UMTRI-2008-52 OCTOBER 2008

UPDATED RATIO OF CRASH SEVERITIES REPORTABLE TO THE MCMIS CRASH FILE

PAUL E. GREEN
DANIEL BLOWER

UNIVERSITY OF MICHIGAN
TRANSPORTATION RESEARCH INSTITUTE

UMTRI-2008-52

Updated Ratio of Crash Severities Reportable to the MCMIS Crash File

Paul E. Green
Daniel Blower

The University of Michigan
Transportation Research Institute
Ann Arbor, MI 48109-2150
U.S.A.

October 2008

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
UMTRI-2008-52		

4. Title and Subtitle	5. Report Date
Updated Ratio of Crash Severities Reportable to the MCMIS Crash File	October 2008
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Green, Paul E., and Blower, Daniel	UMTRI-2008-52
9. Performing Organization Name and Address	10. Work Unit no. (TRAIS)
The University of Michigan Transportation Research Institute 2901 Baxter Road Ann Arbor, Michigan 48109-2150 U.S.A.	052702 11. Contract or Grant No. DTMC75-06-H-00003
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
U.S. Department of Transportation Federal Motor Carrier Safety Administration 400 Seventh Street, SW Washington, D.C. 20590	Special report 14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

The Motor Carrier Management Information System (MCMIS) Crash file has been developed by the Federal Motor Carrier Safety Administration (FMCSA) to serve as a census file of trucks and buses involved in traffic crashes meeting a specific crash severity threshold. Each state is responsible for identifying cases that meet the MCMIS Crash file criteria and reporting the required data through the SafetyNet system. UMTRI has completed a set of evaluations of state reporting and found that reporting rates range from over 80 percent to less than 10. The present report provides an updated model to predict, given a known number of fatal involvements, the number of crash involvements a state should be reporting. Additional observations became available since the previous report. These observations are incorporated and the model re-estimated.

In each state, the number of fatal involvements is well-known, so all states will start with a known quantity, the number of fatal truck and bus crash involvements. It is then hypothesized that the ratio of reportable crash severities, that is, the ratio of fatal involvements to nonfatal involvements, will apply across all the states.

Data from 11 states (representing 13 observations) that provide all the information necessary to identify MCMIS-reportable cases were used. A weighted log-linear model is fit to MCMIS data for the states that have information recorded for both fatal and nonfatal crashes. The model is then used to estimate the number of nonfatal crashes for a new state in which only the number of fatal crashes is known. Prediction intervals are presented.

17. Key Words	18. Distribution Statement
MCMIS, crash severity, crash ratios	Unlimited

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	24	

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 yard	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 degrees)		
^oF	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	^oC
		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	35.314	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 degrees)		
^oC	Celsius	1.8C+32	Fahrenheit	^oF
		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 March 2003)

Table of Contents

List of Tables

List of Figures

Updated Ratio of Crash Severities Reportable to the MCMIS Crash File

1 Introduction

This report is part of a series of reports developed by the University of Michigan Transportation Research Institute (UMTRI) to assist the Federal Motor Carrier Safety Administration (FMCSA) in its efforts to improve truck and bus crash data reporting by the States. One part of the effort is to evaluate the completeness and accuracy of reporting by the States to the Motor Carrier Management Information System (MCMIS) Crash file. UMTRI has issued a series of evaluations of individual States.[1] Another part of the effort is to develop a method of predicting the number of total reportable cases from each State, to help individual States estimate the number of cases they should be reporting, based on the number of fatal involvements in the state, which is known with good reliability. The present work reports on further development of the methodology to predict total reportable crash involvements qualifying for the MCMIS Crash file from the number of fatal involvements. Five additional States with suitable crash data have been evaluated since the last report on the ratio. Data from those States are incorporated into the methodology and a new prediction equation is developed. This report represents an update of the previous reports.[1, 2]

The Motor Carrier Management Information System (MCMIS) Crash file has been created and compiled by the Federal Motor Carrier Safety Administration (FMCSA) to serve as a census file of trucks and buses involved in traffic crashes meeting a specific crash severity threshold. FMCSA maintains the MCMIS file to support its mission to reduce crashes, injuries, and fatalities involving large trucks and buses. Designing effective safety measures requires accurate and complete crash data to understand the dimensions of the crash problem. The data are used to monitor the safety performance of carriers and to identify crash safety trends. The usefulness of the MCMIS Crash file depends upon individual states transmitting a standard set of data items on all trucks and buses involved in traffic crashes that meet the file's crash severity threshold.

The University of Michigan Transportation Research Institute (UMTRI) has prepared a series of reports evaluating the completeness of reporting from selected states. As of this report, UMTRI has completed 29 evaluations, covering 27 states. Reporting rates have ranged from less than ten percent of reportable cases to over eighty percent.

Each state is responsible for identifying cases that meet the MCMIS Crash file criteria and reporting the required data through the SafetyNet system. The MCMIS selection criteria are clearly stated and, in theory, easily applied. To comply with the reporting requirements, states have adapted existing systems, (which were developed for their own safety and enforcement purposes), to identify and capture the correct cases. Each state has its own threshold for reportable crashes, its own system for classifying vehicles, and its own set of information that it collects on the crashes. Given the multiple purposes for which crash data are collected by states,

---

[1] These State reports may be found at the MCMIS Evaluation Reports at this website: http://www.umtri.umich.edu/divisionPage.php?pageID=4.

it is often not clear which crashes should be reported and which should not. This is especially true for crashes that do not involve a fatality. Many states do not regularly collect the specific information needed to determine if a nonfatal case is reportable—e.g., whether an injured person was transported for medical attention or whether a vehicle was towed due to disabling damage.

