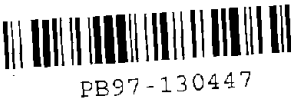
U.S. Department
of Transportation
Federal Highway
Administration



Publication No. FHWA-RD-96-040
December 1996

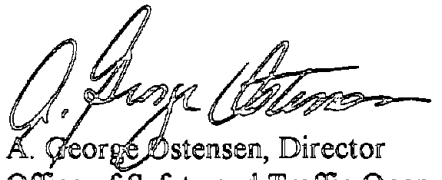
Measuring the Goodness-of-Fit of Accident Prediction Models

Research and Development
Turner-Fairbank Highway Research Center
6300 Georgetown Pike
McLean, Virginia 22101-2296

FOREWORD

This report contains research findings on the strengths and limitations of using several popular statistical measures to evaluate the goodness-of-fit of accident prediction models. The results in this report will be of interest to those concerned with quantifying the relationship between safety and highway design and operations.

The study demonstrated the pitfalls of using the coefficient of determination (R^2) to determine the quality of accident prediction models. Other popular goodness-of-fit measures were also introduced and evaluated. Alternative measures for safety research were developed, tested, and recommended. In addition, the interrelationship between the accident-based approach and the encroachment-based approach for developing the association between run-off-the-road accidents and roadside hazards was studied. It was shown that exploring the complementary nature of these two approaches could be a viable avenue to reducing data collection cost.



A. George Ostensen, Director
Office of Safety and Traffic Operations
Research and Development

NOTICE

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-RD-96-040		

4. Title and Subtitle	5. Report Date
Measuring the Goodness-of-Fit of Accident Prediction Models	December 1996
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Shaw-Pin Miaou	

9. Performing organization Name and Address	10. Work Unit No. (TRAIS)
Center for Transportation Analysis Oak Ridge National Laboratory P.O. Box 2008, MS-6366 Oak Ridge, Tennessee 37831	A3b
	11. Contract or Grant No. DTFH-61-94-Y-00107

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Office of Safety and Traffic Operations R&D Federal Highway Administration 6300 Georgetown Pike McLean, VA 22101-2296	Final Report July 1994 - November 1995
	14. Sponsoring Agency Code

15. Supplementary Notes

Contracting Officer's Technical Representative - Harry Lum (Jul. 1994-Oct. 1994) and Mike Griffith (Nov. 1994-Jan. 1996) (HSR-20)

16. Abstract

In developing accidents-flow-roadway design models, the R^2 goodness-of-fit measure has been used by traffic safety engineers and researchers for many years to (1) determine the quality and usability of a model; (2) select covariates (or explanatory variables) for inclusion in the model; (3) make a decision as to whether it would be worthwhile to collect additional covariates; and (4) compare the relative quality of models from different studies. Through computer simulations, this study demonstrated the pitfalls of using R^2 to make these decisions and comparisons. Other goodness-of-fit criteria such as the Akaike Information Criterion, scaled deviance, and Pearson's X^2 statistics were also introduced and evaluated. Based on limited simulation results, one of the alternative criteria called R^2_a was recommended for evaluating and comparing the quality of accident prediction models when sample size is large. Finally, the interrelated and complementary nature of two approaches that have traditionally been used to develop the relationship between run-off-the-road accident frequency and roadside hazards (i.e., accident-based approach and encroachment-based approach) were studied and demonstrated using data from an Federal Highway Administration and Transportation Research Board roadway cross-section design data base. It was suggested that exploring the complementary nature of these two approaches could be a viable avenue to reduce data collection cost.

17. Key Words	18. Distribution Statement
Accident prediction model, Poisson regression, Negative binomial regression, Goodness-of-fit, Coefficient of determination, Highway geometric design	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of pages	22. Price
Unclassified	Unclassified	130	

FORM DOT F 1700.7(8-72)Reproduction of completed page authorized

SI* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
	LENGTH			
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
	AREA			
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yards	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
	VOLUME			
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
NOTE: Volumes greater than 1000 l shall be shown in m^3.				
	MASS			
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
	TEMPERATURE (exact)			
°F	Fahrenheit temperature	5(F-32)/9 or (F-32)/1.8	Celcius temperature	°C
	ILLUMINATION			
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	candela/m^2	cd/m^2
	FORCE and PRESSURE or STRESS			
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

APPROXIMATE CONVERSIONS FROM SI UNITS

Symbol	When You Know	Multiply By	To Find	Symbol
	LENGTH			
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
	AREA			
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
	VOLUME			
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.71	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
	MASS			
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
	TEMPERATURE (exact)			
°C	Celcius temperature	1.8C + 32	Fahrenheit temperature	°F
	ILLUMINATION			
lx	lux	0.0929	foot-candles	fc
cd/m^2	candela/m^2	0.2919	foot-Lamberts	fl
	FORCE and PRESSURE or STRESS			
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

* SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380.

(Revised September 1993)

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AADT	Average Annual Daily Traffic
AASHTO	American Association of State Highway and Transportation Officials
AIC	Akaike Information Criterion
CAIC	Corrected Akaike Information Criterion
EB	Empirical Bayes
FHWA	Federal Highway Administration
GLM	Generalized Linear Models
HPMS	Highway Performance Monitoring System
HSIS	Highway Safety Information System
iid	Independent and Identically Distributed
IRLS	Iteratively Reweighted Least Squares
K-L	Kullback-Leibler
ML	Maximum Likelihood
MQL	Maximum Quasi-Likelihood
NB	Negative Binomial
NCHRP	National Cooperative Highway Research Program
OLS	Ordinary Least Squares
PDO	Property Damage Only
RORA	Run-Off-the-Road Accidents
SD	Scaled Deviance
SV	Single Vehicle
TRB	Transportation Research Board

1. INTRODUCTION

Traffic accidents-flow-roadway geometric design relationships have been modeled for over 30 years by traffic safety engineers and researchers to estimate and predict the traffic accident frequency or rate under different traffic flow and geometric design conditions. The coefficient of determination, or R^2, has traditionally been used as a criterion to determine how well the developed models fit the observed accident data [e.g., see Roy Jorgensen Associates, 1978, for many earlier studies; Council et al., 1980; Zegeer et al., 1987; Okamoto and Koshi, 1989; Zegeer et al., 1990; Joshua and Garber, 1990; Mohamedshah et al., 1992; and Belanger, 1994]. It is commonly believed that R^2 is always bounded between 0 and 1 and the higher the R^2 value, the better the fit; conversely, the lower the R^2 value, the poorer the fit.

Specifically, in developing accidents-flow-roadway design models, R^2 has been used by traffic safety engineers and researchers to:

1. Determine the quality and usability of a model:

It is thought by many that, for any given data set, the R^2 value of the developed accident model always has a lower bound of 0 and a upper bound of 1. Based on this common notion, a model with a R^2 value of, say, 0.7 or less is often considered as a poor model and not recommended for use.

2. Select covariates (or explanatory variables) for inclusion in the model:

Conceptually, for a given data set, we would like to have the R^2 value of a model be reduced when an unworthy covariate, which explains very little variation of the accident frequencies among studied sites, is added to the model. This property of a goodness-of-fit measure would enable us to distinguish the unworthy covariates from the rest of the covariates. Statistically speaking, dropping the unworthy covariates from the model would enable us to reduce the uncertainty of our predictions when the model is used. The adjusted R^2, denoted by $\tilde{R}^2$, is a modified measure of R^2 which allows the total number of degrees of freedom in the model to be reflected in R^2. In developing accident prediction models, $\tilde{R}^2$ has been used to decide which covariates should be included in a model. Typically, the model which includes a subset of all candidate covariates and gives the largest $\tilde{R}^2$ value is considered the best model.

3. Make a decision as to whether it would be worthwhile to collect additional covariates:

Based on the notion that the upper bound of R^2 for accident prediction models is always 1, many would use $(1-R^2)$ as a measure of potential improvement that one might be able to achieve by collecting additional covariates. For example, if the current model has an R^2 value of 0.95, then based on $(1-R^2)=0.05$, one might decide that collecting additional covariates will not pay off. On the other hand, if the current model has an R^2 value of 0.45, based on $(1-R^2)=0.55$, one may be led to think that there is still a lot of room for improvement and collecting additional covariates is likely to pay off. Of course, the

decision discussed here is purely based on statistical consideration. Engineering judgment usually plays an important role in the decision-making process as to whether the current model makes good engineering sense and which additional covariates should be collected if a data collection effort is to be made.

4. Compare the relative quality of models from different studies:

Many accident prediction models and their R^2 values have been reported in the last 30 years. When comparing the relative quality of these models, many traffic safety engineers and researchers tend to favor the models with high R^2 values regardless of the fact that different localities, accident types, time periods, sample size, and covariates have been considered in different studies.

The pitfalls of using R^2 (or $\bar{R}^2$) to assess the quality of a model and to make the decisions and comparisons discussed above have been discussed in some statistical literature [e.g., Barrett, 1974; Kvålseth, 1985; Scott and Wild, 1991; Willett and Singer, 1988; Cox and Wermuth, 1992; Anderson-Sprecher, 1994] and a recent safety research paper [Brüde and Larsson, 1993]. To the best of this author's knowledge, no systematic demonstration of these pitfalls has been reported, especially for the applications in highway safety research. Also, many traffic safety engineers and researchers are not aware of these pitfalls and have continued to use R^2 as a main goodness-of-fit measure for accident prediction models.

To give an example of the pitfalls, Cox and Wermuth [1992] showed in the context of binary response regression models that, for an exemplary data distribution, the upper bound of the R^2 value for a perfect model could be much lower than 1. One important implication of their example is that a model with a low R^2 value does not necessarily mean that the fit is a poor one. *[Note that in this report a perfect model is referred to as a model that: (1) has specified a correct probability distribution for the dependent variable; (2) has chosen a correct functional form which describes the relationship between the expected number of accidents and associated covariates; (3) has included all necessary covariates; and (4) has correctly estimated each model parameter.]*

Another example of the pitfalls can be found in Brüde and Larsson [1993] in which they showed that the R^2 value of the Poisson regression models is dependent on the mean level of the dependent variable, i.e., the mean level of accident frequency. Essentially, it was shown that higher mean accident levels would result in higher R^2 values regardless of the quality of the model. This is one of the reasons why the R^2 values of accident prediction models for urban areas were usually reported to be higher than those for rural areas. Also, this is a main reason why the R^2 values of those models developed for data with high aggregation levels (with respect, e.g., to accident type, the length-of-time periods, or the length-of-road sections) were reported to be higher than those models developed for disaggregate data. Because traffic accident events are known to follow some Poisson type of distributions, this example suggests that, in general, we would not be able to compare the quality of models from different studies using R^2 values if these studies were performed for different localities, accident types, or length-of-time periods.

OBJECTIVES

The objective of this study was threefold. The main objective was to demonstrate to traffic safety engineers and researchers the potential pitfalls of using R^2 (or $\tilde{R}^2$) to determine the goodness-of-fit of accident prediction models. This objective was accomplished through computer simulations of previously used accident prediction models, including normal linear regression, lognormal regression, Poisson regression, and negative binomial regression models. The last two types of regression models are commonly used in recent studies to describe traffic accidents-flow-roadway geometric design relationships.

In the last 20 years, an alternative model selection criterion called the Akaike Information Criterion (AIC) has been developed by statisticians [see e.g., Bozdogan, 1987]. The capability of this criterion to select the correct models has traditionally been shown in a linear regression or time series context in statistical literature [e.g., Hurvich and Tsai, 1989] and just recently in a logistic regression context [Hurvich and Tsai, 1994]. Also, this criterion has been coded as one of the outputs in some of the new statistical software packages. However, few traffic safety engineers and researchers are aware of the development of this criterion. The second objective of this study was, therefore, to bring the latest development of AIC to the attention of traffic safety engineers and researchers. This objective was achieved through some illustrations of the power of AIC-based criteria in model selection. Again, the illustrations were carried out using computer simulations. It was hoped that through the simulation studies the strengths and limitations of AIC-based criteria in evaluating the goodness-of-fit of accident prediction models could become clearer to traffic safety engineers and researchers. In addition to AIC, other criteria such as likelihood-ratio based criterion and Pearson's X^2 statistics were also considered in the illustration.

Based on lessons learned from the simulations above, the third objective was to suggest the type of models that is appropriate for the prediction of run-off-the-road accidents, and to discuss the merits and shortcomings of the model as applies to the prediction of run-off-the-road accidents and vehicle roadside encroachments that may lead to run-off-the-road accidents. This study relates to one of the major tasks in developing accident prediction models, which is to determine an appropriate functional form that describes the accidents-flow-roadway design relationship.

SCOPE OF WORK

The focus of this study was on two types of statistics: R^2 and AIC-based criteria. There are, however, many statistical tests and criteria other than R^2 and AIC that are available for testing the quality of a model. For example, the t-statistic of an estimated parameter can be used to assess the significance of the association between a covariate and traffic accidents, and some score test statistics developed in recent statistical literature can be used to decide the adequacy of Poisson assumption [see e.g., Miaou et al., 1993]. All these tests and criteria, when appropriate, can be and should be used to better determine the quality of accident prediction models.

Because the resources available for this study were limited, the research performed was intended to be exploratory and illustrative in nature. It is hoped, however, that this study could shed some light on future research needs in accident prediction modeling and more indepth follow-on studies will be conducted in the near future.

REPORT ORGANIZATION

Chapter 2 gives a review of the state of the development of accident prediction models in describing traffic accidents-flow-roadway geometric design relationships. Chapter 3 discusses the concept of R^2 and illustrates the pitfalls of using the R^2 values to make the decisions and comparisons described earlier. These pitfalls are illustrated using computer simulations. Chapter 4 introduces AIC and other criteria, such as the likelihood-ratio based criterion and Pearson's X^2 statistics. The capability and limitations of these criteria to select the correct model is demonstrated through various computer simulations. Chapter 5 examines the property of three possible alternative measures that have the potential of overcoming some of the limitations of R^2 and AIC. Chapter 6 investigates the relationship between two modeling approaches that have traditionally been used to evaluate roadside safety and discusses the merits and shortcomings of the models as applies to the prediction of run-off-the-road accidents and vehicle roadside encroachments that may lead to run-off-the-road accidents. Run-off-the-road accidents and roadway data for rural two-lane undivided roads from a roadway cross-section design data base [Hummer, 1986], administered by the Federal Highway Administration (FHWA) and the Transportation Research Board (TRB), were used to facilitate the discussion. Chapter 7 concludes this study by providing some suggestive directions for future research.

2. ACCIDENT PREDICTION MODELS

This chapter gives a review of the current state of the development of accident prediction models in describing traffic accidents-flow-roadway geometric design (or simply accidents-flow-roadway design) relationships. The purpose of this review is to set the stage for the discussion of the concept of R^2 and AIC and their simulation studies in the next few chapters. First, the concept of variation in accident frequency is introduced. Second, major statistical and engineering considerations in developing accident prediction models are discussed. Third, commonly used accident prediction models, as presented in recent research papers, are introduced. Fourth, potential areas where accident prediction models may be misspecified are discussed. Finally, some additional observations on the current status of the development of accident prediction models for roadway planning and design are offered.

For ease of exposition, in the following discussion, "accidents" refers to vehicle accidents of a particular vehicle and severity type, "accident frequency" refers to either the number of accidents or the number of vehicles involved in accidents (depending on interest), and "site" refers to a road section (including both mainline and roadside), an intersection, or a ramp. In addition, for the interest of this study, "accident prediction models" refers only to those statistical models that describe accidents-flow-roadway geometric design relationships.

VARIATION OF ACCIDENT FREQUENCY

Vehicle accidents are complex events involving the interactions of five major factors: the drivers, the traffic, the road, the vehicles, and the environment (e.g., weather and lighting conditions). Developing accident prediction models is a means of summarizing these complicated interactive effects based on information contained in the data, as well as our engineering judgment and analytical assumptions about the accident process. Essentially, through modeling, we attempt to explain why accident frequency is different from one site to another (the so-called between-site variation) and from one time interval to another (the so-called between-time variation). It is believed that a significant part of the variations in accident frequency is due to the differences of these five major factors among sites and time intervals. Conceptually, once the relationships between the variation of accident frequency and these five factors are established through models, one can use the models to devise cost-effective means or regulatory policies to affect or change some of these five factors in such a way that accidents will be reduced in the long run.

Figure 1 illustrates the concept of variation of accident frequency. This concept is helpful in understanding the construction of R^2 and in interpreting R^2 values for real-world problems. Figure 1(a) shows that the total variation of accident frequency, Y, can be decomposed into two components: systematic variation and random variation (or, mathematically, the variation of $f(X, Z, U; \beta)$ and the variation of ϵ, respectively). The between-site and between-time variations discussed above are part of the systematic variation. Conceptually, we can think of systematic variation as the variation of long-term means among different sites and time intervals. Hypothetically speaking, if we can repeat the accident process over and over again while keeping

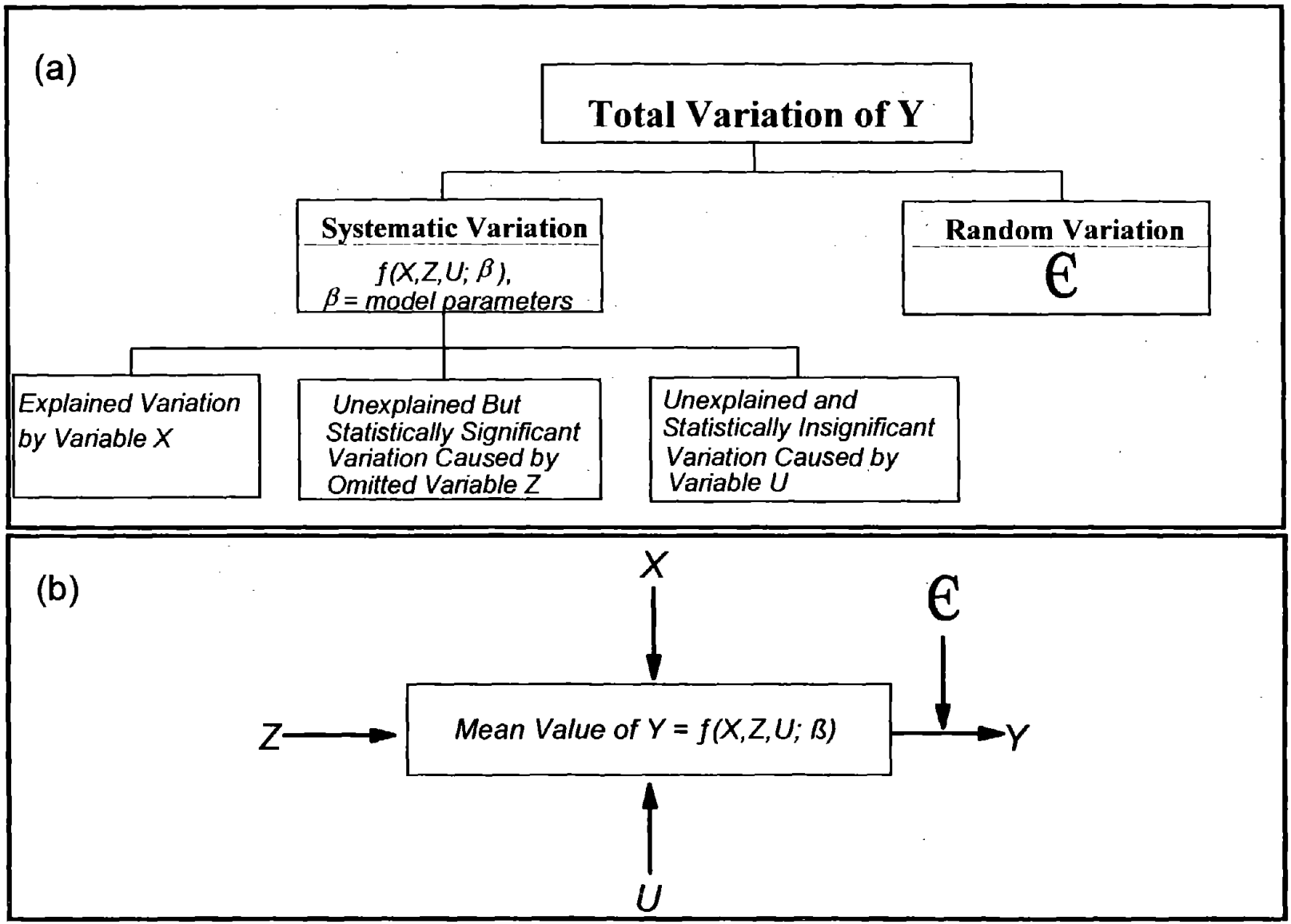


Figure 1. Total explained, unexplained, and random variations of accident frequencies among sites and time intervals.

the five major factors the same for each site and time interval, we can observe accident frequency for each site and time interval over and over again. This replication would then allow us to compute the long-term mean value of accident frequency for each site and time interval. The variation of these mean values among sites and time intervals is the systematic variation. The variation of accident frequencies observed from various replications about the "long-term mean" at each site and time interval is the random variation (which can also be called within-site/within-time variation). Of course, this replication cannot be conducted in real-world and these long-term means and within-site/within-time variation cannot really be observed and are traditionally estimated using the regression analysis.

The random variation (or within-site/within-time variation) can be thought of as the variation that is beyond our explanation. Statistically, it is believed that the random variation follows certain probability law and can be characterized by a probability (mass) function. In traffic accident modeling, through analytical argument and analysis of accident data, it is commonly believed that this random variation can be well characterized by a Poisson probability mass function. Note that the analytical argument can be found in most of the statistics textbooks where the derivation of the Poisson distribution from the binomial distribution is discussed.

The systematic variation in figure 1(a) is further decomposed into three components: (1) explained variation by variable X; (2) unexplained but statistically significant variation caused by omitted or missing variable Z; and (3) unexplained and statistically insignificant variation caused by variable U. This decomposition is a recognition of the reality of accident modeling where we usually do not have all the variables or information that we need on the five major factors to explain the variation of accident frequencies among sites and time intervals. For example, in developing accidents-flow-roadway models, for each investigated site we may be able to collect some traffic and vehicle variables, such as traffic volume, truck-car mix, turning movements (for intersections), and roadway geometric design variables for mainlines and roadside, such as lane width, horizontal curvature, vertical grade, shoulder width and type, median width and type, sideslope, and clear roadside recovery distance. Often, these data are available over several time intervals (typically recorded by year). It is unlikely, however, that one will be able to know whether there are more drivers that are accident-prone or careless on some of the sites than other sites (except in some cases where we may have data on the locations of, e.g., some night clubs that serve alcohol). It is also unlikely that we will be able to know whether vehicles of some sites are better equipped with safety equipments than vehicles of other sites. Similarly, there are many other human factors that are difficult to quantify by site, e.g., the effect of law enforcement level, the familiarity of the drivers with the sites, driver's age distribution, etc. Environmental conditions to some extent can be assumed to be the same at different sites for each time interval if the area under consideration is not too wide. (Otherwise, area dummy variables can be used to capture the area variations or effects.) However, these environmental conditions can vary significantly from one time interval to another. For example, the number of snow storms and rainy days can vary significantly from one year to another. Generally speaking, even though data availability varies from study to study, in developing accident prediction models, major traffic and roadway variables and some of the environmental variables are available (the X's and U's) and the driver and vehicle variables which are known to be important variables are largely unavailable (the omitted variables Z's).

What follows is an illustration of the concept of variation using a conventional normal linear regression model. It should be emphasized that normal linear regression models have been shown to be inadequate for modeling accident-flow-roadway design relationships [see e.g., Miaou and Lum, 1993]. Here it is used purely for illustration purpose. As will be discussed later, accident prediction models are discrete, non-normal, nonlinear, and interactive in nature.

Let's consider a process which is generated from a normal linear regression model as follows:

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \epsilon_i \qquad (1)$$

where Y_i are dependent variables; $x_{i1}, x_{i2}, \ldots x_{i6}$ are observed values of covariates (or explanatory variables) X_{ij}, $j=2,3,\ldots,6$, respectively; ϵ_i are independent and identically distributed (*iid*) normal random variable with mean 0 and constant variance σ^2; and β's are constant regression parameters. Collectively, x_{ij}, $j=1,2,\ldots,6$, will be denoted by x_i and X_{ij}, $j=2,3,\ldots,6$, denoted by X_i. Other common assumptions associated with linear regression models, such as x_i's are measured without errors, ϵ_i's do not depend on x_i, and x_i's do not collinear with each another, are also assumed for the model in Eq. (1). For ease of exposition, let's further simplify the model as follows: (1) set all β_j, j = 1,2,3, ..,6, equal to 1, denoted simply by β=1; (2) set the first covariate x_{i1}=1 (i.e., X_{i1} is a dummy variable equal to 1 and β_1 is an intercept); and (3) assume that X_{ij}, $j=2,3,\ldots,6$, are independent normal random variables, each of which has zero mean and constant variance of 1, denoted as $X_{ij} \sim iid\ N(0,1)$ for $j=2,3,4,5,6$ or collectively, without confusion, simply $X_i \sim iid\ N(0,1)$. Under these conditions, the model can now be expressed in a simpler form as:

$$Y_i = 1 + x_{i2} + x_{i3} + x_{i4} + x_{i5} + x_{i6} + \epsilon_i \qquad (2)$$

For this illustration, the variance of a random variable is used to measure its variation. As will be seen later, variance is the underlying measure of variation in R^2. Note that conditional and unconditional expectation and variance of random variables will be used quite often to compute the R^2 values in this illustration and other illustrations in the next chapter. For reference on the definition and statistical properties of conditional and unconditional expectation and variance, the readers are referred to basic statistics and probability textbooks, such as Ross [1989] and Rohatgi [1976]. For example, one important relationship between conditional and unconditional variance that has been shown in many textbooks is: the unconditional variance of Y_i can be broken down into two components, $Var[Y_i] = E[Var[Y_i|X_i]] + Var[E[Y_i|X_i]]$. When X_i represents all covariates necessary to explain Y_i, then the first component, $E[Var[Y_i|X_i]]$, is the random variance, i.e., the variance that is beyond the explanation of covariates, and the second component, $Var[E[Y_i|X_i]]$, is the systematic variance.

For the particular model in Eq. (2), the conditional mean of Y_i given the observed values x_i is $E[Y_i|x_i] = E[1+x_{i2}+x_{i3}+x_{i4}+x_{i5}+x_{i6}+\epsilon_i] = 1+x_{i2}+x_{i3}+x_{i4}+x_{i5}+x_{i6}$, where the

expectation is taken over ϵ_i. The conditional variance of Y_i given the observed values x_i is $Var[Y_i | x_i] = E[(Y_i - E[Y_i | x_i])^2 | x_i] = E[\epsilon_i^2 | x_i] = \sigma^2$. That is, the conditional mean of Y_i varies with the observed values x's, while the conditional variance is a fixed constant σ^2.

The unconditional mean of Y_i, denoted by $E[Y_i]$, is equal to 1. That is, $E[Y_i] = E[E[Y_i|X_i]] = E[E[1+X_{i2} + X_{i3} + X_{i4} + X_{i5} + X_{i6} + \epsilon_i | X_i]] = 1$, where the expectation is first taken over ϵ_i and then over all X_i's. Using the conditional-unconditional variance relationship described above, one can show that the unconditional variance of Y_i, denoted by $Var[Y_i]$, in Eq. (2) is composed of a random variance σ^2 and a systematic variance caused by the variation of X through the long-term mean $f(X_i; \beta=1)$ which is equal to $1+X_{i2}+X_{i3}+X_{i4}+X_{i5}+X_{i6}$. To elaborate, the systematic variance is computed as: $Var[f(X_i; \beta=1)] = Var[E[Y_i|X_i]] = Var[1+X_{i2}+X_{i3}+X_{i4}+X_{i5}+X_{i6}] = 5$, where the variance is taken over all X_i's; while the random variance is computed as: $E[Var[Y_i|X_i]] = E[\epsilon_i^2|X_i] = \sigma^2$.

In Eq. (2), if all X's are available (no omitted variables) and all parameters correctly estimated with no sampling variations (i.e., the estimated parameters $\hat{\beta}$ are 1 with no uncertainty), then the total explained variance of the model is the systematic variance which is equal to 5. Now, consider a situation in Eq. (2) where there is one omitted variable. Without loss of generality, let's say X_{i6} is not available for developing the model, i.e., $Z=X_{i6}$. The model under consideration is now:

$$Y_i = 1 + x_{i2} + x_{i3} + x_{i4} + x_{i5} + \epsilon_i' \qquad (3)$$

where ϵ_i' is the sum of X_{i6} and random error ϵ_i which are now indistinguishable from each other. Again, assume that all parameters of the available covariates are correctly estimated with no uncertainty. The total explained variance by available X is now $Var[E[Y_i|X_{i2},X_{i3},X_{i4},X_{i5}]] = Var[1+X_{i2}+X_{i3}+X_{i4}+X_{i5}] = 4$. In the same token, if there are two, three, four, and five omitted variables, the total explained variances become 3, 2, 1, and 0, respectively.

Let's now adopt the traditional definitions of R^2 as follows:

$$R^2 = \frac{\textit{Total Explained Variance of Y by Available X}}{\textit{Total Unconditional Variance of Y}} = 1 - \frac{\textit{Total Unexplained Variance of Y}}{\textit{Total unconditional Variance of Y}} \qquad (4)$$

Under Eq. (2), for a perfect model, which has correct probability function (i.e., normal probability density function), correct functional form, no omitted variable, and all parameters correctly estimated with no uncertainty, we have $R^2 = 5/(5+\sigma^2)$. Furthermore, when $\sigma^2=1$, 5, and 20, R^2 values for the perfect model are 5/6, 1/2, and 1/5, respectively. This means that for a perfect model the best R^2 value that can be achieved is always less than 1 unless there is no

random error (i.e., $\sigma^2 = 0$). Figure 2 gives an illustration of different R^2 values under Eq. (2) with different numbers of omitted variables and σ^2 values. Again, it is assumed in all cases that all parameters of the available covariates are correctly estimated. One observation that can be made from this figure is that the R^2 value increases linearly as the covariate is added to the model one at a time. As will be seen later, this is not the case in nonlinear models such as the typical Poisson and negative binomial regression models.

The normal linear regression illustration above ignores the effect of sampling variations (or sampling errors) caused by the use of finite samples. In practice, the number of data available for developing a model is always finite. Under a finite sample, variable U may exist. In addition, even if we could specify probability function and functional form of a long-term mean correctly and include all the necessary covariates, it is not possible to estimate parameters precisely. The only assurance the analysts have when using the so-called consistent estimators, such as the maximum likelihood (ML) estimator, is that as sample size increases, the uncertainty of the estimated parameters decreases. Chapter 3 will illustrate the effect of sampling errors on R^2 values and extend the illustration above to more commonly used accident prediction models that are non-normal in probability function and nonlinear in functional form. Note that the role of statistically insignificant variables *U's* in developing accident prediction models will be discussed in the next section.

MAJOR STATISTICAL AND ENGINEERING CONSIDERATIONS

In developing accident prediction models, there are five main tasks that require both statistical and engineering considerations: (1) find a good probability (mass) function, P(.), to describe the random variation of ϵ; (2) determine an appropriate functional form and parameterization, i.e., $f(.\ ;\ \beta)$, to describe the effects of variables *X's* and *U's* on the long-term means; (3) select the right variables to include in $f(.\ ;\ \beta)$, i.e., to find appropriate *X's* and eliminate *U's*; (4) estimate the regression parameters (or coefficients) β in $f(.\ ;\ \beta)$ and obtain good statistical inferences for the estimated parameters based on available data; and (5) assess the quality of the model, judge whether the model make good engineering sense, decide whether the developed model meets the planning and design requirements, and identify cost-effective ways to improve the model. Note that even though the role of sample size, *n*, is not specifically indicated in these tasks, it is crucial in every aspect of the modeling process. Also, in accomplishing these five tasks, the potential impact of omitted variables *Z's* should always be kept in mind. Although data quality issues, especially underreporting and misrecording of accidents, are out of the scope of this study, they are as important as developing models.

In this report, accident prediction models refer to the totality of the model which includes the probability function, P(.), the functional form and regression parameters, $f(.\ ;\ \beta)$, and the variables, *X's*, which are selected for inclusion in $f(.\ ;\ \beta)$.

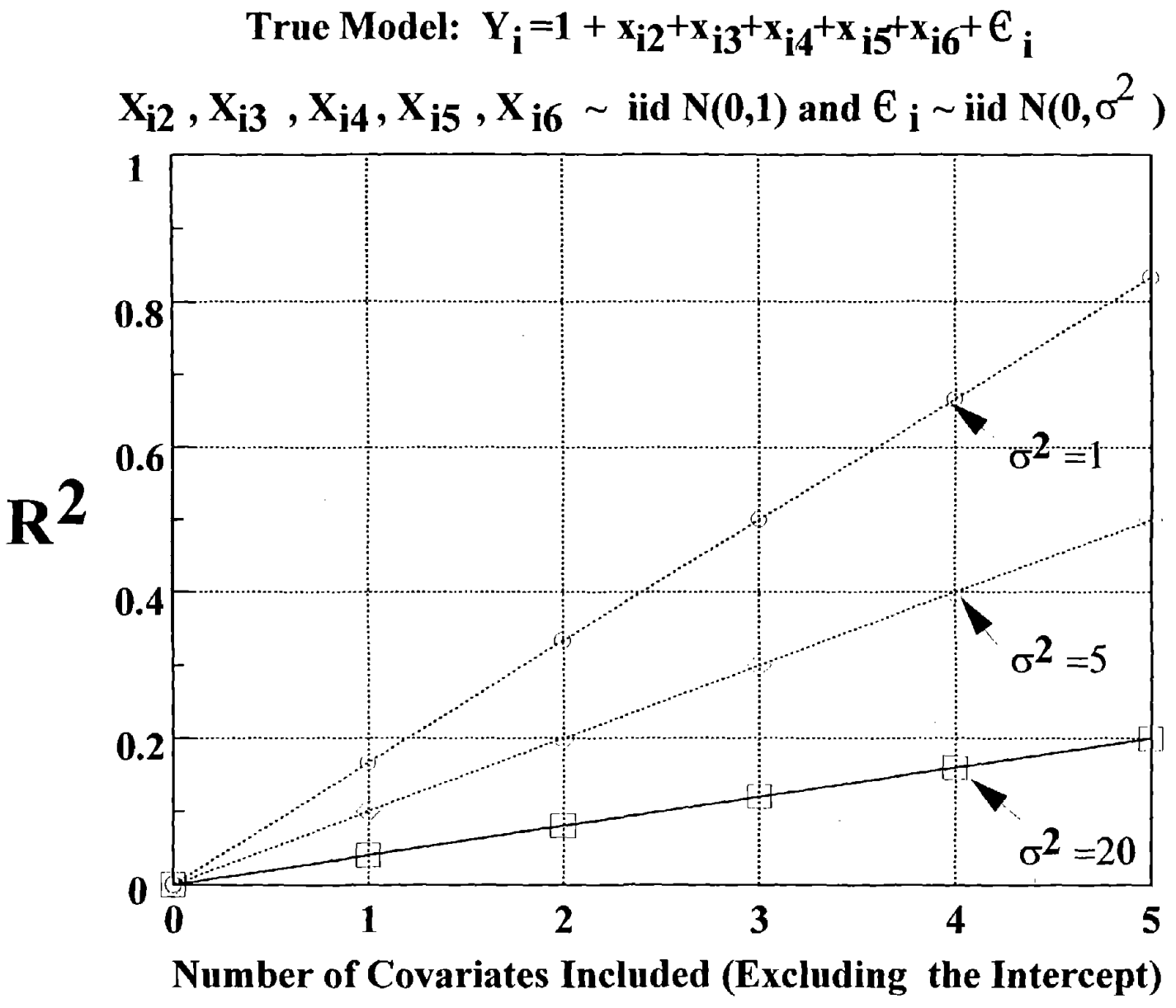


Figure 2. R^2 values of a simple normal linear regression model at three variance levels.

The purpose of this section is to review the current status of research and thinking in accident prediction modeling with respect to the five tasks mentioned above. More detailed mathematical description of some of these tasks will be provided in the next section. The order of these five tasks is not particularly important because in practice some tasks can very well be conducted concurrently or iteratively depending on the state of knowledge of the problem and of studied sites.

Task 1. Find a good probability (mass) function, P(.), to describe the random variation of ϵ.

In the last decade or so, there has been a steady realization in accident prediction modeling research that the conventional normal or lognormal regression models simply do not have the necessary statistical property to adequately describe the vehicle accident events on the road. Better choices of P(.) are the Poisson and negative binomial (NB) distributions [Maycock and Hall, 1984; Jovanis and Chang, 1986; Joshua and Garber, 1990; Miaou and Lum, 1993; and Miaou et al., 1993]. It is also a common view among statisticians, econometricians, and some safety researchers in analyzing discrete rare events that if we are able to include all the necessary variables to explain the systematic variation of dependent variable Y, then the Poisson distribution will be the best choice [Lawless, 1987; Cameron and Trivedi, 1990; and Miaou et al., 1992]. (Note that the underlying assumption of the Poisson distribution is that the variance of the data is equal to the mean.) In cases where we have omitted variables Z's, the use of Poisson model causes a phenomenon called overdispersion, meaning that the variance of the accident data is greater than the Poisson model indicated. Furthermore, if the exponent of Z's, i.e., $exp(Z)$, follows a distribution that can be represented by a gamma distribution, then the NB model is a good candidate probability function for Y given available X's. More detailed mathematical description of the Poisson and NB regression models will be given in the next section.

The discussion of the choice of P(.) above may be oversimplified both statistically and conceptually. Nevertheless, it points out the importance of considering one critical problem in accident prediction modeling, i.e., the omitted variable problem. As mentioned earlier, driver and vehicle variables are largely unavailable in developing accident prediction models. It has been suggested that the omission of driver variables is perhaps one of the main sources of overdispersion in accident prediction models [Miaou et al., 1993; Miaou and Lum, 1993]. In view of this inevitable omitted variable problem, the NB distribution seems to be a favored choice over the Poisson distribution. However, as will be discussed later, the omitted variable problem is by no means the only source of overdispersion in accident prediction models. Also, one should know that there is more than one NB probability function that the analysts can choose from, each of which implies a specific mean-variance relationship. The most commonly used NB distribution is the one with a quadratic variance function, i.e., the variance is a quadratic function of the mean.

