	Predictor variable value							Y'		λ'	
	Sex	Age	License class	Restriction status	P & M status	1986-88 total citations	1986-88 total accidents	Ordinary least squares estimate[a]	Weighted least squares estimate[b]	Poisson estimate[c]	Negative binomial estimate[d]
A	1.48	45.67	0.033	0.34	0.01	0.64	0.17	0.1513	0.1481	0.1422	0.1421
B	1	35	1	1	0	5	2	0.5228	0.5121	0.7531	0.7601
C	1	45	1	0	0	6	2	0.5328	0.5291	0.7299	0.7385
D	1	59	0	1	1	3	1	0.3289	0.3396	0.3053	0.3082
E	1	60	0	1	0	2	0	0.1788	0.1856	0.1556	0.1558
F	1	70	0	0	1	0	0	0.1607	0.1653	0.1425	0.1428
G	2	29	0	0	0	3	1	0.2679	0.2699	0.2346	0.2358
H	2	30	0	1	0	0	0	0.1268	0.1200	0.1327	0.1324
I	2	35	1	1	0	5	2	0.4946	0.4829	0.6058	0.6118

Note. For license class, 0 = Class 3/C or motorcycle; 1 = Class 1/A or 2/B. For restriction status, 0 = no license restriction on record; 1 = one or more license restrictions on record. For P&M status, 0 = no P&M condition on record; 1 = one or more P&M conditions on record. For sex, 1 = male; 2 = female. Values in row "A" represent sample averages.

[a]Equation for ordinary least squares estimate:

Y' = 0.21060199 + (0.10759473 x License class) + (0.00769221 x Restriction status) + (0.05990415 x P&M status) + (-0.00116449 x Age) + (0.02933590 x Total citations) + (0.0596493 x Total accidents) + (-0.02827030 x Sex)

[b]Equation for weighted least squares regression:

Y' = 0.204585 + (0.092161 x License class) + (0.006735 x Restriction status) + (0.066815 x P&M status) + (-0.001100 x Age) + (0.034696 x Total citations) + (0.051395 x Total accidents) + (-0.029154 x Sex)

[c]Equation for Poisson estimate:

λ' = exp $[$-1.3517 + (0.4583 x License class) + (0.0491 x Restriction status) + (0.2789 x P&M status) + (-0.0094 x Age) + (0.1119 x Total citations) + (0.2739 x Total accidents) + (-0.2176 x Sex)$]$

[d]Equation for negative binomial estimate:

λ' = exp $[$-1.3550 + (0.4582 x License class) + (0.0488 x Restriction status) + (0.2840 x P&M status) + (-0.0094 x Age) + (0.1139 x Total citations) + (0.2751 x Total accidents) + (-0.2170 x Sex)$]$

13

<u>Classification and Prediction Accuracy</u>
Two measures of performance were used to compare the adequacy of the different
regression techniques. The first measure selects the group of drivers with the most
prior total accidents in 1986-88, another group with the most prior total citations in
1986-88, and five more groups estimated from the predicted scores in the multiple
regression models as having the highest accident potential. Next, a count was made
of the number of subsequent accidents in which the drivers of these seven groups
were involved during 1989-91. The model or scheme that identified drivers who in
1989-91 had the most accidents was deemed best. All models were compared at Y'
thresholds which produced equal numbers of high-risk drivers.

The second approach evaluated the accuracy of the models in terms of predicting the
subsequent accident status of the subjects (accident-involved versus accident-free).
The false-negative and false-positive rates produced by the models were compared at
a variety of "cut points" in order to evaluate the respective sensitivity and specificity
of the equations in predicting subsequent accident involvement.

The performance of each of the seven schemes (prior citations, prior accidents, and
the five regression models) is presented in Table 8. Several conclusions emerge from
these results. First, to identify drivers with high accident potential, one can do better
than to use either prior citations or prior accidents alone. Second, no one multiple
regression procedure substantially outperforms the others. Third, the larger the pool
of drivers that are considered, the lower is the "yield." For example, among drivers
selected by the Poisson regression model, the 1,000 highest accident-risk drivers
have, on the average, about 0.47 accidents over the subsequent 3-year period, which
is 2.76 times the average (0.17) for the total sample; the next 4,000 have about 0.35
accidents per driver over the subsequent 3-year period; the next 5,000 have about
0.27 accidents per driver over the subsequent 3-year period; and drivers ranking
between 20,000 and 120,000 have 0.16 accidents per driver over the 3-years.