As a consequence of the mismatch between MCMIS requirements and the methodology of states' established systems, there is often no easy way for the states, or FMCSA, to know if the right number of cases is being reported. Some states have thought that they were in compliance and fully reporting to the MCMIS Crash file, but the UMTRI evaluation found significant underreporting. Until all states can be directly evaluated, a method of predicting, within a reasonable range, the number of cases that each state should report could serve as a guidepost or benchmark to the states on where they stand. Such a benchmark can alert a state that a problem exists, and motivate a process to identify a solution.

This paper develops a method to estimate the number of involvements each state should report to the MCMIS Crash file, even if the state's data system cannot readily identify all the crashes that meet the Crash file criteria. It is based on developing a ratio of fatal to nonfatal reportable crashes, using data from states that have sufficiently complete information to identify with good confidence. In each state, the number of fatal involvements is well-known, so all states will start with a known quantity, the number of fatal truck and bus crash involvements. It is then hypothesized that the ratio of fatal involvements to nonfatal reportable involvements will apply across all the states. If this is true, it is possible, knowing the number of fatal involvements that occurred in a state, to predict the number of involvements of lesser severity, and thus predict the total number of cases that the state should report to the MCMIS Crash file.

During the process of evaluating state reporting, UMTRI identified a number of states whose data systems provide the data necessary to apply the MCMIS reporting criteria completely. Virtually all states can identify trucks and buses reasonably well, and all states can identify fatal crashes cleanly. However, most states do not regularly collect the information needed to identify reportable nonfatal crashes: crashes in which an injured person was transported for immediate medical attention or crashes in which at least one vehicle was towed due to disabling damage. States identify persons injured in a crash, and even nominally use the same system to classify injury severity, but not all have taken the next step to capture if the person was transported for medical attention. Similarly, many states record if a vehicle was towed, but not whether the reason for the tow was disabling damage.

UMTRI identified eleven states whose existing crash data systems can identify all levels of the reporting criteria. Two of the states were evaluated for two different data years, providing a total of thirteen observations. These data can be used to estimate the ratio of reportable crash severities, that is, the ratio of fatal involvements to nonfatal involvements. [See evaluations in references 3 through 14.]

The General Estimates System (GES) file from the National Highway Traffic Safety Administration (NHTSA) may be considered a source of crash data for this work. GES is a nationally-representative sample of police-reported crashes. GES cases are sampled from primary sampling units (PSUs) around the country and a standard set of data are coded from the sampled police-reports. The GES data can be used to identify trucks and buses, and crash detail includes whether an injured person was transported for immediate medical attention or a vehicle towed due to disabling damage. Thus, GES nominally can be used to cleanly estimate the number of reportable involvements to the MCMIS Crash file.

However, there are a number of reasons why the GES file is not satisfactory for this application. Though the GES variable recording whether a vehicle was towed has a level for towed due to damage, we know through doing the state reporting evaluations that many state police reports do not include that information, so this variable likely underestimates the number of vehicles towed due to disabling damage. In addition, the GES file is a sample drawn through a complex stratified, hierarchical sampling system. Truck and bus crashes are a small sample, relative to automobile crashes. The standard errors for small subsets of the file, such as trucks, are relatively large. Moreover, it is known that GES underestimates the number of fatal truck and bus involvements. For the years 2000 through 2005, GES estimated 2,903 to 5,819 trucks involved in a fatal crash, with an average of about 4,100. The average from UMTRI's Trucks Involved in Fatal Accidents file is over 5,100. Moreover, the 95 percent confidence interval for an estimate in GES of 4,100 fatal truck involvements is ±1,800, or about 40 percent. Since the basis of the ratio is the number of fatal involvements, which it is assumed can be identified precisely in state crash data, the GES estimate is too uncertain to be reliable.

2 Problem statement

The purpose of this report is to develop a method that can be used to assist states in determining if they are in compliance with FMCSA's requirement to report all crash involvements reportable to the MCMIS Crash file. This method is not intended to identify a precise number of reportable cases for each state, but to give guidance as to whether a state's reporting is within an expected range. It is assumed that states can identify fatal involvements with reasonably good confidence. Virtually all the states can identify trucks and buses readily. Crashes in which a fatality occurred are equally clearly identifiable. All states identify fatal injuries, and the definition of a fatal traffic accident—death within 30 days of the crash—is standard. Accordingly, one level of the hypothesized ratio, i.e., fatal involvements, should be well established in all states.

It is assumed that the relationship of fatal to nonfatal <u>reportable</u> involvements exists independent of any particular state system. That is, the ratio does not depend on a state's definitions or system of collecting data, so the ratio established in one set of states should hold true for other states. As a counter-example, consider the common system for classifying injury severity. Most states use the KABC0 system, which classifies injuries as fatal, incapacitating, non-incapacitating but evident, complaint of pain, and no injury. Fatal injuries are clear and not subject to much interpretation. But the other injury levels are more difficult to classify consistently and can be subject to more interpretation. As a consequence, states vary widely in the relative proportions of A-, B-, and C-injuries.[2] The differing proportions are related not to some underlying difference

[2] See O'Day, J., *Accident Data Quality*. National Cooperative Highway Research Program Synthesis of Highway Practice, No. 192. Federal Highway Administration, Washington, DC, 1993. O'Day found that the proportion of A injuries varied from 4.9% to 23.8% in a sample of about 20 states. The findings were for 1990-1991 data, but illustrate the point.

in the severity of crashes in different states but to variations in the interpretation and application of standard definitions.