The aggregation level used to define a site, e.g., the length of road sections or radius of intersections, and the associated length-of-time intervals, also play some role in the choice of P(.). Ideally, each site should be homogeneous in the sense that its attributes regarding the five major factors discussed earlier can be uniquely and clearly defined. In practice, this is, however,

unrealistic because too many attributes need to be considered and as a result it is almost impossible to find a sufficient number of sites that are totally homogeneous in all considered attributes. A good example would be the attributes associated with roadside hazards that vary considerably from site to site by type, shape, size, and their positions from the travelway.

As an alternative, it has been suggested that major traffic and roadway design attributes of each site should be as homogeneous as possible [e.g., Miaou et al., 1993; Miaou and Lum, 1993]. For example, in studying accidents on road sections, major attributes for each road section, such as average annual daily traffic (AADT), percent trucks, access control, number of lanes, lane width, horizontal curvature, vertical grade, shoulder width, median type and width, roadside clear recovery distance (from shoulders), should be as homogeneous as possible. This suggestion was based on both theoretical and roadway design considerations. On the theoretical consideration, the higher the level of aggregation, the less the site can be considered homogeneous and the further the accident frequency of each site deviates from the Poisson distribution. Especially, the conditional variance of accident frequencies will be exceeding the conditional mean. Now, let's take a highly aggregate and nonhomogeneous site for example. Basically, one can argue that the accident frequency at the site can be considered as the sum of the accident frequency at numerous homogeneous sub-sites within the site, and, therefore, by the Central Limit Theorem, the accident frequency at the site is about normally distributed. On the roadway design consideration, as the level of aggregation increases, the effects of many roadway design elements become diluted and the developed model will be less useful for engineering design [Miaou and Lum, 1993]. Obviously, highly disaggregate data are too expensive to collect, while highly aggregate data have almost no value for engineering design. Some engineering judgments seem inevitable in setting the priorities of the engineering design questions that need to be addressed and in developing criteria for use to determine whether a site can be considered as relatively homogeneous.

Task 2: Determine an appropriate functional form and parameterization , i.e., $f(\,.\,;\,\beta)$, to describe the effects of variables X's and U's on the long-term means.

The function $f(\,.\,;\,\beta)$ represents the long-term mean and is a function of the five major factors discussed earlier. For simplicity, it will now be called the mean function of the model. To the best of this author's knowledge, there is no comprehensive statistical study aimed at identifying appropriate mean function $f(\,.\,;\,\beta)$ in the area of accident prediction modeling. However, there seems to be a consensus among traffic safety engineers and researchers that the effects of X's on accident frequency Y are interactive in nature [e.g., Maycock and Hall, 1984; Miaou and Lum, 1993]. This suggests that some sort of multiplicative functional forms should be used to describe the effects. Another important consideration about the choice of $f(\,.\,;\,\beta)$ is that it always has to be nonnegative. A natural candidate function for describing the interactive effects which at the same time ensures that the function values are always nonnegative, is the exponential function. It has been widely used by statisticians and econometricians and found to be very flexible in fitting different types of count data [e.g., Cox and Lewis, 1966; Cameron and Trivedi, 1986]. In a study by Miaou and Lum [1993], they compared the performance of two multiplicative functions, one of which is the exponential function, for truck accidents under the Poisson model and found that the exponential form was indeed a better choice.

The selection of right variables to include in $f(\,.\,;\,\beta)$, i.e., to identify appropriate X's and eliminate U's, requires both engineering judgment and statistical consideration. The initial variable selection is always based on engineering judgment. Once the candidate variables are selected, the next step is to derive a theory (again using engineering judgment and knowledge) as to how each variable is likely to affect accident frequencies and what the relative importance of these variables are. This exercise gives the analysts some idea of the expected (algebraic) signs of each parameter, β, in the model, and their relative order of magnitudes. Instead of using the engineering judgment, some have relied on sample correlation, r_{xy}, between each individual variable X and Y to get some idea of the association [see e.g., Roy Jorgensen Associates, 1978]. It should be emphasized, however, that sample correlation is good only for measuring a close-to-linear relationship and is computed one variable at a time. Since many studies have shown that the effects of some variables are highly nonlinear, e.g., AADT, and are highly interactive, sample correlation could be misleading and should be used with extreme caution.

It is worth mentioning that studies to determine the mean function from engineering viewpoints do exist. These studies typically focused on the use of traffic flow theory, geometry, vehicle dynamics, and probability theory to describe traffic conditions and events that may lead to an accident, and very little attention was given to driver behavior. For example, for roadway mainline accidents on road sections, intersections, and interchanges, Council et al. [1983] conducted a very comprehensive study under the title of *Exposure Measures for Evaluating Highway Safety Issues*. For roadside accidents, there has been a constant effort attempting to refine the mean function on the basis of engineering judgment or knowledge. These studies used a series of conditional probabilities to describe the sequence of events resulting in a roadside accident. An example sequence of events would be (1) an errant vehicle leaves the traveled way and encroaches on the shoulder; (2) the location of encroachment is such that the path of travel is directed towards a potentially hazardous object; (3) the hazardous object is sufficiently close to the travel lanes that control is not regained before encounter or collision between vehicle and object; and (4) the collision is sufficiently severe enough to result in an accident. These types of models have traditionally been called roadside encroachment models [Glennon, 1974; TRB, 1987; Mak and Sicking, 1992; and Daily et al., 1994]. Generally speaking, these types of engineering studies have ignored the statistical side of the problem and paid very little attention to the other tasks described in this section. In addition, for many years these studies have been criticized as being full of wishful assumptions and lack supporting data. Roadside encroachment models are discussed in more detail in chapter 6.

Task 3: Select the right variables to include in $f(\,.\,;\,\beta)$, i.e., to find appropriate X's and eliminate U's.

Given a set of candidate variables, several statistical criteria have been used to select variables, including $\tilde{R}^2$ and AIC. Typically, the model which contains a particular subset of candidate variables that has the highest $\tilde{R}^2$ value or lowest AIC value is considered the best model. However, it happens quite often that the analysts may find several models to be equally good (i.e., their $\tilde{R}^2$ and AIC values are very close). In such cases, all these models need to be selected and subjected to further study, including case analysis, questions of interpretability, tests

of prediction capability, and so on [Weisberg, 1985]. In addition, as suggested in Miaou et al. [1992], it is important to check whether all estimated parameters of selected models have expected signs and high t-statistics.

For problems with only a few candidate variables, say eight or less, it is not difficult to consider all possible models, each of which contains a subset of candidate variables. For example, for a problem with eight candidate variables, there are 2^8=256 possible subset models that need to be estimated and compared, which is computationally quite manageable even with a 486-based personal computer. Now, for a problem with 15 candidate variables, which is not uncommon in accident studies, there are 2^{15} = 32,768 possible subset models that need to be estimated and examined which is almost unmanageable even with today's powerful workstations. *[Note that because accident prediction models have mean functions f(. ; β) that are nonlinear and the probability functions P(.) that are non-normal, the estimation of model parameters require extensive iterative searching procedures.]*

At present, there are several methods that are available in most of the statistical software for alleviating such a massive computational problem. These methods include forward selection, backward elimination, and stepwise procedure [see e.g., Weisberg, 1985]. Collectively, these methods are called stepwise regressions. The statistical criterion used in stepwise regressions to add or delete variables is either t- or F-statistic. Although stepwise regressions are widely used in accident studies, they should be used with caution. The reason is that these methods do not check whether the parameters in the model have the expected signs and whether some combinations of candidate variables make good engineering sense. For example, in developing accident prediction models for trucks, AADT and percent-truck should always be included together to account for truck exposure. It does not make sense to keep only one of them and leave the other one out of the model. Also, several problems regarding stepwise regressions were pointed out by Weisberg [1985]: (1) the model selected in stepwise fashion need not optimize any reasonable criterion function for choosing a model; (2) the ordering of candidate variables (or predictors) is an artifact of the method and need not reflect relationships of substantive interest; and (3) stepwise regression may seriously overstate the significance of the results.

Those variables that are considered statistically insignificant and eliminated from the final model, i.e., the *U's*, do not necessarily have no effect on accident frequency. Oftentimes, variables are excluded from the model simply because they do not have enough variation in the available data, and at the same time the sample size is small. To use an extreme example, if a variable (say lane width) has no variation in the data (say that all are 3.66 m in lane width), then regardless of the sample size there is no way that any statistical model could be developed from the data to describe the effect of such variable on accident frequency [see e.g., Miaou et al., 1993]. As in many problems, if a point is reached at which it is necessary to quantify the effect of such variables with a limited range of variations, then two approaches may be taken: (1) if feasible, collect additional data with a wider range of variations in these variables and redevelop the model; and (2) use results from other studies with similar road environment. The second approach requires a need to combine two or more models developed for different data sets. Engineers' discretion under approach two is essential. Before developing any models, it is

always helpful to generate simple descriptive statistics, including the coefficient of variation, to identify variables that either have no variation or may have very limited variations.

In sum, there is no good way of substituting statistical methods for good engineering judgment. Statistical methods are analytical tools and should be used as such to deepen our understanding of the problem, enhance our engineering knowledge, and help us make more intelligent decisions.

Task 4: Estimate the regression parameters (or coefficient) β in $f(.\ ;\ \beta)$ and obtain good statistical inferences for the estimated parameters based on available data.

Once the probability function P(.) and mean function $f(.\ ;\ \beta)$ are determined and the data of candidate variables (*X's* and *U's*) are collected and checked, the next step is to estimate the parameters β in the model. Because accident prediction models are Poisson-based and nonlinear models, the estimation methods are all recursive types of estimators, e.g., iteratively reweighted least squares (IRLS) method, ML method, and maximum quasi-likelihood (MQL) method. The derivation of statistical inferences, such as t-statistics of estimated parameters, requires the use of asymptotic theory [e.g., Cramer, 1989]. This task is, in general, quite statistically involved. Some discussion on parameter estimation for lognormal, Poisson, and NB regression models will be given in the next section. Here are some useful references: Myers [1990], Cramer [1989], and McCullagh and Nelder [1983]. One of the key issues that need to be addressed in this task is to develop an efficient and numerical stable algorithm to estimate model parameters and produce reliable statistical inferences. The quality of the estimates and statistical inferences, however, depends heavily on how good the selected probability function P(.) and mean function $f(.\ ;\ \beta)$ represents the true model and, of course, the quality of data.

Task 5: Assess the quality of the model, judge whether the model make good engineering sense, decide whether the developed model meets the planning and design requirements, and identify cost-effective ways to improve the model.

As discussed earlier, in assessing the quality of a model, it is important to check whether all estimated parameters of the model have expected signs. Further investigations are required if the estimated parameters of some covariates do not have expected signs. It is also important to determine if all the estimated parameters have high t-statistics (e.g., > 2.0, when sample size is large). This check allows the analysts to assess whether these parameters are well determined and whether there is a need to increase sample size or to collect additional data for increasing the range of variations of some covariates .

Another question that needs to be addressed is the overall quality of the developed model. For example, How close is the developed model from the true model? Obviously, the closer the developed model from the true model, the more confident the analysts are of using the developed model in engineering practice. Two other related questions mentioned in chapter 1 are Do we need to collect additional variables? and Would it be cost-effective to collect additional variables? As mentioned earlier, many traffic safety engineers and researchers have been using the closeness of R^2 value to 1 as a yardstick to address these questions. The next chapter will

show that R^2 is not an appropriate measure to address these questions for accident prediction models and will examine three alternative criteria to address these questions in chapter 5.

There are three basic concepts involved when examining these alternative criteria: (1) [0,1] bound concept; (2) proportional increase concept; and (3) invariant with respect to the mean concept. Basically, [0,1] bound concept says that the analysts would like to have a value of 0 if no covariate (other than the intercept term) is included in the model, and a value of 1 if all necessary covariates are included. Proportional increase concept says that if all covariates are independent and equally important, then when the analysts select and add these covariates to the model one at a time, the increase in value should be the same for each covariate regardless of their order of selection. Invariant with respect to the mean concept says that the value of the criterion will not change by simply increasing or decreasing the value of the intercept term in the model.

Finally, all developed models should be reviewed by traffic and roadway design engineers and subjective to case analysis, questions of interpretability, tests of prediction capability, and so on.

ACCIDENT PREDICTION MODELS USED IN RECENT STUDIES

Significant progress has been made in the development of accident prediction models over the last decade, largely due to the use of the so-called Generalized Linear Models [McCullagh and Nelder, 1983]. The most promising models have been the Poisson and NB regression models [e.g., Maycock and Hall, 1984; and Miaou, 1994]. These models have been developed for planning, preliminary design, and evaluation purposes.

In this section, three types of accident prediction models that have been used in recent studies are presented: lognormal regression, Poisson regression, and NB regression models. The lognormal regression model has been shown to be inadequate for developing accidents-flow-roadway design relationships and is presented only for illustration and comparison purposes [Miaou and Lum, 1993].

The presentation below uses road sections as examples. The same models with slight modifications can be used for intersections and ramps [see e.g., Maycock and Hall, 1984]. The models described in this section can be applied to any roadway class, vehicle configuration, and accident severity type of interest. For ease of exposition, the following presentation focuses on accidents of all severity types involving all types of vehicles on a particular roadway class.

Consider a set of n road sections of a particular roadway type, say, rural Interstate. Let Y_i be a random variable representing the number of vehicle accidents on road section i during a period of, say, 1 year, where $i=1,2,...,n$. (The same road section in different sample periods are considered as separate road sections.) Further, the actual observation of Y_i during the period is denoted as y_i, where $y_i=0,1,2,3,...$ and $i=1,2,...,n$. Let the amount of vehicle travel (or vehicle exposure) during the sample year on this road section be v_i. The vehicle travel, v_i, is usually

computed as $365 \times AADT_i \times \ell_i$, where $AADT_i$ is the number of vehicles and ℓ_i is the length (in miles or kilometers) of road section i. [If only accidents involving trucks are of interest, then a typical exposure measure is truck miles, computed as $365 \times AADT_i \times (T\%_i/100) \times \ell_i$, where $T\%_i$ is the average percentage of trucks in the traffic stream (or percent trucks) on road section i, e.g., 15, and $AADT_i \times (T\%_i/100)$ is the truck AADT of road section i during the observed year.] Associated with each road section i, there is a $k \times 1$ covariate vector, x_i, describing its geometric characteristics, traffic conditions, and other relevant attributes. The transpose of the covariate vector is denoted by $x_i' = (x_{i1}, x_{i2}, ..., x_{ik})$. Without loss of generality, let the first covariate x_{i1} be a dummy variable equal to one for all i (i.e., $x_{i1}=1$). Some of the covariates can be 0,1 dummy variables, indicating the presence or absence of a condition. For example, to assess the effects of terrain on vehicle accidents, a dummy variable is set equal to zero if a road section is located in level terrain and set equal to 1 if located in rolling or mountainous terrain.

In this presentation, all three models are formulated under the assumption that (1) vehicle miles data and other covariates are free from measurement and recording errors; and (2) the occurrences of vehicle accidents on different road sections are independent. Many of the materials presented in this section are adopted directly from the papers by Miaou and Lum [1993] and Miaou [1994]. The detailed discussion of each model now follows.

Lognormal Regression Model

$$Y_i + \delta = v_i \left[\beta_1 \left(\prod_{j=2}^{k} (1 + x_{ij})^{\beta_j} \right) \right] e^{\epsilon_i}, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i=1,2,3,...,n. \tag{5}$$

or

$$\log\left(\frac{Y_i + \delta}{v_i}\right) = \beta_1' + \sum_{j=2}^{k} \beta_j \log(1 + x_{ij}) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2), \qquad i=1,2,3,...,n. \tag{6}$$

where ϵ_i is an error term, and $\epsilon_i \sim iid\ N(0,\sigma^2)$. The parameter δ is a preselected small constant (e.g., 0.5, 0.01, 0.001), which is added to Y_i to avoid the log(0) problem, and $\beta_1^* = \log(\beta_1)$. (Note that log(0) is undefined.) This model requires that $x_{ij} \geq 0$ for all i and j. One has been added to each covariate (except x_{i1}) to avoid the log(0) problem in Eq. (6). Without the transformation, the right hand side of Eq. (5) will be rendered zero as long as one covariate has value zero, regardless of what the values of other covariates are. Basically, this transformation shifts the origin of the covariates from zero to one. Since $(1+0)^{\beta_j} = 1$ and $\beta_j \log(1+0) = 0$, these covariates with values of zero do not contribute to the occurrences of accidents at this new origin in Eqs. (5) and (6). This is a desirable property. For example, horizontal curvature should not be a contributing factor to the occurrences of accidents on tangent road sections. Note that this transformation can be applied to 0,1 dummy variables without problem. Other transformations are, of course, possible.

The model can also be reexpressed as

$$\log(Y_i + \delta) \sim N(\mu_i', \sigma^2)$$
$$\text{where } \mu_i' = E[\log(Y_i + \delta) \mid v_i, x_i] = \log(v_i) + \beta_1' + \sum_{j=2}^{k} \beta_j \log(1 + x_{ij}) \qquad i=1,2,\ldots,n. \tag{7}$$

The underlying mean, variance, and coefficient of skewness of the distribution of Y_i when conditional on x_i are given in table 1. Two interesting statistical properties can be observed: (1) the conditional variance is a function of the conditional mean; and (2) the coefficient of skewness is dependent on the constant σ^2, which does not go to zero as the conditional mean μ_i approaches ∞.

It is important to point out that the expected value of Y_i under the model is:

$$\mu_i = E[Y_i \mid v_i, x_i] = -\delta + v_i \left[\beta_1 \left(\prod_{j=2}^{k} (1 + x_{ij})^{\beta_j} \right) \right] e^{\frac{1}{2}\sigma^2} \tag{8}$$

not

$$-\delta + v_i \left[\beta_1 \left(\prod_{j=2}^{k} (1 + x_{ij})^{\beta_j} \right) \right] \tag{9}$$

Equation (8) includes an adjustment factor, $\exp(\sigma^2/2)$, which is multiplicative and grows exponentially with the variance σ^2. The estimator in Eq. (9), without the adjustment factor $\exp(\sigma^2/2)$, can be shown to provide an estimator for the median, rather than the mean, of the conditional probability distribution of Y_i, and this estimator systematically underestimates the conditional mean of Y_i [see e.g., Miller, 1984]. In a typical vehicle accident study, the magnitude of this adjustment factor is quite large because of large σ^2. Therefore, ignoring this adjustment factor would seriously understate the expected number of vehicle accidents. The truck accident study of Miaou and Lum [1993] showed that without this adjustment factor the underestimations were over 80 percent.

Using the linearized model in Eq. (6), the least squares estimates of the regression parameters denoted by $\hat{\beta}_1^*$ and $\hat{\beta}_j$, $j=2,3,\ldots,k$, the estimated variance of the residuals denoted by $\hat{\theta}^2$, as well as the estimated variance and t-statistics of the estimated parameters, can be obtained with a standard linear regression computer program. Substituting the parameters and residual variance σ^2 in Eq. (8) with the least squares estimates gives an estimate of μ_i. That is,

$$\hat{\mu}_i = -\delta + v_i \left[\hat{\beta}_1 \left(\prod_{j=2}^{k} (1 + x_{ij})^{\hat{\beta}_j} \right) \right] e^{\frac{1}{2}\hat{\theta}^2} \tag{10}$$

where $\hat{\beta}_1 = \exp(\hat{\beta}_1^*)$. This estimate is somewhat biased because $\exp(\hat{\beta}_1^*)$ is a biased estimate of β_1. However, the bias of this estimate is much smaller than that using Eq. (9) as an estimator [Miller, 1984].

Table 1. The underlying distributions of accident frequency, Y_i, for the three commonly used accident prediction models and their conditional mean, variance, and coefficient of skewness.

Model	Conditional Distribution of Y_i	$E[Y_i\|x_i,v_i] = \mu_i$	$Var[Y_i\|x_i,v_i]$	$Skew[Y_i\|x_i,v_i]$
Lognormal Regression	Lognormal	$-\delta + v_i\left[\beta_1\left(\prod_{j=2}^{k}(1+x_{ij})^{\beta_j}\right)\right] e^{\frac{1}{2}\sigma^2}$	$(\mu_i+\delta)^2(e^{\sigma^2}-1)$	$(e^{\sigma^2}-1)^{1/2}(e^{\sigma^2}+2) \quad (>0)$
Poisson Regression	Poisson	$v_i[e^{x_i'\beta}]$	μ_i	$\mu_i^{-1/2} \quad (>0)$
Negative Binomial Regression	Negative Binomial	$v_i[e^{x_i'\beta}]$	$\mu_i+\alpha\mu_i^2$	$(1+2\alpha\mu_i)/(\mu_i+\alpha\mu_i^2)^{1/2} \quad (>0)$

The estimated variance of $\hat{\beta}_1$ can be approximated by [e.g., Ratkowsky, 1983, page 23]

$$\hat{Var}(\hat{\beta}_1) \approx \hat{Var}(\hat{\beta}'_1)\, e^{2\hat{\beta}'_1} \tag{11}$$

and its t-statistic can be computed as

$$t(\hat{\beta}_1) = \frac{\hat{\beta}_1}{[Var(\hat{\beta}_1)]^{1/2}} \approx \frac{1}{[\hat{Var}(\hat{\beta}'_1)]^{1/2}} \tag{12}$$

The expected number of vehicle accidents per vehicle mile (or kilometer) of travel, i.e., $E(Y_i|x_i,v_i)/v_i$, is called the rate function of the model and is typically denoted by λ_i. This lognormal regression model implies that the rate function, i.e., the expected number of vehicle accidents, μ_i, in Eq. (8) divided by vehicle travel v_i, is somewhat complicated because of the small constant δ and the exponential adjustment factor. This model, however, implies that μ_i is linearly related to the amount of vehicle travel v_i.

These types of models have been used in several recent studies, e.g., Zegeer et al. [1987], Zegeer et al. [1990], and Mohamedshah et al. [1992]. However, to this author's knowledge, most of these studies did not consider the adjustment factor when using the model to estimate or predict accidents. In addition, t-statistics of the estimated parameters were rarely reported. In a case study by Miaou and Lum [1993], it was shown that the selection of the small constant δ can have significant effects on the parameter estimation when the mean level is low. In addition, the statistical inferences, e.g., the α level of the estimated parameters, obtained from the lognormal regression models can be very different from those obtained from the Poisson regression models.

Poisson Regression Model

$$Y_i \sim Poisson(\mu_i) \quad or \quad P(Y_i = y_i) = P(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \qquad i=1,2,3,...,n. \tag{13}$$

where

$$\mu_i = E[Y_i | v_i, x_i] = v_i[e^{x_i'\beta}] = v_i[e^{\sum_{j=1}^{k} x_{ij}\beta_j}] \qquad i=1,2,3,...,n. \tag{14}$$

and β is a $k \times 1$ vector of unknown regression parameters, the transpose of which is denoted by $\beta' = (\beta_1, \beta_2, ..., \beta_k)$. This model assumes that Y_i, $i=1,2,...,n$, are independently and Poisson distributed with conditional mean μ_i. As in the lognormal regression model, the expected number of vehicle accidents μ_i in this model is proportional to vehicle travel v_i. The model, however, assumes an exponential rate function, $\lambda_i=\exp(x_i'\beta)$, which ensures that the accident rate is always nonnegative. This type of rate function has been widely employed in statistical literature and found to be very flexible in fitting different types of count data [e.g., Cox and Lewis, 1966; Cameron and Trivedi, 1986; and Frome et al., 1990]. Note that whenever

appropriate higher order and interaction terms of covariates can be included in Eq. (14) without difficulties.

The underlying distribution of Y_i and its conditional mean, variance, and coefficient of skewness are presented in table 1. Two interesting statistical properties can be observed: (1) the conditional variance is equal to the conditional mean; and (2) the coefficient of skewness is a function of the conditional mean μ_i, and as μ_i approaches ∞, the coefficient of skewness goes to zero. In previous studies, the Poisson distributional assumption is used to obtain tests and confidence statements about the estimated regression parameters and, unlike the lognormal regression model, this distribution can also be used to make reasonable probabilistic statements about Y_i in many cases.

The regression parameters of this model can be estimated using the ML method presented in Cramer [1989], or the MQL method presented in McCullagh and Nelder [1983], or the IRLS method described in Carroll and Ruppert [1988].

This model has been used to develop truck accidents and highway geometric design relationships in Miaou et al. [1992, 1993]. It has also been used in other areas of highway safety studies. For example, Jovanis and Chang [1986] used the model to examine the relationship between vehicle accidents and vehicle miles of travel, and Saccomanno and Buyco [1988] applied the model to relate vehicle accident rates with different traffic volumes, truck types, hour of day, and driver ages.

A limitation of using the Poisson regression model, which is well-known in statistical literature [e.g., Cox, 1983; Dean and Lawless, 1989], is that the conditional variance of the data is restrained to be equal to the conditional mean. In many applications, count data were found to display extra variation or overdispersion relative to a Poisson model [e.g., Dean and Lawless, 1989]. That is, the variance of the data was greater than the Poisson model indicated.

In vehicle accidents-geometric design studies, the overdispersion could come from several possible sources. Some sources were identified in Miaou and Lum [1993]: omitted variables, uncertainties in vehicle exposure data and traffic variables, nonhomogeneous roadway environment related to the level of aggregation, and correlation between accident events. The effects of omitted variables will be discussed in the next section.

The consequences of ignoring the extra variations in the Poisson regression models are that consistent estimates, such as the ML estimates of the regression parameters under the Poisson model, are still consistent; however, the variances of the estimated parameters would tend to be underestimated. In other words, the significance levels of the estimated parameters may be overstated [Cameron and Trivedi, 1990]. Following Wedderburn [1974], to correct the overdispersion problem for the Poisson regression model, one can assume that the variance of Y_i is $\tau\mu_i$ instead of μ_i as that originally assumed in the Poisson model, where τ is called the overdispersion parameter. Furthermore, the overdispersion parameter τ can be estimated by $X^2/(n-k)$, where X^2 is the Pearson's chi-square statistic, n is the number of observations (i.e., the number of road sections), and k is the number of regression parameters in the Poisson regression

model. The Pearson's X^2 statistic is computed as $\Sigma_i(y_i-\hat{\mu}_i)^2/\hat{\mu}_i$, where $\hat{\mu}_i = v_i \exp(x_i\hat{\beta})$ and $\hat{\beta}$ is the ML estimates of the regression parameters β. A better estimate of the asymptotic t-statistic for each regression parameter is $\tau^{-1/2}$ times that obtained from the original Poisson regression model based on the ML method [Agresti, 1990]. The Poisson regression model coupled with such an adjustment for overdispersion is sometimes referred to as the quasi-Poisson regression model.

Other ways to overcome this overdispersion problem are available. A simple way is to use the NB distribution assumption. The NB distribution allows the conditional variance to exceed the conditional mean. More discussion on NB models will be given later. Currently, many statistical studies are under way attempting to modify existing probability functions within the exponential family to allow more flexible mean-variance relationships [e.g., Efron, 1986; Gelfand and Dalal, 1990].

To help the discussion on the relationship between the Poisson and NB regression models, the following is a brief discussion of the effect of omitted variables on the Poisson regression. As indicated earlier, in developing accident prediction models, driver and vehicle variables are largely unavailable by site. To reflect this in the Poisson model, Eq. (14) can be rewritten to make omitted variables explicit as follows:

$$\mu_i = E[Y_i | v_i, x_i, z_i] = v_i[e^{x_i'\beta + z_i'\gamma}] = v_i[e^{\sum_{j=1}^{k} x_{ij}\beta_j + \sum_{j=1}^{k'} z_{ij}\gamma_j}] \qquad i=1,2,3,...,n. \tag{15}$$

where z_{ij}, j=1,2,...,k' are values of omitted variables Z_{ij}; γ_j is the regression parameter associated with omitted variable Z_{ij}; and z_i and γ are vector representations of z_{ij} and γ_j, respectively. It can be shown that if $\exp(\Sigma Z_{ij}\gamma_j)$ follows a gamma distribution, then Y_i given x_i is NB distributed [Cameron and Trivedi, 1986]. More discussion will be given later.

Negative Binomial Regression Model

To deal with the overdispersion problem in count data, one commonly used distribution is the NB distribution. The typical NB regression model has the following form:

$$P(Y_i=y_i) = \frac{\Gamma(y_i+\frac{1}{\alpha_i})}{\Gamma(y_i+1)\,\Gamma(\frac{1}{\alpha_i})}\left(\frac{1}{1+\alpha_i\mu_i}\right)^{\frac{1}{\alpha_i}}\left(\frac{\alpha_i\mu_i}{1+\alpha_i\mu_i}\right)^{y_i}, \qquad y_i=0,1,2,... \tag{16}$$

where

$$\mu_i = E[Y_i | v_i, x_i] = v_i[e^{x_i'\beta}] = v_i[e^{\sum_{j=1}^{k} x_{ij}\beta_j}] \qquad i=1,2,3,...,n. \tag{17}$$

and the conditional variance of Y_i is a quadratic function of conditional mean:

$$Var[Y_i | v_i, x_i] = \mu_i + \alpha_i\mu_i^2 \tag{18}$$

where $\alpha_i \geq 0$ and are usually referred to as dispersion parameters. From Eq. (18) one can see that this model allows the conditional variance to exceed the conditional mean. Also, the Poisson regression model can be regarded as a limiting model of the NB regression model as all α_i approach zero. This model will be denoted as $NB(\mu_i, \alpha_i)$.

The standard NB regression model that has been used in accident prediction modeling (and areas such as econometrics and biometrics) is a special case of Eq. (16) where α_i, $i=1,2,..,n$, are set equal to a constant α (> 0). Note that such standard NB models will be denoted as $NB(\mu_i ; \alpha)$. The conditional mean, variance, and coefficient of skewness for such standard NB regression models are given in table 1.

A wide range of variance-mean relationships can be obtained by letting, e.g., $\alpha_i = \alpha(\mu_i)^{-c}$ for $\alpha > 0$ and an arbitrary constant c [Cameron and Trivedi, 1986]. One can show that, under such a choice of α_i, the conditional variance becomes

$$Var[Y_i | v_i, x_i] = \mu_i + \alpha \mu_i^{2-c} \qquad (19)$$

The variance-mean relationship of the standard NB regression model is, of course, a special case of Eq. (19) with c=0. In addition, the variance-mean relationship of the quasi-Poisson regression model discussed earlier can also be represented by Eq. (19) with c=1, i.e., $\tau = 1+\alpha$. Even though Eq. (19) provides a more flexible variance-mean relationship, to the best of this author's knowledge, no accident prediction models have been reported using such a relationship because of its computational difficulty. Note that this author is aware of some ongoing research aimed at identifying an appropriate range of c values for use in developing accident prediction models. In this study, only standard NB regression models are considered.

The NB regression model can be derived from the Poisson regression model under the assumption that the exponent of the omitted variables in Eq. (15), i.e., $\exp(\Sigma Z_{ij} \gamma_j)$, follows a gamma distribution [Cameron and Trivedi, 1986]. The implication of this assumption is that even if the observed values of the available covariates, x_i, are the same for different sites, the conditional means of the Poisson, μ_i, will be different from site to site and follow a gamma distribution because of the differences in unobserved covariates Z_i among these sites. Even though the gamma probability function is quite a flexible function, there is still some arbitrariness involved here. For example, if $(\Sigma Z_{ij}\gamma_j)$ is normally distributed, then $\exp(\Sigma Z_{ij}\gamma_j)$ is lognormally distributed. In many cases, the gamma distribution is a good approximation for the lognormal distribution. Therefore, in such cases, the use of NB distribution is justified. However, if $(\Sigma Z_{ij}\gamma_j)$ happens to be, e.g., uniformly distributed with a narrow range of variation, then the gamma distribution is simply not a good approximation for the distribution of $\exp(\Sigma Z_{ij}\gamma_j)$ and the use of NB distribution cannot be justified. For real-world problems, since Z_i are omitted variables, their exact distributions are unknown. However, using the Central Limit Theorem, one can argue that if $(\Sigma Z_{ij}\gamma_j)$ is made up of the sum of many small effects, then $(\Sigma Z_{ij}\gamma_j)$ is likely to be normally distributed and NB distribution assumption is likely to be justified. Obviously, more research in this area is needed.

The ML estimation of the NB regression model and the calculation of associated statistics are described in detail by Lawless [1987]. The moment estimation, which was first suggested by Breslow [1984], is also commonly used for estimating the parameters in the NB model. Another estimation method for estimating the NB regression model is a regression-based estimation method suggested by Cameron and Trivedi [1986, 1990]. Miaou [1994] conducted a study to compare the estimation results from these three estimation methods using 5 years of truck accident data collected on rural Interstate highway sections in Utah. On average, the studied road sections had a relatively low mean accident frequency per site (about 0.2 truck accident involvements per site). It was found that the estimated dispersion parameters were quite different from the three methods and suggested that the moment method and regression-based method should be used with caution.

The NB regression model using the ML method has been used in Miaou et al. [1993] to establish truck accidents-geometric design relationships for three different roadway classes. Although the NB regression model is more general than the Poisson regression model, it requires more extensive computations to estimate model parameters and to generate inferential statistics than the Poisson regression model. Furthermore, the statistical property of different estimators, e.g., the ML and moment estimators, of the NB regression model under different mean levels and sample sizes have not yet been fully investigated.

POTENTIAL MISSPECIFICATIONS OF ACCIDENT PREDICTION MODELS

The number of possible misspecifications that could be made when developing accident prediction models is very large, but most of them fall into one of the following categories: (1) sampling and nonsampling errors in collected data for dependent and independent variables; (2) omitted variables; (3) non-Poisson or non-NB distribution; (4) imprecise mean function; and (5) correlation. Some of these possible misspecifications have been mentioned or discussed earlier. The following is a summary of the effects of these potential misspecifications in developing accident prediction models.

(1) Sampling and nonsampling errors in collected data for dependent and independent variables.

For many years, accident prediction models have been proposed for evaluating the relative safety of roadway design alternatives and for identifying high risk locations for reconditioning and safety improvements [TRB, 1987; AASHTO, 1989]. Despite significant progress in the development of accident prediction models over the last decade (largely because of the use of the Generalized Linear Models (GLM) [McCullagh and Nelder, 1983]), many roadway design engineers and safety planners are still not confident with these models. One of the reasons is the lack of confidence in the quality of the reported accident data, from which accident prediction models are developed. One quality issue of major concern is the underreporting of accidents, especially, the underreporting of minor injury and property damage only (PDO) accidents. Because of this underreporting problem, questions have been raised regarding (1) the statistical

validity of the developed accident prediction models; and (2) the potential of accident prediction models to understate the benefits of various improvements in roadway design.

It has long been established in statistical literature that, under a mild assumption, the underreporting of accidents would not affect the statistical validity of the accident prediction models [see e.g., Hauer and Hakkert, 1988]. However, to account for the underreporting, these accident prediction models would have to be adjusted by a factor which, in essence, indicates the rate of underreporting. In other words, the development of accident prediction models using reported accident data is valid, provided that the underreporting of accidents be adjusted outside of the models. Depending on the extent of underreporting, without proper adjustments, the use of accident prediction models would indeed understate the benefits of roadway improvements.

In practice, to adjust for underreporting, the underreporting rates have to be estimated for various classes of accident prediction models, such as those considered in FHWA's "Interactive Highway Design Models" [Harwood et al., 1994]. Specifically, it is necessary to know the underreporting rate by roadway type (e.g., rural Interstates, urban Interstates and freeways, urban principal multi-lane arterial, and rural two-lane arterial), by highway geometric type and feature (e.g., road sections, intersections, ramps, and roadsides), by vehicle type (e.g., light vehicles and large trucks), by accident severity type (e.g., fatal, injury, and PDO accidents), and perhaps by lighting and weather conditions.

In addition to the underreporting problem, the location of accidents is also subject to errors. The location of an accident is often estimated by the police, and occasionally it is roughly assigned to the nearest milepost of the route where it occurred. Therefore, assigning vehicle accidents to very short road sections are more susceptible to locational error than assigning them to longer road sections. This uncertainty in accident location creates a so-called errors in dependent variable problem. As a result, extra model uncertainties are introduced. No study has been found in this particular area.

Another source of uncertainty (or error) is the measurement of traffic volume by vehicle type and time of day. Vehicle exposure data, computed from AADT and percent vehicles, come primarily from FHWA's Highway Performance Monitoring System (HPMS), a highway sampling system statistically designed to obtain physical, traffic, and operational information on national highways from a small portion of selected highway sections [FHWA, 1987]. Both AADT and percent vehicles are subject to sampling errors (e.g., daily, day-of-week, seasonal, and spatial variations) and nonsampling errors (e.g., vehicle axle counting and vehicle classification errors). This area is rarely studied. Miaou et al. [1993] attempted to include the sampling errors of vehicle exposure data in their NB regression models.

It should be emphasized that, despite the data quality problem mentioned above, the quality of accident, traffic, and roadway design data has improved significantly over the years because of the use of better computer and other electronic technologies. This trend is expected to continue in the future.

(2) Omitted variables.

As discussed earlier, in developing accident prediction models, driver and vehicle variables are largely unavailable by site. The effects of omitted variables on the selection of probability function P(.) have been discussed earlier and will not be repeated here.

If some of the available covariates *X's* are correlated with some of the omitted variables *Z's*, then the estimated parameters for these *X's* are biased. This bias is sometimes referred to as omitted variable bias. It should be noted, however, that the predictions from the estimated model will still be good if the correlations between these *X's* and *Z's* persist into the predicted future.

When time-series data are available, the effects of those omitted variables which are constant for individual sites or groups of sites could sometimes be captured using the so-called fixed effect models (as opposed to random effect models) in econometrics literature. For a discussion on fixed effect and random effect models, the book *Analysis of Panel Data* by Hsiao [1986] is a good starting reference. Generally speaking, there are two types of fixed effects: fixed time and fixed site. Ideally, fixed time and fixed site effects account for immobile factors specific to individual time intervals or sites. (Note that, by extending this immobile factors concept, fixed effects can very well be applied to a group of time intervals or sites.) For example, an abnormal weather condition at all sites under study has a fixed time effect; while unknown driver and vehicle factors, to the extent that they do not vary over time, have fixed site effects. Most of the accident prediction models that this author is aware of are developed with panel data with a relative short time series, and are typically developed without fixed site effects. At present, it is not clear whether fixed site effects could be well determined using such short time series data. If not, it would be interesting to see if the fixed site effect models can be used to reinforce our overall confidence in the model specification. The author of this report is not aware of any comprehensive study that examined the relative strengths and limitations of using fixed effect and random effect models under the Poisson and NB distributional assumptions in accident prediction modeling. Some research in this area should be useful. For example, the use of empirical Bayes (EB) method to understand and quantify the effects of omitted variables. Note that, traditionally, EB method has been used only to alleviate the so-called regression-to-the-mean site selection bias problem [Morris et al., 1991; Christiansen et al., 1992]. More experimental research on drivers' response to roadway and roadside designs and weather conditions may also be useful.