Table 8

Number of Drivers Identified in Each 3-Year (1989-91)
Accident-Risk Strata by Each Model

Model	Drivers estimated by model to be in:					
	Top 1,000	Next 4,000	Next 5,000	Next 10,000	Next 100,000	Total
Prior accidents (1986-88)	406	986	1,234	2,048	14,258	18,932
Prior citations (1986-88)	377	1,210	1,326	2,260	14,124	19,297
Ordinary least squares	473	1,380	1,385	2,317	14,783	20,338
Poisson	471	1,388	1,345	2,307	14,764	20,275
Negative binomial	469	1,390	1,343	2,310	14,770	20,282
Linear probability	467	1,381	1,389	2,338	14,753	20,328
Logistic	463	1,380	1,372	2,311	14,769	20,295

<u>Note</u>. Entries for prior accidents and citations represent the numbers of drivers having the highest
counts of incidents during 1986-88.

The following section provides a comparison of the models in terms of "hits," "false alarms," and "misses" in estimating individual accident involvement.

Predicting individual accident involvement. Multiple regression equations can be used to predict whether or not a driver will be accident-involved in a subsequent period of time. The accuracy of the classification can be summarized in Table 9.

Table 9

Contingency Table of Predicted vs. Actual Outcomes

	Predicted state	
Actual state	Accident-involved	Accident-free
Accident-involved	a (true positive)	b (false negative)
Accident-free	c (false positive)	d (true negative)

This classification table is obtained by accumulating the number of observations for each category. Sensitivity is the proportion of the event (i.e., accident-involved) outcomes that were predicted to be accident involved. Specificity is the proportion of no event (i.e., accident-free) outcomes that were predicted to be no event. The false-positive rate is the proportion of predicted accident outcomes that were observed as no accidents. The false-negative rate is the proportion of predicted no accident outcomes that were observed as accidents.

With perfect prediction, all drivers would be counted in cells a and d, and none would be counted in cells b and c. Drivers counted in cell c are false positives. They are predicted to be accident-involved, but are actually accident-free. Drivers counted in cell b are false negatives. They are predicted to be accident-free, but are actually accident-involved. The desired outcome is to minimize the proportion of drivers in cells b and c and to make fewer errors than would be made in classifying drivers without the prediction equation. To be of any use, the equation must result in more classification accuracy than could be expected by chance alone.

To illustrate the accuracy of the regression equations in predicting the future accident expectancy of individual drivers, a series of four-fold contingency tables were generated displaying the relationship between each individual's predicted and actual accident-involvement frequency.

Tables 10-13 were constructed, each differing in the predicted accident score used for predicting whether a given driver will have an accident. These cutoff scores were selected by generating predicted accident scores from the different equations and then iteratively tabulating the sample using different scores of the predicted values until nearly equal marginal proportions were obtained. The tables summarize the results for the ordinary least squares, Poisson, linear probability, and logistic regression

15

procedures. The cutoff score used in each analysis also produced approximately equal numbers of false-negative and false-positive predictions, as would be expected from the equality of the marginal distributions. The use of equal marginals assigns equal weights to both types of errors and tends to maximize the overall accuracy of classification as represented by the phi coefficient.

Table 10

Predicted 3-Year Accident-Involvement Frequency and
Percentage Using Ordinary Least Squares Regression

Actual accident status	Predicted accident status		Total
	Accident-involved	Accident-free	
Accident-involved	4,609 (3.01%)	15,766 (10.31%)	20,375 (13.32%)
Accident-free	15,762 (10.31%)	116,794 (76.37%)	132,556 (86.68%)
Total	20,371 (13.32%)	132,510 (86.68%)	152,931 (100.00%)
Percent correctly classified	22.63%	88.11%	

Note. A predicted accident rate cutoff of 0.216 was used to equalize marginals. The odds ratio is 2.2, and the phi coefficient is .11.