In contrast, the MCMIS Crash file criteria do not depend on crash severity standards that are known to be applied unevenly, but instead provide a relatively simple definition that should apply in roughly the same way everywhere. Reportable nonfatal involvements include either an injury transported for immediate medical attention or a vehicle towed due to disabling damage. An injury serious enough to be transported for treatment in Maine likely would also be transported if it occurred in California. There may be some variations from state to state, but they are not expected to be large. Similarly, it is not expected that whether a vehicle is disabled enough to be towed will vary much by region. There may be areas where towing following a crash is more common, but less variation, by state, in judging whether a vehicle has suffered *disabling* damage. In this way, the choice of criteria for the MCMIS Crash file is astute, specifically because the criteria do <u>not</u> depend on how a state may define an injury severity level or train their officers to identify it.

If it is true that there is a fundamental relationship between fatal and nonfatal (injury/transported and towed/disabled) involvements, then the ratio can be discovered by examining data files with enough detail to identify each crash type. In this process, each state and year of data is one observation, one estimate of the underlying ratio. By assembling such observations and fitting a statistical model, it is possible to estimate the true ratio of crash severities that applies across states. This statistical model will allow states to estimate the number of cases that they should be reporting to the MCMIS Crash file, with some margin of error. It is assumed that each state, and FMCSA, will know the correct number of fatal involvements. The model will then predict the number of nonfatal (injury/transported and towed/disabled) involvements that the number of fatal involvements implies.

An earlier report presented a different method using data available at the time. [2] The approach was a simple linear regression model that fit the data well statistically but which produced prediction ranges that were large and did not predict well back to the original data. This report uses the statistical model developed in a second report, updated with observations from additional states.[1] The model here predicts *nonfatal* reportable involvements, the combination of injured/transported, and towed/disabled cases. Predicting a single outcome is more straightforward, produces better results, and meets the basic need to predict the number of cases a state should report to the MCMIS Crash file.

3 Data

We selected states for modeling the distribution of reportable cases that collected the detail necessary to reproduce the MCMIS Crash file reporting criteria. The essential criteria for reporting are displayed in Table 1. Adequate methods of identifying trucks and buses could be developed for all the states evaluated to date (October 2008), with some qualifications. In some states, light vehicles displaying hazardous materials placards were not identified with high confidence, but the number of such vehicles is so small relative to the number of trucks and buses that it should have only an insignificant impact on the analysis.

Table 1 Vehicle and Crash Severity Threshold for MCMIS Crash File

Vehicle	Truck with GVWR over 10,000 or GCWR over 10,000, or Bus with seating for at least nine, including the driver, or Vehicle displaying a hazardous materials placard.
Accident	Fatality, or Injury transported to a medical facility for immediate medical attention, or Vehicle towed due to disabling damage.

Identifying crashes that meet the reporting criteria is the crux of the problem in estimating reportable cases. Table 2 shows the states that were selected for this problem. In all these states, determining the number of reportable fatal involvements can be done fairly cleanly and with minimal ambiguity. Most of the states directly coded the detail needed to identify the different crash severities. An indicator that an injured person was transported for immediate medical attention was critical for selection. The experience of the UMTRI state evaluations showed that whether an injured person was transported for care does not map cleanly to coded injury severities, so injury severity could not be used as a surrogate. With respect to the towed/disabled criteria, the method some states used to code vehicle damage severity could be used as a substitute for a direct indicator that a vehicle was towed due to disabling damage. The severity scale employed by the states here directly indicate whether a vehicle was disabled. This is not a perfect substitute for towed/disabled, but it is a reasonable surrogate.

Table 2 States Selected for Modeling the Ratio of Crash Severities

State	Data Year	Injured	Transported	Towed	Disabled
Florida	2003	Yes	Yes	No*	Yes
Iowa	2004	Yes	Yes	No*	No*
Louisiana	2005	Yes	Yes	Yes	Yes
Missouri (1)	2001	Yes	Yes	Yes	Yes
Missouri (2)	2005	Yes	Yes	Yes	Yes
Ohio (1)	2000	Yes	Yes	Yes	Yes
Ohio (2)	2005	Yes	Yes	Yes	Yes
South Dakota	2005	Yes	Yes	Yes	Yes
Alabama	2006	Yes	Yes	Yes	Yes
Wisconsin	2006	Yes	Yes	Yes	Yes
South Carolina	2006	Yes	Yes	Yes	No*
Arizona	2006	Yes	Yes	Yes	Yes
Pennsylvania	2006	Yes	No	Yes	Yes

* Vehicle damage severity used as surrogate for towed due to damage

Table 3 Counts of Fatal and Nonfatal Reportable Involvements Used in Modeling Crash Severity Ratio