(3) Non-Poisson or non-NB distribution.

As indicated earlier, the aggregation level used to define a site, e.g., length of road sections or radius of intersections, and the associated length-of-time intervals also play some role in the choice of P(.). The higher the level of aggregation, the less the site can be considered homogeneous, and the further the accident frequency of each site deviates from the Poisson distribution. For example, vehicle accident rate during daytime and nighttime may be very different, failing to disaggregate daytime and nighttime vehicle accidents and associated vehicle miles in the analysis may introduce extra unexplainable variations into the data.

The NB regression model is based on the assumption that the exponent of the omitted variables, $\exp(\Sigma Z_{ij}\gamma_j)$, follows a gamma distribution. For real-world problems, since Z_i are omitted variables, the exact distributions of Z_i are unknown. There is always some possibility that this assumption may not be appropriate. In addition, as discussed earlier, there is more than one NB probability function that one can choose from, each of which implies a specific mean-variance relationship. The most commonly used NB distribution is the one with a quadratic variance function, i.e., the variance is a quadratic function of the mean. The actual variance-mean relationship may differ from this assumed quadratic relationship. More research on this subject is not critical in the near future, but it should be encouraged.

(4) Imprecise mean function.

As indicated earlier, the effects of *X's* on accident frequency *Y* are nonlinear and interactive in nature. This, together with the consideration that the mean function has to be nonnegative, suggests the use of the exponential function in conventional Poisson and NB regression models. Although the exponential function has been widely used in many areas, it is possible that the actual relationship may differ somewhat from the exponential function. Although there are some statistical tests that have recently been developed to help assess the goodness-of-fit of different mean functions [e.g., Cheng and Wu, 1994], these tests are still not widely known in most areas and the power of these tests have not been examined in practice. To date, most of the studies on accidents-flow-roadway design relationships seem to have found the exponential mean function to be satisfactory. Therefore, the statistical research in this area is not likely to be particularly beneficial to the overall development of accidents-flow-roadway design models in the near future. On the other hand, the engineering types of research, as described in chapter 2, which attempts to refine mean functions based on the use of traffic flow theory, geometry, vehicle dynamics, driver behavior models, and probability theory is more likely to help us understand the details of accident process and should be encouraged.

(5) Correlation.

The occurrences of vehicle accidents may be correlated between different road sections in different time periods, rather than independent [e.g., Maher, 1990; Black, 1991]. In the paper by Maher [1990], a phenomenon called accident migration was discussed. This is a phenomenon whereby the accident rate rises at sites that are untreated but are neighbors to treated sites. This phenomenon may be related to the change of drivers' expectation after the site is treated. Overall, this correlation problem has not been perceived to be important in most of the accidents-flow-roadway design relationship studies. In addition, the correlation problem is likely to be reduced when time-specific constants (or fixed time effects) are included in the model.

SOME ADDITIONAL OBSERVATIONS

In this section, some additional observations on the current status of the accident and related data and the development of accident prediction models for roadway planning and design are offered:

(1) Accident frequency and severity level are direct and easy to understand measures of roadway safety. In spite of the criticisms made by many regarding the data quality issues on minor accidents, a good cost-effective alternative measure, that is acceptable to both engineers and policy-makers, is still to be found.

(2) In the last few years, the use of computer, telecommunication, and other electronic technologies has begun to make the data collection, linking, and checking much easier and faster.

(3) The development of the Highway Safety Information System (HSIS) by FHWA has made it possible for safety researchers to statistically analyze the complex interactions of the five major factors described earlier that lead to the occurrence of vehicle accidents. As a result, many important and encouraging results, based on the HSIS data, have been reported.

(4) A cost-effective instrument to obtain necessary and quality data at the national level for estimating the underreporting rates discussed earlier has not yet been identified.

(5) The transferability of models developed using data from one location for use in another location has not been clear. Because the HSIS includes data from many States, it provides a unique opportunity for researchers to study the model transferability issue among States. As a start, this author suggests that, for the same roadway class, estimated model parameters (e.g., β) of the same design element (e.g., horizontal curvature, paved shoulder width) be compared for models developed in different States. A quick check made by this author using the Poisson and NB regression models developed for rural two-lane roads in two of the HSIS States suggested that the estimated parameters of paved shoulder width are very consistent between the two States.

3. COEFFICIENT OF DETERMINATION, R^2

The objective of this chapter is to demonstrate the pitfalls of using the R^2 values to determine the goodness-of-fit of accident prediction models that are typically non-normal and nonlinear. The demonstrations will be given using computer simulated data from commonly used accident prediction models, such as Poisson and NB regression models. Sampling variations (or sampling errors) of R^2 values caused by the use of finite samples are also illustrated in these demonstrations.

Only models with an intercept term are considered in the demonstration. The reason is that all the accident prediction models that this author is aware of have an intercept term. Note that models without an intercept term can have very peculiar statistical properties [Kvålseth, 1985].

As in earlier chapters, a perfect model will be referred to as a model that *(1) has specified a correct probability distribution for the dependent variable; (2) has chosen a correct functional form which describes the relationship between the expected number of accidents and associated covariates; (3) has included all necessary covariates; and (4) has correctly estimated each model parameter.*

The notations used in this chapter shall be consistent with those used in chapter 2. For example, Y_i is a dependent variable representing the accident frequency of the *ith* site/time interval, the observation or sample value of which is denoted by y_i; X_i is a collection of all available covariates or independent variables that affect the accident frequency of the *ith* site/time interval, whose sample values are denoted by x_i; and Z_i is a collection of all omitted independent variables associated with the *ith* site/time interval, whose values (if observed) are denoted by z_i.

First, the formulae for R^2 and $\bar{R}^2$ are presented and interpreted using the concept of conditional and unconditional variances. Second, an interesting simple example which uses a normal linear regression model with an intercept and one covariate is demonstrated. Third, a more sophisticated demonstration using the lognormal regression model is presented. Similar demonstrations for two key types of accident prediction models, Poisson and NB regressions, are then given. At the end, the pitfalls of using R^2 values to determine the goodness-of-fit of accident prediction models as demonstrated in this chapter are summarized.

INTERPRETATION AND FORMULATION OF R^2 AND $\bar{R}^2$

The definition of R^2 which was presented earlier in Eq. (4) is repeated here for convenience.

$$R^2 = \frac{\textit{Total Explained Variance of Y by Available X}}{\textit{Total Unconditional Variance of Y}}$$

$$= 1 - \frac{\textit{Total Unexplained Variance of Y}}{\textit{Total Unconditional Variance of Y}} \qquad (20)$$

Many basic statistics and probability textbooks [e.g., Ross, 1989; Rohatgi, 1976] show that the unconditional variance of Y_i can be broken down into two components: $Var[Y_i]=E[Var[Y_i|X_i, Z_i]] + Var[E[Y_i|X_i, Z_i]]$, where X_i and Z_i represent available and omitted covariates, respectively. Following the discussion in chapter 2, the first component, $E[Var[Y_i|X_i, Z_i]]$, is the random variance, and the second component, $Var[E[Y_i|X_i, Z_i]]$, is the systematic variance. Also, in the first component, $E[Var[Y_i|X_i, Z_i]]$, the variance is first taken over Y_i and then the expectation is taken over X_i and Z_i. Similarly, in the second component, $Var[E[Y_i|X_i, Z_i]]$, the expectation is first taken over Y_i and then the variance is taken over X_i and Z_i.

Now, given Y_i, X_i, and Z_i, Eq. (20) can be statistically expressed as

$$R^2 = \frac{Var[E[Y_i|X_i]]}{Var[Y_i]} = \frac{Var[E[Y_i|X_i]]}{Var[E[Y_i|X_i,Z_i]] + E[Var[Y_i|X_i,Z_i]]} = 1 - \frac{Var[Y_i] - Var[E[Y_i|X_i]]}{Var[Y_i]} \qquad (21)$$

It is clear from Eq. (21) that, even for a perfect linear regression model, the R^2 value from Eq. (21) can not reach 1 if the random variance, $E[Var[Y_i|X_i, Z_i]]$, is not zero. This was shown for a normal linear regression model in chapter 2.

The discussion above applies to hypothetical situations where a correctly estimated model that has no sampling errors is obtained. In practice, the number of data available for developing a model is always finite. And, under finite samples, even if one could specify probability function and the form of mean function correctly and include all necessary covariates (i.e., no omitted variables), it is not possible to estimate parameters precisely. The only guarantee that the analysts have when using a consistent estimator, such as the ML method, is that as sample size increases the uncertainty of the parameter estimates decreases. For a finite sample of size n, a sampling estimate of the R^2 value in Eq. (21) is:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (22)$$

where $\bar{y}$ is sample mean and $\hat{y}_i$ is an estimate of the conditional mean of Y_i given available x's. (Note that in this report $\hat{y}_i$ and $\hat{\mu}_i$ are used interchangeably.) This formulation applies to both linear and nonlinear regression models. Kvålseth [1985] presented several other possible forms for R^2 and recommended that Eq. (22) "ought to be used consistently for any type of model- and curve (surface)-fitting techniques."

To adjust for the degrees of freedom available to (or the number of unknown parameters used in) the model, the adjusted R^2 is typically used and computed as:

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (23)$$

where k is the total number of unknown parameters (including the intercept) in the model.

NORMAL LINEAR REGRESSION MODELS

Barrett in his 1974 paper gave an interesting illustration of the property of R^2 in which he used a simple normal linear regression model with an intercept and one covariate [Barrett, 1974]. Basically, it was shown that if the slope parameter in the model is increased continuously, while keeping everything else fixed (including the centroid and residuals), the R^2 value would continue to increase. In this section, the results of a similar illustration is presented and the reason why the R^2 value increases as the slope parameter increases is explained.

The normal linear regression model under consideration is as follows:

$$Y_i = 1.0 + \alpha x_i + \epsilon_i, \qquad i = 1, 2, \ldots, 50. \qquad (24)$$

where Y_i are dependent variables; x_i are the observed values of covariate X_i; the residuals ϵ_i are distributed as *iid N(0, 0.01)*; and α is the slope parameter. Furthermore, it is assumed that X_i are *iid* and are uniformly distributed between -1 and 1, denoted by $X_i \sim iid\ U[-1,1]$. As before, the observations of Y_i will be denoted by y_i. In addition, the sample value of the residuals ϵ_i will be denoted by e_i. The sample size n, as indicated in Eq. (24), is 50.

The simulation is conducted as follows:

Step 1. Generate ϵ_i from $N(0, 0.01)$ for $i=1,2,\ldots,50$ (denoted by e_i);
Step 2. Generate X_i from $U[-1,1]$ for $i=1,2,\ldots,50$ (denoted by x_i);
Step 3. Set α equal to 0.1;
Step 4. Compute $y_i = 1.0 + \alpha x_i + e_i$ for $i=1,2,\ldots,50$;

Step 5. Compute estimates $\hat{y}_i=1.0+\alpha x_i$ for $i=1,2,...,50$ (best estimates);
Step 6. Compute R^2 using Eq. (22); and
Step 7. Stop, if $\alpha=0.8$; otherwise set α equal to 2α and return to step 4.

Figure 3 shows the simulation results for $\alpha=0.1$, 0.2, 0.4, and 0.8. It shows that the R^2 value increases from 0.439 to 0.966 as α increases from 0.1 to 0.8. Essentially, in this simulation, the angle of the regression line is rotated counter-clockwise from 5.7 degrees to 38.7 degrees around the centroid $(x=0, y=1)$ [see also table 2]. Both e_i and x_i are fixed when rotating. In step 5, $\hat{y}_i$ $=1.0+\alpha x_i$ is the best estimate of the mean of Y_i one can obtain given x_i (which has no sampling errors).

Conceptually, all four estimated models are perfect and should all have a R^2 value of 1. The reason why the R^2 value increases as α increases can be explained using Eq. (21). Basically, in this illustration the random variance, $E[Var[Y_i|X_i]]$, is fixed and is equal to $\sigma^2=0.01$; while the systematic variance, $Var[E[Y_i|X_i]]$, increases as α increases. Note that there is no omitted variables (Z's) in this case. In an illustration like this, since the random variance σ^2 is known, the R^2 measure can easily be adjusted by removing the random variance from the total variance of Y_i as follows:

$$R_0^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 - \sigma^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2 - \sigma^2} \qquad (25)$$

where $\sigma^2=0.01$. Now, by applying Eq. (25), it was found that all four fitted lines in figure 3 indeed have R_0^2 values very close to 1. Unfortunately, for most of the real-world problems with multiple covariates, this adjustment can not be made because it is not possible to distinguish the part of systematic variance that is caused by omitted variables from the random variance.

LOGNORMAL REGRESSION MODELS

As indicated earlier, the lognormal regression model has been shown to be inadequate for developing accidents-flow-roadway design relationships. It is presented here only for illustration purpose.

The simulated lognormal regression model has the following form:

$$Y_i' = log(Y_i) \sim N(\mu_i', \sigma^2) \qquad (26)$$
$$where\ \mu_i' = E[Y_i'|x_i] = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}\ ,\quad i=1,2,...,n.$$

The covariate X_{i1} is a dummy variable equal to 1 (i.e., β_1 is the intercept); x_{i2} is the observed value of random variable X_{i2} which is *iid* as $U[-1,1]$; and x_{i3} is the observed value of random variable X_{i3} which is *iid* and has probabilities of 0.3, 0.4, and 0.3 of being observed to be

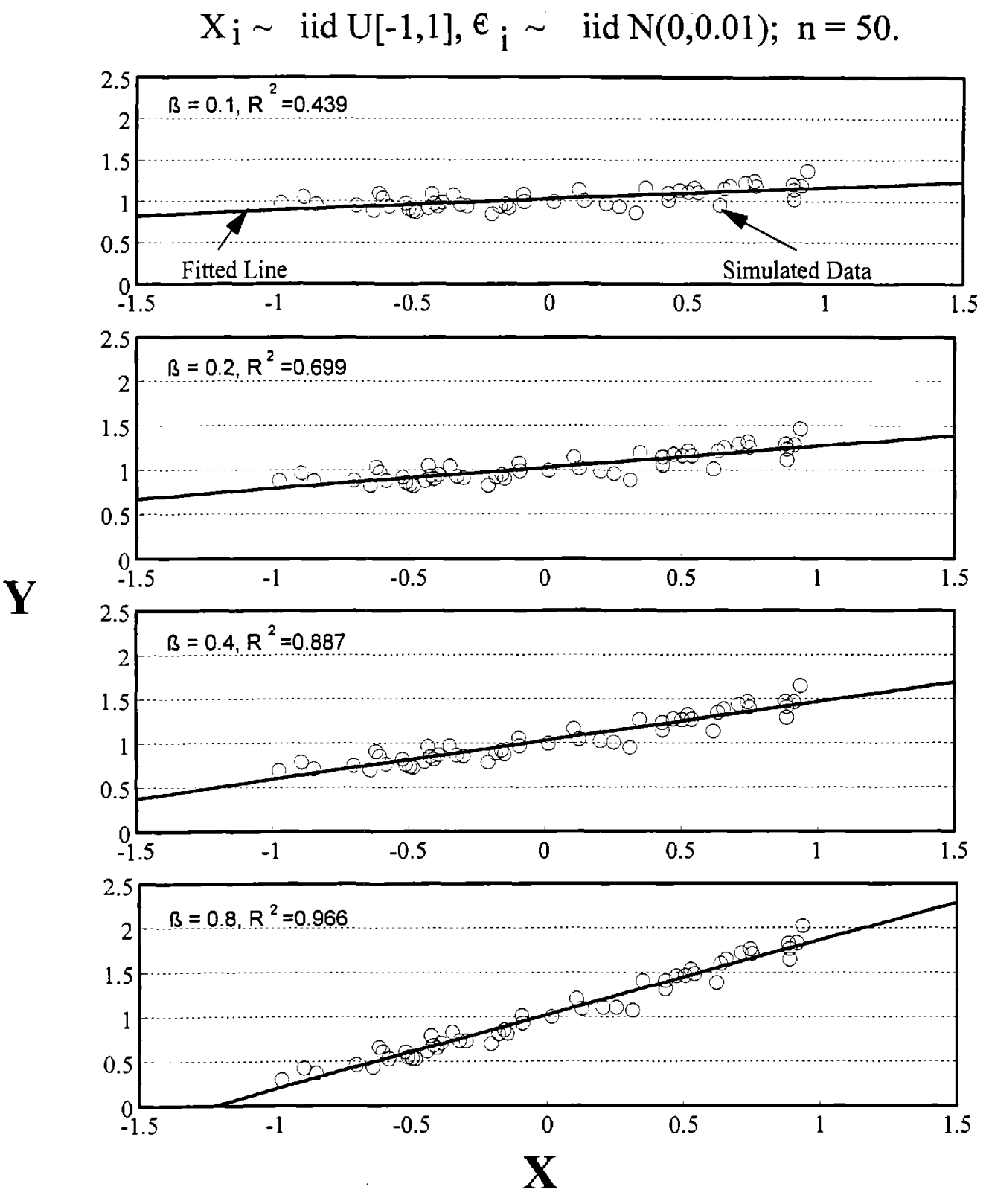


Figure 3. R^2 values of a linear regression model under different slope parameters.

Table 2. R^2 values of a linear regression model under different slope parameters.

Linear Regression Model: $Y_i = 1.0 + \alpha x_i + \varepsilon_i$
$X_i \sim iid\ U[-1,1]$, $\varepsilon_i \sim iid\ N(0, 0.01)$, sample size $n=50$

Slope Parameter (α)	Slope Angle (in degrees)	R^2
$\alpha = 0.0$	0	0
$\alpha = 0.1$	5.7	0.439
$\alpha = 0.2$	11.3	0.699
$\alpha = 0.4$	21.8	0.887
$\alpha = 0.8$	38.7	0.996

-1, 0, 1, respectively. The conditional variance of Y_i^* (or $log(Y_i)$) given x_i is σ^2=0.1. The regression parameters β are set as follows: β_1=0.1, β_2=0.25, and β_3=0.25. Note that X_{i2} and X_{i3} are independent.

The purpose of this simulation is to show that R^2 is subject to sampling error and its value can be substantially less than 1 even if all covariates are included. The simulation is conducted for different sample sizes (*n= 50, 100, 200, 500, 1,000, 2,000, 3,500,* and *5,000*). For each sample size, 10,000 data sets (each with a sample size of *n*) are generated using Eq. (26), then regression parameters are estimated and R^2 values computed for each data set. For example, for *n=50*, the following steps are taken in each simulation run:

Step 1. Set *n=50* and *m=0* (m=data set);
Step 2. Set *m=m+1*;
Step 3. Generate ϵ_i from *N(0, 0.1)* for *i=1,2,...,n* (denoted by e_i);
Step 4. Generate X_{i2} from *U[-1,1]* for *i=1,2,...,n* (denoted by x_{i2});
Step 5. Generate a uniformly distributed deviate *u* from *U[0,1]*; if $u<0.3$, $x_{i3}=-1$; if $0.3 \le u<0.7$, $x_{i3}=0$; and if $u \ge 0.7$, $x_{i3}=1$; for *i=1,2,...,n*;
Step 6. Compute $y_i^*=log(y_i) = 0.1+0.25x_{i2}+0.25x_{i3}+ e_i$ for *i=1,2,...,n*;
Step 7. Generate $y_i = exp(y_i^*)$ for *i=1,2,...,n*;
Step 8. Estimate parameters β's for the lognormal regression model: $Y_i^* = \beta_1+\beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$, with the least squares method using the entire sample of size *n* (the estimated parameters are denoted by $\hat{\beta}_j, j=1,2,3$);
Step 9. Estimate the variance of model residual ϵ_i, denoted by $\hat{\sigma}^2$;
Step 10. Compute the prediction $\hat{y}_i=exp(\hat{\beta}_1+\hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} +0.5\hat{\sigma}^2)$ for *i=1,2,...,50* (best prediction of the conditional mean of Y_i with no omitted variables);
Step 11. Compute R^2 using Eq. (22) and save the value; and
Step 12. Stop, if *m=10,000*; otherwise return to step 1.

This distribution of the 10,000 R^2 values is recorded and its mean, 0.5 percentile, and 99.5 percentile are computed. Figure 4 shows the mean, 0.5 percentile, 99.5 percentile of the 10,000 R^2 values from the simulation for different sample sizes. (Note that figure 4 has three pages. More detailed distribution for each examined sample size is shown on the second and third pages.) For each sample size *n*, the mean of the 10,000 R^2 values is about 0.35. The interval between 0.5 percentile and 99.5 percentile contains 99 percent of the 10,000 R^2 values. The size of the interval decreases rather quickly as the sample size increases. For example, for a sample size of 50, there is a 99 percent probability that the R^2 value will fall within the interval between about 0.12 and 0.63. As the sample size increases to 500, the interval falls between 0.27 and 0.44. Table 3 shows more detailed descriptive statistics of the distribution of these R^2 values by sample size.

In theory, as sample size *n* approaches ∞, the sampling error reduces to zero and thus the parameters are correctly estimated. That is, at $n=\infty$, the perfect model as defined earlier is obtained. The R^2 value under the perfect model can be computed using Eq. (21) as follows:

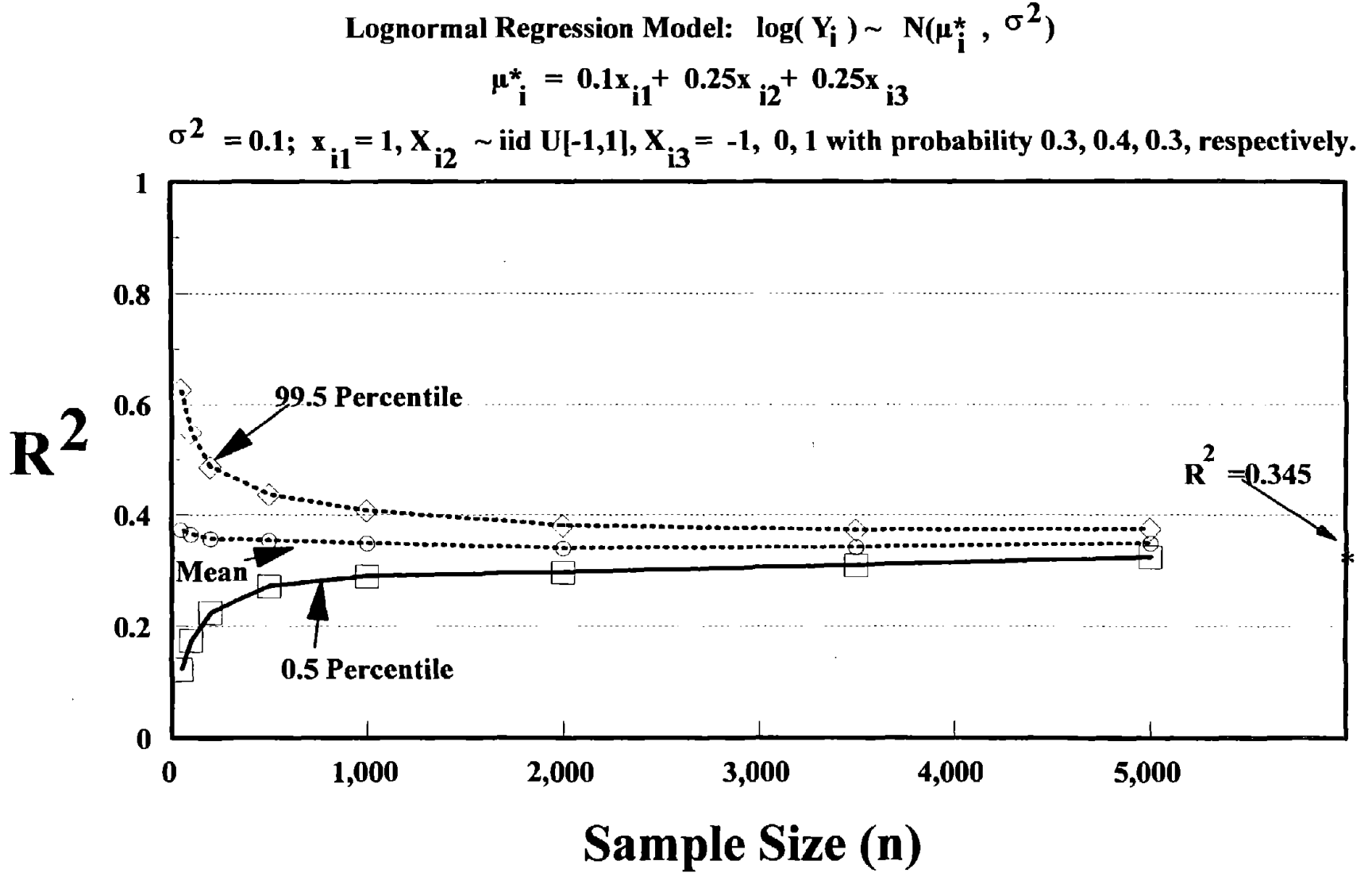


Figure 4. Distribution of R^2 values of 10,000 simulation runs from a lognormal regression model at different sample sizes.

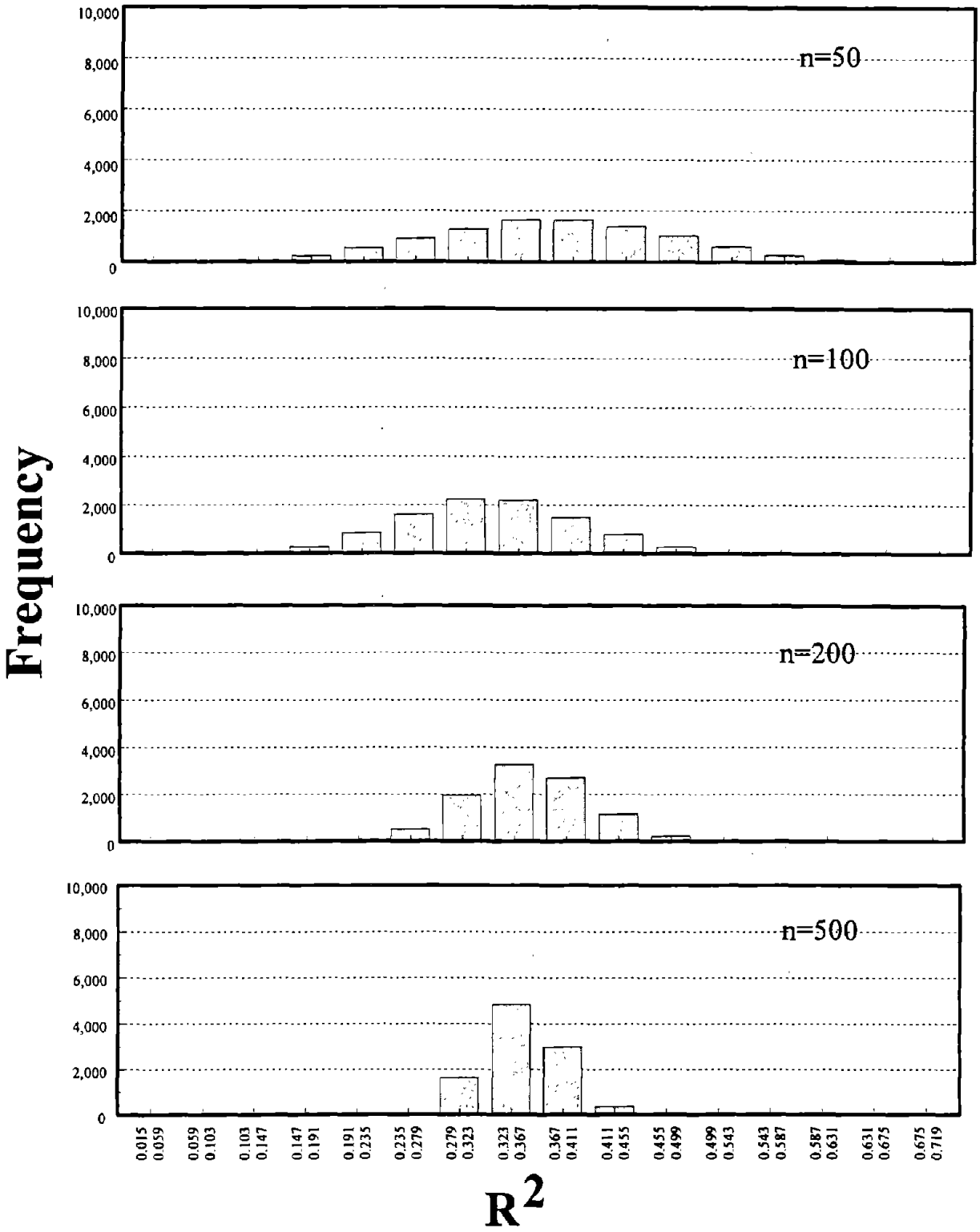


Figure 4. Distribution of R^2 values of 10,000 simulation runs from a lognormal regression model at different sample sizes (Continued).

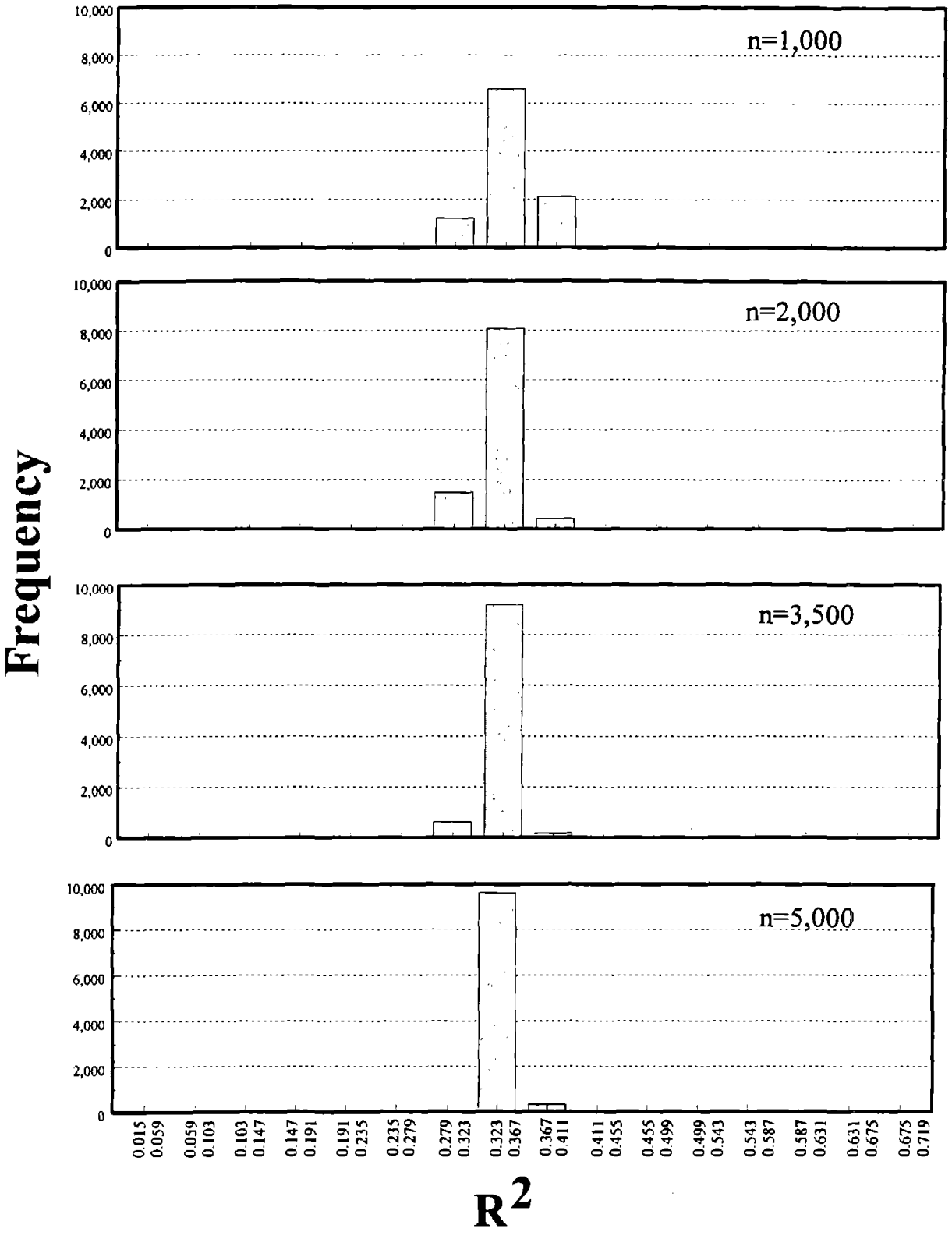


Figure 4. Distribution of R^2 values of 10,000 simulation runs from a lognormal regression model at different sample sizes (Continued).

Table 3. Statistics of R^2 values of 10,000 simulation runs from a lognormal regression model at different sample sizes.

Lognormal Regression Model: $log\ (Y_i) \sim N(\mu_i^*, \sigma^2)$

$\mu_i^* = 0.1x_{i1} + 0.25x_{i2} + 0.25x_{i3}$

$x_{i1} = 1$, $X_{i2} \sim iid\ U[-1,1]$, $X_{i3} = -1, 0, 1$ with probability $0.3, 0.4, 0.3$, respectively. $\sigma^2 = 0.1$

Statistics of R^2 Values	*n=50*	*n=100*	*n=200*	*n=500*	*n=1,000*	*n=2,000*	*n=3,500*	*n=5,000*
Mean	0.373	0.365	0.357	0.354	0.349	0.340	0.342	0.349
Standard Deviation	0.101	0.074	0.050	0.033	0.023	0.016	0.012	0.010
Skewness Coeff.	0.021	-0.001	0.029	0.019	-0.022	-0.030	-0.028	0.007
0.5 Percentile	0.124	0.175	0.225	0.272	0.290	0.297	0.310	0.324
99.5 Percentile	0.628	0.550	0.487	0.437	0.408	0.381	0.373	0.375

(1) Random variance, $E[Var[Y_i|X_i]]$:

Since the conditional probability $P(Y_i|x_i)$ is lognormal, $E[Y_i|X_i]=exp(\beta_1+\beta_2 X_{i2}+\beta_3 X_{i3}+0.5\sigma^2)$ and $Var[Y_i|X_i] = \{E[Y_i|X_i]\}^2 \{exp(\sigma^2)-1\}$ [see e.g., Lindgren, 1976, page 191]. Therefore, the random variance $E[Var[Y_i|X_i]] = E[\{E[Y_i|X_i]\}^2\{exp(\sigma^2)-1\}] = E[\{exp(\beta_1+\beta_2 X_{i2}+\beta_3 X_{i3}+0.5\sigma^2)\}^2\{exp(\sigma^2)-1\}]$. Since X_{i2} is U[-1,1], it can be shown that $E[exp(\beta_2 X_{i2})]=\{exp(\beta_2)-exp(-\beta_2)\}/2\beta_2$. In addition, it can be shown that $E[exp(\beta_3 X_{i3})]=0.3exp(-\beta_3)+0.4+0.3exp(\beta_3)$. Therefore, $E[Var[Y_i|X_i]]= exp(2\beta_1)\{\{exp(2\beta_2)-exp(-2\beta_2)\}/4\beta_2\}\{0.3exp(-2\beta_3)+0.4+0.3exp(2\beta_3)\}exp(\sigma^2)\{exp(\sigma^2)-1\}= 0.1593$.

(2) Systematic variance, $Var[E[Y_i|X_i]]$:

$Var[E[Y_i|X_i]]=Var[exp(\beta_1+\beta_2 X_{i2}+\beta_3 X_{i3}+0.5\sigma^2)]= exp(2\beta_1+\sigma^2)Var[exp(\beta_2 X_{i2}+\beta_3 X_{i3})]$ where $Var[exp(\beta_2 X_{i2}+\beta_3 X_{i3})] =E[exp(2\beta_2 X_{i2}+2\beta_3 X_{i3})]-\{E[exp(\beta_2 X_{i2}+\beta_3 X_{i3})\}^2$. Using the relationships that $E[exp(\beta_2 X_{i2})]=\{exp(\beta_2)-exp(-\beta_2)\}/2\beta_2$ and $E[exp(\beta_3 X_{i3})]=0.3exp(\beta_3)+0.4+0.3exp(\beta_3)$, it can be shown that $Var[E[Y_i|X_i]] =0.0839$.

Therefore, for the perfect model, $R^2=0.0839/(0.0839+0.1593)=0.345$. This R^2 value is consistent with the simulation results shown in figure 4, which converges to a value of about 3.5 as the sample size increases.

POISSON REGRESSION MODELS

The simulated Poisson regression model has the following form:

$$Y_i \sim Poisson(\mu_i) \quad \text{where} \quad \mu_i = E[Y_i|x_i] = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}), \quad i=1,2,...,n. \tag{27}$$

Same as in the lognormal regression illustration above, the covariate X_{i1} is a dummy variable equal to 1 (i.e., β_1 is the intercept); x_{i2} is the observed value of random variable X_{i2} which is distributed as *iid U[-1,1]*; and x_{i3} is the observed value of random variable X_{i3} which is *iid* and has probabilities of 0.3, 0.4, and 0.3 of being observed to be -1, 0, 1, respectively. Since this is a Poisson model, the conditional variance of Y_i given x_i is equal to its conditional mean: $Var[Y_i|x_i] =E[Y_i|x_i]$. The regression parameters β are set as follows: $\beta_1=4$ or 7, $\beta_2=0.1$, and $\beta_3=0.1$. Again, X_{i2} and X_{i3} are independent.