Table 11

Predicted 3-Year Accident-Involvement Frequency
and Percentage Using Poisson Regression

Actual accident status	Predicted accident status		Total
	Accident-involved	Accident-free	
Accident-involved	4,551 (2.98%)	15,824 (10.35%)	20,375 (13.32%)
Accident-free	15,787 (10.32%)	116,769 (76.35%)	132,556 (86.68%)
Total	20,338 (13.30%)	132,593 (86.70%)	152,931 (100.00%)
Percent correctly classified	22.38%	88.07%	

Note. A predicted accident rate cutoff of 0.200 was used to equalize marginals. The odds ratio is 2.1, and the phi coefficient is .10.

Table 12

Predicted 3-Year Accident-Involvement Frequency
and Percentage Using Linear Probability Regression

| Actual accident status | Predicted accident status | | Total |
	Accident-involved	Accident-free	
Accident-involved	4,587 (3.00%)	15,788 (10.32%)	20,375 (13.32%)
Accident-free	15,709 (10.27%)	116,794 (76.41%)	132,556 (86.68%)
Total	20,296 (13.27%)	132,635 (86.73%)	152,931 (100.00%)
Percent correctly classified	22.60%	88.10%	

Note. A predicted accident rate cutoff of 0.182 was used to equalize marginals. The odds ratio is 2.2, and the phi coefficient is .11.

Table 13

Predicted 3-Year Accident-Involvement Frequency
and Percentage Using Logistic Regression

| Actual accident status | Predicted accident status | | Total |
	Accident-involved	Accident-free	
Accident-involved	4,576 (2.99%)	15,799 (10.33%)	20,375 (13.32%)
Accident-free	15,796 (10.33%)	116,760 (76.35%)	132,556 (86.68%)
Total	20,372 (13.32%)	132,559 (86.68%)	152,931 (100.00%)
Percent correctly classified	22.46%	88.08%	

Note. A predicted accident rate cutoff of 0.175 was used to equalize marginals. The odds ratio is 2.1, and the phi coefficient is .11.

17

Using Table 10 as an example, this table shows a statistically significant association ($p < 0.001$) between predicted and actual accident involvement. Persons predicted to have accidents are approximately 2 times more likely to have accidents than are those predicted to be accident-free ($3.0 \div 13.3 = 22.6\%$ vs. $10.3 \div 86.7 = 11.9\%$). However, the equation failed to correctly predict the majority of accident-involved drivers, as evidenced by the low true-positive rate of 22.6%. Although the false-negative rate ($10.3 \div 86.7 = 11.9\%$) appears low, this percentage of misclassification represents the majority of the 13.3% of the total sample who were truly accident involved.

The phi coefficient and odds ratio, shown at the bottom of each table, are commonly used indices for quantifying the degree of association in contingency tables. The phi coefficient is simply the Pearson correlation coefficient between the actual and predicted accident-status categories. The odds ratio refers to the relative odds of being accident-involved as a function of a predicted accident category. More specifically, the odds ratio is equal to $(P_a \div P_c) \div (P_b \div P_d)$, where P_a, P_b, P_c, and P_d represent the grand percentages in the respective cells.

In Table 10, the odds of predicted accident-involved subjects actually having an accident as opposed to not actually having an accident, are $(3.0\% \div 10.3\%) = 0.2919$. The same odds for the predicted accident-free group are $(10.3\% \div 76.4\%) = 0.1350$. The ratio of these two odds (i.e., the odds ratio) is 2.2. If the odds of having an accident did not vary as a function of the sample's predicted score, the odds ratio would be 1. This would indicate no relationship between the categories. An odds ratio exceeding 1 indicates some relationship between the categories. However, the index has no upper limit and is not a measure of correlation as is the phi coefficient. The fact that the odds ratio and phi coefficient are of modest size in Table 10 indicates that the degree of individual predictive accuracy is low. This is demonstrated by the previously-discussed high false-positive rate and the fact that the equation misclassifies the majority of the accident-involved drivers.

As demonstrated in Tables 10-13, all four multiple regression techniques are almost identical in accuracy of individual prediction. For example, the percent correctly classified as accident-involved ranges between 22.6% for ordinary least squares regression and 22.4% for Poisson regression. Although not shown here, additional contingency tables were produced for the four regression techniques using cutoff-score values that would predict accident involvement for all drivers with accident expectancies of first two or more, and then three or more, standard deviations above the mean. As was the case with equal marginals, the four regression methods produced almost identical results in correctly classifying accident-involved and accident-free drivers.