State	Crash severity		Total
	Fatal	Nonfatal	
S Dakota	19	434	453
Iowa	68	1,974	2,042
Louisiana	147	4,250	4,397
Missouri 1	155	6,002	6,157
Missouri 2	186	5,946	6,132
Ohio 1	205	8,840	9,045
Ohio 2	211	9,489	9,700
Florida	444	13,353	13,797
Alabama	128	4,383	4,511
Wisconsin	95	3,773	3,868
South Carolina	102	3,260	3,362
Arizona	128	4,283	4,411
Pennsylvania	217	7,381	7,598

Table 3 shows the data used in modeling the crash severity ratio. Each state and crash year is one observation. It is desirable that the data used in the model covers the range in the number of cases expected to be reported from the fifty states. There was a reasonable range of data available for the modeling effort, including a relatively small state like South Dakota with 19 annual fatal truck and bus crash involvements to Florida, which reported 444 fatal involvements for the data year used. Most of the states available, however, fell into the range between 150 and 200 annual fatalities. The very small states and the very large states were underrepresented in the data available.

UMTRI’s Trucks Involved in Fatal Accidents (TIFA) and Buses Involved in Fatal Accidents (BIFA) can be used to provide an accurate distribution of the annual expected reportable truck and bus fatal involvements for each state. The two files include all truck and bus fatal involvements that are reportable to the MCMIS Crash file, since the definitions of a reportable truck or bus are compatible with the MCMIS definitions. Only light vehicles transporting hazardous materials that require a placard are not included. Since there is only a small number of such vehicles each year so they would not affect the overall distribution.

Figure 1 shows the distribution of states by the annual average number of truck and bus fatal involvements. There are about eleven states each year with 25 or fewer MCMIS-reportable fatal involvements, six with between 26 and 50, and five with between 51 and 75. Almost 60 percent of the states have 100 or fewer fatal involvements annually. The circled numbers in the figure show the number of states used in the analysis within each interval. The available states cover the range, but the smaller states are underrepresented. Where there are eleven states with 25 or fewer truck or bus fatal crash involvements, there is only one observation in the data set for those states. There are 30 states with fewer than 100 involvements annually, but only three states in the data. However the range between 100 and 225, which includes 17 states, is covered well with nine observations.

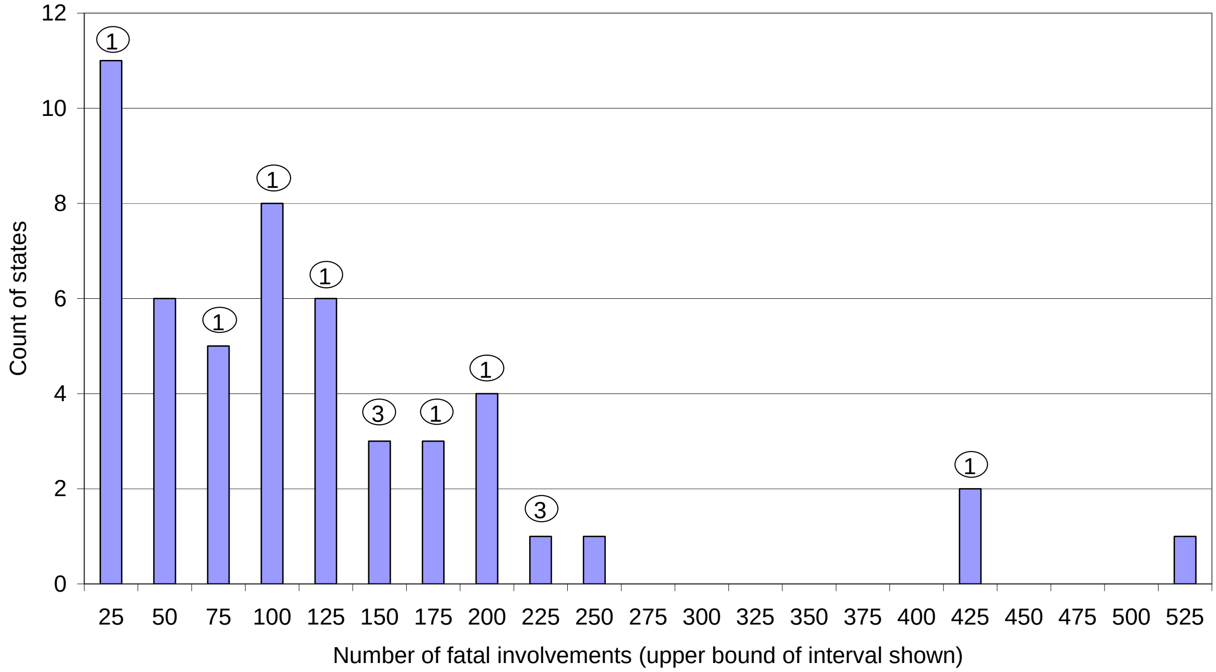

Figure 1 Counts of States by Average Annual Number of Fatal Truck and Bus Involvements TIFA 1999-2006, BIFA 1999-2006

Note also in Figure 1 that there are three states which annually average over 400 fatal involvements, and none between 250 and 400. The very large states are represented only by Florida.