Besides showing that R^2 is subject to sampling errors and its value for a perfect model can be substantially less than 1, this illustration further shows that the R^2 value is dependent on the mean level of the Poisson model. Specifically, higher mean levels result in higher R^2 values. Again, this simulation is conducted for different sample sizes (*n= 50, 100, 200, 500, 1,000, 2,000, 3,500,* and *5,000*). For example, for *n=50* and $\beta_1=4$, the following steps are taken in each simulation run:

Step 1. Set $n=50$ and $m=0$ (m=data set);
Step 2. Set $m=m+1$;
Step 3. Generate x_{i2} from $U[-1,1]$ for $i=1,2,...,n$;
Step 4. Generate a uniformly distributed deviate u from $U[0,1]$; if $u<0.3$, $x_{i3}=-1$; if $0.3 \le u<0.7$, $x_{i3}=0$; and if $u \ge 0.7$, $x_{i3}=1$; for $i=1,2,...,n$;
Step 5. Compute $\mu_i = exp(4.0+0.1x_{i2}+0.1x_{i3})$ for $i=1,2,...,n$;
Step 6. Generate y_i from $Poisson(\mu_i)$ for $i=1,2,...,n$;
Step 7. Estimate parameters β for the Poisson regression model: $Y_i \sim Poisson(\mu_i)$ where $\mu_i = \exp(\beta_1+\beta_2 x_{i2} + \beta_3 x_{i3})$. The ML estimation method is used for the entire sample of size n; (the estimated parameters are denoted by $\hat{\beta}_j, j=1,2,3$);
Step 8. Compute the ML estimate of the conditional mean $\hat{y}_i = \hat{\mu}_i = exp(\hat{\beta}_1+\hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3})$ for $i=1,2,...,50$ (best estimate of the conditional mean);
Step 9. Compute R^2 value using Eq. (22) and save the value; and
Step 10. Stop, if $m=10,000$; otherwise return to step 2.

The same simulation steps were performed for $\beta_1=7$, which has a much higher mean level than when $\beta_1=4$. Figure 5 shows the mean, 0.5 percentile, 99.5 percentile of the 10,000 R^2 values from the simulation for different sample sizes. (Note that figure 5 has three pages. For $\beta_1=4$, more detailed distribution for each examined sample size is shown on the second and third pages.) For each sample size n, the mean of the 10,000 R^2 values is about 0.34 and 0.91 for $\beta_1=4$ and $\beta_1=7$, respectively. The interval between 0.5 percentile and 99.5 percentile contains 99 percent of the 10,000 R^2 values. The size of the interval decreases rather quickly as the sample size increases. For example, for a sample size of 50 and $\beta_1=4$, there is a 99-percent probability that the R^2 value will fall within the interval between about 0.12 and 0.61. As the sample size increases to 500, the interval falls between 0.26 and 0.41. Table 4 shows more detailed descriptive statistics of the distribution of these R^2 values by the sample size.

In theory, as the sample size n approaches ∞, the sampling error reduces to zero and thus the parameters are correctly estimated. That is, at $n=\infty$, the perfect model as defined earlier is obtained. Using the same procedure as in the lognormal regression model, the R^2 value under the perfect model can be computed using Eq. (21) as follows:

(1) Random variance, $E[Var[Y_i|X_i]]$:

Since the conditional probability $P(Y_i|x_i)$ is Poisson, $Var[Y_i|X_i] = E[Y_i|X_i] = exp(\beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3})$. Therefore, the random variance $E[Var[Y_i|X_i]] = E[exp(\beta_1+\beta_2 X_{i2} + \beta_3 X_{i3})]$. Since X_{i2} is U[-1,1], it can be shown that $E[exp(\beta_2 X_{i2})] = \{exp(\beta_2) - exp(-\beta_2)\}/2\beta_2$. It can also be shown that $E[exp(\beta_3 X_{i3})] = 0.3exp(-\beta_3)+0.4+0.3exp(\beta_3)$. Therefore, $E[Var[Y_i|X_i]] = exp(\beta_1)\{\{exp(\beta_2) - exp(-\beta_2)\}/2\beta_2\}\{0.3exp(-\beta_3)+0.4+0.3exp(\beta_3)\}$ which is equal to *54.8534* and *1,101.76* for $\beta_1=4$ and $\beta_1=7$, respectively.

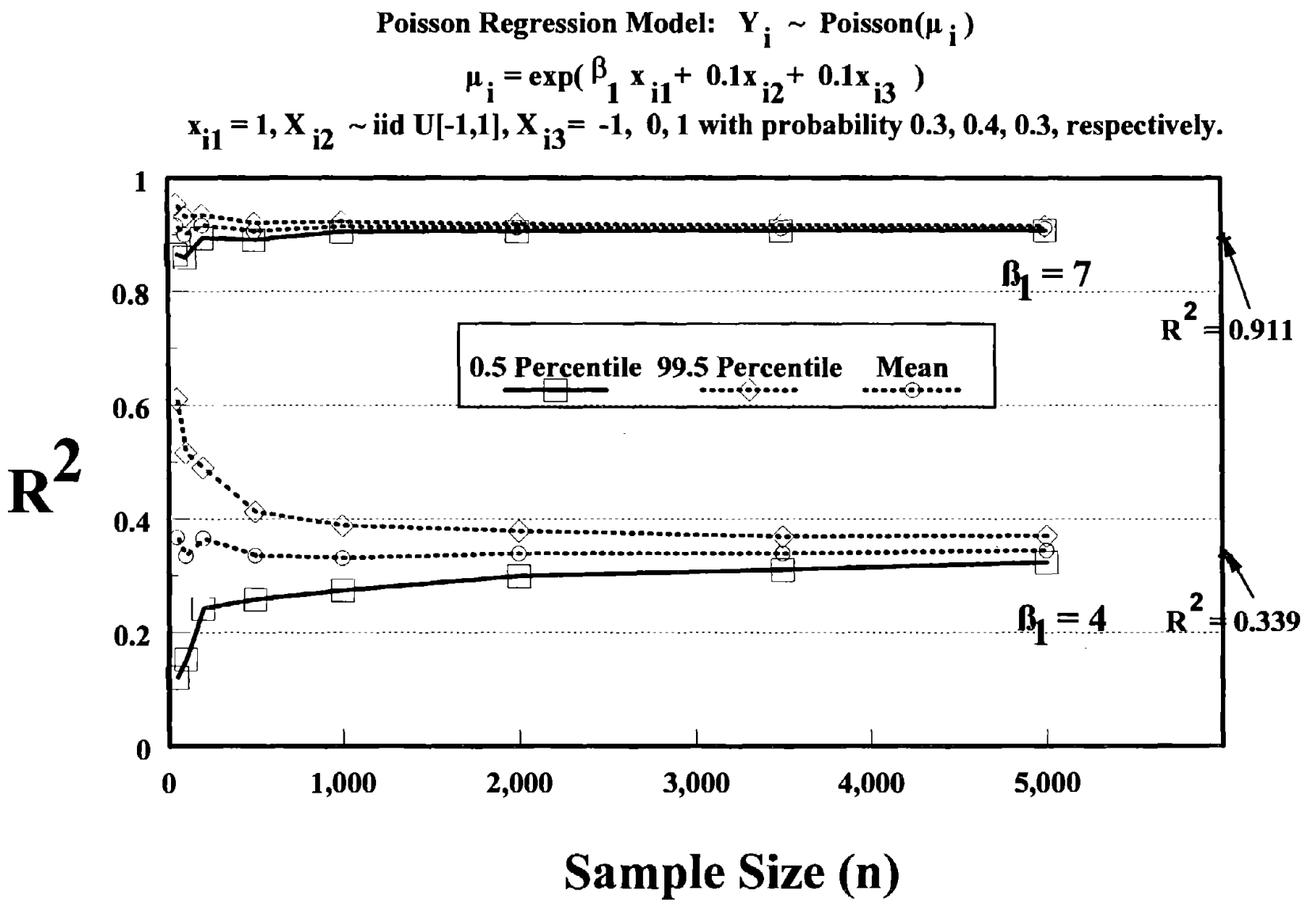


Figure 5. Distribution of R^2 values of 10,000 simulation runs from two Poisson regression models at different sample sizes.

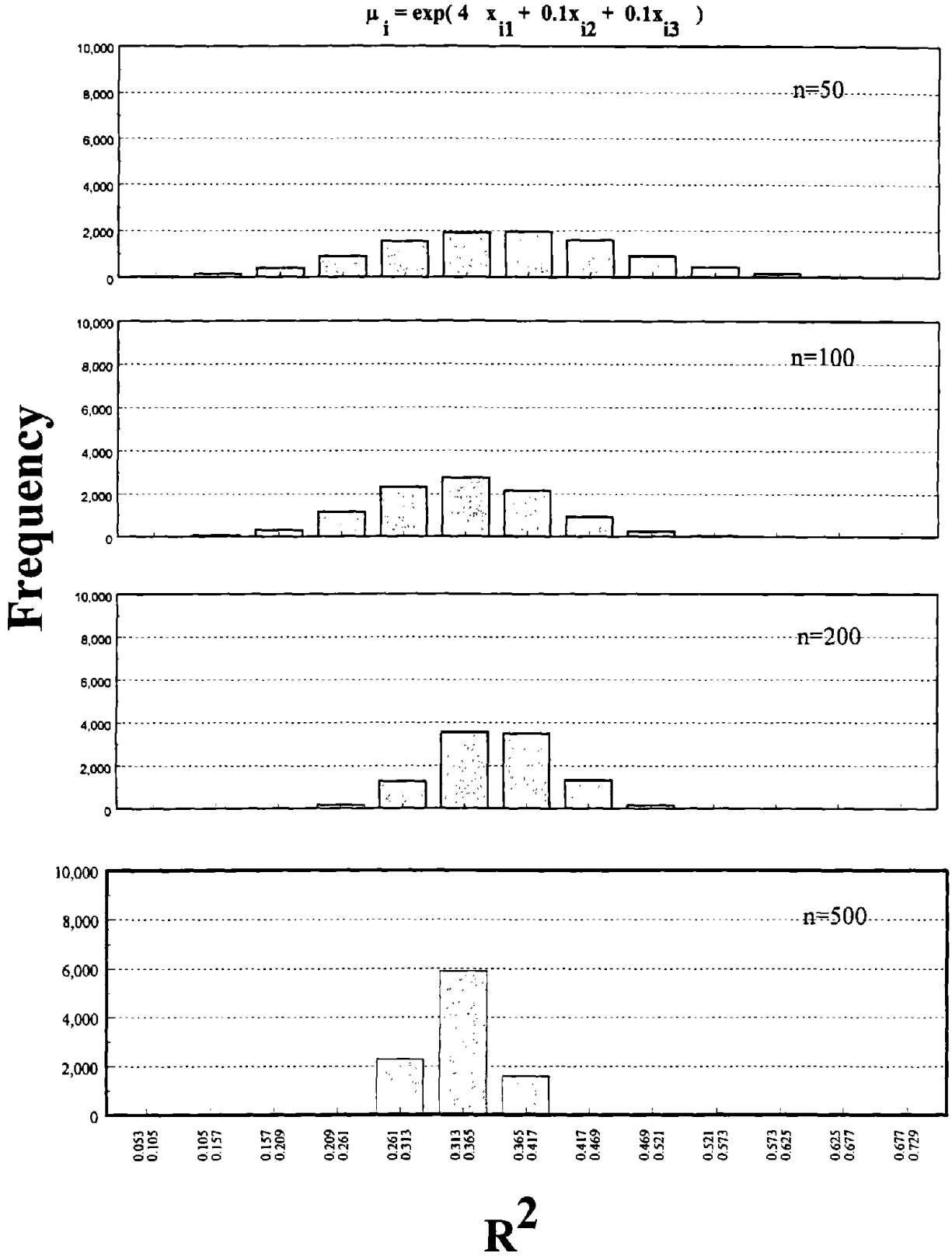


Figure 5. Distribution of R^2 values of 10,000 simulation runs from two Poisson regression models at different sample sizes (Continued).

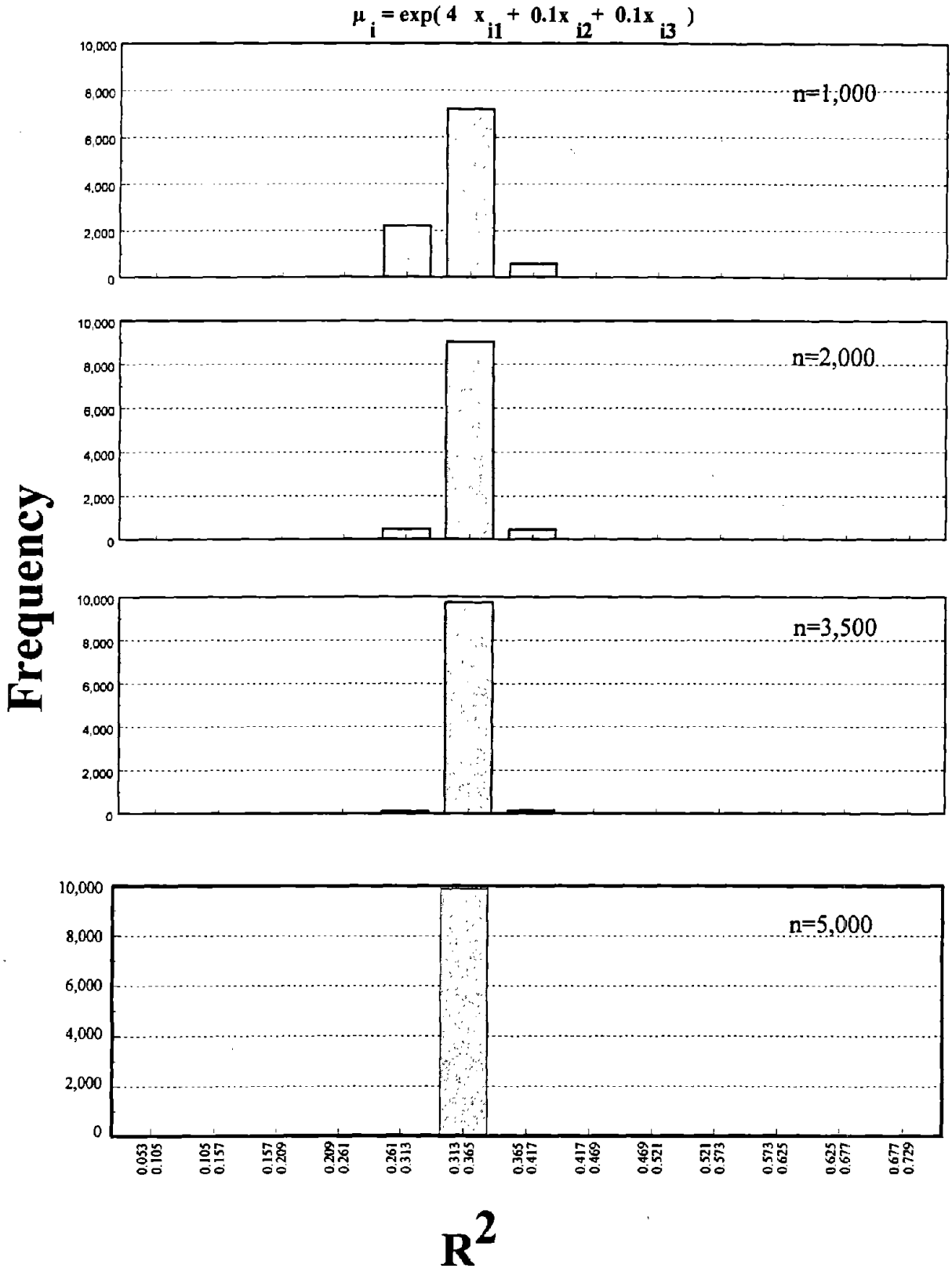


Figure 5. Distribution of R^2 values of 10,000 simulation runs from two Poisson regression models at different sample sizes (Continued).

Table 4. Statistics of R^2 values of 10,000 simulation runs from two Poisson regression models at different sample sizes.

(a) Poisson Regression Model: $Y_i \sim Poisson(\mu_i)$

$\mu_i = exp(4x_{i1} + 0.1x_{i2} + 0.1x_{i3})$

$x_{i1} = 1$, $X_{i2} \sim iid\ U[-1,1]$, $X_{i3} = -1, 0, 1$ with probability $0.3, 0.4, 0.3$, respectively.

Statistics of R^2 Values	*n=50*	*n=100*	*n=200*	*n=500*	*n=1,000*	*n=2,000*	*n=3,500*	*n=5,000*
Mean	0.367	0.335	0.366	0.335	0.330	0.339	0.339	0.344
Standard Deviation	0.099	0.071	0.050	0.031	0.022	0.016	0.012	0.010
Skewness Coeff.	-0.004	0.004	0.023	-0.029	0.013	-0.008	-0.007	-0.004
0.5 Percentile	0.121	0.153	0.242	0.258	0.274	0.299	0.310	0.323
99.5 Percentile	0.611	0.517	0.491	0.413	0.388	0.379	0.369	0.370

(b) Same model as in (a) except that $\mu_i = exp(7x_{i1} + 0.1x_{i2} + 0.1x_{i3})$

Statistics of R^2 Values	*n=50*	*n=100*	*n=200*	*n=500*	*n=1,000*	*n=2,000*	*n=3,500*	*n=5,000*
Mean	0.916	0.900	0.916	0.907	0.914	0.913	0.912	0.912
Standard Deviation	0.017	0.014	0.008	0.006	0.004	0.003	0.002	0.002
Skewness Coeff.	-0.415	-0.306	-0.193	-0.110	-0.082	-0.119	-0.019	0.095
0.5 Percentile	0.865	0.860	0.894	0.891	0.905	0.906	0.907	0.908
99.5 Percentile	0.954	0.932	0.935	0.921	0.923	0.919	0.917	0.916

(2) Systematic variance, $Var[E[Y_i|X_i]]$:

$Var[E[Y_i|X_i]]=Var[exp(\beta_1+\beta_2 X_{i2}+\beta_3 X_{i3})]= exp(2\beta_1)\ Var[exp(\beta_2 X_{i2}+\beta_3 X_{i3})]$ where $Var[exp(\beta_2 X_{i2}+\beta_3 X_{i3})] =E[exp(2\beta_2 X_{i2}+2\beta_3 X_{i3})] - \{E[exp(\beta_2 X_{i2}+\beta_3 X_{i3})\}^2$. Using the relationships that $E[exp(\beta_2 X_{i2})]=\{exp(\beta_2)-exp(-\beta_2)\}/2\beta_2$ and $E[exp(\beta 3Xi3)]=0.3exp(-\beta_3)+0.4+0.3exp(\beta_3)$, it can be shown that $Var[E[Y_i|X_i]]$ is equal to *28.1063* and *11,338.9* for β_1=4 and β_1=7, respectively.

Therefore, for the perfect model, the R^2 values are 0.339 and 0.911, respectively, for β_1=4 and β_1=7. In the last section, although it was not shown that the R^2 value of lognormal regression models also depends on its mean level, from the derivation of R^2 value for the perfect model, it can be seen that it is indeed the case.

The second illustration performed under the Poisson regression model uses the following form:

$$Y_i \sim Poisson(\mu_i)$$
$$\text{where } \mu_i = E[Y_i \mid x_i] = exp(\sum_{j=1}^{6} \beta_j x_{ij}) \quad , \quad i=1,2,...,n. \tag{28}$$

where the covariate x_{i1} is a dummy variable equal to 1(i.e., β_1 is the intercept); $x_{ij}, j=2,...,6$, are the observed values of random variable $X_{ij}, j=2,...,6$, which are *iid* as $N(0,1)$. The parameters β_j is equal to 0.4 for j=2,3,...,6, which means that $x_{ij}, j=2,...,6$, have the same effect on the conditional mean of Y_i and in that sense are equally important. In addition, three different values of β_1 (β_1 = -2.0, -1.0, and 1.0) are used to show the effect of different mean levels on R^2 under the Poisson model.

To see how the R^2 value increases when the covariate is added to the model one at a time, Eq. (21) is used to compute the R^2 value for models with (intercept only), (intercept + 1 covariate), (intercept + 2 covariates), (intercept + 3 covariates), (intercept + 4 covariates), and (intercept + 5 covariates). Note that parameters under each model are assumed to be correctly estimated with no sampling errors or bias. The R^2 values under different numbers of covariates and mean levels are shown in figure 6 and table 5. Using the fact that $exp(X_i)$ are lognormally distributed, the unconditional mean levels of Y_i for β_1 of -2.0, -1.0, and 1.0, are 0.2, 0.55, and 4.1, respectively. Specifically, these means are computed as follows: $E[Y_i]=E[E[Y_i|X_i]]= exp(\beta_1+(5/2)\times(0.4)^2)$. For any number of covariates, as mean level increases, the R^2 value increases. Also, for a particular mean level, as the number of covariates increases, the R^2 value increases. However, it can be seen from figure 6 that the increase is not linear. Since $x_{ij}, j=2,...,6$, in this simulation are independent and equally important, it is desirable that these increases be linear. The reason that the increase is nonlinear is that the mean function μ_i is a nonlinear (exponential) function of the covariates. [Recall that, in chapter 2, the normal linear regression example which has a linear, additive mean function showed a linear increase in the R^2 value when an additional covariate is added to the model (figure 2).] Therefore, this demonstration shows that R^2 has the desirable linear increase property only under linear and additive mean functions. Accident prediction models are, however, nonlinear and interactive in nature.

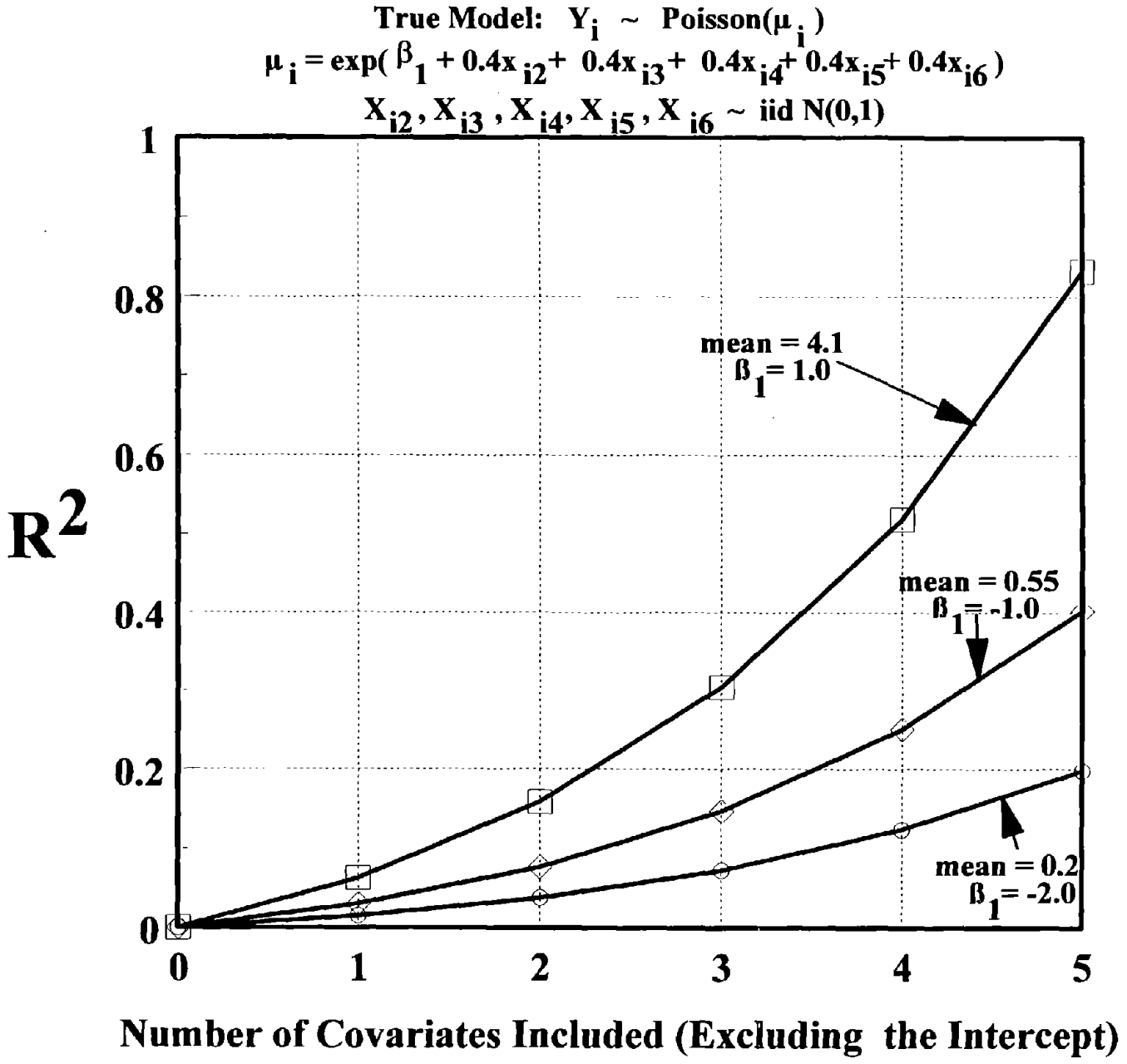


Figure 6. R^2 values of a Poisson model at three mean levels.

Table 5. R^2 values of a Poisson regression model at three mean levels.

True Model: $Y_i \sim Poisson(\mu_i)$

$\mu_i = exp(\beta_1 + 0.4x_{i2} + 0.4x_{i3} + 0.4x_{i4} + 0.4x_{i5} + 0.4x_{i6})$

$X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} \sim iid\ N(0,1)$

Number of Covariates Included (except the intercept)	β_1=1	β_1=-1	β_1=-2
0	0	0	0
1	0.062	0.030	0.015
2	0.159	0.077	0.038
3	0.304	0.147	0.072
4	0.519	0.251	0.124
5	0.832	0.402	0.198

NEGATIVE BINOMIAL REGRESSION MODELS

Same as in the Poisson regression illustration above, it can also be shown that the R^2 value of NB regression models is dependent on mean level and is subject to sampling errors. Here, another factor, overdispersion parameter α, that also influences the R^2 value of a NB regression model is considered. The simulated NB regression model has the following form:

$$Y_i \sim NB(\mu_i; \alpha)$$
$$\text{where } \mu_i = E[Y_i \mid x_i] = \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}), \quad i=1,2,...,n. \tag{29}$$

The covariate x_{i1} is a dummy variable equal to 1 (i.e., β_1 is the intercept); x_{i2} is the observed value of random variable X_{i2}, where $i=1,2,...,n$, which are *iid* as $U[-1,1]$; and x_{i3} is the observed value of random variable X_{i3}, where $i=1,2,...,n$, which are *iid* and have probabilities of 0.3, 0.4, and 0.3 of being observed to be -1, 0, 1, respectively. The conditional variance of Y_i given x_i is a quadratic function of its conditional mean: $\mu_i + \alpha\mu_i^2$ (see table 1). The regression parameters β are set as follows: $\beta_1=1.0$, $\beta_2=0.5$, and $\beta_3=0.5$. Again, X_{i2} and X_{i3} are independent of each other. Two levels of overdispersion parameter α are considered: $\alpha=0.1$ and $\alpha=1.0$.

Using the same simulation procedures as in the Poisson illustration earlier, figure 7 shows the distribution of 5,000 R^2 values for different sample sizes from the NB regression models. Detailed statistics of the R^2 values are given in table 6. On average, the model with a higher overdispersion parameter value has lower R^2 values. The R^2 values as $n=\infty$, computed from Eq. (21), are 0.345 and 0.132 for $\alpha=0.1$ and $\alpha=1.0$, respectively.

The second illustration is based on simulated data obtained from two actual NB regression models that were previously developed in an FHWA project for truck accidents [Miaou et al., 1993]. These two models are presented in table 7. These models were developed using data from Utah, one of the HSIS States. Both models are developed for road sections: one for rural Interstate highway and one for urban Interstate and freeway. Number of large trucks involved in accidents, traffic condition (e.g., AADT and percent trucks), and roadway geometric design data from 1985 to 1989 were used to develop the models. The time period considered in this study was 1 year, which means that the same road section, even if nothing had changed, was considered as five independent sections: one for each year from 1985 to 1989. This allowed the year-to-year changes on highway geometric design and traffic conditions to be considered in the model. There were a total of 8,263 and 2,810 section-years, respectively, for the rural Interstate and urban Interstate and freeway. The average number of trucks involved in accident per section-year is about 0.2 and 0.7 truck accident involvements per section-year, respectively. Detailed description of the data, descriptive statistics of each variable, and variable definitions can be found in Miaou et al. [1993]. Note that the fixed-time effects described earlier in chapter 2 were considered in both models.

The Poisson regression models developed from the same data sets are also presented in table 7 for comparison purpose. It was suggested that NB regression models were more appropriate for the studied data because of the omitted variable problem.

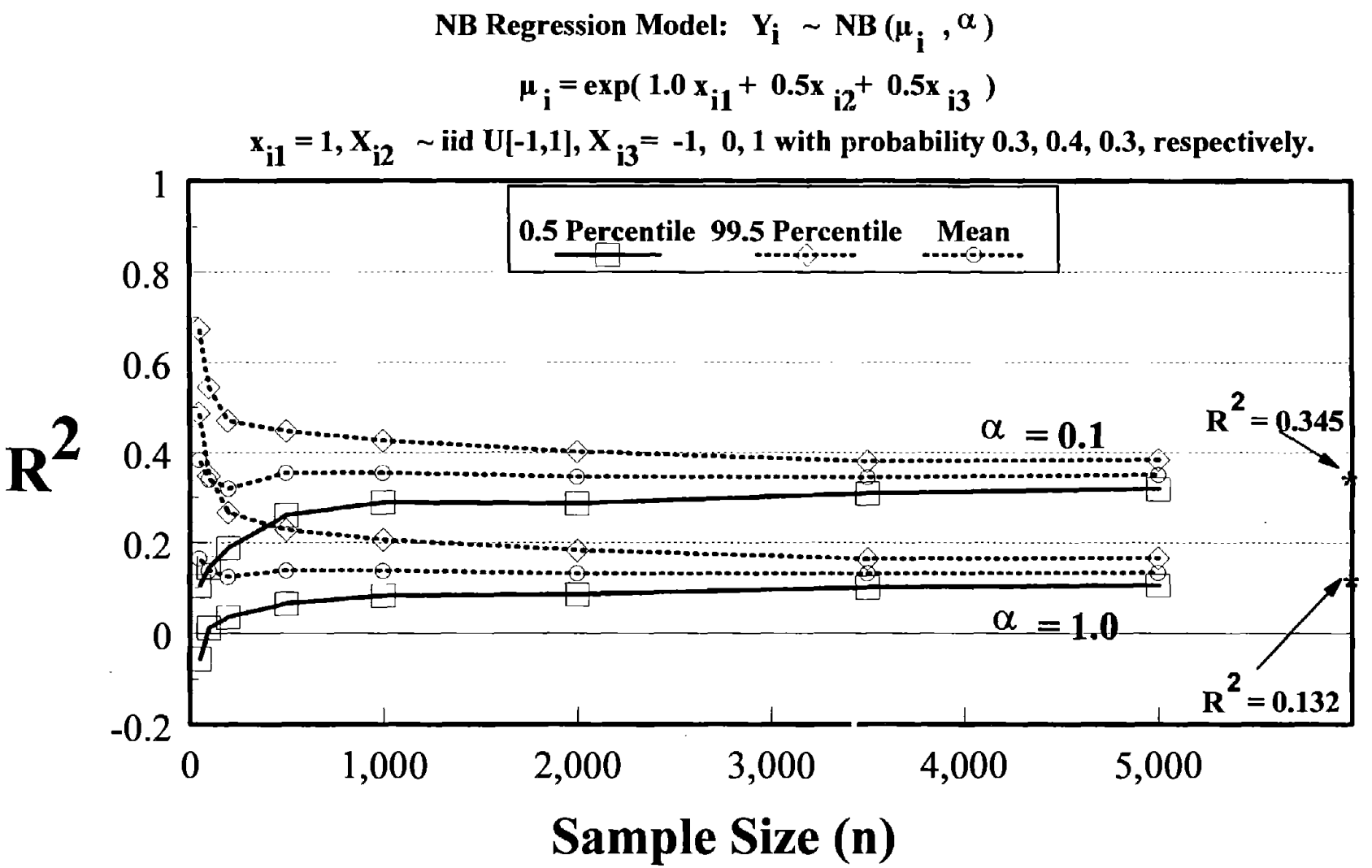


Figure 7. Distribution of R^2 values of 5,000 simulation runs from two NB regression models at different sample sizes.

Table 6. Statistics of R^2 values of 5,000 simulation runs from two negative binomial regression models at different sample sizes.

(a) Negative Binomial Regression Model: $Y_i \sim NB(\mu_i, \alpha)$

$\mu_i = exp(1.0x_{i1} + 0.5x_{i2} + 0.5x_{i3})$, α =0.1

$x_{i1} = 1$, $X_{i2} \sim iid\ U[-1,1]$, $X_{i3} = -1, 0, 1$ with probability *0.3, 0.4, 0.3,* respectively.

Statistics of R^2 Values	n=50	n=100	n=200	n=500	n=1,000	n=2,000	n=3,500	n=5,000
Mean	0.383	0.339	0.319	0.355	0.355	0.346	0.345	0.350
Standard Deviation	0.117	0.079	0.056	0.036	0.029	0.025	0.014	0.017
Coeff. of Skewness	0.072	0.100	0.232	0.057	2.214	7.011	1.110	11.942
0.5 Percentile	0.104	0.145	0.188	0.261	0.289	0.287	0.309	0.319
99.5 Percentile	0.674	0.545	0.470	0.448	0.427	0.402	0.380	0.384

(b) Same model as in (a) except that α =1.0

Statistics of R^2 Values	n=50	n=100	n=200	n=500	n=1,000	n=2,000	n=3,500	n=5,000
Mean	0.165	0.138	0.124	0.138	0.138	0.132	0.132	0.134
Standard Deviation	0.112	0.069	0.047	0.031	0.027	0.026	0.013	0.018
Skewness Coeff.	0.527	0.653	0.611	0.221	5.329	13.944	2.449	21.269
0.5 Percentile	-0.056	0.012	0.035	0.065	0.083	0.085	0.101	0.106
99.5 Percentile	0.488	0.348	0.266	0.228	0.207	0.183	0.164	0.166

Table 7. Estimated parameters of the Poisson and negative binomial regression models for truck accident involvements.

	Rural Interstate (8,263 Section-Years)		Urban Interstate & Freeway (2,810 Section-Years)	
Model Parameter	Poisson	Negative Binomial	Poisson	Negative Binomial
β_1 Dummy intercept	-0.431762 (±0.360;-1.20)	-0.265214 (±0.349;-0.76)	-0.947077 (±0.665;-1.42)	-0.221901 (±0.496;-0.45)
β_2 Dummy variable for 1986	-0.183853 (±0.108;-1.71)	-0.204387 (±0.104;-1.96)	-0.385215 (±0.125;-3.08)	-0.389702 (±0.102;-3.85)
β_3 Dummy variable for 1987	-0.161461 (±0.106;-1.52)	-0.139613 (±0.104;-1.35)	-0.582372 (±0.130;-4.51)	-0.551033 (±0.103;-5.34)
β_4 Dummy variable for 1988	-0.111511 (±0.106;-1.05)	-0.083996 (±0.104;-0.80)	-0.292152 (±0.118;-2.48)	-0.231267 (±0.096;-2.39)
β_5 Dummy variable for 1989	-0.311155 (±0.110;-2.83)	-0.311454 (±0.108;-2.90)	-0.273012 (±0.115;-2.39)	-0.230436 (±0.095;-2.43)
β_{16} Dummy variable for urban freeways	------	------	1.40603 (±0.210;6.68)	1.30802 (±0.161;8.14)
β_6 AADT per lane (10^3)	0.024400 (±0.019;1.27)	0.024621 (±0.020;1.22)	0.046010 (±0.010;4.75)	0.050161 (±0.008;6.30)
β_{17} Number of lanes (4 to 8 lanes)	------	------	0.124950 (±0.053;2.34)	0.088493 (±0.040;2.23)
β_7 Horizontal curvature	0.088861 (±0.035;2.51)	0.073650 (±0.032;2.31)	0.016375 (±0.062;0.26)	0.053897 (±0.044;1.24)
β_{13} Horizontal curvature×Length of original curve	0.234209 (±0.105;2.22)	0.277068 (±0.100;2.77)	0.128738 (±0.152;0.85)	0.049554 (±0.112;0.44)
β_9 Vertical grade	0.077815 (±0.035;2.25)	0.086784 (±0.032;2.72)	0.101143 (±0.056;1.78)	0.093379 (±0.036;2.63)
β_{14} Vertical grade × Length of original grade	0.033973 (±0.019;1.81)	0.027904 (±0.019;1.45)	------	------
β_{11} Deviation of inside shoulder width per direction from 12 ft	0.085763 (±0.045;1.90)	0.070920 (±0.044;1.61)	0.153900 (±0.070;2.20)	0.083181 (±0.052;1.59)
β_{12} Percent Trucks (e.g., 10)	-0.025233 (±0.005;-4.70)	-0.026532 (±0.005;-4.96)	-0.093899 (±0.014;-6.82)	-0.084985 (±0.010;-8.57)
Overdispersion Parameter (α)		0.94652 (±0.107;8.89)		0.58397 (±0.064;9.07)
T_{score} (Score test for overdispersion)	13.49 » 1.96		20.12 » 1.96	
L(β) (=loglikelihood function)	-3771.0	-3682.4	-2741.9	-2620.8
Akaike Information Criterion Value	7566.0	7390.7	5509.7	5269.5
Expected vs. Observed Total Truck Accident Involvements (5-year)	1,644.3 1,643.0	1,702.6 1,643.0	1,903.9 1,904.0	2,039.5 1,904.0

Notes: (1) From Miaou et al. [1993].
(2) Values in parentheses are (adjusted) standard deviation and t-statistics of the parameters above.
(3) ----- Not included in the model.
(4) 1 mi = 1.61 km, 1 ft = 0.3048 m.

For each of the two roadway classes, by assuming that the developed NB regression model is the true model, simulated truck accident involvement frequencies are generated for each section-year. Model parameters are then reestimated using the simulated frequencies and associated covariates. Estimated conditional mean of each section-year is computed from the reestimated model, and the R^2 value is calculated. This process is repeated 5,000 times, and 5,000 R^2 values are obtained. The descriptive statistics of these R^2 values are presented in table 8. The average R^2 value is 0.230 and 0.412 for the rural Interstate and urban Interstate and freeway, respectively. As expected, models for the urban Interstate and freeway have consistently higher R^2 values than those of the rural Interstate. This is mainly because of the fact that the urban Interstate and freeway has a higher overall mean level than that of the rural Interstate (about 0.7 vs. 0.2 truck accident involvements per section-year).

SUMMARY

The R^2 statistic is a measure of the percentage of the unconditional variance of dependent variable that is explainable by the available covariates. It is a meaningful goodness-of-fit measure for normal regression models that have a linear and additive mean function and that the conditional variance of the dependent variable is not a function of the conditional mean. Even under such linear models, a perfect model can have an R^2 value that is substantially less than 1, if the random variation of the data is relatively large when compared to the systematic variation. In theory, if one can estimate the random variance with reasonable accuracy, one can compute R^2 statistic with the random variance being removed from the total unconditional variance of the dependent variable. In such cases, one can indeed reach an R^2 value close to 1 when a perfect model is obtained. Unfortunately, for most of the real-world problems with multiple covariates, the estimate of random variance can not be made because it is not possible to distinguish the systematic variance caused by omitted variables from those caused by the random variance.