Sampling Validation Study

In an attempt to investigate the dependence of the preceding results on sample size, an additional study was performed by selecting a 10% (n = 15,348) random sample of the drivers used in the above analyses. For purposes of this additional analysis, equations were produced for the ordinary least squares, Poisson, and logistic regression techniques.

Table 14 displays descriptive statistics for the biographical and driver record variables for the total sample and the 10% sample.

Table 14

Descriptive Statistics for the Total Sample and the 10% Sample

Variable	Total sample (n = 152,931)	10% sample (n = 15,348)
Total accidents (1989-91)		
$\overline{X}$	0.1517	0.1533
SD	0.4138	0.4152
Variance	0.1713	0.1724
Total accidents (1986-88)		
$\overline{X}$	0.1706	0.1680
SD	0.4380	0.4353
Variance	0.1918	0.1895
Total citations		
$\overline{X}$	0.6414	0.6409
SD	1.1964	1.1859
Variance	1.4313	1.4064
Age		
$\overline{X}$	45.67	45.44
SD	15.33	15.17
Variance	234.96	230.26
% class 1/A or 2/B	3.3	3.4
% one or more P&M conditions	1.4	1.4
% one or more restrictions	34.0	33.8
% male	52.4	52.6

In comparing the samples, it is evident that differences between the total and 10% samples on the biographical and driver record variables are very small (less than 4% in absolute value).

Table 15 presents a summary of the regression equations for the reduced sample study. As was the case with the previous analyses, subsequent total accidents was associated with:

19

- Increased prior citation frequency
- Increased prior accident frequency
- Having a commercial driver license
- Being young
- Being male
- Having one or more P&M conditions on record
- Having one or more driver license restrictions on record

Note from Table 15 that the *p* values for the first six coefficients are identical through three digits. Only the *p* values for P&M and restriction status differ.

Table 15

Summary of Nonconcurrent 6-Year (1986-88; 1989-91) Multiple Regression
Equation for Predicting Total Accidents within the 10% Sample Using Ordinary
Least Squares, Poisson, and Logistic Regression Models ($n = 15,348$)

Predictor variable	Ordinary least squares regression				Poisson regression				Logistic regression			
	Regression coefficient	Standard error	F	p	Regression coefficient	Standard error	χ^2	p	Regression coefficient	Standard error	Wald χ^2	p
Constant	0.230	0.016	207.09	.000	-1.245	0.099	158.20	.000	-1.241	0.115	115.62	.000
Prior total citations	0.026	0.003	75.15	.000	0.1023	0.014	56.07	.000	0.133	0.018	54.60	.000
Prior total accidents	0.062	0.008	63.84	.000	0.287	0.037	59.36	.000	0.319	0.048	44.60	.000
License class	0.126	0.019	46.52	.000	0.524	0.083	40.33	.000	0.527	0.108	23.81	.000
Age	-0.001	0.000	37.14	.000	-0.011	0.002	50.00	.000	-0.012	0.002	41.86	.000
Gender	-0.034	0.007	24.66	.000	-0.253	0.044	32.79	.000	-0.241	0.050	22.93	.000
P&M indicator	0.051	0.028	3.30	.069	0.238	0.143	2.75	.097	0.249	0.180	1.91	.166
Restriction status	0.019	0.007	6.42	.011	0.120	0.046	6.98	.008	0.133	0.053	6.31	.012
	F for the equation = 57.06 $p = .000$ $R^2 = .025$				Log likelihood for intercept only = -6765.0565 Log likelihood for full model = -6583.0565 Likelihood ratio test = 365.021, $df = 7$, $p < .000$				-2 Log L for intercept only = 12111.80 -2 Log L for intercept and covariates = 11811.97 Chi-square for covariates = 336.94, $df = 7$, $p < .000$			

Note. The criterion measure, total accidents during 1989-91, had a mean of 0.153 and a standard deviation of 0.415.

Table 16 presents the number of accidents for drivers selected by the various models. Tables 17-19 present the classification of individual drivers into accident-status categories as determined by the regression models. Prediction cutoff scores were selected in order to nearly equalize the marginals.