4 Model and Methods

The goal is to predict the number of nonfatal crashes from the number of fatal crashes for states that only have data recorded for the number of fatal crashes. Ninety-percent prediction intervals for the estimates are also desired. To accomplish this goal, a weighted log-linear model is fit to MCMIS data for eleven states that have information recorded for both fatal and nonfatal crashes. (Thirteen observations are available for the eleven states, because two states were evaluated twice.) The model is then used to estimate the number of nonfatal crashes for a new state in which only the number of fatal crashes is known. Prediction intervals are presented, instead of confidence intervals, because data from the state to be predicted were not used to estimate the regression line in the modeling process. Since there is more uncertainty in predicting nonfatal crash involvements for a state not used in the modeling process, prediction intervals are wider than confidence intervals. Table 4 shows data for the thirteen observations that have numbers of crashes recorded for both fatal and nonfatal crashes. The logs of the count of fatals and nonfatals are also shown.

Table 4 Data Used in the Modeling Process

State	Fatal	Non Fatal	Log fatal	Log nonfatal
S Dakota	19	434	2.9444	6.0730
Iowa	68	1,974	4.2195	7.5878
Louisiana	147	4,250	4.9904	8.3547
Missouri 1	155	6,002	5.0434	8.6998
Missouri 2	186	5,946	5.2257	8.6905
Ohio 1	205	8,840	5.3230	9.0870
Ohio 2	211	9,489	5.3519	9.1579
Florida	444	13,353	6.0958	9.4995
Alabama	128	4,383	4.8520	8.3855
Wisconsin	95	3,773	4.5539	8.2356
South Carolina	102	3,260	4.6250	8.0895
Arizona	128	4,283	4.8520	8.3624
Pennsylvania	217	7,381	5.3799	8.9067

Two decisions were made with respect to the modeling procedure: the data were analyzed on the log scale, and weights were incorporated to reflect the idea that larger states should receive more weight than smaller states. The decision to analyze the data on the log scale is based on a scatterplot of the log of the nonfatal crashes by the log of the fatal crashes. The scatterplot shows a strong linear association between the logs of the two variables, with a correlation of about 0.97. Crash numbers are often in the thousands, and crash data are often analyzed using log-linear models. The decision to incorporate weights into the regression model is based on the realization that there is more uncertainty in the number of crashes when it is small, while the variance in the number of crashes when it is large is relatively smaller. States with larger numbers of crashes should receive more weight than states with fewer numbers of crashes, because their annual number of fatal crash involvements is more stable.

Figure 2 shows a scatter plot of the coefficient of variation against the mean number of annual fatal involvements for each state. The plot shows that states with smaller number of annual fatal involvements have more variability in the annual number than larger states. The coefficient of variation (CV) is calculated by dividing the standard deviation by the mean number of fatal involvements. States with high CV's have more variation in the count of fatal involvements than states with low CVs. In other words, where there are many fatal involvements annually, the number does not fluctuate from year to year as much relatively as states with few fatal involvements. Because of the variation associated with the size of the state, it is not desirable to treat each state as an observation with equal weight in fitting the model. On the other hand, most of the states have fewer than 100 fatal involvements annually, and using the count of fatals to weight the states would give the few large states excessive leverage in the model. According, we weighted the states by the square root of the number of fatal involvements. This choice gives more weight to the large states, which have less variability in the number of fatals, but still gives appropriate weight to the smaller states, which is where most of the data are.

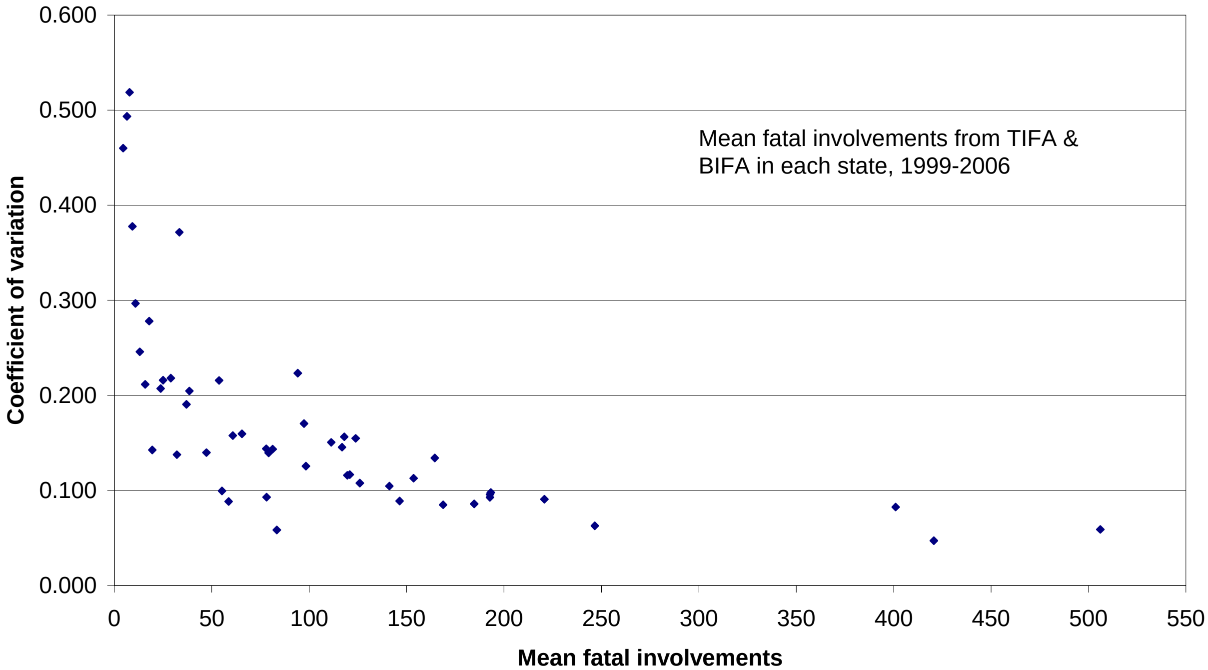

Figure 2 Variation in the Annual Number of Fatal Involvements Reportable From the States