As in all statistics, R^2 statistic is subjected to sampling variation (or sampling error) because of the use of finite samples. The larger the sample size, the smaller the sampling error can be expected. This means that the larger the sample size, the better one can trust the R^2 value to decide the goodness-of-fit of the developed models. This fact has, however, not been emphasized by traffic safety engineers and researchers when assessing the goodness-of-fit of accident prediction models using R^2.

The functional form of the commonly accepted candidate accident prediction models, such as the Poisson and NB regression models, are nonlinear and multiplicative. In addition, the Poisson and NB distributional assumptions imply that the conditional variance of the dependent variable is a function of its conditional mean. Under such models, the R^2 statistic has the following three undesirable properties: (1) As in normal linear regression models, a perfect model can have an R^2 value that is substantially less than 1; (2) the R^2 statistic is a function of the mean of the dependent variable; a high mean level would automatically result in a high R^2 value regardless of the goodness-of-fit of the model; and (3) when independent and equally important covariates are selected and added to the model one at a time, the increase in the R^2 value is not linear. These undesirable properties have created potential pitfalls for using R^2 (or $\tilde{R}^2$) to assess

Table 8. Statistics of R^2 values of 5,000 simulation runs using two actual negative binomial regression models for truck accidents.

Statistics of R^2 Values	Model for Rural Interstate (n=8,263 section-years)	Model for Urban Interstate & Freeway (n=2,810 section-years)
Mean	0.230	0.412
Standard Deviation	0.019	0.028
Skewness Coeff.	-0.147	-0.070
0.5 Percentile	0.174	0.339
99.5 Percentile	0.278	0.485

the quality of a model and to make the kind of decisions and comparisons discussed in chapter 1 by traffic safety engineers and researchers.

In light of the fact that appropriate accident prediction models are typically nonlinear and interactive, there is a need to develop alternative measures to evaluate the goodness-of-fit of accident prediction models. Three desirable properties of alternative measures are (1) [0,1] bound property; (2) proportional increase property; and (3) invariant with respect to the mean property. Basically, [0,1] bound property says that one would like to have a value of 0 if no covariate (except the intercept term) is included in the model and a value of 1 if all the necessary covariates are included. Proportional increase property says that, if all covariates are independent and equally important, then when one selects and adds these covariates to the model one at a time, the increase in value should be the same for each covariate regardless of their order of selection. Invariant with respect to the mean property says that the value of the criterion will not change by simply increasing or decreasing the value of the intercept term in the model.

In chapter 5, simulations will be conducted to examine the performance of three alternative goodness-of-fit measures in terms of these three properties. These three alternative criteria are developed on the basis of the Poisson concept described in chapter 2. This concept essentially says that if one can collect all the necessary variables to explain all the variation of accident frequencies among sites and time intervals, then conditional on these variables, the accident frequencies are Poisson distributed. Using this concept, one can first estimate the contribution of random variation in accident data and then remove the estimated random variation from the total variation to obtain an estimate of the total systematic variation which is explainable.

4. AIC AND OTHER GOODNESS-OF-FIT CRITERIA

In the last 20 years, a model selection criterion called AIC has been developed by statisticians [see e.g., Sakamoto, et al., 1986; Bozdogan, 1987]. The capability of this criterion to select the correct models has traditionally been shown in a linear regression or time series context in statistical literature [e.g., Hurvich and Tsai, 1989] and recently in a logistic regression context [Hurvich and Tsai, 1994]. Also, this criterion has been coded as one of the outputs in several statistical software packages. However, few traffic safety engineers and researchers are aware of the development of this criterion.

The second objective of this study was therefore to bring the development of AIC to the attention of traffic safety engineers and researchers. This objective was to be achieved through some illustrations of the power of AIC-based criteria in model selection. Again, the illustrations were to be carried out using computer simulations. It was hoped that through the simulation studies the strengths and limitations of AIC-based criteria in evaluating the goodness-of-fit of accident prediction models could become clearer to traffic safety engineers and researchers. Specifically, the Poisson regression model was used as the ground truth accident prediction model, based on which simulated data were generated and which tests of model selection capability were conducted. In addition to AIC, other criteria such as likelihood-ratio based criterion and Pearson's X^2 statistics were also considered in the illustration for comparison purpose. To the best of this author's knowledge, tests of model selection capability of AIC have not been conducted specifically for the Poisson and NB regression models.

As in earlier chapters, accident prediction models refer to the totality of the model, which includes the probability function P(.), the form of the mean function (i.e., functional form) $f(.)$, regression parameters β, and the covariates X's which are selected for inclusion in the mean function $f(.)$. Conceptually, the probability function, functional form, regression parameters, and covariates can be regarded as four key elements that characterize a model. Candidate models can be different from each other and from the true model in any of these four elements. Ideally, tests of model selection capability can be conducted for candidate models that are different in any of these four elements. In practice, these tests have typically been limited to the comparison of models with the same type of probability function and functional form. Under such tests, a model selection test is reduced to a variable selection test. (Recall that, for a given data set and a selected set of variables, the parameters of each candidate model are estimated based on a predetermined statistical estimation method such as the ML estimation method.) In this study, tests were performed for the Poisson and NB regression models with the same functional form, but different numbers of covariates. Since the Poisson distribution can be considered as a limiting distribution of the NB distribution (as discussed in chapter 2), the tests conducted in this study can also be regarded as comparing models of the same type of probability function. However, this should not be considered a limitation of the tests conducted in this study because the Poisson and NB distributions have been widely accepted in recent years for developing accident prediction models.

Most of the variable selection tests have been conducted by statisticians under the situation where the correct model, which has all the correct covariates in the model, is among a

set of candidate models considered by the modeler. The test is conducted by first generating a data set of a specific sample size, say $n=100$, from the true model using computer simulation techniques. Given the simulated data, the parameters of each candidate model, including the correct model, are estimated based on an appropriate statistical estimation method, e.g., the ML estimation method. The value of each criterion is then calculated for each estimated candidate model using the simulated data. For a specific criterion, the best fitted model among all estimated candidate models is the one which has the best value calculated under the criterion, e.g., the one with the highest value in adjusted R^2 when adjusted R^2 is the criterion considered or the lowest value in AIC when AIC is considered. If the best fitted model identified by the criterion is the correct model, then the use of this criterion to select variable is considered a success. The same experiment with the same sample size is repeated for a number of times, e.g., m=100. Thus, each experiment has a unique data set of the same sample size which represents one possible realization generated from the true model. Finally, the performance of each criterion is measured in terms of the relative frequency the correct model is identified as the best fitted model, e.g., 70 successes out of 100 experiments. In general, the best criterion among all the criteria considered is the one that has the highest relative frequency of successes. In statistical literature, there has been a lot of emphasis on evaluating the performance of these criteria under small sample sizes, e.g., n = 10, 30, 50, 100, or 200. The difficulty in developing accident prediction models is, however, not due to a small sample size, but rather low means and unavailability of certain important covariates.

It is important to point out that the experiment conducted under the situation where the correct model is among a set of candidate models considered by the modeler is not very realistic for many real-world observational studies. The reason is that omitted variables, either unavailable or unobservable, are almost inevitable in these studies. For accident prediction models, as discussed in chapter 2, variables pertaining to site-specific driver and vehicle characteristics are unlikely to be available to the modeler and are almost always omitted from the model. Therefore, the results of the model selection tests performed under such a situation do not seem to be particularly meaningful for the purpose of this study. What the analysts are interested in knowing is the ability of these criteria to select the best model(s) under the situation where some of the relevant variables are omitted from all candidate models. In this chapter, variable selection tests under both situations will be illustrated.

This chapter is organized as follows: First, a brief description of the concept behind AIC is given. Second, a list of criteria that are used in the variable selection tests in the following two sections is presented. Third, the results of the first variable selection test where the correct model is among a set of candidate models considered by the modeler are described. Fourth, the results of the second test where some relevant variables are omitted from all candidate models are presented. Based on the simulation results, the last section gives some recommendations for future research in variable selection.

CONCEPT OF AIC

Let the true model of the *ith* observation be expressed as a probability distribution $P(Y_i = y_i)$, which has a mean function $\mu_i^* = f(X_i, Z_i; \beta^*)$ where X_i and Z_i are covariates and β^* is the true parameter vector associated with the covariates. Further, suppose that a given data set, $y_i, i=1,2,...,n$, which are generated from the true model, are available for analysis. Now, let a candidate model be expressed as $Q(Y_i = y_i)$ which has a mean function $\mu_i = g(X_i, W_i; \beta)$ where X_i are available covariates, W_i are irrelevant covariates, and β is the parameter vector associated with covariates X_i, W_i, and any other parameters that may be required. Note that Z_i represent the variables omitted from the candidate model. In practice, the parameter vector β needs to be estimated from the data. The estimates of β and the mean function evaluated under the estimated parameter will be denoted by $\hat{\beta}$ and $\hat{\mu}_i = g(X_i, W_i; \hat{\beta})$, respectively. As indicated earlier, the true model $P(Y_i = y_i)$ and a candidate model $Q(Y_i = y_i)$ can be different in many ways, e.g., the probability function ($P(.)$ vs $Q(.)$), form of mean function ($f(.)$ vs $g(.)$), regression parameter values (β^* vs $\hat{\beta}$) and covariates (X_i, Z_i vs X_i, W_i). Consequently, for a given observation, say the *ith* observation, the two probability models can be different in their mean (i.e., μ_i^* vs $\hat{\mu}_i$), variance, skewness, and other higher moments.

It is important to point out that the true model is usually unknown. For certain problems, the analysts may have some knowledge of it. For example, as discussed in chapter 2, in developing accident prediction models, the analysts generally feel comfortable to assume that $P(Y_i = y_i)$ is Poisson distributed. Also, from engineering judgment and previous experience with the accident data, the analysts have some idea on the relative importance of certain traffic and geometric design variables. In addition, there seems to be an agreement between safety engineers and researchers that the functional form should be of muliplicative types, indicating the interactive nature of the effects among the five major factors discussed earlier. As discussed in chapter 2, engineering knowledge should play a key role in the determination of functional form and initial candidate variable selection. More discussion and an example on roadside accidents will be given in chapter 6.

Conceptually, the R^2 goodness-of-fit measure discussed earlier can be regarded as a measure of the average distance between y and $\hat{\mu}$, i.e., averaged over all i. Geometrically, it can also be viewed as a measure of the average distance between y and μ^*, plus the average distance between μ^* and $\hat{\mu}_i$. The former distance relates to the random variance referred to in chapter 3, while the latter relates to the unexplained systematic variance. To be more precise, recall that the R^2 measure is further normalized so that the value is bounded between 0 and 1. Models with shorter average distances or with higher R^2 values are favored. Although it is not known, the amount of random variance for a given data set is fixed. Therefore, the R^2 measure is essentially a measure of the average distance between μ^* and $\hat{\mu}_i$, i.e., between the mean of the true model and that of the estimated candidate model. Thus, the R^2 measure ignores model differences in variance, skewness, and other higher moments.

AIC is originated from an information measure called the Kullback-Leibler (K-L) information measure [Sakamoto et al., 1986]. The basic idea of the K-L measure is to measure the closeness of the entire distribution of the true model with respect to that of the estimated

candidate model. In other words, instead of measuring the distance between two means (i.e., two points, geometrically), the K-L information measure attempts to measure the distance between two distributions (which are essentially two curves in the continuous distribution case). It goes without saying that the closer between the two distributions, the better the fit. Of course, there are many ways of measuring the closeness of two distributions. The reasons that the K-L information measure is widely accepted by statisticians include: (1) it is derivable from minimal assumptions; and (2) it possesses several desirable properties [Bozdogan, 1987].

The derivation of AIC from the K-L information measure requires the use of sophisticated large sample theory in statistics. This author will not attempt to derive AIC from the K-L information measure in this report. Instead, only key concepts and assumptions pertaining to the derivation will be presented. The readers who are interested in the derivation can consult Sakamoto, et al. [1986] and Bozdogan [1987] for details.

K-L Information Measure

In this subsection, the K-L information measure will be introduced for continuous distributions. The introduction can be easily extended to discrete distributions.

Let $P(y)$ be the true probability density function and $Q(y)$ be a candidate probability model under consideration, where y is a vector of observations: $y=(y_1, y_2, ..., y_n)'$. The K-L information measure is defined by:

$$I(P,Q) = \int_{-\infty}^{\infty} P(y) \log\left(\frac{P(y)}{Q(y)}\right) dy \qquad (30)$$

where log denotes the natural logarithm.

$I(P,Q)$ has three important and basic properties. The first two are:

(1) $I(P,Q) \geq 0$; and
(2) $I(P,Q) = 0$ if and only if $P(y)=Q(y)$.

That is, $I(P,Q)$ will always be non-negative and $I(P,Q)$ is equal to zero if and only if the candidate model $Q(y)$ is the same as the true model $P(y)$. The third property of the K-L information measure is that when Y_i are independent, $I(P,Q)$ is additive. Note that even if Y_i are not independent, an additive property still exists after appropriately conditioning on one another. In general, the closer the candidate distribution $Q(y)$ is to the true distribution $P(y)$, the smaller the value of $I(P,Q)$. Thus, under the K-L information measure the best candidate model is the one that has the smallest value of $I(P,Q)$.

Equation (30) can be rewritten as:

$$\begin{aligned} I(P,Q) &= \int_{-\infty}^{\infty} P(y)\, log\left(\frac{P(y)}{Q(y)}\right) dy \\ &= \int_{-\infty}^{\infty} P(y)\, log(P(y))\, dy - \int_{-\infty}^{\infty} P(y)\, log(Q(y))\, dy \\ &= Constant \quad - \quad E[\, log(Q(Y))] \end{aligned} \tag{31}$$

The first term on the right hand side of the second equality is a constant that depends on the true distribution $P(y)$ only; while the second term is simply the expression for the expected value of the random variable $log(Q(Y))$. Therefore, to minimize the K-L information measure is equivalent to maximizing the expected value of the loglikelihood of $Q(Y=y)$, i.e., $E[log(Q(Y))]$, where the expectation is taken over Y.

To evaluate $E[log(Q(Y))]$, one needs to know the true distribution $P(Y=y)$. In other words, the K-L information measure in its exact form can be used only if the analysts know the true model. This is, of course, unrealistic. Another unknown which is not explicitly expressed in the discussion is the parameter vector β embedded in the candidate model $Q(y)$. To facilitate the following discussion, the parameter vector β will be made explicit by using $Q(y|\beta)$ in place of $Q(y)$. In the same vein, $E[log(Q(Y|\beta))]$ will be used instead of $E[log(Q(Y))]$.

Akaike Information Criterion

To make the K-L information measure operational, the first step taken by statisticians is to evaluate the goodness-of-fit of models based on $E[log(Q(Y|\hat{\beta}))]$ instead of $E[log(Q(Y|\beta))]$, where $\hat{\beta}$ is the ML estimate of the unknown parameter vector β using observations y. To evaluate $E[log(Q(Y|\hat{\beta}))]$, one still needs the true model. The second step is to estimate the true model from the observations. Akaike's work relates mainly to the second step.

To derive AIC, Akaike relied on the following two key assumptions: (1) the sample size is very large so that asymptotic statistical theory can be applied; and (2) the parameter vector of the candidate model β is a restricted vector of the true parameter vector β^*, i.e., one can obtain β from β^* by setting some of the elements in β^* to zero. The crux of Akaike's finding is that the loglikelihood function of the candidate model evaluated under $\hat{\beta}$ using observations, i.e., $log(Q(y|\hat{\beta}))$, is a biased estimate of $E[log(Q(Y|\hat{\beta}))]$. On average, $log(Q(y|\hat{\beta}))$ is larger than $E[log(Q(Y|\hat{\beta}))]$ by an amount that is about equal to the number of unknown parameters in β, say k. Based on this finding, Akaike defined his AIC as follow:

$$AIC = -2 \times loglike(\hat{\beta}) + 2 \times k \tag{32}$$

where $\hat{\beta}$ is the ML estimate of β, k is the number of unknown parameters in the candidate model, and $loglike(\hat{\beta}) = log(Q(y|\hat{\beta}))$, which is the loglikelihood function of the candidate model evaluated under $\hat{\beta}$ using observations. The best candidate model, according to AIC, is

the one that has the smallest AIC value. The second term on the right hand side can be considered as a penalty for using more parameters (as a result of, e.g., using more covariates). Conceptually, AIC can be regarded as a criterion that strikes a balance between bias and random errors [Bozdogan, 1987]. Specifically, increasing the number of parameters will reduce bias, but it will increase random error. AIC seeks to choose the model that reduces the total error.

Corrected Akaike Information Criterion (CAIC)

Through simulation studies, there has been a steady realization for many years that AIC has a tendency to select overfitted models as the best models, especially when the sample size is small. In the last 10 years or so, there has been a constant effort by statisticians attempting to develop a correction for AIC so that it can perform better under small samples [e.g., Bozdogan, 1987; Hurvich and Tsai, 1989 and 1994]. One version of the correction proposed by Hurvich and Tsai [1989, 1994] has been demonstrated to have better model selection capability than AIC under small samples. Their corrected AIC has the following form:

$$CAIC = AIC + \frac{2k(k+1)}{n-k-1} \qquad (33)$$

One can see from Eq. (33) that for a fixed k and large n, the second term on the right hand side is small and the CAIC is close to the AIC. It is when n is small and k is large that CAIC differs from AIC.

VARIABLE SELECTION CRITERIA CONSIDERED

The criteria that were used to test variable selection capability in this study include AIC and CAIC under both the Poisson and NB regression models, scaled deviance (SD), and Pearson's chi-square statistics. In addition, for the NB model, two modified versions of CAIC were also considered. The modified versions attempted to take into account the size of the dispersion parameter in the NB model. The following is a list of criteria tested:

(1) $\tilde{R}^2$: adjusted R^2 (see chapter 3).
(2) AIC_{PO} : AIC under the Poisson model.
(3) $CAIC_{PO}$: CAIC under the Poisson model.
(4) SD_{PO}: SD under the Poisson model [see McCullagh and Nelder, 1983]; models with lower scaled deviance are preferred.

Scaled deviance is a loglikelihood-ratio based criterion. According to standard statistical theory, for a well fitted model, the value of SD should come from a central χ^2 (i.e., chi-square) distribution with n-k degrees of freedom, denoted by $\chi^2(n-k)$. In this study, the use of this criterion for variable selection is as follows: For any two candidate models M_{j1} and M_{j2}, where j1 and j2 are integers indicating the numbers of covariates included in the two models and j2 is greater than j1, M_{j2} is declared a better model than M_{j1} only if

the SD_{PO} of M_{j2} is less than that of M_{j1} by at least $\chi^2_{0.05}(j2\text{-}j1)$, the critical value of a chi-square distribution with *(j2-j1)* degrees of freedom at a 5 percent significance level. Note that $\chi^2_{0.05}(1)=3.84$, $\chi^2_{0.05}(2)=5.99$, $\chi^2_{0.05}(3)=7.81$, etc.

The steps taken in this study to select the best model among a set of candidate models are as follows:

Step 1. Group the candidate models by the number of covariates included in the model;

Step 2. Within each group, find the best model, i.e., the one that has the lowest value in SD_{PO}; denote these best models by $M_1, M_2, \cdots, M_k$; and

Step 3. Compare SD_{PO} of M_1 and M_2, and the better of the two is kept and then compared with M_3 and so on.

(5) X^2: Pearson's chi-square statistic (see chapter 3 for definition).

The same procedure as in SD_{PO} is used to select the best candidate model.

(6) $CAIC_{NB}$: CAIC under the NB model.

(7) $CAIC_{NB}$-$log(n)$: CAIC under the NB model with a modification on the penalty term:

$$CAIC_{NB}\text{-}log(n) = CAIC_{PO} + log(n) \times \hat{\alpha} \qquad (34)$$

where $\hat{\alpha}$ is the estimated dispersion parameter in the NB model. The additional penalty term is intended for penalizing models with large overdispersions and the penalty is increased as sample size n increases. When $\hat{\alpha}$ is equal to zero, indicating no overdispersion, this criterion is the same as $CAIC_{PO}$.

(8) $CAIC_{NB}$-2: Same as the previous criterion with a different penalty term as follows:

$$CAIC_{NB}\text{-}2 = CAIC_{PO} + 2 \times \hat{\alpha} \qquad (35)$$

The additional penalty term is now independent of sample size n.

VARIABLE SELECTION CAPABILITY TEST: ILLUSTRATION ONE

In this simulation test, the experiment is set up in such a way that the correct model is among a set of candidate models considered. Recall that the correct model is the one that has all the right covariates in the model. The ground truth model is assumed to be conventional Poisson regression models that have exponential mean functions as presented in chapter 2. Candidate models include both the Poisson and NB regression models. There are five covariates in the true

models; the first covariate being a constant (i.e., an intercept). The four nonconstant covariates are assumed to be independent from one another and are simulated as both normal and uniform random variables. Candidate models considered include all possible subsets of these four covariates. For each true model, 100 realizations (or replications) with different mean levels and three sample sizes (n = 50, 100, and 1,000) are simulated and tests are conducted. The parameters of all candidate models are estimated using the ML estimation method.

The model structure of the true models considered is as follows:

$$Y_i \sim Poisson(\mu_i) \quad or \quad P(Y_i = y_i) = P(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \qquad i=1,2,3,\ldots,n. \tag{36}$$

and

$$\mu_i = E[Y_i | x_i] = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5}) \qquad i=1,2,3,\ldots,n. \tag{37}$$

Based on this model structure, different true models are considered and their test results are presented as follows:

True Model #1: x_{i1} =1, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $N(0,1)$; β_1 =1, β_2=β_3=1, β_4=β_5=0.

Essentially this true model has only three covariates, x_{i1}, x_{i2}, and x_{i3} (since β_4=β_5=0). Also, X_{i2} and X_{i3} are equally important in the sense that they are *iid* as $N(0,1)$ and have the same parameter value of 1. It can be shown that the unconditional mean and variance of Y_i under this true model are E[Y_i]=7.39 and Var[Y_i] =348.83, respectively.

The test results from the simulations are presented by criterion in table 9. The first column on the left is a list of all 15 candidate models for each sample size. For example, the model denoted as 1,2,3 is a candidate model that includes covariates X_{i1}, X_{i2}, and X_{i3}, and model 1,2,3,4,5 is a model which includes all covariates. Since the true model includes covariates x_{i1}, x_{i2}, and x_{i3}, the correct model is candidate model 1,2,3. The numbers in columns 2 to 9 indicate the number of times (out of 100 replications) the candidate model on the first column was selected as the best fitted model under each criterion. For example, for sample size n=50, the criterion AIC_{PO} has identified model 1,2,3 as the best fitted model 67 times out of 100 tests. For this example, since model 1,2,3 is the correct model, the success rate is therefore 67 percent.

Several observations can be made from this table:

(1) Under this particular true model, SD_{PO} is a clear winner for all three sample sizes. The success rates are between 88 percent and 90 percent.

(2) X^2, AIC_{PO}, and $CAIC_{PO}$ are all performing reasonably well, with success rates above 70 percent in most cases.

Table 9. Frequencies of models selected by various criteria in 100 replications: True Model #1.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = E[Y_i|x] = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; $x_{i1}=1$, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $N(0,1)$; and $\beta_1=1$, $\beta_2=\beta_3=1$, $\beta_4=\beta_5=0$. ($E[Y_i]=7.39$, $Var[Y_i]=348.83$)

Models & Sample Size (n)	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n = 50								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	24	67	79	90	81	47	47	47
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	28	16	9	5	7	16	16	16
1,2,3,5	27	15	11	4	12	36	36	36
1,2,4,5								
1,3,4,5								
1,2,3,4,5	21	2	1	1		1	1	1
n = 100								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	18	71	74	88	86	40	40	40
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	19	8	8	4	4	7	7	7
1,2,3,5	31	18	15	7	9	48	48	48
1,2,4,5								
1,3,4,5								
1,2,3,4,5	32	3	3	1	1	5	5	5
n = 1,000								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	23	70	70	89	72	24	24	24
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	27	12	12	4	13	5	5	5
1,2,3,5	20	13	13	5	9	60	60	60
1,2,4,5								
1,3,4,5								
1,2,3,4,5	30	5	5	2	6	11	11	11

(3) For a small sample size of 50, $CAIC_{PO}$ does perform better than AIC_{PO}, but as the sample size increases the improvement in performance over AIC_{PO} decreases rather quickly, which is expected.

(4) When the correct model is not correctly identified, all criteria have the tendency to favor larger models that include more covariates than necessary.

(5) $\tilde{R}^2$ performs very poorly. (This is in part because the ML estimation method is used instead of the OLS method. Nevertheless, it has been shown clearly in the literature that the ML estimation method is a much better method to estimate the parameter under the Poisson and NB regression models than the OLS method [McCullagh and Nelder, 1983].)

(6) There is no difference in performance among the three CAIC criteria under the NB model.

True Model #2: x_{i1} =-2, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $N(0,1)$; β_1=1, β_2=β_3=1, β_4=β_5=0.

The only difference between this model and the previous model is on x_{i1}, the value of which drops from 1 in the previous model to -2 in this model. Since x_{i1} is the intercept of the model, this change reduces overall mean and variance of Y_i. It can be shown that the unconditional mean and variance of Y_i under this true model are now E[Y_i]=0.37 and Var[Y_i] =0.86, respectively, which are much smaller than the previous model. In developing accident prediction models, the analysts are often faced with data sets with very low overall means and, therefore, from this perspective this model is more realistic than the previous one.

The test results from the simulations are presented in table 10. Several interesting observations can be made from the table and from the comparison with the previous table. They are listed as follows:

(1) Again, SD_{PO} is a clear winner for all three sample sizes. The success rate increases from 79 percent to 94 percent as the sample size increases from 50 to 1,000. This indicates the importance of having large samples when the overall mean is low, which is common in developing accident prediction models.

(2) The performance of X^2 drops significantly, even under the large sample size of 1,000. This is an indication that X^2 should not be used for variable selection when the average accident frequency per site is low (e.g., less than 1).

(3) AIC_{PO} and $CAIC_{PO}$ are all still performing reasonably well, with success rates above 70 percent in most cases. This indicates that the performance of AIC_{PO} and $CAIC_{PO}$ is relatively unaffected by the mean level of the data.

(4) Except X^2 under $n=50$, when the correct model is not correctly identified, all other criteria again have the tendency to favor larger models which include more covariates than necessary. This indicates that there are some risk of selecting a model which includes unrelated covariates when these criteria are applied. Again, this points up the importance of exercising engineering knowledge to identify candidate variables in the beginning of the model development (as emphasized in chapter 2).

Table 10. Frequencies of models selected by various criteria in 100 replications: True Model #2.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; x_{i1}=-2, X_{i2}, X_{i3}, X_{i4},X_{i5} are *iid* as $N(0,1)$; and β_1=1, β_2=β_3=1, β_4=β_5=0. ($E[Y_i]$=0.37, $Var[Y_i]$=0.86)

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2								
1,3		3	4	7	20	5	4	4
1,4								
1,5								
1,2,3 (Correct Model)	25	66	70	79	47	64	64	64
1,2,4								
1,2,5					3			
1,3,4	2			2	3			
1,3,5	2	2	3	3	5	3	3	3
1,4,5								
1,2,3,4	20	14	12	3	12	14	15	15
1,2,3,5	34	12	9	5	7	13	13	13
1,2,4,5								
1,3,4,5	1		1	1		1	1	1
1,2,3,4,5	16	3	1		3			
n=100								
1,2				2	4			
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	29	74	78	89	54	64	63	63
1,2,4						1	1	1
1,2,5					3			
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	25	6	4	3	8	8	8	8
1,2,3,5	14	16	15	5	25	25	26	26
1,2,4,5								
1,3,4,5								
1,2,3,4,5	29	4	2	1	6	2	2	2
n=1,000								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	21	70	70	94	48	41	41	41
1,2,4								
1,2,5								
1,3,4								
1,3,5					1			
1,4,5								
1,2,3,4	26	15	15	2	24	17	17	17
1,2,3,5	24	14	14	3	15	37	37	37
1,2,4,5								
1,3,4,5								
1,2,3,4,5	29	1	1	1	12	5	5	5

(5) $\tilde{R}^2$ still performs very poorly.

(6) Again, there is no difference in performance among the three CAIC criteria under the NB model. However, their performance improves quite significantly. At present, it is not clear why lowering the overall mean and variance would improve the success rate of these three criteria.

True Model #3: x_{i1} =1, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $N(0,1)$; β_1=1, β_2=1, β_3=0.5, β_4=0.25, and β_5=0.

This model is similar to True Model #1. The four parameters associated with the nonconstant covariates are however changed. The values of these parameters are set in such a way that they decrease from 1 to 0: β_2=1, β_3=0.5, β_4=0.25, and β_5=0. This choice of parameter values allows the relative importance of these four covariates in explaining the variations of Y_i to decrease from x_{i2} to x_{i5}, with x_{i2} being the most important covariate, x_{i3} the second important, x_{i4} the third, and x_{i5} of no importance. The correct model in this simulation test is model 1,2,3,4. However, since β_4 is quite small when compared to β_2 and β_3, it is reasonable to accept model 1,2,3 as a good model. Therefore, in this simulation, a success is declared if model 1,2,3,4 or model 1,2,3 is selected. It can be shown that the unconditional mean and variance of Y_i under this true model are now $E[Y_i]$ =5.24 and Var[Y_i] =74.55, respectively. Therefore, this model has about the same overall mean as in the True Model #1, but with significantly lower overall variance.

The test results from the simulations are presented in table 11. One observation can be made from the table is that while SD_{PO} is a winner again, X^2 is performing equally well. Except $\tilde{R}^2$, every criterion performs quite well. The main reason is that the true model has a smaller random variance.

True Model #4: x_{i1} =-2, X_{i2}, X_{i3}, X_{i4},X_{i5} are *iid* as $N(0,1)$; β_1=1, β_2=1, β_3=0.5, β_4=0.25, and β_5=0.

The only difference between this model and the previous model is on x_{i1}, the value of which drops from 1 in the previous model to -2 in this model. This change reduces overall mean and variance of Y_i. It can be shown that the unconditional mean and variance of Y_i under this true model are now $E[Y_i]$=0.26 and $Var[Y_i]$ =0.18, respectively, which are much smaller than the previous model.

The test results from the simulations are presented in table 12. No criterion is performing well for n=50 and 100. When n=1,000, SD_{PO} , AIC_{PO}, and $CAIC_{PO}$ perform very well. Again, this indicates the importance of having a large sample size when the overall mean is very low. Also, as in the second model, the performance of X^2 is very poor even under the large sample size of 1,000. Again, this clearly suggests that X^2 should not be used for variable selection when the average accident frequency per site is very low.

Table 11. Frequencies of models selected by various criteria in 100 replications: True Model #3.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; $x_{i1}=1$, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $N(0,1)$; and $\beta_1=1$, $\beta_2=1$, $\beta_3=0.5$, $\beta_4=0.25$, $\beta_5=0$. $(E[Y_i]=5.24, Var[Y_i]=74.55)$

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Good Model)		1	1	3	4	2	2	2
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	54	86	91	94	93	85	85	85
1,2,3,5								
1,2,4,5								
1,3,4,5								
1,2,3,4,5	46	13	8	3	3	13	13	13
n=100								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Good Model)	3							
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	62	90	92	97	98	62	62	62
1,2,3,5								
1,2,4,5								
1,3,4,5								
1,2,3,4,5	35	10	8	3	2	38	38	38
n=1,000								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Good Model)								
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	44	83	83	96	92	69	69	69
1,2,3,5								
1,2,4,5								
1,3,4,5								
1,2,3,4,5	56	17	17	4	8	31	31	31

Table 12. Frequencies of models selected by various criteria in 100 replications: True Model #4.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; x_{i1}=-2, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $N(0,1)$; and β_1=1, β_2=1, β_3=0.5, β_4=0.25, β_5=0. ($E[Y_i]$=0.26, $Var[Y_i]$=0.18)

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2	7	22	26	40	30	26	25	26
1,3	1	7	8	12	16	8	8	8
1,4		1	1	1	1			
1,5				3	5	1	1	1
1,2,3 (Good Model)	21	27	27	26	19	28	29	28
1,2,4	2	6	6	1	5	7	6	7
1,2,5	8	7	9	4	5	9	9	9
1,3,4	1			1	1			
1,3,5		2	1	1	1	2	1	2
1,4,5								
1,2,3,4 (Correct Model)	23	20	16	6	7	14	15	14
1,2,3,5	14	4	3	3	3	3	3	3
1,2,4,5	5	2	1	1	2	1	1	1
1,3,4,5	1				1			
1,2,3,4,5	17	2	2	1	4	1	2	1
n=100								
1,2	4	10	13	27	17	16	13	13
1,3	1				1			
1,4					1			
1,5					2			
1,2,3 (Good Model)	23	44	45	43	22	37	37	38
1,2,4	9	13	13	15	22	13	14	14
1,2,5	2	2	2	1	3	2	2	2
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	18	25	21	12	21	25	26	26
1,2,3,5	14	2	2	0	2	3	3	3
1,2,4,5	3	2	2	2	5	2	2	2
1,3,4,5					1			
1,2,3,4,5	26	2	2	0	3	2	3	2
n=1,000								
1,2					1			
1,3								
1,4								
1,5								
1,2,3 (Good Model)	1				21			
1,2,4					4			
1,2,5					1			
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	54	87	87	96	48	71	70	70
1,2,3,5		1	1	1	6	1	1	1
1,2,4,5					1			
1,3,4,5								
1,2,3,4,5	42	12	12	3	18	28	29	29

True Models #5 to 8: These models correspond to True Models #1 to 4. The only difference is that the four nonconstant covariates are now *iid* as *U[-1,1]*, instead of *iid* as *N(0,1)*. The four covariates are still independent from one another.

The test results from the simulations are presented in tables 13-16. The observations made from these tables are consistent with those observed in tables 9-12. There is however an interesting exception that deserves some discussion. In tables 9-12, the performance of the three CAIC criteria under the NB model is very bad when compared with the performance of AIC_{PO} and $CAIC_{PO}$. Under these new true models, however, their performances are as well as those of AIC_{PO} and $CAIC_{PO}$. This change in performance comes mainly from the change of the distribution of covariates. This author's preliminary analysis suggests that it has something to do with the difference between the distribution of *exp(X)* when *X* is *N(0,1)* and when *X* is *U[-1,1]*. Figure 8 illustrates the difference of these two distributions. One plausible reason for this change is that the distribution of *exp(X)* when *X* is *N(0,1)* can be better approximately by a gamma distribution than when *X* is *U[-1,1]*. Recall the discussion on the relationship between NB model and gamma distribution in chapter 2. However, more analytical research will be needed for a thorough understanding of this observation.

VARIABLE SELECTION CAPABILITY TEST: ILLUSTRATION TWO

As indicated earlier, the experiment under the situation that the correct model is among a set of candidate models considered by the modeler may not be realistic in developing accident prediction models. What would be of more interest to our study is the ability of these criteria to select the best model(s) under the situation where some of the relevant variables are omitted from all candidate models. That is, the correct model is not among the candidate models considered by the modeler.