Table 16

Number of Drivers Identified in Each 3-Year (1989-91) Accident
Risk Strata by Each Model for the 10% Sample

Model	Drivers estimated by model to be in:					
	Top 500	Next 500	Next 1,000	Next 4,000	Next 5,000	Total
Prior accidents (1986-88)	140	124	215	657	639	1,775
Prior citations (1986-88)	145	132	223	729	621	1,850
Ordinary least squares	177	141	230	774	621	1,943
Poisson	177	147	230	750	644	1,948
Logistic	179	148	226	768	623	1,944

Note. Entries for prior accidents and citations represent the numbers of drivers having the highest counts of incidents during 1986-88.

Table 17

Predicted 3-Year Accident Involvement Using
Ordinary Least Squares Regression for the 10% Sample

Actual accident status	Predicted accident status		Total
	Accident-involved	Accident-free	
Accident-involved	461 (3.00%)	1,601 (10.43%)	2,062 (13.43%)
Accident-free	1,601 (10.43%)	11,685 (76.13%)	13,286 (86.57%)
Total	2,062 (13.43%)	13,286 (86.57%)	15,348 (100.00%)
Percent correctly classified	22.36%	87.95%	

Note. A predicted accident rate cutoff of 0.218 was used to equalize marginals. The odds ratio is 2.1, and the phi coefficient is .10.

Table 18

Predicted 3-Year Accident Involvement Using
Poisson Regression for the 10% Sample

| Actual accident status | Predicted accident status | | Total |
	Accident-involved	Accident-free	
Accident-involved	464 (3.02%)	1,598 (10.41%)	2,062 (13.43%)
Accident-free	1,599 (10.42%)	11,687 (76.15%)	13,286 (86.57%)
Total	2,063 (13.44%)	13,285 (86.56%)	15,348 (100.00%)
Percent correctly classified	22.49%	76.15%	

Note. A predicted accident rate cutoff of 0.204 was used to equalize marginals. The odds ratio is 2.1, and the phi coefficient is .11.

Table 19

Predicted 3-Year Accident Involvement Using
Logistic Regression for the 10% Sample

| Actual accident status | Predicted accident status | | Total |
	Accident-involved	Accident-free	
Accident-involved	462 (3.01%)	1,600 (10.42%)	2,062 (13.43%)
Accident-free	1,596 (10.40%)	11,690 (76.17%)	13,286 (86.57%)
Total	2,058 (13.41%)	13,290 (86.59%)	15,348 (100.00%)
Percent correctly classified	22.45%	7.96%	

Note. A predicted accident rate cutoff of 0.178 was used to equalize marginals. The odds ratio is 2.1, and the phi coefficient is .10.

As was the case with the previous analyses, the regression methods produced similar results in terms of driver selection and percent correctly classified into accident-involved and accident-free categories. The results of this validation analysis closely parallel the previous findings, providing substantiation for the robustness and reliability of the findings with sample sizes much smaller than the original N.

DISCUSSION

The results of the present analyses are consistent with those of prior research (e.g., Gebers & Peck, 1994; Peck & Gebers, 1992; Peck & Kuan, 1983). For example,

it was shown in all the models that increased accident involvement was associated with increased prior citation and accident frequencies, possessing a commercial driver license, being young, being male, having a medical condition on record, and having a driver license restriction on record.

Any generalization about driving performance from the present analyses is limited by the absence of exposure data (e.g., miles driven) and territorial data (e.g., driver record by ZIP Code and U.S. census variables). Exposure and territorial variables not available from the driver record file have been collected and will be analyzed in the next report.

Results presented in this paper indicate that with these data, the use of different regression techniques do not lead to any greater increase in individual accident prediction beyond that obtained through application of ordinary least squares regression. It therefore appears safe to employ OLS multiple regression techniques on driver accident-count distributions of the type represented by California driver records, at least when Ns are extremely large. This conclusion is consistent with those contained in Peck, McBride, and Coppin (1971) and Peck and Kuan (1983). Further asymptotic justifications for the use of parametric models on highly skewed accident count data can be found in DeYoung (1995) and Gebers, DeYoung, and Peck (1997). In fact, a series of follow-up analyses to the present findings provide support for the robustness of the parametric ordinary least squares technique in the presence of extreme skewness with Ns as small as 2,500 (Gebers, in press). The results of these analyses indicate that the use of different regression techniques on smaller sample sizes do not lead to any lessor or greater of an increase in individual accident prediction beyond that obtained through application of ordinary least squares regression.