The model for analyzing the data shown in Table 4 takes the form

$$\log y_i = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2 / \sqrt{x_i}), \qquad i = 1,\ldots,13$$

where y_i is the number of nonfatal crashes for state i, x_i is the number of fatal crashes, β_0 and β_1 are the intercept and slope parameters, respectively, and ε_i are the error terms. The random component of the model is contained in the error terms. These terms are modeled as Normal random variables with mean 0 and variance $\sigma^2 / \sqrt{x_i}$. In standard regression models the variance is constant at σ^2, but since this is a weighted regression, each case receives a different weight. In this regression model each state is being weighted by the square root of the number of fatals. This ensures that states with more fatals get more weight than states with fewer fatals. In other words, the variances of the error terms are smaller for states with more fatals. Weighting states by the square root was arrived at by trial and error. Preliminary results showed that weighting by the number of fatals was too severe and weighting on the log scale was too mild. Weighting by the square root appears to be a reasonable compromise.

After fitting this model, the estimated equation is

$$\log \hat{y}_i = 3.2143 + 1.0631 \log(x_i).$$

Note that the slope parameter, 1.0631, is very close to one. This means that for a unit increase in the log fatals, the estimated log nonfatals increase by a little bit more than one unit. Suppose now that it is desired to estimate the number of nonfatal crashes for a new state with 100 fatal crashes. Applying the fitted equation to the new state gives

$$3.2143 + 1.0631 \times \log(100) = 8.1098$$

and a 90 percent prediction interval is (7.600, 8.4596). Exponentiating these results back to the original scale gives an estimated exp(8.1098) = 3,327 nonfatal crashes and the 90 percent prediction interval is (2345, 4720).

An Excel spreadsheet is provided that allows the user to input the number of fatal crashes for a new state. The output produced consists of the estimated number of nonfatal crashes and the 90 percent prediction interval. Using the present model, if the number of fatals for a new state is 200, the estimated number of nonfatals is 6,951 and the 90 percent prediction interval is (4616, 10,469). Figure 3 below shows the scatterplot, the fitted regression line, and 90 percent prediction intervals based on the fit to 13 observations in eleven states.

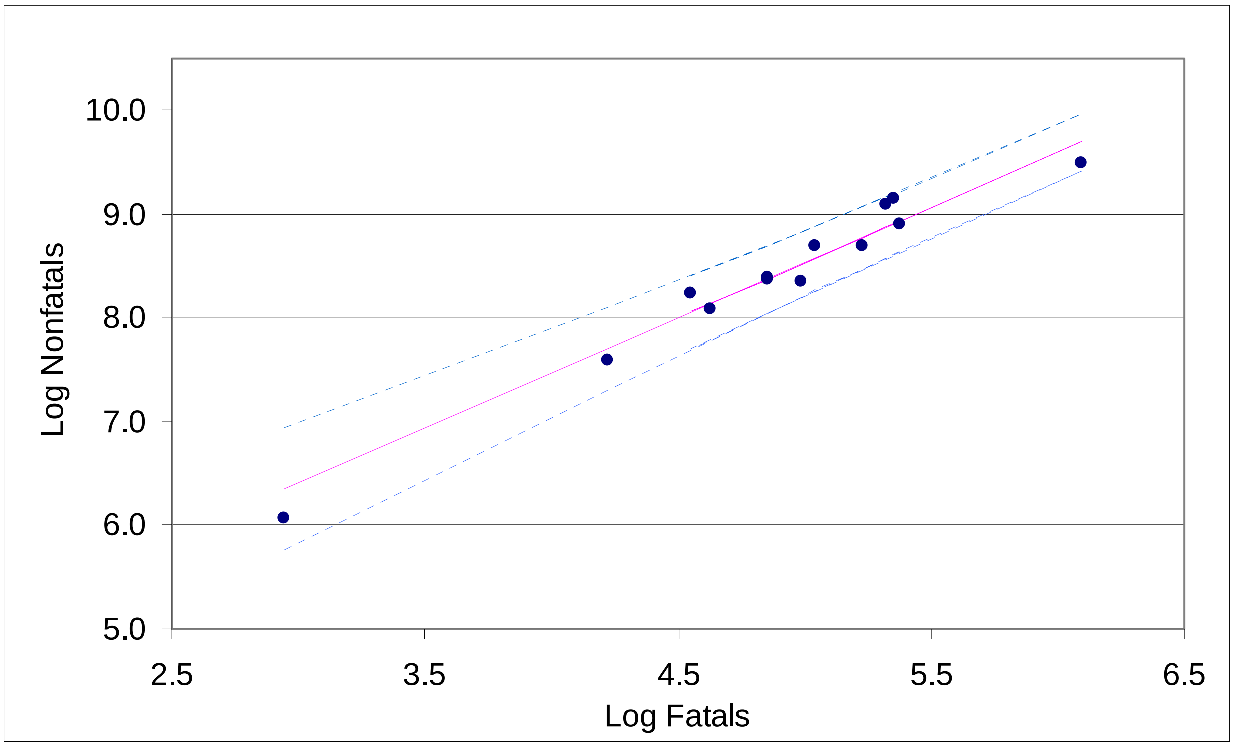

Figure 3 Scatterplot, Fitted Regression Line, and 90 Percent Prediction Intervals for Eight States

5 Application

The model can be applied to individual states to estimate reportable cases, although it is not valid to predict outside of the range of values covered in the model. However, the model is based on a range that covers most of the states: There are only eight states[3] that average fewer than 19 truck or bus fatal involvements annually, which is the smallest observation used in the model. Only one (California) that averages over 444 (the largest observation used in the model). Thus, the model covers 42 of the 51 (including the District of Columbia) units that supply data to the MCMIS Crash file.