It is however not easy to set up a simulation study for such a situation. One has to determine, e.g., the importance of the omitted variables relative to other variables in explaining the variations of Y_i, the degree of correlation between omitted variables and available covariates, and the distribution of omitted variables. From this author's previous experience in developing accident prediction models for road segments, site-specific driver and vehicle variables may be responsible for 10 to 40 percent of the variation of the accident frequencies among sites. (Of course, traffic volume is the key determinant of the variation.) This author has to admit at this point however that there is quite limited data and research on the relative importance of these driver and vehicle omitted variables in developing accident prediction models. One thing is certain is that for the simulation results to be useful for accident prediction modeling, omitted variables should be explaining a significant portion of the variation of Y_i in the true model. As to the correlation issue, it is this author's judgment at this time that the correlation between site-specific driver and vehicle variables and traffic and geometric design variables are very weak, and therefore the use of independent assumption between the available and omitted covariates in the simulation is reasonable. As to the distribution of site-specific driver and vehicle variables, this author does not have a good idea at this time what the appropriate distributions are to simulate these variables. However, it is unlikely to be perfectly normal or uniform. Probably,

Table 13. Frequencies of models selected by various criteria in 100 replications: True Model #5.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; $x_{i1}=1$, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $U[-1,1]$; and $\beta_1=1$, $\beta_2=\beta_3=1$, $\beta_4=\beta_5=0$. $(E[Y]=3.75, Var[Y]=10.21)$

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	39	73	83	94	93	84	84	84
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	17	12	10	4	1	9	9	9
1,2,3,5	26	12	6	2	5	6	6	6
1,2,4,5								
1,3,4,5								
1,2,3,4,5	18	3	1		1	1	1	1
n=100								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	27	70	71	93	89	74	74	74
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	27	12	11	2	3	10	10	10
1,2,3,5	26	17	17	5	7	15	15	15
1,2,4,5								
1,3,4,5								
1,2,3,4,5	20	1	1		1	1	1	1
n=1,000								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Correct Model)	33	69	70	90	84	72	72	72
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	26	19	18	7	11	16	16	16
1,2,3,5	25	10	10	3	3	10	10	10
1,2,4,5								
1,3,4,5								
1,2,3,4,5	16	2	2		2	2	2	2

Table 14. Frequencies of models selected by various criteria in 100 replications: True Model #6.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; $x_{i1}=-2$, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $U[-1,1]$; and $\beta_1=1$, $\beta_2=\beta_3=1$, $\beta_4=\beta_5=0$. ($E[Y_i]=0.19$, $Var[Y_i]=0.025$)

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2	7	17	19	31	20	22	20	21
1,3	10	22	25	36	16	26	22	24
1,4	2	7	8	9	7	8	8	8
1,5	1	1	1	1	6	1	1	1
1,2,3 (Correct Model)	19	31	31	16	19	29	32	30
1,2,4	11	9	8	3	5	7	8	8
1,2,5	3			1	1	1	1	1
1,3,4	2	4	2	1	2	2	2	2
1,3,5	4	2	3	1	4	3	3	4
1,4,5								
1,2,3,4	9							
1,2,3,5	15	6	2		10			
1,2,4,5	5				7			
1,3,4,5	1							
1,2,3,4,5	11	1	1	1	3	1	2	1
n=100								
1,2	5	9	11	26	15	13	10	10
1,3	7	13	14	22	16	14	12	12
1,4	1	1	1	3	2	1	1	1
1,5				1	3			
1,2,3 (Correct Model)	15	48	50	37	27	49	50	51
1,2,4	4	1	2	1	2	2	2	2
1,2,5	2	3	3		3	2	3	3
1,3,4	3	2	3	3	3	3	3	3
1,3,5			1		2	1	1	1
1,4,5					2			
1,2,3,4	20	7	4	2	10	4	5	4
1,2,3,5	22	10	6	5	11	6	7	7
1,2,4,5	1				1			
1,3,4,5	2	3	2		1	2	2	2
1,2,3,4,5	18	3	3		2	3	4	4
n=1,000								
1,2					4			
1,3					2			
1,4								
1,5								
1,2,3 (Correct Model)	35	75	75	92	48	75	74	75
1,2,4					1			
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	20	9	9	3	20	9	9	9
1,2,3,5	24	13	13	4	18	13	14	13
1,2,4,5					1			
1,3,4,5					1			
1,2,3,4,5	21	3	3	1	5	3	3	3

Table 15. Frequencies of models selected by various criteria in 100 replications: True Model #7.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; $x_{i1}=1$, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $U[-1,1]$; and $\beta_1=1$, $\beta_2=1$, $\beta_3=0.5$, $\beta_4=0.25$, $\beta_5=0$. ($E[Y_i]=3.36$, $Var[Y_i]=5.09$)

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Good Model)	27	36	44	62	65	45	45	45
1,2,4	2		2	7	5	3	3	3
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	42	49	44	28	25	42	42	42
1,2,3,5	10	9	7	2	3	7	7	7
1,2,4,5	1							
1,3,4,5								
1,2,3,4,5	18	6	3	1	2	3	3	3
n=100								
1,2					1			
1,3								
1,4								
1,5								
1,2,3 (Good Model)	6	13	18	33	29	19	19	19
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	56	72	67	64	67	68	68	68
1,2,3,5	3	2	2	1		1	1	1
1,2,4,5								
1,3,4,5								
1,2,3,4,5	35	13	13	2	3	12	12	12
n=1,000								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Good Model)								
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4 (Correct Model)	63	82	82	96	92	82	82	82
1,2,3,5								
1,2,4,5								
1,3,4,5								
1,2,3,4,5	37	18	18	4	8	18	18	18

Table 16. Frequencies of models selected by various criteria in 100 replications: True Model #8.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5})$; $x_{i1}=-2$, X_{i2}, X_{i3}, X_{i4}, X_{i5} are *iid* as $U[-1,1]$; and $\beta_1=1$, $\beta_2=1$, $\beta_3=0.5$, $\beta_4=0.25$, $\beta_5=0$. $(E[Y_i]=0.17, Var[Y_i]=0.0183)$

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2	14	43	48	56	21	50	43	43
1,3	4	10	11	14	18	11	10	11
1,4		5	6	9	12	6	4	4
1,5		6	6	6	7	6	6	6
1,2,3 (Good Model)	13	14	11	7	15	9	13	12
1,2,4	11	10	7	4	3	7	11	11
1,2,5	2	2	4	2	7	3	2	2
1,3,4	2	1	1		2	1	1	1
1,3,5	3	3	3	1	2	3	3	3
1,4,5	2	1	1		1	1	1	1
1,2,3,4 (Correct Model)	11	2	1	1	5	2	2	2
1,2,3,5	17	2	1		3	1	4	4
1,2,4,5	8	1			4			
1,3,4,5	2							
1,2,3,4,5	11							
n=100								
1,2	12	38	41	55	25	42	37	40
1,3	1	5	6	12	16	7	6	6
1,4	1	2	2	3	2	2	2	2
1,5	1	7	7	8	7	7	6	6
1,2,3 (Good Model)	19	20	18	7	11	16	18	19
1,2,4	6	5	5	1	4	5	5	5
1,2,5	11	5	7	6	8	6	6	7
1,3,4	3	3	3	3	4	3	3	3
1,3,5					2			
1,4,5					1			
1,2,3,4 (Correct Model)	9	3	2	1	5	2	3	2
1,2,3,5	15	7	5	1	5	6	9	6
1,2,4,5	5	2	1		5	1	2	1
1,3,4,5	1			1	2			
1,2,3,4,5	16	3	3	2	3	3	3	3
n=1,000								
1,2					14			
1,3					1			
1,4								
1,5								
1,2,3 (Good Model)	7	26	27	48	21	27	27	27
1,2,4				3	6			
1,2,5					3			
1,3,4					2			
1,3,5					1			
1,4,5								
1,2,3,4 (Correct Model)	59	62	62	46	27	62	62	62
1,2,3,5	8	7	6	2	10	6	6	6
1,2,4,5					2			
1,3,4,5					3			
1,2,3,4,5	26	5	5	1	10	5	5	5

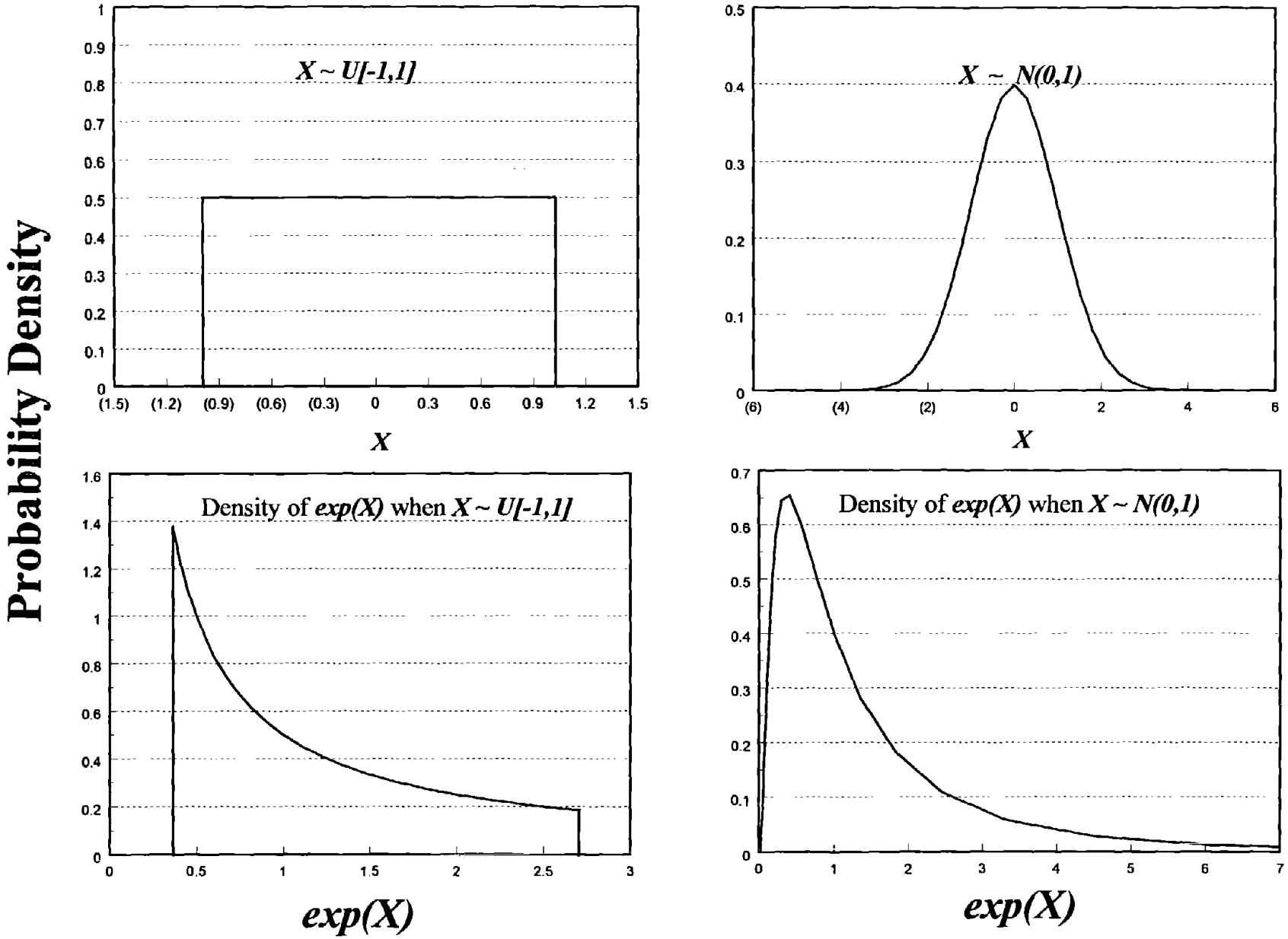


Figure 8. Probability densities of *exp(X)* when X is uniformly and normally distributed.

appropriate distributions would fall inbetween normal and uniform distributions. In the following study, independent normal and uniform distributions will continue to be used to simulate covariate values.

The model structure of the true models considered is as follows:

$$Y_i \sim Poisson(\mu_i) \quad or \quad P(Y_i = y_i) = P(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \qquad i=1,2,3,...,n. \tag{38}$$

and

$$\mu_i = E[Y_i | x_i] = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}) \qquad i=1,2,3,...,n. \tag{39}$$

Based on this model structure, two true models are considered. In these two true models, X_{i6} is treated as an omitted variable, i.e., none of the candidate models considered includes X_{i6}. The simulation results are presented as follows:

True Model #9: x_{i1} =1, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} are *iid* as $N(0,1)$; β_1=1, β_2=β_3=1, β_4=β_5=0, β_6=1.

Essentially this true model has four covariates, X_{i1}, X_{i2}, X_{i3}, and X_{i6} (since β_4=β_5=0). Also, X_{i2}, X_{i3}, and X_{i6} are equally important in the sense that they are identically distributed as $N(0,1)$ and have the same parameter value of 1.

The test results from the simulations are presented in table 17. Only the three CAIC criteria under the NB model perform well under small sample sizes of 50 and 100. Performance of all criteria decreases as the sample size increases, and all criteria tend to over-select covariates.

True Model #10: Same as True Model #9 except that X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} are *iid* as $U[-1,1]$.

The test results from the simulations are presented in table 18. Again, only the three CAIC criteria under the NB model perform well. However, this time, the performance of the three CAIC criteria under the NB model improves as the sample size increases. Again, this may have something to do with the distribution of $exp(X)$ discussed earlier.

Based on these two limited simulations, this author's current recommendation is to use $CAIC_{NB}$ as a variable selection criterion for developing accident prediction models. However, engineering judgment should be exercised to identify candidate variables. The key is not to include variables that are likely to be unrelated with Y_i from the engineering standpoint. In addition, use other statistics, such as the t-statistic mentioned in chapters 1 and 2, to help determine the importance of each variable. Another suggestion is that instead of selecting the candidate model that has the lowest $CAIC_{NB}$, one should also consider those candidate models that are compatible. For example, one should also look into those models that have $CAIC_{NB}$ values not greater than three to five of the $CAIC_{NB}$ value of the best model.

Table 17. Frequencies of models selected by various criteria in 100 replications: True Model #9.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6})$; $x_{i1}=1$, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} are *iid* as $N(0,1)$; $\beta_1=1$, $\beta_2=\beta_3=1$, $\beta_4=\beta_5=0$, $\beta_6=1$; and x_{i6} is an omitted variable.

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Best Model)	40	19	24	39	12	86	82	85
1,2,4								
1,2,5								
1,3,4					5			
1,3,5								
1,4,5								
1,2,3,4	1	27	28	20	76	13	17	14
1,2,3,5	56	28	35	34				
1,2,4,5								
1,3,4,5					6			
1,2,3,4,5	3	26	13	7	1	1	1	1
n=100								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Best Model)						78	77	77
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4						9	9	9
1,2,3,5	98	83	84	96	30	13	14	14
1,2,4,5								
1,3,4,5								
1,2,3,4,5	2	17	16	4	70			
n=1,000								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Best Model)						27	36	29
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	99				1			
1,2,3,5						73	64	71
1,2,4,5								
1,3,4,5								
1,2,3,4,5	1	100	100	100	99			

Table 18. Frequencies of models selected by various criteria in 100 replications:True Model #10.

True Model: $Y_i \sim$ Poisson(μ_i), where $\mu_i = exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6})$; $x_{i1}=1$, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} are *iid* as $U[-1,1]$; $\beta_1=1$, $\beta_2=\beta_3=1$, $\beta_4=\beta_5=0$, $\beta_6=1$; and x_{i6} is an omitted variable.

Models & Sample Size *(n)*	Adj R^2	AIC_{PO}	$CAIC_{PO}$	SD_{PO}	X^2	$CAIC_{NB}$	$CAIC_{NB}$-log(n)	$CAIC_{NB}$-2
n=50								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Best Model)				1	7	36	35	36
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	1	5	8	12	34	22	23	22
1,2,3,5	2	11	17	24	26	31	31	31
1,2,4,5								
1,3,4,5					6			
1,2,3,4,5	97	84	75	63	33	11	11	11
n=100								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Best Model)	23	26	29	52	43	57	57	57
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	26	15	14	6	17	4	4	4
1,2,3,5	28	40	43	39	23	39	39	39
1,2,4,5								
1,3,4,5								
1,2,3,4,5	23	19	14	3	17			
n=1,000								
1,2								
1,3								
1,4								
1,5								
1,2,3 (Best Model)	38	34	34	57	46	71	71	71
1,2,4								
1,2,5								
1,3,4								
1,3,5								
1,4,5								
1,2,3,4	46	45	45	35	32	23	23	23
1,2,3,5	8	10	10	6	13	5	5	5
1,2,4,5								
1,3,4,5								
1,2,3,4,5	8	11	11	2	9	1	1	1

RECOMMENDATIONS FOR FUTURE RESEARCH

Although the simulations conducted in this study were quite limited in its coverage, several interesting results have been found and reported. The simulations confirmed this author's suspicion that the results of statisticians' experiments under the situation where the correct model is among a set of candidate models considered by the modeler are not appropriate for use in developing accident prediction models. A more appropriate situation that needs to be tested is the ability of the goodness-of-fit criteria to select the best model(s) when some of the relevant variables are omitted from all candidate models. This suggests the importance of furthering the research along the line of the experiments conducted in the last section. More systematic simulation studies than those reported in the last section should be planned. In addition, tests should include data sets with very low means, and distributions other than normal and uniform should be considered for simulating omitted variables. Another dimension of the simulation that may be of interest to this study is to allow different degrees of correlation among simulated covariates.

An interesting observation made in this simulation study is that the performance of the three CAIC criteria under the NB model is very dependent on the distribution of covariates. An analytical study is recommended for a thorough understanding of this observation.

5. SOME ALTERNATIVE GOODNESS-OF-FIT CRITERIA

As suggested in chapter 3, a desirable goodness-of-fit criterion, which can be used to make some of the decisions and comparisons described in chapter 1, should have at least the following three properties: (1) [0,1] bound property; (2) proportional increase property; and (3) invariant with respect to the mean property. Specifically, one would like to have a criterion that is bounded between zero and 1, a value of zero indicating no covariate is included and a value of 1 indicating all necessary covariates are included. In addition, if all covariates are independent and equally important, then when one selects and adds these covariates to the model one at a time, the increase in the value of this criterion should be the same for each covariate regardless of their order of selection. Furthermore, one would like the value of the criterion to be unaffected by simply increasing or decreasing the intercept term of the model.

As shown in chapter 3, R^2 does not possess any of these properties in the context of the Poisson and NB regression models. The AIC presented in chapter 4 is not normalized and therefore unbounded. It is not clear whether AIC has the second and third properties under the Poisson and NB regression models. At the time of this study, there was a paper published by Fridstrøm et al. [1995] in which five criteria were proposed for evaluating the goodness-of-fit of accident prediction models. The proposed criteria were developed by modifying existing criteria, such as R^2 and scaled deviance, to achieve some degree of [0,1] bound property. It is not clear how these criteria would actually perform in terms of the three properties discussed under the type of Poisson and NB regression models that are typically used in developing accident prediction models. Also, the sampling property of these modified criteria was not studied, especially, sampling errors under small and medium samples.

In this study, the performance of three alternative goodness-of-fit criteria are examined in terms of the three properties discussed above. Simulation method is again used as a tool to gain some insights on their performance. The author did not have the opportunity to examine the criteria proposed by Fridstrøm et al. [1995]. However, it appears that the first two alternative criteria that this author examined are very similar to some of the criteria they proposed. Again, as in other chapters, the study reported here is intended to be exploratory and illustrative in nature. Furthermore, because of resource limitations, this study only examined the performance of these three alternative criteria under large samples.

ALTERNATIVE CRITERIA CONSIDERED

The main idea behind the development of the three alternative criteria is the Poisson concept discussed in chapter 2. The concept says that if the analysts are able to collect all the necessary covariates to explain the variation of accident frequencies among sites and time intervals, then conditional on these covariates, the accident frequencies are Poisson distributed. Another idea used in developing these criteria is the worst model concept. The worst model is the one that includes no covariate and therefore has no explaining power on the variation of Y_i. Using these two ideas, one can derive lower and upper bounds for any criterion of interest, e.g., R^2 and AIC. These bounds are then used to normalize the criterion of interest so that the

normalized criterion would possess the [0,1] bound property. This normalization scheme was briefly alluded to in chapter 3 when linear regression models were discussed (see Eq. (25)).

The three alternative criteria considered in this study are as follows:

(1) R^2_n: normalized adjusted R^2.

The concept behind the development of this criterion is to estimate the amount of random variance from the data and then remove it from the total variance. Using the Poisson concept, where the conditional variance is equal to the conditional mean, a consistent estimate of random variance is the sample mean: $(1/n)\sum_{i=1}^{n} y_i$. Note that for a Poisson regression model under the ML estimation method, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$ is guaranteed, where $\hat{y}_i$ is the estimate of the conditional mean of Y_i from an estimated candidate model (which was denoted as $\hat{\mu}_i$ in some of the chapters). Similar to Eq. (25), we can define a normalized adjusted R^2 as follow:

$$R_n^2 = 1 - \frac{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 - \frac{1}{n}\sum_{i=1}^{n} y_i}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 - \frac{1}{n}\sum_{i=1}^{n} y_i} \qquad (40)$$

Note that on the right hand side of the equation, the denominator is an estimate of the total systematic variance (or explainable variance) and the numerator is the total systematic variance unexplained by the candidate model.

(2) R^2_α: a dispersion parameter-based R^2.

This criterion uses the size of dispersion parameter in the conventional NB regression model as a yardstick to determine how well the variance of the data is explained. For a given data set, the largest dispersion parameter value is first estimated by fitting the observed data Y_i with an NB distribution (which includes no covariate). The estimated dispersion parameter, denoted as $\hat{\alpha}_{max}$, is the upper bound of the dispersion parameter for this particular data set. Since the Poisson regression model is a limiting model of the NB regression model as α approaches zero, under the Poisson concept a perfectly specified NB regression model should have an α value very close to zero. Therefore, the lower bound of the dispersion parameter is zero for any accident data set of interest. Now, for a candidate NB regression model, which has some number of covariates in the model, the regression parameters as well as the dispersion parameter can be estimated using the ML estimation method. Let the estimated dispersion parameter for the candidate model be denoted as $\hat{\alpha}$. Given the upper bound $\hat{\alpha}_{max}$, the lower bound zero, and the estimate from a candidate model $\hat{\alpha}$, a natural dispersion parameter-based criterion can be devised as:

$$R_{\alpha}^{2} = 1 - \frac{\hat{\alpha}}{\hat{\alpha}_{max}} \qquad (41)$$

This criterion appears to be extremely simple. It has a value of zero when no covariate is included in the model and a value of 1 when covariates are perfectly specified. Note however that this criterion does not seem to have a good statistical interpretation other than the fact that it is indirectly associated with the proportion of unexplained systematic variance. Another limitation of this criterion is that it does not reflect the number of covariates included in the candidate model.

(3) R^2_{AIC}: AIC-based R^2.

This alternative criterion is based on the $CAIC_{NB}$ criterion, i.e., the CAIC under the NB model, discussed in the last chapter. For a given data set, let the value of the $CAIC_{NB}$ criterion under the worst model (with no covariate) be $CAIC_{NB}(\bar{y})$, which is the largest possible CAIC value. In addition, let the value under an ML estimated candidate model be $CAIC_{NB}(\hat{y}_i)$. Furthermore, let the value under a perfect model be $CAIC_{PO}(\mu_i^*)$, where μ_i^* is the unknown true mean as before. In theory, $CAIC_{PO}(\mu_i^*)$ is the smallest possible CAIC value achievable as the sample size approaches ∞. Using these $CAIC_{NB}$ values, an AIC-based alternative criterion can be defined as:

$$R_{AIC}^{2} = \frac{CAIC_{NB}(\hat{y}_i) - CAIC_{NB}(\bar{y})}{CAIC_{PO}(\mu_i^*) - CAIC_{NB}(\bar{y})} \qquad (42)$$

Of course, this criterion is not really usable because the true mean μ_i^* is unknown. To make the alternative criterion operational, one needs a good approximation for μ_i^*. One plausible suggestion is to use the EB estimate as follows:

$$\hat{\mu}_i^* = \left(\frac{1}{1+\hat{\alpha}\hat{y}_i}\right)\hat{y}_i + \left(1 - \frac{1}{1+\hat{\alpha}\hat{y}_i}\right)y_i \qquad (43)$$

where $\hat{\alpha}$ and $\hat{y}_i$ are estimates from the candidate model under consideration. Note that in the simulation studies presented in the next section, $CAIC_{PO}(\mu_i^*)$ is computable since μ_i^* is known.

SIMULATION RESULTS

The model structure of the true models considered is as follows:

$$Y_i \sim Poisson(\mu_i) \quad or \quad P(Y_i = y_i) = P(y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!} \qquad i=1,2,3,\ldots,n. \tag{44}$$

and

$$\mu_i = E[Y_i | x_i] = exp(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}) \qquad i=1,2,3,\ldots,n. \tag{45}$$

Based on this model structure, different true models are considered and their test results are presented as follows:

True Models #1 and #2: X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} are *iid* as $N(0,1)$; β_1 =1 for True Model #1 and β_1 =-4.8 for True Model #2; and $\beta_2=\beta_3=\beta_4=\beta_5=\beta_6$ =1 for both models.

Essentially these two true models have five independent and equally important covariates: X_{i2}, X_{i3}, X_{i4}, X_{i5}, and X_{i6}. Because of the difference in the intercept term, these two true models have very distinct mean levels: E[Y_i] is about 33 for True Model #1 and about 0.1 for True Model #2. As indicated before, most of the accident data sets used in developing accident prediction models have mean levels within this range. As in chapter 3, one can compute random variance and systematic variance for these true models. One can find that, for these two models, their random variance is extremely small when compared to their systematic variance. This means that, under these true models, a perfect model can achieve an R^2 value of close to 1. In developing accident prediction models, one does expect the accident data to contain a significant amount of random variance. Therefore, these models are not particularly realistic models for this study. They can be used however for comparison purposes and as a basis for developing more realistic example models.

As indicated earlier, in this study, the performance of the three alternative criteria are studied under large samples only. A sample size of 10,000 is used in the simulation.

To see how the values of these three alternative criteria increase when the covariate is added to the model one at a time, the ML estimation method is used to estimate candidate models with (intercept only), (intercept + 1 covariate), (intercept + 2 covariates), ... , and (intercept + 5 covariates). Under these candidate models, an ideal criterion that possesses the three properties discussed earlier would have values of 0, 0.2, 0.4, 0.6, 0.8, and 1.0 as the number of covariates increases from zero to five. From the simulation study, the R^2 values as well as the values of the three alternative criteria under these candidate models are shown in figure 9.

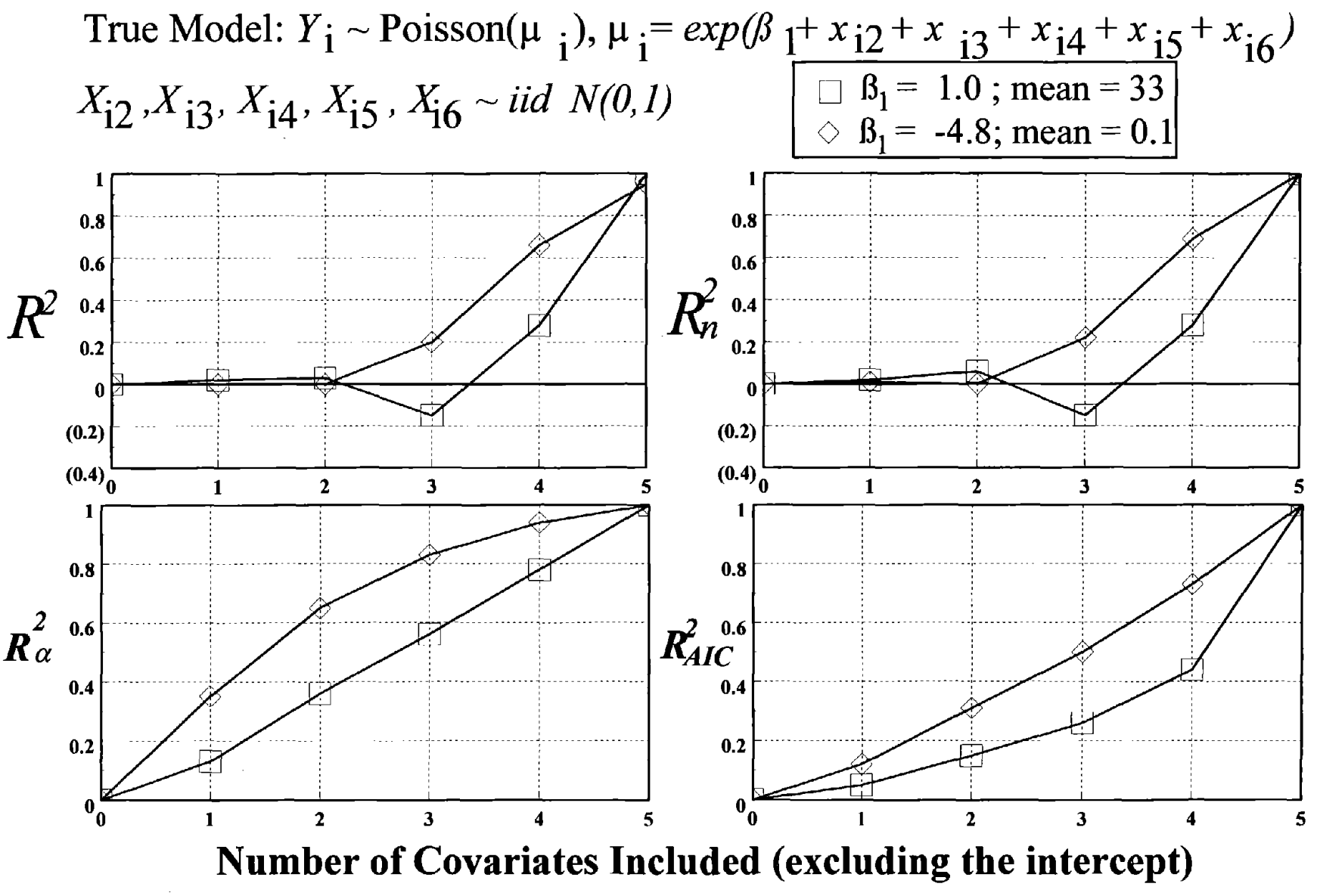


Figure 9. Values of R^2 and alternative R^2 under a Poisson model with two different mean levels.

The following observations can be made from figure 9 with respect to the three properties of interest:

(1) [0,1] bound property: For these two particular true models, all criteria have good [0,1] bound property.
(2) Proportional increase property: R^2 and R^2_n do not have this desired property at both high and low mean levels. R^2_α performs very well when the mean level is high and reasonable when the mean level is low. On the other hand, R^2_{AIC} performs reasonably well when the mean level is low, and reasonable at high mean level.
(3) Invariant with respect to the mean property: To some extent, all criteria are affected by the mean level. R^2_α tends to overstate the performance of candidate models when the mean level is low, while R^2_α tends to understate the performance of candidate models when the mean level is high.

True Models #3 and #8: X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} are *iid* as $N(0,1)$; $\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=0.4$ for all models; and β_1 =-3, -2, -0.4, 1, 2, and 3 for True Model #3 through #8, respectively.

Again, these true models have five independent and equally important covariates: X_{i2}, X_{i3}, X_{i4}, X_{i5}, and X_{i6}. The only difference among these models is in the intercept term, which allows them to have different mean levels varying from 0.07 to 30. Also, as a result of their difference in the intercept term, they are also different in terms of their random variance over total variance ratio. Under a perfect model, these models would have R^2 values ranging from about 0.1 to 0.95. These models allow one to examine and compare the performance of the three alternative criteria under different mean and random variance levels.

As in the first two simulations, a sample size of 10,000 is used and candidate models with a different number of covariates under the ML estimation method are evaluated. The R^2 values as well as the values of the three alternative criteria under these candidate models are shown in figures 10-12. In these figures, the results from two models of very different mean levels are plotted in one figure for comparison purposes. The following observations can be made with respect to the three properties of interest (under large samples):

(1) [0,1] bound property: Except R^2, all criteria have good [0,1] bound property.
(2) Proportional increase property: Overall, R^2_n performs quite well at all mean levels. But it tends to slightly understate the performance of all candidate models. R^2_α performs very well at all mean levels. It has a tendency to slightly overstate the performance of candidate models when the mean levels are low. As in the first two simulations, R^2_{AIC} performs reasonably well when the mean level is low and the performance deteriorates as the mean level increases.
(3) Invariant with respect to the mean property: R^2_n performs very well at all mean levels; R^2_α also performs quite well; and R^2_{AIC} does not do well.

True Model: $Y_i \sim \text{Poisson}(\mu_i)$, $\mu_i = exp(\beta_1 + x_{i2} + x_{i3} + x_{i4} + x_{i5} + x_{i6})$

$X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6} \sim iid\ N(0,1)$

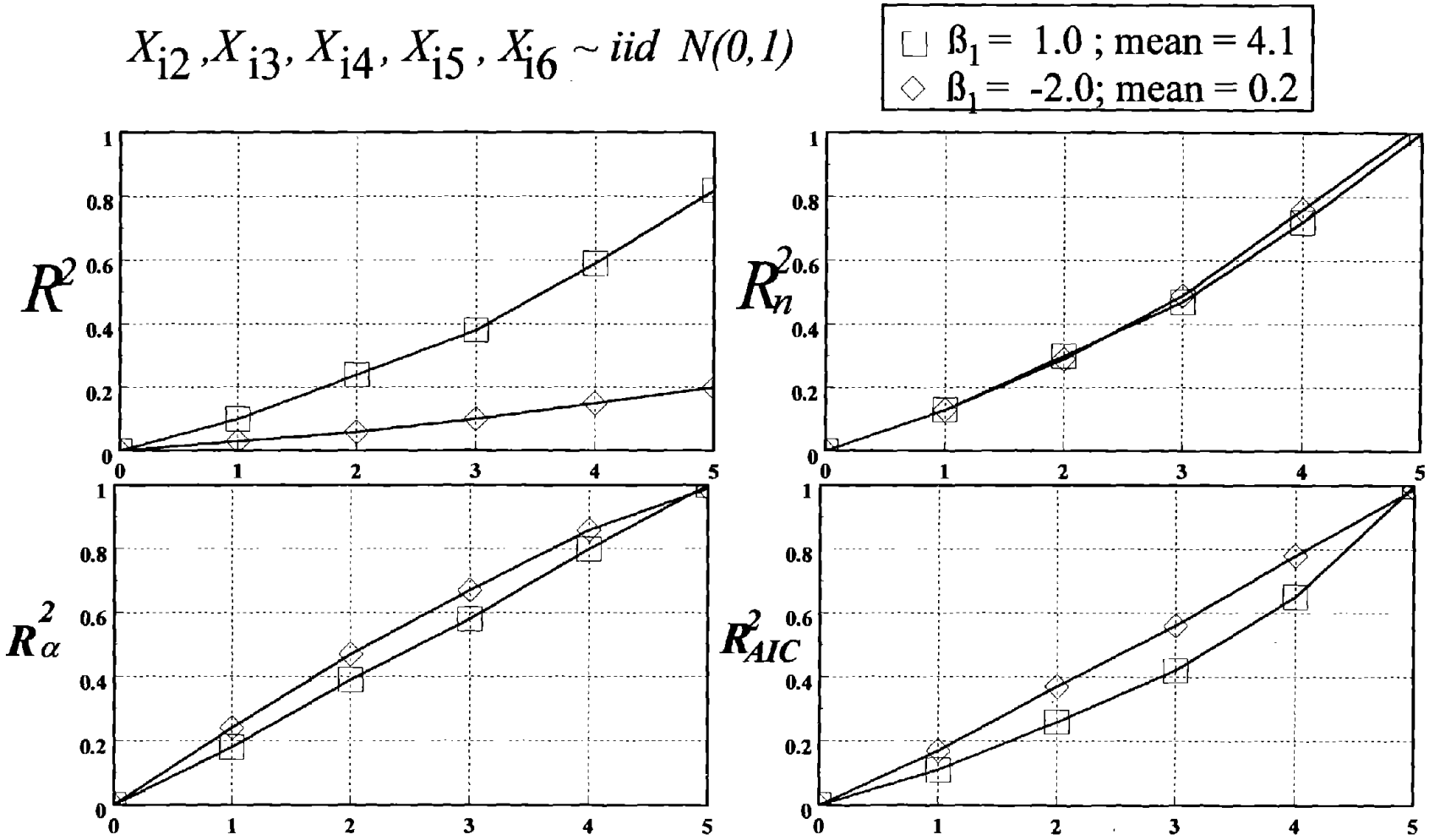


Figure 10. Values of R^2 and alternative R^2 under a second Poisson model with two different mean levels.

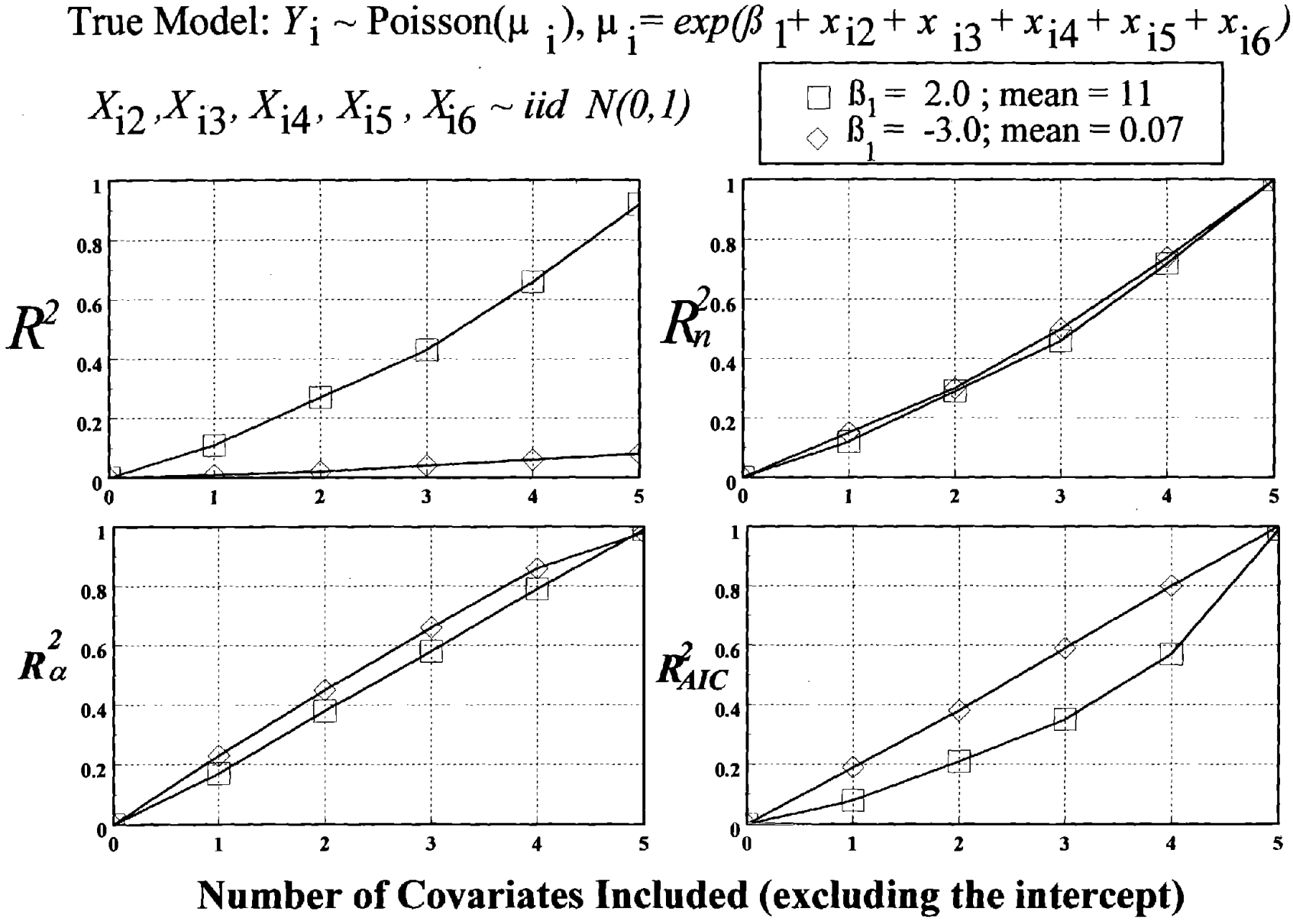


Figure 11. Values of R^2 and alternative R^2 under a third Poisson model with two different mean levels.