In future reports, the statistical interaction between predictor variables (e.g., how the relationship between subsequent accidents and age varies as a function of the prior number of citations) will also be examined. The subsequent analyses will also include the following:

- Regressions on concurrent and nonconcurrent 3, 6, 9, 14, and 20-year samples.

- Regressions using accident sub-type criteria, such as single-vehicle accidents, fatal/injury accidents, police-reported accidents, and culpable accidents.

- Regressions using an expanded set of predictors that will include individual violation types (e.g., speeding, DUI, following too close) and additional driver licensing variables (e.g., type of restriction, limited-term license, vision referral, and number of months the driver license is suspended or revoked).

The objective of these future studies will be to provide driver license officials, epidemiologists, traffic safety researchers, and organizations involved in risk management and assessment with actuarial data on driver accident risk profiles.

REFERENCES

Boyer, M., Dionne, G., & Vanasse, C. (1990). *Econometric models of accident distributions.* University of Montreal: Center for Research on Transportation.

Davis, C. S. (1990). *The driver education demonstration project: Analysis of its long-term effect.* Georgia: DeKalb County.

Dean, C., & Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association, 84*, 467-472.

DeYoung, D. J. (1995). *An evaluation of the effectiveness of California drinking driver programs.* (Report No. 146). Sacramento: California Department of Motor Vehicles.

Draper, N. R., & Smith, H. (1966). *Applied regression analysis: Wiley series in probability and mathematical statistics.* John Wiley & Sons, Inc.

Gebers, M. A. (In press). *Follow-up analysis to paper exploratory multivariable analyses of California driver record accident rates.* Sacramento: California Department of Motor Vehicles.

Gebers, M. A., DeYoung, D. J., & Peck, R. C. (1997). The impact of mail contact strategy on the effectiveness of driver license withdrawal. *Accident Analysis and Prevention, 29,* 65-77.

Gebers, M. A., & Peck, R. C. (1994). *An inventory of California driver accident risk factors.* (Report No. 144). Sacramento: California Department of Motor Vehicles.

Grogger, J. (1990). The deterrent effect of capital punishment: An analysis of daily homicide counts. *Journal of the American Statistical Association, 85*, 295-302.

Hilbe, J. (1994). *Log negative binomial regression using the GENMOD procedure SAS/STAT software, proceedings of SUGI 19.* Cary, NC: SAS Institute Inc.

Kleinbaum, D. G., Kupper, L. L., & Muller, K. E. (1988). *Applied regression analysis and other multivariable methods.* Boston: PWS - KENT Publishing Company.

Maddala, G. S. (1991). *Limited dependent and qualitative variables in econometrics.* Cambridge: Cambridge University Press.

Miaou, S., Lu, A., & Lum, H. S. (1996). Pitfalls of using R^2 to evaluate goodness of fit of accident prediction models. *Transportation Research Record, 1542*, 6-13.

Peck, R. C., & Gebers, M. A. (1992). The California driver record study: *A multiple regression analysis of driver record histories from 1969 through 1982.* Sacramento: California Department of Motor Vehicles.

Peck, R. C., & Kuan, J. (1983). A statistical model of individual accident risk prediction using driver record, territory and other biographical factors. *Accident Analysis and Prevention, 15*, 371-393.

Peck, R. C., McBride, R. S., & Coppin, R. S. (1971). The distribution and prediction of driver accident frequencies. *Accident Analysis and Prevention, 2*, 243-299.

Rawlings, J. O. (1988). *Applied regression analysis: A research tool.* Pacific Grove, CA. Wadsworth & Brooks/Cole Advanced Books & Software.

SAS Institute Inc. (1993). *SAS/STAT® technical report P-243, SAS/STAT® software: The GENMOD procedure, release 6.09.* Cary, NC: SAS Institute Inc.

Stock, J. R., Weaver, J. K., Ray, H. W., Brink, J. R., & Sadox, M. G. (1983). *Evaluation of safe performance secondary school driver education curriculum demonstration projects.* Washington, DC: National Highway Traffic Safety Administration (NTIS No. DOT-HS-6-01462).