Table 5 compares the predicted and actual values for nonfatal involvements for the states used in generating the model. Generally, the model estimates are reasonably close to the observed values. Almost all the predictions are within 20 percent of the actual number, and Alabama, South Carolina, Arizona, and Pennsylvania are within a few percent. Agreement is less good for South Dakota (differs by 31 percent) and Florida (22 percent) but all the predicted numbers are within the 90 percent prediction interval, as shown in Figure 3.

Table 5 Comparison of Actual and Nonfatal and 90 Percent Prediction Intervals

	Nonfatal involvements		90% Prediction Interval	
State	Actual	Predicted	Lower bound	Upper bound
S Dakota	434	569	315	1,029
Iowa	1,974	2,208	1,487	3,278
Louisiana	4,250	5,011	3,661	6,858
Missouri 1	6,002	5,301	3,889	7,226
Missouri 2	5,946	6,435	4,779	8,665
Ohio 1	8,840	7,136	5,328	9,557
Ohio 2	9,489	7,358	5,502	9,840
Florida	13,353	16,228	12,254	21,489
Alabama	4,383	4,325	3,123	5,990
Wisconsin	3,773	3,150	2,208	4,494
South Carolina	3,260	3,398	2,400	4,810
Arizona	4,283	4,325	3,123	5,990
Pennsylvania	7,381	7,581	5,676	10,125

Figure 4 displays the data in Table 5 graphically. Note that all the observed values for nonfatal involvements are within the 90 percent prediction intervals. The range is relatively large for Florida. Both predictions for Ohio fall substantially below the actual number of nonfatal involvements, but within the range predicted by the model. Only the predictions for Florida and the two for Ohio differ substantially from the observed values.

---

[3] The District of Columbia, Rhode Island, Alaska, Hawaii, Vermont, New Hampshire, North Dakota, and Delaware.