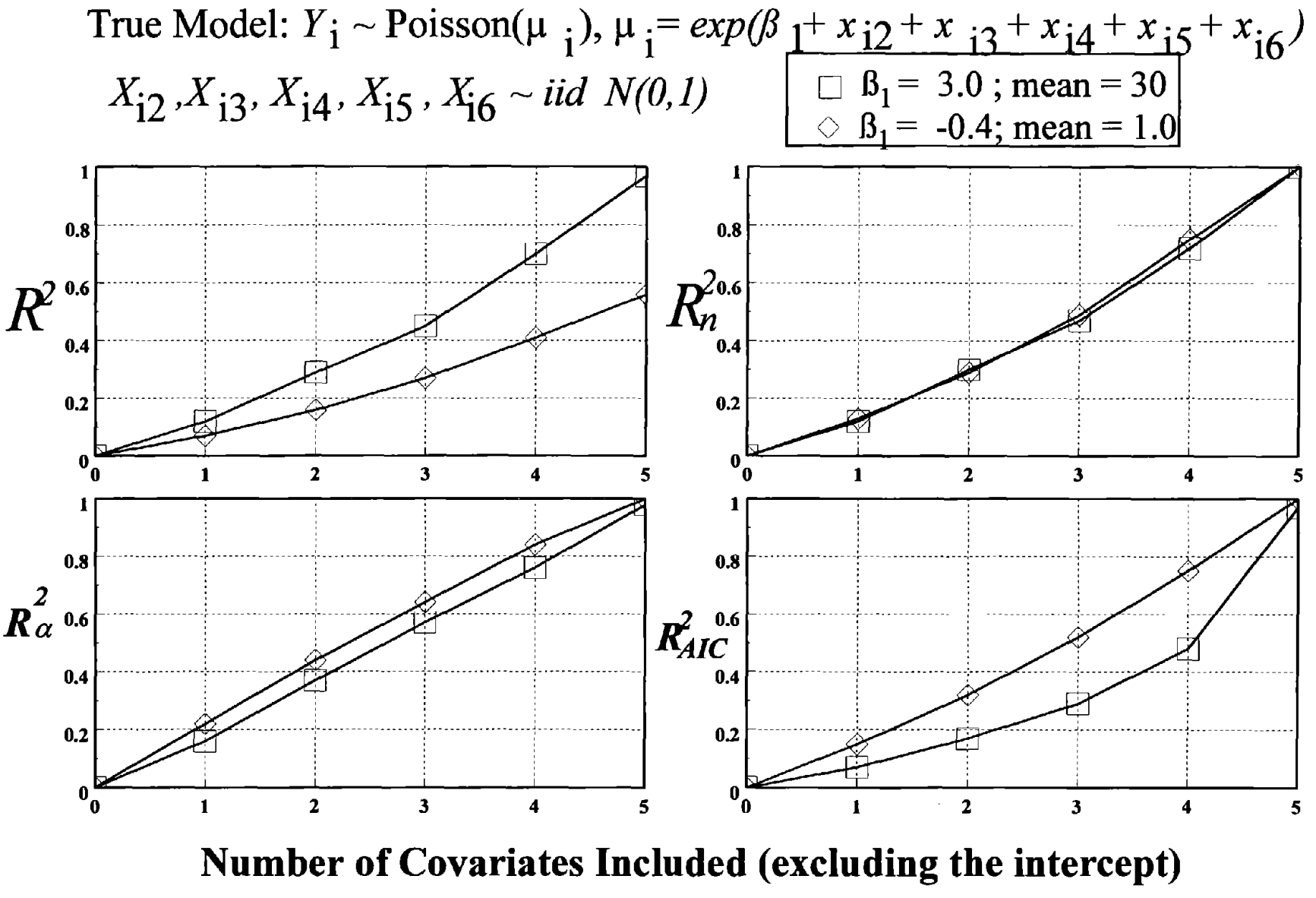


Figure 12. Values of R^2 and alternative R^2 under a fourth Poisson model with two different mean levels.

RECOMMENDATIONS

Based on the limited simulation results discussed above, R^2_α seems to be a reasonable choice, among the three alternative criteria considered, for large samples at all mean and random variance levels of interest. It is also quite easy to compute. For the two NB regression truck accident models in table 7, the R^2_α values are both about 0.74.

More systematic simulation studies than those reported in this chapter should be planned and carried out in the future. The study should include situations where covariates are not normally distributed. Also, simulation studies could allow different degrees of correlation among simulated covariates. Finally, sampling property of these three alternative criteria should be carefully examined under medium and small samples.

One final note for this chapter is that, throughout the simulation study, the uncertainties or measurement errors of the covariates have been ignored. In practice, important covariates such as AADT are subjected to both sampling and nonsampling errors.

6. ROADSIDE ENCROACHMENT AND RUN-OFF-THE-ROAD ACCIDENTS

The third objective of this study was to suggest a model that is appropriate for the prediction of run-off-the-road accidents (RORA) and to discuss the merits and shortcomings of the model as it applies to the prediction of RORA and vehicle roadside encroachments which may lead to RORA. This chapter provides a summary of the research results and discusses key findings. Because of the limited resources available for this study, the research reported in this chapter is again exploratory and illustrative in nature.

The first section of this chapter gives a review of two approaches that have traditionally been used in previous studies to develop the relationship between roadside accident frequency and roadside hazards, such as embankments, utility poles, trees, luminaries, guardrail, median barriers, traffic sign posts, mailboxes, culverts, bridge piers, etc. Using the model development principles described and simulation experiences learned in earlier chapters, an accident prediction model was developed using single vehicle (SV) RORA and roadway data (including mainline and roadside data) for rural two-lane undivided road and is presented in the second section. The third section provides estimates of roadside encroachment frequency using the accident prediction model presented in the second section. The last section concludes the chapter by offering some recommendations to enhance the RORA prediction model presented in the second section and for future research in roadside safety.

In the following discussion, a roadside encroachment is said to occur when an errant vehicle crosses the outside edges of the travelway and encroaches on the shoulder, including both inside and outside shoulders. Thus, for a two-lane undivided road that has no inside shoulder, the total number of roadside encroachments includes departures of vehicles from the near-side and far-side edges of the travelway in both directions. A lane encroachment, on the other hand, describes an errant vehicle that travels onto an adjacent lane or shoulder. Thus, for a two-lane undivided road with no inside shoulder, a lane encroachment on an adjacent shoulder is considered a roadside encroachment (which will be called near-side or right-side roadside encroachment). However, a lane encroachment onto an adjacent lane may or may not lead to a roadside encroachment, depending on the lateral distance the vehicle travels before regaining control by the driver. Such a lane encroachment will be considered a roadside encroachment only if the errant vehicle crosses the entire width of the adjacent lane and encroaches on the far-side edge of the travelway (which will be called far-side or left-side roadside encroachment).

It is important to note that roadside or lane encroachments refer only to unintentional encroachments. In other words, the intentional encroachments as a result of vehicles being intentionally driven outside of the travel lane, e.g., on adjacent lane (in the same or opposite direction), shoulders, and traversable medians, are not counted as encroachments.

ACCIDENT-BASED AND ENCROACHMENT-BASED APPROACHES

Models used in previous studies to describe the relationship between roadside accident frequency and roadside hazards have traditionally been categorized either as an accident-based

approach or encroachment-based approach. The first approach uses accident prediction models such as those presented in chapter 2. The second approach uses a series of conditional probabilities to describe the process of an encroaching vehicle that reaches a certain lateral displacement from the travelway and results in a collision with a roadside hazard. A good review of these two approaches can be found in a recent study by Daily et al. [1994] for FHWA. Since accident-based prediction models have been described in some detail in earlier chapters, this section will focus on describing encroachment-based models and their relationships with accident-based models. The strengths and weaknesses of each approach will be discussed.

Encroachment-Based Approaches

The main purpose of the roadside encroachment model is to estimate the annual number of RORA by severity level, e.g., fatal, injury, and PDO accidents. The basic concept involved is that the consequences of an errant vehicle leaving the roadway depend not only on the encroachment speed and driver's reaction (e.g., steering and braking) but also on the roadside design. For example, at a site with good roadside design, a roadside encroachment may result in no injury or property damage; while the same encroachment at a location with poor roadside design could result in a severe accident.

Over the years, roadside encroachment models have been developed for predicting roadside accident frequency, e.g., the National Cooperative Highway Research Program (NCHRP) Report 77 [1969], NCHRP Report 148 [1974], the American Association of State Highway and Transportation Officials' (AASHTO) *Roadside Design Guide* [1988], and the appendix F of the Transportation Research Board's Special Report 214 (SR214) [1987]. The roadside safety model included in the AASHTO *Roadside Design Guide*, which was based on NCHRP Reports 77 and 148, is known as the ROADSIDE model. The model has been coded as a microcomputer program that can be used to compare roadside accident rates by accident severity type for different roadside designs. By far, the most sophisticated encroachment model that has been introduced is the one presented in the TRB SR214. The ROADSIDE model can be considered as a simpler version of the SR214 model.

To illustrate the concept behind an encroachment-based model, SR214 model will be used in the following discussion. For a particular type of roadside hazard (e.g., utility poles), a basic roadside encroachment model of a road section with ℓ mile (or $\ell \times 1.6$ km) in length can be conceptually represented by the following equation:

$$\mu_s = V \times P(Ln\ Encro) \times \ell \times P(In\ Impact\ Envelope \mid Ln\ Encro) \times P(Collide\ with\ Hazard \mid In\ Impact\ Envelope) \times P(RORA \mid Collide\ with\ Hazard) \times P(RORA_s \mid RORA) \quad (46)$$

where

μ_s = expected number of RORA involving a specific roadside hazard per year with severity level of s;

V = number of vehicles per year passing the road section (=365 ×AADT);

$P(Ln\ Encro)$ = probability of having a lane encroachment by a vehicle that travels 1 mi or 1 km of such road (assuming that the probability is the same for vehicles traveling in both directions); (Note that it is also assumed here that the probability of having more than one lane encroachment by a vehicle is zero);

$V \times P(Ln\ Encro)$
= expected number of lane encroachments per mi (or per km) per year in both directions;

ℓ = length of the road section (in mi or km);

$V \times P(Ln\ Encro) \times \ell$ = expected number of lane encroachments on the road section per year in both directions;

$P(In\ Impact\ Envelope|Ln\ Encro)$
= conditional probability that, given a lane encroachment, its location is such that an impact with the hazard is possible;

$P(Collide\ with\ Hazard|In\ Impact\ Envelope)$
= conditional probability that, given an encroachment in potential impact envelope, a collision between vehicle and hazard will occur;

$P(RORA|Collide\ with\ Hazard)$
= conditional probability that, given a collision, its severity will be so great as to result in a RORA; and

$P(RORA_s|RORA)$ = conditional probability that, given a RORA, an accident of severity level of s will occur.

In principle, Eq. (46) can be developed for each type of roadside hazard and then summed over all types of roadside hazards along the road section to estimate the total number of RORA by severity type for the road section.

Each conditional probability in Eq. (46) is dependent on a number of factors. Table 19 gives a list of potential factors that may need to be considered when developing roadside encroachment models. For example, it is expected that (1) $V \times P(Ln\ Encro)$ is dependent on mainline traffic and geometric design characteristics, such as AADT, number of lanes, lane width, horizontal curvature, and vertical grade; (2) $P(In\ Impact\ Envelope|Ln\ Encro)$ is a function of the size, shape, and density of the roadside hazard considered, and the width and encroachment angle of the encroaching vehicle; (3) $P(Collide\ with\ Hazard|In\ Impact\ Envelope)$ is a function of the lateral distance of the encroaching vehicle, which is associated with factors such as encroachment speed, friction between the vehicle tires and the surface, and driver's reaction (e.g., steering and braking); and (4) $P(RORA|Collide\ with\ Hazard) \times P(RORA_s|RORA)$ is a function of many factors, including impact conditions (i.e., impact speed, impact angle, and

Table 19. Potential factors that may affect the conditional probabilities of the roadside encroachment model in Eq. (46) for two-lane undivided roads.

Conditional Probability	Potential Factors, X	Remarks
P[Ln Encro | X]	Traffic Density (e.g., AADT/Lane), Lane Width, Horizontal Curvature, and Vertical Grade.	1. Mainly associated with mainline traffic and geometric design characteristics. 2. Other factors: shoulder width and types (affecting drivers' attentiveness), car-truck mix, traffic signs (i.e., divers' behavior in response to the presence or absence of different traffic signs).
P[Rdside Encro | X]	Same As Above.	For left-side encroachments, encroachment angle and speed are relevant.
P(In Impact Envelope|Ln Encro, X)	Size and Shape (or effective width and length) of the Hazard, Density of the Hazard, Width of the Vehicle, Encroachment Angle.	1. Strictly speaking, encroachment angle and speed could also be a function of traffic density, horizontal curvature, vertical grade, etc. 2. When calculating the effective impact envelope, one needs to consider two possible situations: (a) impact envelopes of separate objects may overlap; and (b) the object of interest may be behind other objects, which can be breakaway or nonbreakaway objects.
P(Collide with Hazard | In Impact Envelope, X)	Lateral Offset of the Hazard, Roadside Slope, Encroachment Speed, Traveled Path, Traffic Density (for left-side encroachments), Type and Width of Shoulder (paved, stabilized, presence of rumble strips), Friction between Tires and Surface (weather).	Traditionally, the path of the encroaching vehicle is assumed to be straight which is a bold assumption that ignores drivers' reaction (e.g., steering and braking).
P(RORA|Collide with Hazard, X) × *P(RORA$_s$ | RORA, X)*	Impact Speed, Angle, and Vehicle Orientation, Size and Weight of the Errant Vehicle.	Driver and Other Occupants Ages, Location of the Accident (e.g., rural vs. urban), Accident Reporting Threshold (varies by State).
P(Reported RORA$_s$ | RORA$_s$, X)	Driver's Demographic and Socioeconomic Conditions, Driver's Previous Accident Record, Location of Accidents, Single-Vehicle or Multiple-Vehicle Accidents	Not all accidents are reported; especially minor injury and PDO accidents.

vehicle orientation), the size and weight of the errant vehicle, and the feature of the impacted roadside hazard (e.g., breakaway vs. nonbreakaway roadside devices).

Not all accidents are reported, especially minor injury and PDO accidents. To model the reported RORA, Eq. (46) can be modified by multiplying μ_s with $P(Reported\ RORA_s|\ RORA_s)$, the reporting probability of an RORA with a severity level of s. This reporting probability is expected to be a function of a couple of factors, including the driver's demographic and socioeconomic characteristics as well as the driver's previous accident record.

In appendix F of SR214, utility pole accidents of all severity levels for two-lane undivided roads were used as an example, and the reported encroachment model can be outlined as follows:

(1) $V \times P(Ln\ Encro) = a(AADT)^b$, where *AADT* is the annual two-directional average daily traffic volume, and *a, b* are model parameters. Essentially, the total number of lane encroachments per 1 mi or 1.6 km (at both sides of the road and in both directions) is assumed to be related only to traffic volume. The equation can also be reexpressed as $P(Ln\ Encro) = a(AADT)^b / V = a(AADT)^b/(365 \times AADT) = a(AADT)^{b-1}/365$. Thus, one can say that the probability of lane encroachments per vehicle per mi or 1.6 km, *P(Ln Encro),* is conditional on *AADT*, which can be symbolized as *P(Ln Encro|AADT)*. This conditional probability is likely to be oversimplified. As indicated earlier, *P(Ln Encro)* is expected to be conditional on other variables, such as lane width, horizontal curvature, and vertical grade.

The relationship between the expected annual number of roadside encroachments and the expected annual number of lane encroachments is as follows: First, let the expected annual number of roadside encroachments per mile be mathematically expressed as $V \times P(Rdside\ Encro)$, where *P(Rdside Encro)* is the probability of having a roadside encroachment by a vehicle that travels 1 mi or 1 km of such road (assuming that the probability is the same for vehicles traveling in both directions). Second, using the assumption that, for an errant vehicle, the lane encroachment is equally likely to occur on the left and right sides of the lane, the relationship can be reexpressed as $V \times P(Rdside\ Encro) = V \times P(Ln\ Encro) \times 0.5 \times [1+exp(d \times LW)] = a(AADT)^b \times 0.5 \times [1+exp(d \times LW)] = 0.5 \times a(AADT)^b + 0.5 \times a(AADT)^b exp(d \times LW)$, where *LW* is lane width, and *a, b, d* are model parameters. The constant 0.5 is used because of the assumption that lane encroachment is equally likely to occur on the left and right sides of the lane. Essentially, the first term (i.e., $0.5 \times a(AADT)^b$) represents the total number of near-side roadside encroachments per mile per year, while the second term (i.e., $0.5 \times a(AADT)^b exp(d \times LW)$) represents the total number of far-side roadside encroachment per mile per year. Note that the exponential function $exp(d \times LW)$ represents the probability of a far-side encroaching vehicle that, without colliding with vehicles traveling in the opposite direction, will cross the entire adjacent lane (with a lane width of *LW*) before regaining control; the parameter *d* is less than zero which indicates that the probability of far-side encroachments is reduced exponentially as *LW* increases. Strictly speaking, the chance of not colliding with vehicles traveling in the opposite

direction is a function of *AADT*, encroachment angle, encroachment speed, etc., and therefore the parameter *d* does not have to be a constant.

(2) $P(\text{In Impact Envelope}|\text{Ln Encro}) \times P(\text{Collide with Hazard}|\text{In Impact Envelope}) = 0.5 \times \Sigma_j [c_r \times exp(d \times LE_{r_j}) + c_l \times exp(d \times LE_{l_j})]$, where LE_{r_j} is the lateral offset of the *jth* utility pole from the edge of the traveled lane of a right-side or near-side encroaching vehicle; similarly, LE_{l_j} is the lateral offset of the *jth* utility pole from the edge of the traveled lane of a left-side or far-side encroaching vehicle; the summation is taken over all utility poles on both sides of the road; the constant 0.5 indicates that, for an errant vehicle, the lane encroachment can occur only on one side of the lane; and c_r, c_l, and d are model parameters. Essentially, $exp(d \times LE_{r_j})$ represents the probability of a right-side encroaching vehicle that, in the absence of roadside obstacles, will leave the traveled lane by a distance of greater than LE_{r_j} before regaining control. Similarly, $exp(d \times LE_{l_j})$ represents the probability of a left-side encroaching vehicle that, in the absence of roadside obstacles and without colliding with the vehicles traveling in the opposite direction, will leave the traveled lane by a distance of greater than LE_{r_j} before regaining control. Strictly speaking, the parameter d does not have to be the same for left- and right-side encroachments.

The parameters c_r and c_l, respectively, represent the portion of the road section (as a percentage of the section length) along the roadway within which a right-side and a left-side encroachment, if continued sufficiently far, will result in an impact with the utility pole. These portions of road section are typically called impact envelopes. By assuming that the path of the encroaching vehicle is straight, parameters c_r and c_l are associated with the width of the encroaching vehicle, encroachment angle, and the length (parallel to the roadway) and width (perpendicular to the roadway) of the utility pole as follows:

$$c_r = \frac{PoleLength + [VehWidth \times csc(\phi_r)] + [PoleWidth \times \cot(\phi_r)]}{\ell \times 5280}$$
$$c_l = \frac{PoleLength + [VehWidth \times csc(\phi_l)] + [PoleWidth \times \cot(\phi_l)]}{\ell \times 5280} \quad (47)$$

where *PoleLength* and *PoleWidth* are respectively the length and width of the utility pole in ft; ϕ_r and ϕ_l are respectively the right-side and left-side roadside encroachment angles. In SR214, the parameters c_r and c_l are estimated using assumptions and limited empirical data as follows: (a) each utility pole has a square cross section with 8-in (20.32-cm) sides; (b) the encroachment angle is taken to be 6.1 degrees for near-side departures and 11.5 degrees for far-side departures; and (c) the width of the encroaching vehicle is 6-ft (1.83m). An additional assumption used (but not indicated) in the study was that the distance between utility poles are far away from one another so that the impact envelopes of utility poles do not overlap.

Note that the actual calculations carried out in the study were somewhat more sophisticated than what is described above. Specifically, for each utility pole the potential impact envelope was divided into impact zones and each impact zone was further subdivided into 1-ft (0.3048 m) strips. This division allows more accurate estimates of the lateral encroachment distances $LE_{r,j}$ and $LE_{l,j}$. Incidentally, because the dimension of utility poles is quite small, $LE_{l,j}$ is approximately equal to $LE_{r,j}+LW$, where LW is the lane width. The readers are referred to the original report for more detailed description.

This conditional probability is perhaps one of the most salient features of the roadside encroachment model. It touches several key factors of the RORA problem, including geometric factors, vehicle factors, driver behavior factors, and their interactions. It is in part because of this feature that SR214 suggested that "A primary advantage of the roadside encroachment model over the regression type model ... is its potential applicability to hazards other than utility poles." Note that, as will be discussed later, this statement is not true in this author's view.

There are several assumptions used in SR214 that require further validation or refinement: (1) the assumption that the path of the encroaching vehicle is straight completely ignores driver's reaction, e.g., steering and braking; (2) the encroachment angles of 6.1 degrees for right-side departures and 11.5 degrees for left-side departures require further validation; and (3) the assumption that parameter d is constant and is the same for left- and right-side encroachments can be refined.

(3) $P(RORA|Collide\ with\ Hazard) \times P(Reported\ RORA|RORA) = \psi = 0.9$. That is, the model uses an assumption that there is a 90-percent probability that when a vehicle collides with a utility pole it will result in a reported RORA (of any severity level). As indicated earlier, this probability is a function of many factors, including impact conditions (i.e., impact speed, angle, and vehicle orientation), the size and weight of the errant vehicle, and the feature of the impacted roadside hazard (e.g., material and shape). The constant probability is obviously a crude assumption.

In sum, for a road section with ℓ mile (or $\ell \times 1.6$ km) in length, the final encroachment model used in SR214 has the following form: $\mu = a(AADT)^b \times \ell \times 0.5 \times \Sigma_j [c_r \times exp(d \times LE_{r,j}) + c_l \times exp(d \times LE_{l,j})] \times 0.9$, in which the unknown parameters are a, b, and d. Other parameters such as c_r, c_l, and ψ were estimated based on engineering judgment and limited empirical data as discussed earlier. The uncertainty of these estimates could not be quantified. Of course, the encroachment model can be formulated in a much more complicated form by using fewer assumptions. But, it will require the collection of a lot more variables and detailed data.

The basic idea of the encroachment modeling is to collect encroachment frequency data, lateral encroachment distance, encroachment angle and speed, etc. to estimate the parameters associated with each conditional probability in the encroachment model. In practice, these data are very difficult and expensive to collect. In SR214, an attempt was made to validate encroachment frequency and rate using reported accident data. This was accomplished by

estimating parameters *a, b,* and *d* (< 0) using the utility pole accident data of Zegeer and Parker [1985]. The data included over 2,500 mi (4,025 km) of rural and urban roads from four States. An ad hoc OLS procedure was used for parameter estimation after log-transformations were taken on both side of the equation. In addition to the statistical limitation of using the lognormal distributional assumption in accident prediction modeling that was discussed in chapter 2, the procedure overlooked the need for an adjustment factor as presented in Eq. (8). Furthermore, no attempt was made in SR214 to assess the quality of the estimated parameters and the goodness-of-fit of the overall model.

The estimated parameter values for *a, b,* and *d* in SR214 were, respectively, 0.07285, 0.5935, and -0.08224. This model implies that the relationship between roadside encroachment frequency and AADT for a two-lane undivided road with 12-ft (3.66-m) lane width is as follows: roadside encroachment frequency per mile per year = $0.07285(AADT)^{0.5935} \times 0.5 \times [1 + \exp(-0.08224 \times 12)]$. For *AADT's* of 1,000, 2,000, 5,000, and 10,000 vehicles, the estimated roadside encroachment frequencies are, respectively, 3.02, 4.55, 7.84, and 11.83 encroachments per mi per year. As a consequence of overlooking the adjustment factor discussed above, these estimates are higher than they should be. As will be seen later, these estimates are also found to be considerably higher than the observations made by Hutchinson and Kennedy [1966] and Cooper [1980].

Relationship with Accident Prediction Models

Fundamentally, the principles behind the roadside encroachment model is consistent with the accident prediction model. Let's recall the five major tasks that are required to develop accident prediction models (which was presented in chapter 2):

Task 1. Find a good probability (mass) function to describe the random variation of accident frequency.

Task 2. Determine an appropriate functional form and parameterization for the mean function which describes the effect of key variables on accident frequency.

Task 3. Select the variables that have statistically significant effects on accident frequency for inclusion in the mean function.

Task 4. Estimate the regression parameters in the mean function and obtain good statistical inferences for the estimated parameters based on available data.

Task 5. Assess the quality of the model, judge whether the developed model makes good engineering sense, decide whether the developed model meets the planning and design requirements, and identify cost-effective ways to improve the model.

The important feature of the encroachment model has been on task 2, i.e., on determining the appropriate functional form and parameterization for the mean function. Specifically, the encroachment model in Eq. (46) deals mainly with the mean function described in earlier chapters. That is, instead of the regular exponential mean function that is commonly used in the Poisson and NB regression based accident prediction models, the mean function form in the encroachment model is developed on the basis of geometry, vehicle dynamics, and driver

behavior, in conjunction with engineering judgment. It should be noted that, as indicated in chapter 2, exercising engineering judgment to formulate the mean function is in fact highly recommended in accident prediction modeling. In addition, it can be shown that the encroachment model described in SR214 is consistent with the exponential mean function that has been used in the Poisson and NB regression models:

$$\mu = a(AADT)^b \times \ell \times 0.5 \times \Sigma_j [c_r \times exp(d \times LE_{r_j}) + c_l \times exp(d \times LE_{l_j})] \times 0.9$$

$$= (365 \times AADT \times \ell) \times \{(a/365) \times (AADT)^{b-1} \times 0.5 \times \Sigma_j [c_r \times exp(d \times LE_{r_j}) + c_l \times exp(d \times LE_{l_j})] \times 0.9\}$$

$$= v \times exp\{log(a/365) + (b-1) \times log(AADT) + log(0.5) + log(\Sigma_j [c_r \times exp(d \times LE_{r_j}) + c_l \times exp(d \times LE_{l_j})]) + log(0.9)\},$$

where v is the exposure measure equal to $365 \times AADT \times \ell$. As discussed in chapter 2, this exponential functional form is essentially multiplicative (which implies interactive effects).

It is formally straightforward to use the above mean function for the accident prediction models described in chapter 2. One can in fact estimate not only parameters *a, b, d* as in SR214, but also right-side and left-side encroachment angles (i.e., ϕ_r and ϕ_l which are imbedded in parameters c_r and c_l) using the usual accident prediction model building procedure. (Note that additional work is needed to develop statistical inferences for the estimated parameters since the mean function is not in the exact exponential form presented in chapter 2. But the work will be relatively straightforward.) In addition, this simple mean function can easily be extended to include, when available, lane width, horizontal curvature, and vertical grade as determinants of roadside encroachment frequency.

It appears that, at the current stage of the development, both approaches suffer from three common criticisms: (1) potentially serious underreporting of minor injury and PDO accidents; (2) too data intensive; and (3) questionable model transferability from one location to another. In addition to these criticisms, there are a number of unanswered questions regarding the encroachment-based approach, e.g., the validity of existing encroachment data, including the inability to distinguish intentional from unintentional encroachments [Mak and Sicking, 1992; Daily et al., 1994].

Although the encroachment-based approach has been criticized as having a lack of sound empirical basis, being unrealistically data intensive, and full of unvalidated assumptions. This approach, however, in this author's view is one way of obtaining a solid and scientific understanding of the nature of RORA events and of devising effective countermeasures to reduce these events. Work in this area using a combination of traffic flow theory, geometry, vehicle dynamics, probability theory, and driver behavior theory will eventually lead to a better determination of the form and parameterization of the mean function in accident prediction models. Using economic theory as an analogy, accident prediction models are like macroeconomic models, while the encroachment models are like microeconomic models. The interrelated and complementary nature of these two approaches indicates the need for both approaches in studying accident-flow-roadside design relationships.

To illustrate the complementary nature of these two approaches, it will be shown in the next two sections how a RORA prediction model can be developed for estimating roadside encroachment frequency and deriving the probability distribution of the lateral extent of encroachment when encroachment occurs.

RUN-OFF-THE-ROAD ACCIDENT PREDICTION MODEL

RORA and roadway data for rural two-lane undivided roads from a roadway cross-section design data base [Hummer, 1986], administered by FHWA and TRB, were used to develop an accident prediction model. One of the important feature of this particular data base is that it contains a rather detailed description of key design elements of various roadside obstacles. The roadway data used in this study include traffic and geometric design data of 596 road sections in three States: Alabama, Michigan, and Washington. The total length of these sections is 1,788 mi (2,861 km). About 5 years of SV RORA data from 1980 to 1984 were available for analysis. During the 5-year period, there were 4,632 SV reported to be involved in RORA on these road sections, regardless of vehicle and accident severity type. With the total vehicle miles estimated to be 7,639 million vehicle mi (14,514 million vehicle km), the overall SV RORA rate was 0.61 SV RORA per million vehicle mi (0.38 SV RORA per million vehicle km). Note that Alabama has incomplete accident data. For example, accidents occurred in icy or snowy conditions were not recorded and some injury accidents were not available [Hummer, 1986].

One important note for this data set is that none of the 596 sections contains continuous roadside objects such as a guardrail or a group of trees. The same data set has been used in Zegeer et al. [1987] to evaluate the effects of sideslope on the rate of SV RORA. Detailed descriptions and statistics of these road sections can be found in Hummer [1986] and Zegeer et al. [1987].

In addition to vehicle miles traveled, the covariates considered for individual road sections are presented in table 20. They include (1) dummy variables for Michigan and Washington to capture the overall difference in SV RORA rate among States, because of differences in omitted variables such as weather, socioeconomic and geographic variables, accident reporting threshold, and incomplete accident records described above; (2) AADT per lane, used as a surrogate measure for traffic density; (3) lane width; (4) median clear roadside recovery distance, measured from the right edge of the shoulder; (5) paved shoulder width; (6) earth, grass, gravel, or stabilized shoulder width; (7) median sideslope; (8) terrain type; (9) posted speed limit; (10) number of intersections per mile; (11) number of driveways per mile; and (12) number of bridges per mile. Many of these covariates were also considered by Zegeer et al. [1987]. Horizontal curvature and vertical grade data were not used in this exercise because 147 sections (about 25 percent) were found to have no curvature data and 341 sections (about 57 percent) did not have grade information. To some extent, terrain type is used as a surrogate measures for horizontal curvature and vertical grade;

The NB regression model, as described in chapter 2, was selected over the Poisson regression models because the estimated overdispersion parameters were found to be statistically

significant for all developed models. For illustration, four of the estimated NB regression models are presented in table 20, which shows the estimated parameters as well as their associated standard deviations and t-statistics. All covariates in all four models have the expected effects. Based on the AIC criterion, Model 3 was the final selected model. Using the goodness-of-fit measure R_α^2 as presented in chapter 5, the final selected model has a R_α^2 value of 0.62. That is, about 62 percent of the explainable variance were explained by the covariates included in this model. It is expected that a higher R_α^2 value could be achieved if horizontal curvature and vertical grade were available.

The posted speed limit was not found to be significant because of the lack of variation; 530 out of the 596 sections had a posted speed limit of 55 mi/h. Although the number of intersections per mile had the expected effect, it was not found to be statistically significant (at a 20 percent α level) and was removed from the final model.

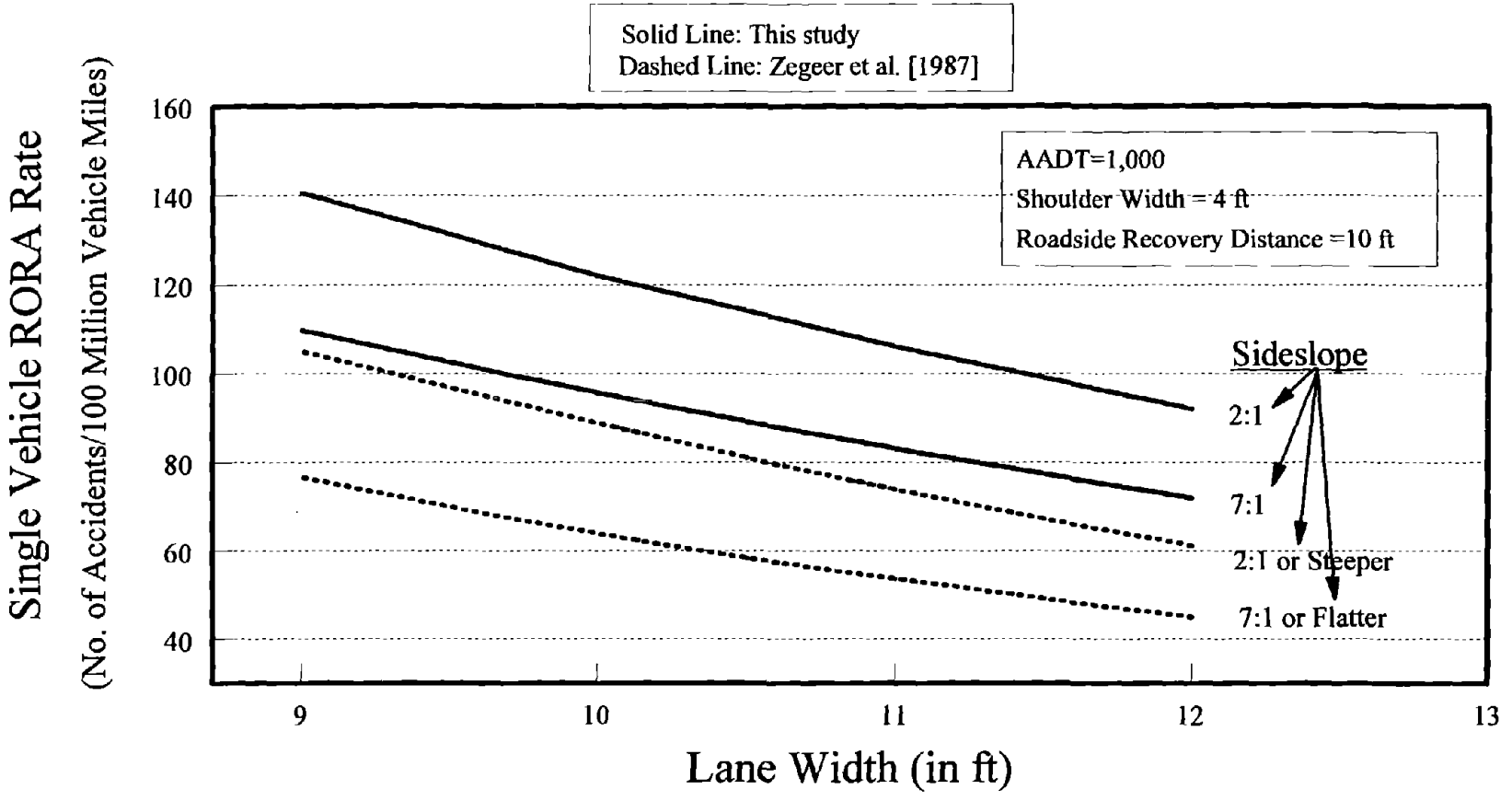
Major findings from Model 3 are presented as follows:

- If all considered variables have the same values, Michigan has the highest SV RORA rate and Alabama has the lowest rate. Michigan's rate is about 20 percent higher than Washington because of the difference in weather and socioeconomic conditions; while Alabama is about 34 percent lower than Washington mainly because of the incomplete Alabama accident data and difference in weather and other factors.
- AADT per lane shows a negative effect. One plausible explanation is that, all else being equal, higher vehicle density results in higher multiple-vehicle accident rate and lower SV accident rate.
- All else being equal, increasing lane width is expected to reduce SV RORA rate. Figure 13 gives an illustration of the SV RORA rates for various lane widths and sideslopes. In addition, this figure shows the same rates derived by Zegeer et al. [1987]. It can be seen that the rates from this study are much higher than those from Zegeer et al.'s study. The main reason is that there is a fundamental problem in the method used by Zegeer et al. to compute the mean rate. This problem is pertaining to the use of lognormal distributional assumption and has been pointed out in Miaou and Lum [1993].
- The effect of paved shoulder width was not found to be significantly different from the effect of stabilized shoulder width. All else being equal, increasing shoulder width by 1 ft (0.3048 m) is expected to reduce SV RORA rate by about 9 percent.
- Steeper sideslope is associated with a higher SV RORA rate. Figure 14 shows the relative rates for various sideslope ratios when compared to the rate of a sideslope of 7:1. This figure also shows that the same relative rates derived from Zegeer et al's model. It can be seen that this study shows lower relative rates than those from Zegeer et al's study. The t-statistic of the estimated parameter in table 20 shows that the sideslope was not as well determined as other variables. One possible reason is that for each road section, the median (i.e., 50th percentile) sideslope measurement was used as the most representative sideslope, but the actual sideslope may vary considerably within a given section [Zegeer et al. [1987].
- As expected, all else being the same, higher numbers of driveways and bridges per mile result in higher SV RORA rates.

Table 20. Estimated regression coefficients of some tested negative binomial regression models and associated statistics for single-vehicle run-off-the-road accidents.

Model Parameter	Model 1	Model 2	Model 3	Model 4
β_1 Dummy intercept (=1)	0.85487 (±0.71;1.20)	1.20021 (±0.46;2.61)	1.20043 (±0.46;2.62)	1.39669 (±0.45;3.09)
β_2 Dummy variable for Michigan (1=Michigan; 0=otherwise)	0.6170 (±0.12;4.97)	0.6075 (±0.12;4.91)	0.6076 (±0.12;4.92)	0.5744 (±0.12;4.66)
β_3 Dummy variable for Washington (1=Washington; 0=otherwise)	0.4429 (±0.15;2.98)	0.4218 (±0.14;3.10)	0.4218 (±0.13;3.16)	0.4799 (±0.12;3.95)
β_4 AADT per lane (in 10^3)	-0.1773 (±0.04;-4.46)	-0.1787 (±0.04;-4.52)	-0.1783 (±0.04;-4.57)	-0.1731 (±0.04;-4.40)
β_5 Lane width (in ft)	-0.1462 (±0.04;-3.51)	-0.1433 (±0.04;-3.49)	-0.1411 (±0.04;-3.43)	-0.1380 (±0.04;-3.35)
β_6 Median clear roadside recovery distance (in ft)	-0.01525 (±0.007;-2.162)	-0.01472 (±0.007;-2.11)	-0.01375 (±0.007;-1.97)	-0.01758 (±0.007;-2.59)
β_7 Paved shoulder width (in ft)	-0.0921 (±0.016;-5.685)	-0.0893 (±0.014;-6.47)	-0.0881 (±0.014;-6.38)	-0.0938 (±0.014;-6.88)
β_8 Earth, grass, gravel, or stabilized shoulder width (in ft)	-0.0894 (±0.015;-6.00)			
β_9 Median sideslope (e.g., 3:1 and 7:1 slopes are recorded as 1/3=0.33 & 1/7=0.14, respectively.)	0.6842 (±0.45;1.50)	0.6920 (±0.45;1.53)	0.6920 (±0.45;1.54)	----
β_{10} Terrain type (0=flat; 1=mountainous+rolling)	0.2973 (±0.09;3.37)	0.2937 (±0.09;3.34)	0.2939 (±0.09;3.35)	0.3123 (±0.09;3.54)
β_{11} Posted speed limit (in mi/h)	0.0070 (±0.01;0.63)	----	----	----
β_{12} Number of intersections per mile	0.0409 (±0.034;1.22)	0.0405 (±0.034;1.21)	----	----
β_{13} Number of driveways per mile	0.0102 (±0.006;1.78)	0.0102 (±0.006;1.77)	0.0129 (±0.006;2.33)	0.0126 (±0.006;2.27)
β_{14} Number of bridges per mile	0.2050 (±0.09;2.16)	0.2016 (±0.095;2.13)	0.2016 (±0.095;2.13)	0.2138 (±0.095;2.25)
Dispersion parameter (α)	0.3992 (±0.037;10.9)	0.3987 (±0.037;10.9)	0.3988 (±0.036;11.0)	0.4065 (±0.037;11.0)
$L(\alpha,\beta)$ (=loglikelihood function)	-1645.8	-1646.0	-1646.8	-1647.9
AIC value	3321.5	3317.9	3317.5	3317.8
Expected vs. observed total number of accidents	4,719.9 4,632.0	4,710.7 4,632.0	4,709.0 4,632.0	4,713.1 4,632.0

Notes: (1) 596 rural two-lane undivided road sections; total length=1,788 mi; about 5 years of accident data (1980-1984).
(2) Values in parentheses are asymptotic standard deviation and t-statistics of the coefficients above.
(3) ----- indicates "not included in the model."
(4) 1 mi = 1.61 km, 1 ft = 0.3048 m.



(1 mi = 1.61 km; 1 ft = 0.3048 m)

Figure 13. Illustration of single-vehicle run-off-the-road accident rates for various lane widths and sideslopes.

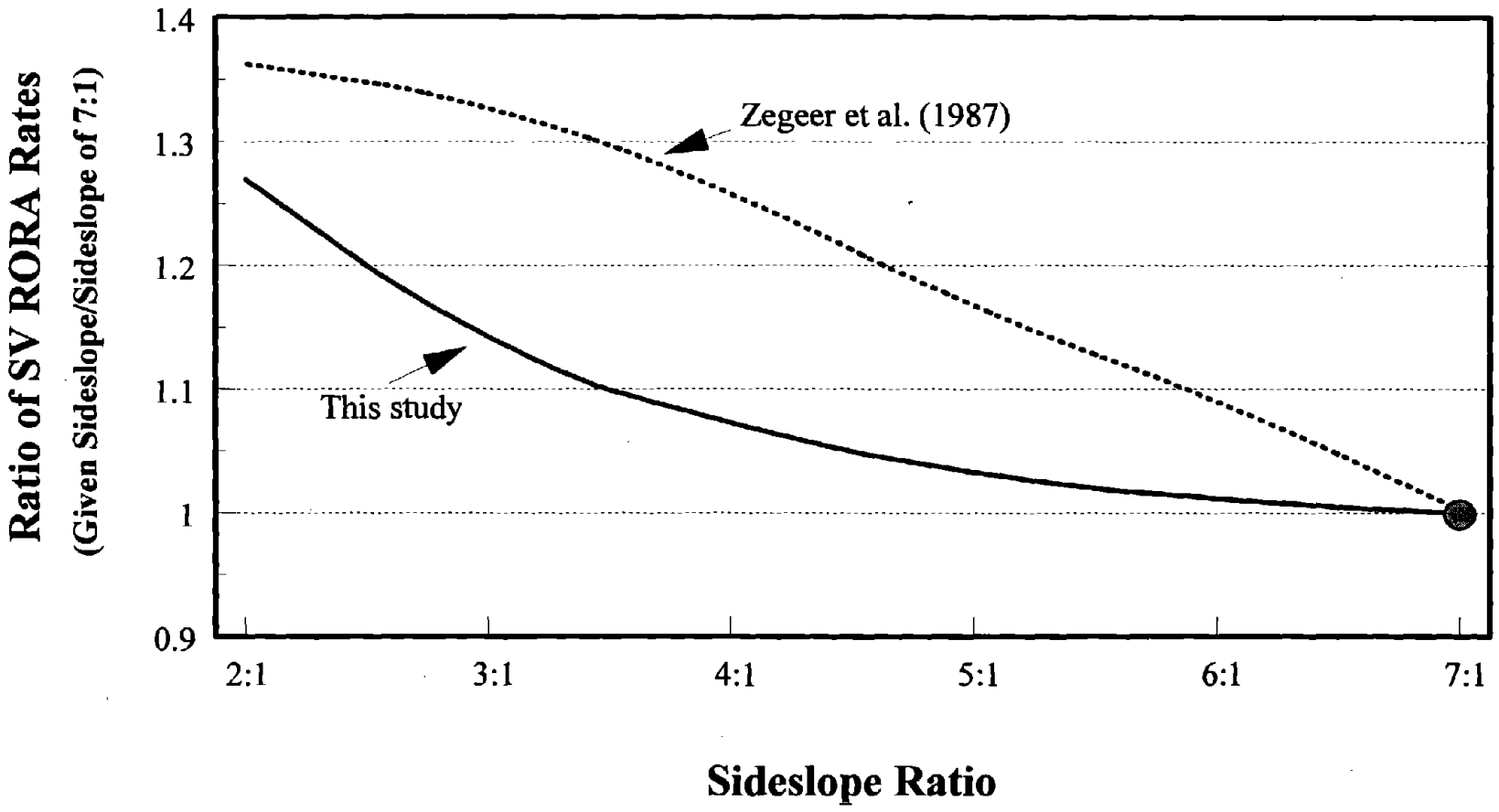


Figure 14. Single-vehicle run-off-the-road accident rates for a given sideslope versus single-vehicle run-off-the-road accident rate for a sideslope of 7:1

In the next section, Model 3 will be used to illustrate how an accident prediction model can be used to estimate roadside encroachment frequency and to derive the probability distribution of the lateral extent of encroachment when encroachment occurs.

ESTIMATING ENCROACHMENT FREQUENCY WITH ACCIDENT PREDICTION MODELS

The relationship between SV RORA probability and SV roadside encroachment probability for a vehicle traveling through a 1-mi or 1-km road section can be mathematically expressed as follow:

$$P(\text{SV RORA} \mid \text{Mainline, Rdside Design}) = P(\text{Rdside Encro} \mid \text{Mainline, Rdside Design}) \times P(\text{SV RORA} \mid \text{Rdside Encro, Mainline, Rdside Design}) \quad (48)$$

where

Mainline = Mainline traffic and geometric design variables;

Rdside Design = Rdside design variables;

P(SV RORA|Mainline, Rdside Design)
= conditional probability of being involved in a SV RORA when a vehicle travels through a 1-mi or 1-km road section that has a given geometric design and traffic characteristics as described in *Mainline* and *Rdside Design*; (Note that it is assumed here that the probability of having more than one SV RORA by a vehicle is zero);

P(Rdside Encro|Mainline, Rdside Design)
= conditional probability of having an SV roadside encroachment when a vehicle travels through a 1-mi or 1-km road section that has a given geometric design and traffic characteristics as described in *Mainline* and *Rdside Design*; (Note that it is assumed here that the probability of having more than one SV roadside encroachment by a vehicle is zero);

P(SV RORA|Rdside Encro, Mainline, Rdside Design)
= conditional probability of being involved in an SV RORA when a vehicle travels on a 1-mi or 1-km road section that has a given geometric design and traffic characteristics as described in *Mainline* and *Rdside Design* and has encroached on the roadside.

By assuming that *Rdside Design* has a very small and negligible effect on roadside encroachment probability, Eq. (48) can be rewritten as:

$$P(SV\ RORA \mid Mainline, Rdside\ Design) = P(Rdside\ Encro \mid Mainline) \times P(SV\ RORA \mid Rdside\ Encro, Mainline, Rdside\ Design) \quad (49)$$

Now, let's picture a condition where there exists an extremely bad roadside design such that when a vehicle encroaches on the roadside at any point on the road section it is 100 percent sure that the vehicle will result in a RORA. For example, one can picture a road section which has no shoulders and a ditch with a 1:1 sideslope ratio built right next to the traveled lane. Note that very dense point objects, such as trees and utility poles, alone the roadside would also be good examples. Of course, a road section with such a bad roadside design may not exist in the study area of interest. Under such a bad roadside design condition, *P(SV RORA|Rdside Encro, Mainline, "extremely bad" Rdside Design)* = 1, and therefore Eq. (49) can be reexpressed as:

$$P(SV\ RORA \mid Mainline, \text{"extremely bad" } Rdside\ Design) = P(Rdside\ Encro \mid Mainline) \quad (50)$$

To estimate the expected annual number of RORA on a road section with ℓ miles, one can simply multiply Eq. (50) with ($V \times \ell$), where V is the total number of vehicles traveling through the section per year (=365 × AADT). That is,

$$P(SV\ RORA \mid Mainline, \text{"extremely bad" } Rdside\ Design) \times V \times \ell = P(Rdside\ Encro \mid Mainline) \times V \times \ell \quad (51)$$

In Eq. (51), the right hand side is the annual roadside encroachment frequency of interest, and the left hand side is the expected number of SV RORA per year, which can be estimated using a conventional accident prediction model such as the Model 3 presented in the last section.

To estimate the roadside encroachment frequency using the Model 3, an extremely bad roadside design condition was created by setting shoulder width = 0, median clear roadside recovery distance = 0, and median sideslope = 1. (Note that sideslope ratio of 1:1 is the maximum median sideslope observed in the sample sections.) Except lane width and AADT, other variables were set equal to their average values. Also, because Alabama has incomplete accident data, only Michigan and Washington models are used. Figure 15 shows the estimated roadside encroachment frequencies per mile per year by various lane widths and AADT's using Eq. (51). The encroachment frequencies collected by Kennedy and Hutchinson [1966] and Cooper [1980], and the estimates given in SR214 are also presented in the figure for comparison.

One important observation can be made from figure 15 is that the estimated encroachment frequencies are very compatible with the encroachment data collected by others. Several comments can be made about this particular approach of estimating roadside encroachment frequency:

- One advantage of such an approach is that the encroachment frequency can be estimated for all kind of mainline design and traffic conditions. For example, if horizontal curvature and vertical grade were included in Model 3, the encroachment frequencies could be estimated for various horizontal curvatures and vertical grades as well.

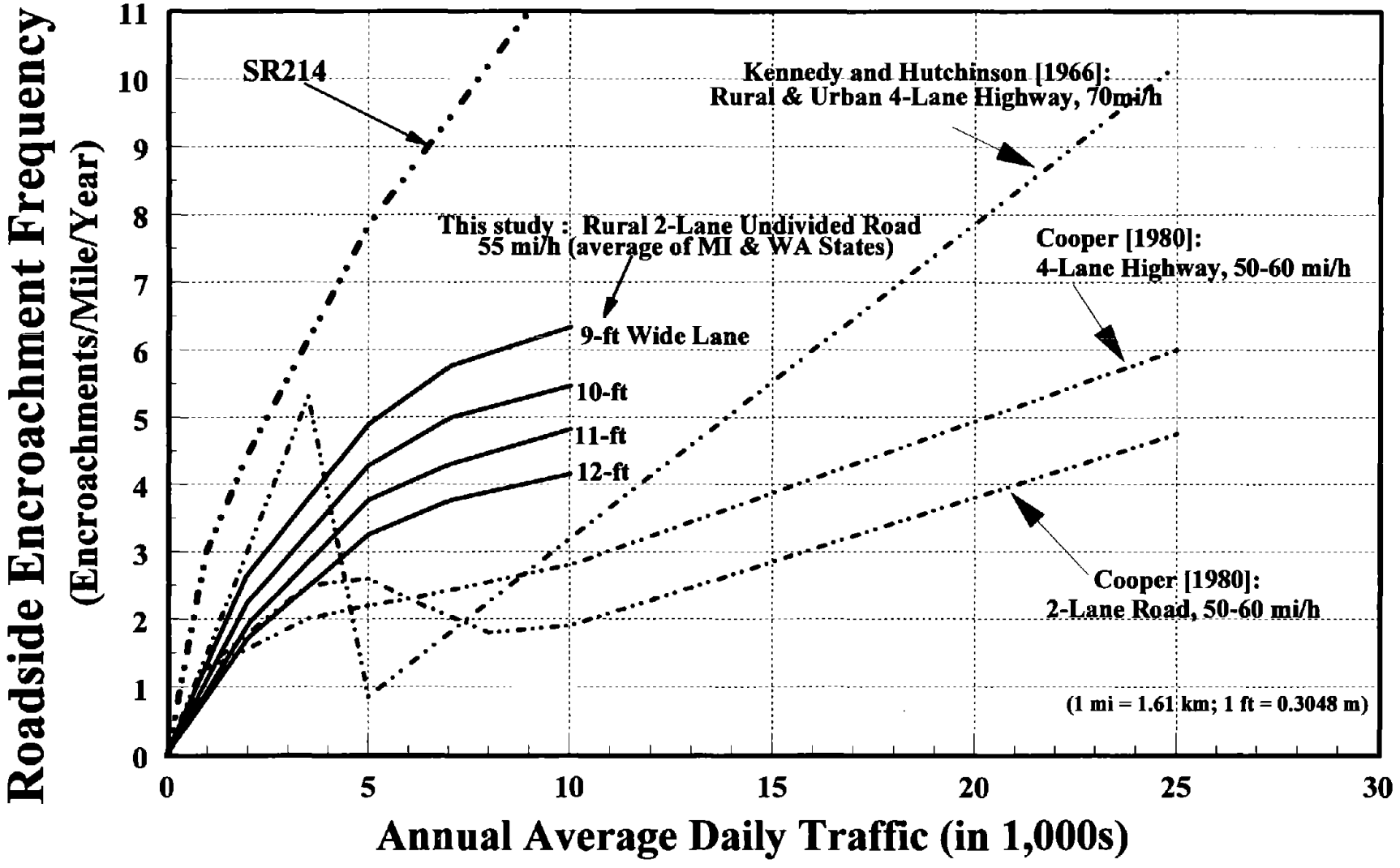


Figure 15. Comparison of the derived roadside encroachment frequency from the accident prediction model developed in this study and observed frequencies from earlier studies.

To actually collect such detailed encroachment data will be very expensive and maybe impractical.

- It has been suggested that "The encroachment frequency estimated in this manner can only be as accurate as the accident data used as input" [Daily et al. , 1994]. The suggestion is mainly related to the concern about the underreporting of minor accidents. This author would like to point out that this concern is not particularly serious for the approach taken in this section. The reason is that under the so-called extremely bad roadside design condition stated above, the resulting RORA is expected to be very severe and underreporting of such accidents is very unlikely. Therefore, provided a flexible mean function form is used in developing accident prediction model, the encroachment frequency estimated from such an approach is relatively unaffected by the underreporting of accidents.

- Another advantage of such an approach is that the estimated encroachment frequency is relatively uncontaminated by intentional encroachments. Again, the reason is that intentional encroachments are not likely to occur under such a bad roadside design condition.

It is important to point out that some degree of extrapolation is used in Eq. (51) because of the assumed extreme roadside conditions where shoulder width = 0, median clear roadside recovery distance = 0, and median sideslope =1. It is this author's judgment that the estimated encroachment frequency from Eq. (51) represents only potentially harmful and unintentional encroachments. In addition, the estimate is expected to be lower than what would actually happen on the roads, especially for those roads with wide shoulders where drivers tend to be more relaxed and harmless and unintentional roadside encroachments do occur quite often.

Another possible use of such an approach is to estimate the probability of the lateral extent of encroachment when a roadside encroachment occurs. That is, given a roadside encroachment has occurred, the approach can be used to estimate the probability that the encroached vehicle, in the absence of roadside obstacles, will leave the traveled lane by at least a distance of, say, L. Conceptually, this estimate can be achieved by a simple extension of the approach described above. Specifically, it can be achieved by setting shoulder width = L, median clear roadside recovery distance = 0, and median sideslope = 1. The other variables can be set in exactly the same way. Figure 16 shows a derived probability distribution of the lateral extent of encroachments using such approach. Since shoulder width is used to estimate the probability, the distribution is good for flat roadside condition (with no slopes). This estimated distribution can be seen to be quite consistent with AASHTO's distributions for roads with a design speed of 50-60 mi/h (80-96 km/h). On the other hand, it is very different from the distributions derived from Hutchinson and Kennedy's encroachment data. Note that the basis of AASHTO's distributions is not clear from its *Roadside Design Guide* [Daily et al., 1994]. In addition, the estimation of a single distribution for a design speed has been controversial; it has been suggested that multiple distributions for different sideslope ratios are necessary. In theory, this distribution could be conditional on sideslope, shoulder type (e.g., paved vs. unpaved,

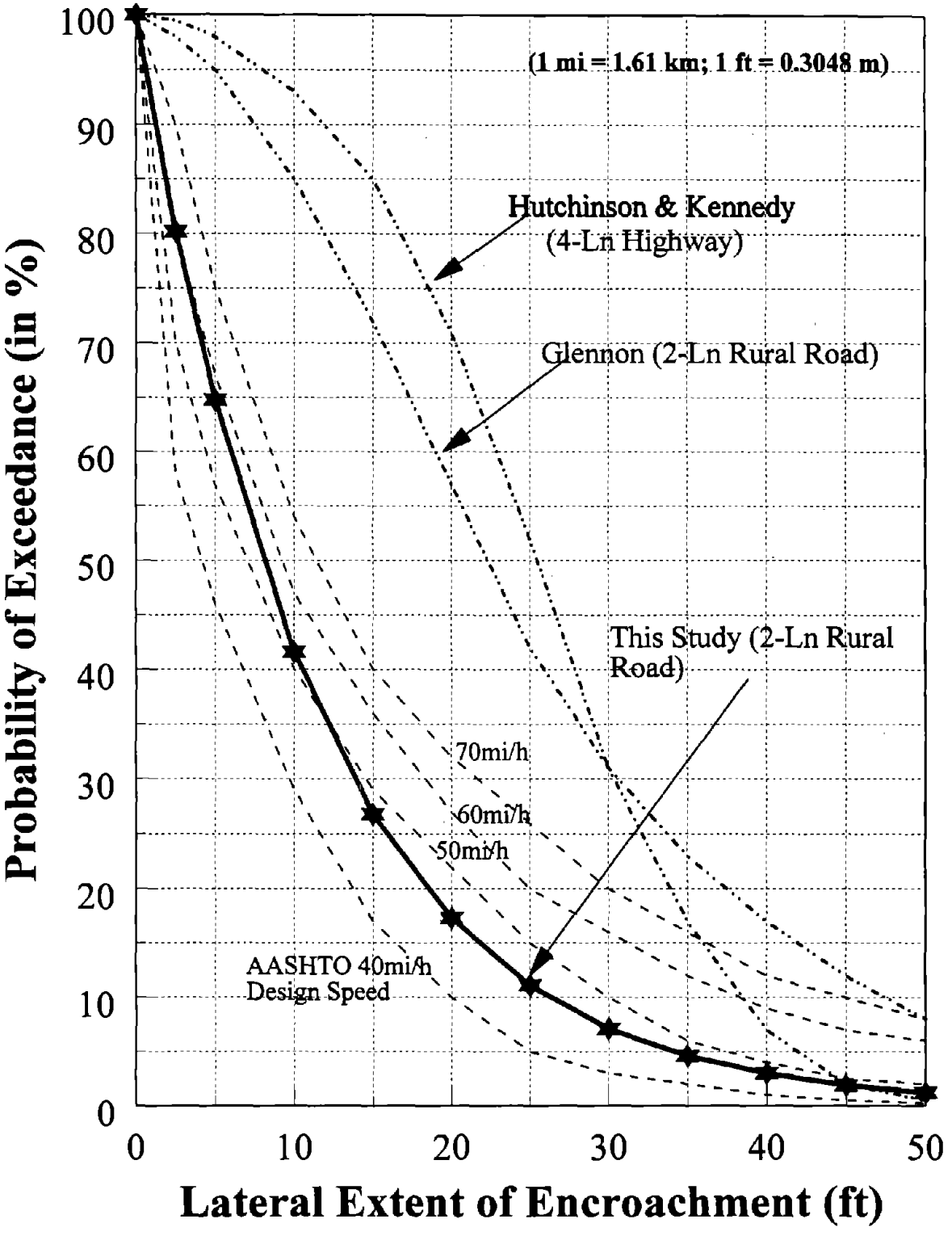


Figure 16. Comparison of various probability distributions of the lateral extent of encroachments.

with or without rumble strips), density of roadside hazards, traveled path, or even encroached angle (table 19). The readers are referred to Daily et al. [1994] for more discussion.

The illustration above shows that the approach described in this section can be a viable approach to estimating encroachment frequency without actually collecting the encroachment data that can be very costly. Most importantly, it is straightforward using such an approach to estimating encroachment frequencies for various mainline traffic and design conditions, e.g., AADT, lane width, horizontal curvature, and vertical grade. The only premise is that a sound accident prediction model be developed. The better the accident prediction model, the better the estimate of roadside encroachment frequency can be expected.

RECOMMENDATIONS FOR FUTURE RESEARCH

- The interrelated and complementary nature of the accident-based approach and encroachment-based approach suggests that both approaches are needed in studying accident-flow-roadside design relationships. It is our view that, given the current state of the art and funding situations, the emphasis of research on the encroachment-based approach should be more scientific-oriented, while the accident-based approach should be more application- or engineering-oriented. In other words, this author recommends that the current goal of encroachment-based approach should be on deepening the scientific understanding of the ROR process that leads to a roadside accident. Each conditional probability shown in table 19 will require a lot of theoretical research, data collection, parameter calibration, and validation. It would be naive to believe that by over-simplifying each conditional probability and then multiplying them together would provide realistic estimates. On the other hand, the accident-based approach should not be limited to conventional Poisson and NB statistical regression models that have a restricted form of mean function. For the accident prediction model to be useful for design and safety engineers, the mean function must have sound engineering basis and interpretation. In this aspect, the accident-based approach can definitely benefit from the engineering-oriented thinking and formulation adopted by the encroachment-based approach. Of course, there are still a lot of room for improvement in the encroachment-based thinking. For example, driver behavior after an errant vehicle leaves the travelway is perhaps one of the missing links in the current encroachment-based model.

- More research to explore the interrelationship between the accident-based approach and encroachment-based approach can help develop viable and cost-effective ways of quantifying roadside safety. The illustration demonstrated in the last two sections is a good example. Another example is the SR214 study on two-lane roads in which both right-side and left-side encroachment angles can be estimated using the usual accident prediction model building procedure without actually collecting them. In studying the interrelationship between these two approaches, one may find that the extension from two-lane undivided road sections to other multiple-lane road sections or intersections is not straightforward.

- Roadside encroachment frequency is expected to be highly dependent on the horizontal curvature of a road section. For the illustration shown earlier, there is a good chance that a more sophisticated statistical method can be used to develop the accident prediction model, which includes horizontal curvature as an explanatory variable and takes into account the 147 sections with missing curvature data. It is this author's judgment that it will be a worthwhile exercise.

- There is a need to develop general roadside design safety indices to measure the relative risk of being involved in an accident under different roadside design conditions when an errant vehicle encroaches on the roadside. Same as most of the economic indices, these indices should have clear concepts and definitions and be objective and quantifiable. For roadway design and safety engineers to accept these indices in their practice, good engineering and statistical interpretations of these indices are essential. The clear zone concept, e.g., is a good starting point of conceptualizing these indices. The roadside hazard ratings developed and used in Hummer [1986] and Zegeer et al. [1987] are subjective and not quantifiable. However, these ratings are useful steppingstones for developing more useful indices in the future.

7. SUMMARY AND FUTURE RESEARCH

In developing accidents-flow-roadway design models, the R^2 goodness-of-fit measure has been used by traffic safety engineers and researchers for many years to (1) determine the quality and usability of a model; (2) select covariates (or explanatory variables) for inclusion in the model; (3) make a decision as to whether it would be worthwhile to collect additional covariates; and (4) compare the relative quality of models from different studies.

The state of the development of accident prediction models was reviewed in chapter 2. In chapter 3, the pitfalls of using R^2 to make these decisions and comparisons were demonstrated through simulation studies for commonly used accident prediction models, such as the Poisson and NB regression models. Because the accident prediction models are non-normal and its functional forms are typically nonlinear and multiplicative (or interactive) in nature, it was shown that R^2 is not an appropriate measure to make any of the decisions and comparisons above. In addition, three properties were identified as essential and desirable for any alternative measures to appropriately evaluate the goodness-of-fit of accident prediction models. These properties are (1) [0,1] bound property; (2) proportional increase property; and (3) invariant with respect to the mean property. Basically, [0,1] bound property says that one would like to have a value of zero if no covariate is included in the model and a value of 1 if all the necessary covariates are included. Proportional increase property says that if all covariates are independent and equally important, then when one selects and adds these covariates to the model one at a time the increase in value should be the same for each covariate regardless of their order of selection. Invariant with respect to the mean property says that the value of the criterion will not change by simply increasing or decreasing the value of the intercept term of the model.

In chapter 4, the concept of a goodness-of-fit criterion called the AIC was introduced. The capability of AIC-based criteria and other criteria, such as scaled deviance and Pearson's X^2 statistics, to select the correct models were then evaluated. Again, the evaluation was carried out using simulations. Although the simulations conducted in this study were limited, several interesting results have been reported. The simulations indicated that the results of statisticians' experiments under the situation where the correct model is among a set of candidate models considered by the modeler are not appropriate for use in developing accident prediction models. A more appropriate situation that needs to be tested is the ability of the goodness-of-fit criteria to select the best model(s) when some of the relevant variables are omitted from all candidate models. In addition, an AIC-based criterion called $CAIC_{NB}$ was recommended for use in variable selection when developing accident prediction models. It was further suggested that engineering judgment should be exercised at every step of the variable selection. Another suggestion was that instead of selecting the candidate model that has the lowest $CAIC_{NB}$ value, one should also consider those candidate models that are compatible. For example, one should also look into those models that have $CAIC_{NB}$ values not greater than 3 to 5 of the $CAIC_{NB}$ value of the best model.

Chapter 5 introduced three alternative goodness-of-fit measures which were developed based on the so-called Poisson concept. Their performances in terms of the three properties discussed above were evaluated again through simulations. Because of limited resources, their

performance was evaluated under large samples only. Based on limited simulation results, one of the alternative criteria called R^2_a was recommended for use to evaluate and compare the quality of accident prediction models if the sample size is large.

The simulation studies in chapters 3 to 5 suggested several factors that make the evaluation of the goodness-of-fit of accident prediction models difficult: (1) highly skewed probability models; (2) nonlinear functional relationships; (3) very low overall means; and (4) omitted variables. It was recommended that more systematic simulation studies than those reported in this report should be planned and carried out in the future.

Chapter 6 reviewed two approaches that have traditionally been used in previous studies to develop the relationship between roadside accident frequency and roadside hazards: accident-based approach and encroachment-based approach. The interrelated and complementary nature of the two approaches as they applied to the prediction of RORA and vehicle roadside encroachments were discussed. To illustrate the complementary nature of these two approaches, RORA and roadway data for rural two-lane undivided roads from an FHWA and TRB roadway cross-section design data base were used. Specifically, it was shown that a RORA prediction model could be developed for use to estimate roadside encroachment frequency and to derive the probability distribution of the lateral extent of encroachment. It was suggested that exploring the complementary nature of these two approaches could be a viable avenue to reduce data collection cost, and more research in this direction was recommended.

REFERENCES

American Association of State Highway and Transportation Officials (AASHTO). *Roadside Design Guide*. Washington, DC; 1989.

Anderson-Sprecher, R. "Model Comparisons and R^2." *The American Statistician* 48(2): 113-117; 1994.

Agresti, A. *Categorical Data Analysis.* New York: John Wiley & Sons; 1990.

Barrett, J. "The Coefficient of Determination – Some Limitations." *The American Statistician* 28(1):19-20; 1974.

Belanger, C. *An Estimation of the Safety of 4 Legged Unsignalized Intersections.* Paper presented at Transportation Research Board 73rd Annual Meeting in Washington, DC; January 9-13, 1994.

Black, W.R. *Highway Accidents: A Spatial and Temporal Analysis*. Paper presented at Transportation Research Board 70th Annual Meeting, Washington, DC; January 13-17, 1991.

Bozdogan, H. "Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions." *Psychometrik* 52(3): 345-370; 1987.

Breslow, N. "Extra-Poisson Variation in Log-Linear Models." *Applied Statistics* 33:38-44; 1984.

Brüde, U.; Larsson, J. "Models for Predicting Accidents at Junctions Where Pedestrians and Cyclists are Involved. How Well Do They Fit?" *Accident Analysis & Prevention* 25(5): 499-509; 1993.

Cameron, A.C.; Trivedi, P.K. "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests." *Journal of Applied Econometrics* 1:29-53; 1986.

Cameron, A.C.; Trivedi, P.K. "Regression-Based Tests for Overdispersion in the Poisson Model." *Journal of Econometrics* 46:347-364; 1990.

Carroll, R.J.; Ruppert, D. *Transformation and Weighting in Regression.* New York: Chapman and Hall; 1988.

Cheng, K.F.; Wu, J.W. "Testing Goodness of Fit for a Parametric Family of Link Functions." *Journal of the American Statistical Association* 89(426): 657-664; 1994.

Christiansen, C.L.; Morris, C.N.; and Pendleton, O.J. *Hierarchical Poisson Model, with Beta Adjustments for Traffic Accident Analysis*. Paper presented at the American Statistical Association Conference in Boston, MA; August 1992.

Cooper, P. *Analysis of Roadside Encroachments–Single Vehicle Run-Off-Road Accident Data Analysis for Five Provinces.* B.C. Research, Vancouver, Canada, March 1980.

Council, F.M.; Stewart, J.R.; Reinfurt, D.W.; and Hunter, W.W. *Exposure Measures for Evaluating Highway Safety Issues.* Volume 1. Final Report. HSRC-PR123. Highway Safety Research Center, University of North Carolina; November 1983.

Council, F.M.; Reinfurt, D.W.; Campbell, B.J.; Roediger, F.L.; Carroll, C.L.; Dutt, A.K.; Dunham, J.R. *Accident Research Manual.* FHWA/RD-80/016. Prepared by the University of North Carolina for FHWA; February 1980.

Cox, D.R. "Some Remarks on Overdispersion." *Biometrika* 70(1):269-274; 1983.

Cox, D.R.; Lewis, P.A.W. *The Statistical Analysis of Series of Events.* London: Chapman and Hall; 1966.

Cox, D.R.; Wermuth, N. "A Comment on the Coefficient of Determination for Binary Responses." *The American Statistician* 46(1): 1-4; 1992.

Cramer, J.S. *Econometric Applications of Maximum Likelihood Methods.* New York: Cambridge University Press; 1989.

Daily, K.; Hughes, W.; McGee, H. *Experimental Plans for Accident Studies of Highway Design Elements: Encroachment Accident Study.* Prepared for FHWA; April 1994.

Dean, C.; Lawless, J.F. "Tests for Detecting Overdispersion in Poisson Regression Models." Journal of the American Statistical Association 84(406): 467-472; June 1989.

Federal Highway Administration. Highway Performance Monitoring System Field Manual. Washington, DC: U.S. Department of Transportation; 1987.

Fridstrøm, L.;. Ifver, J.; Ingebrigtsen, S.; Kulmala, R.; and Thomsen, L.K. "Measuring the Contribution of Randomness, Exposure, Weather, and Daylight to the Variation in Road Accident Counts." *Accident Analysis & Prevention* 27(1): 1-20; 1995.

Frome, E.L.; Cragle, D.L.; McLain, R.W. "Poisson Regression Analysis of the Mortality Among a Cohort of World War II Nuclear Industry Workers." *Radiation Research* 123:138-152; 1990.

Gelfand, A.E.; Dalal, S.R. "A Note on Overdispersed Exponential Families." *Biometrika* 71(1): 55-64; 1990.

Glennon, J.C. *Roadside Safety Improvement Programs for Freeways–A Cost Effectiveness Priority Approach.* NCHRP Report 148, Transportation Research Board. Washington, DC; 1974.

Harwood, D.W.; Mason, J.M.; Graham, J.L. *Conceptual Plan for an Interactive Highway Safety Design Model.* FHWA-RD-93-122. Washington, DC: Federal Highway Administration; February 1994.

Hauer, E.; Hakkert, A.S. "Extent and Some Implications of Incomplete Accident Reporting." *Transportation Research Record* 1185:1-10; 1988.

Hsiao, C., *Analysis of Panel Data.* Cambridge University Press, New York; 1986.

Hummer, J. *Safety Effects of Cross-Section Design for Two Lane Roads – Data Base User's Guide.* Prepared by Goodell-Grivas, Inc., for FHWA and the Transportation Research Board; December 1986.

Hurvich, C.M; Tsai, C.-L. *Model Selection for Extended Quasi-Likelihood Models in Small Samples.* Technical Report; 1994.

Hurvich, C.M; Tsai, C.-L. "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76(2): 297-307; 1989.

Hutchinson, J.W.; and Kennedy, T.W. *Medians of Divided Highways – Frequency and Nature of Vehicle Encroachments.* Engineering Experiment Station Bulletin 487, University of Illinois, June 1966.

Joshua, S.C.; Garber, N.J. "Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models." *Transportation Planning and Technology.* 15: 41-58; 1990.

Jovanis, P.P.; Chang, H.L. "Modeling the Relationship of Accidents to Miles Traveled." *Transportation Research Record* 1068: 42-51; 1986.

Kvålseth, T.O. "Cautionary Note About R^2." *The American Statistician* 39(4): 279-285; 1985.

Lawless, J.F. "Negative Binomial and Mixed Poisson Regression." *The Canadian Journal of Statistics* 15(3):209-225; 1987.

Maher, M.J. "A Bivariate Negative Binomial Model to Explain Traffic Accident Migration." *Accident Analysis & Prevention* 22(5): 487-498; 1990.

Mak, K.K. and Sicking, D.L. *Development of Roadside Safety Data Collection Plan,* Technical Report, Texas Transportation Institute, Texas A&M University System, College Station, Texas; 1992.

Maycock, G.; and Hall, R.D. *Accidents at 4-arm Roundabouts.* Transport and Road Research Laboratory Report 1120; 1984.

McCullagh, P.; Nelder, J.A. *Generalized Linear Models.* London: Chapman and Hall; 1983.

McFarland, W.F.; and Ross, H.E. Jr. *Development of Design Criteria for Safer Luminaire Supports.* National Cooperative Highway Research Program (NCHRP) Report 77, Transportation Research Board. Washington, DC; 1969.

Miaou, S.-P. "The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions." *Accident Analysis & Prevention* 26(4): 471-482; 1994.

Miaou, S.-P.; Hu, P.S.; Wright, T.; Rathi, A.K.; Davis, S.C. "Relationships Between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach." *Transportation Research Record.* 1376:10-18; 1992.

Miaou, S.-P., Hu, P.S., Wright, T., Davis, S.C., Rathi, A.K. *Development of Relationship Between Truck Accidents and Geometric Design: Phase I.* Publication No. FHWA-RD-91-124; 1993.

Miaou, S.-P.; Lum, H. "Modeling Vehicle Accidents and Highway Geometric Design Relationships." *Accident Analysis and Prevention* 25(6):689-709; 1993.

Miller, D.M. "Reducing Transformation Bias in Curve Fitting." *The American Statistician* 38(2): 124-126; May 1984.

Mohamedshah, Y.M.; Paniati, J.F.; Hobeika, A.G. "Truck Accident Models for Interstates and Two Lane Rural Roads." *Transportation Research Record* 1407: 35-41; 1993.

Morris, C.N.; Christiansen, C.L.; and Pendleton, O.J. *Application of New Accident Analysis Methodologies, Volume III – Theoretical Development of New Accident Analysis Methodology.* FHWA-RD-91-015, FHWA; September 1991.

Myers, R.H. *Classical and Modern Regression with Applications.* Second Edition. PWS-KENT Publishing Company, Boston; 1990.

Okamoto, H.; Koshi, M. "A Method to Cope with the Random Errors of Observed Accident Rates in Regression Analysis." *Accident Analysis & Prevention* 21(4): 317-332; 1989.

Ratkowsky, D.A. *Nonlinear Regression Modeling: A Unified Practical Approach.* New York: Marcel Dekker, Inc.; 1983.

Rohatgi, V.K. *An Introduction to Probability Theory and Mathematical Statistics.* John Wiley & Sons, New York; 1976.

Ross, S.M. *Introduction to Probability Models.* Fourth Edition. Academic Press, New York; 1989.

Roy Jorgensen Associates, Inc. *Cost and Safety Effectiveness of Highway Design Elements.* Report 197. Washington, DC: National Cooperative Highway Research Program; 1978.

Saccomanno, F.F.; Buyco, C. *Generalized Loglinear Models of Truck Accident Rates*. Paper presented at Transportation Research Board 67th Annual Meeting, Washington, DC; January 1988.

Sakamoto, Y.; Ishiguro, M.; and Kitagawa, G. *Akaike Information Criterion Statistics*. D. Reidel Publishing Company; 1986.

Scott, A.; Wild, C. "Transformations and R^2." *The American Statistician* 45(2): 127-129; 1991.

Transportation Research Board. *Designing Safer Roads–Practices for Resurfacing, Restoration, and Rehabilitation*. TRB Special Report 214. Washington, DC; 1987.

Wedderburn, R.W.M. "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method." *Biometrika* 54:439-447; 1974.

Weisberg, S. *Applied Linear Regression*. New York: John Wiley; 1985.

Willett, J.B; Singer, J.D. "Another Cautionary Note About R^2: Its Use in Weighted Least-Squares Regression Analysis." *The American Statistician* 42(3): 236-238; 1988.

Zegeer, C.V. and Parker, M.R. Jr. *Cost-Effectiveness of Countermeasures for Utility Pole Accidents*. Publication No. FHWA/RD-83/063, FHWA, Washington, DC; December 1985.

Zegeer, C.V.; Hummer, J.; Reinfurt, D.; Herf, L.; Hunter, W. *Safety Effects of Cross-Section Design for Two-Lane Roads*. Volumes I and II. Chapel Hill: University of North Carolina; 1987.

Zegeer, C.V.; Stewart, R.; Reinfurt, D.; Council, F.; Neuman, T.; Hamilton, E.; Miller, T.; and Hunter, W. *Cost-Effective Geometric Improvements for Safety Upgrading of Horizontal Curves*. Volumes I. Final Report. Chapel Hill: University of North Carolina; 1990.