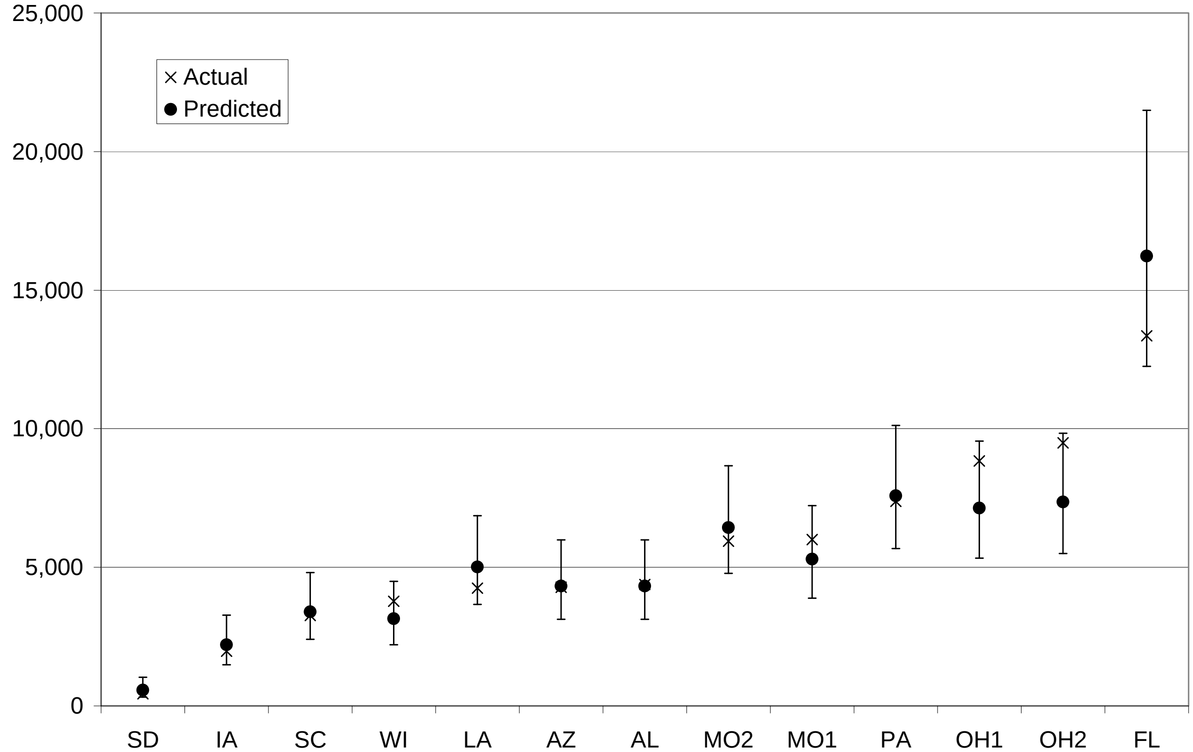

Figure 4 Actual, Predicted and 90 Percent Prediction Interval for Case States

6 Comparison with earlier models

The model presented here is improved from the models in the earlier attempts to use the number of fatal involvements to predict the number of nonfatal involvements reportable to the MCMIS Crash file. There is clearly a relationship between the number of fatal involvements and nonfatal involvements. The first effort split the set of nonfatal involvements into injury/transported and towed/disabled components, and tried to work out the relationship between the counts of fatal involvements and the counts of injury/transported and towed disabled crash involvements. The work showed that the relationship is linear and the association between the counts of fatal and injury/transported or fatal and tow/disabled involvements rather good, with R^2 statistics of 0.87 and 0.85 respectively. But the attempt to create statistical models was not entirely satisfactory because they did not predict back to the original data well, and the confidence intervals were too wide to provide useful guidance to the states.[2]

A revised model simplified the approach by modeling only the nonfatal involvements and using the log transform improved the results.[1] We also provided prediction intervals, rather than confidence intervals. In estimating the regression line for predicting the number of nonfatal crashes from the number of fatal crashes, eight states were used. These eight states represent a sample from a larger population of states. The regression model had an R^2 statistic of 0.94, and all the observations fell within the 90% prediction interval. This model was a significant improvement over the previous approach.

The present effort extends this model with additional observations, adding data from five states. These five additional observations are from states evaluated since the earlier work. In each case, the state data provided apparently reliable information about the number of fatal and nonfatal reportable involvements, and so qualified for inclusion in the model.

Re-estimating the model with these new observations resulted in a new model that was very close to the previous model. The parameters for the new model were quite close to those for the old one. The term for the intercept changed from 3.0983 to 3.2143, and the parameter for the slope changed from 1.0835 to 1.0631. R^2, a measure of fit, was identical at 0.94. Table 6 shows the comparison between the previous and current models of model parameters and measures of fit, and the results for predicting the number of nonfatal involvements for a new state with 100 fatal involvements. The predicted number of nonfatal involvements is quite similar, but there is a substantial reduction in the range of the prediction interval because of the additional observations used in the model.

Table 6 Comparison of Previous Model with the New Model

	Old model	New model
Model parameters		
Slope	3.0983	3.2143
Intercept	1.0835	1.0631
R^2	0.94	0.94
Prediction for new state with 100 fatal involvements		
Nonfatal involvements	3,254	3,327
90% prediction interval, lower bound	1,972	2,345
90% prediction interval, upper bound	5,371	4,720
Range of prediction interval	3,399	2,375

7 Discussion

Since there is uncertainty in a sample, a confidence interval is appropriate for the estimate of the number of nonfatal crashes for a state, given the number of fatal crashes. Often, 90 or 95 percent confidence intervals are calculated for each state. A 90 percent confidence interval is calculated by a procedure, such that if this procedure were repeated over and over again, 90 percent of the confidence intervals would *trap* the true number of nonfatal crashes in the population. Thus, we are 90 percent confident that our estimate traps the true number of nonfatal crashes in the population for a particular state. By collecting a sample and calculating our estimates, we only perform this procedure once. Confidence intervals apply to states that were used in the estimation process.

Prediction intervals, on the other hand, apply to out-of-sample states. In other words, prediction intervals are used for new states that were not used to estimate the regression line. Intuitively, prediction intervals are wider than confidence intervals. Thus, the prediction problem begins by first fitting a model to a sample of states. Once the model is fitted, an estimate of the number of nonfatal crashes can be predicted for a new state not used in the estimation process. Since the new state is out-of-sample, a prediction interval should be reported. The interpretation is similar to that of a confidence interval. We are 90 percent certain that the interval traps the population value for the new state.

The prediction intervals provide reasonable guidance to the states in terms of expected number of nonfatal involvements. The range of the intervals is about ±30 percent. This may be regarded as relatively wide, but it should be recognized that there is a lot of variability in the underlying data. There are several sources of this variability. One source is in the number of fatal involvements. Figure 2 showed how the amount of variability from year to year differs with the mean number of fatal involvements, such that the variability is much higher in states with few fatals. This is because whether a person is killed in a particular crash is highly random. When there are many fatal crashes in a state, the randomness tends to wash out, but when there are only a few, the randomness can have a substantial effect on the absolute number of fatals.

There is also no doubt significant measurement error in the counts of nonfatal involvements determined in the state data. Even though the states used in the statistical model coded all the information needed to identify crash involvements that meet the MCMIS Crash file reporting criteria, it is important to remember that the source of the data is ultimately an individual police officer completing a crash report. The UMTRI evaluations have shown that the accuracy of reported cases vary widely. The reporting officers often work in difficult conditions; protecting life and property, rather than accurate crash data, is their primary mission; and quality control is difficult and expensive. All these factors contribute to variability in the underlying data.

Adding further states to the model may improve the estimates and narrow the prediction intervals. Of particular interest would be to add states in areas not well-covered by the set of states available for the model at this time. States with fewer than 19 fatals, between 70 and 150, and more than 200 would help fill gaps in the range of states covered by the model. However, the prediction intervals available in the current model should provide meaningful guidance to the states.

8 References

1 Green, P., and Blower, D., Revised Ratio of Crash Severities Reportable to the MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. February 2007. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

2 Green, P., and Blower, D., Preliminary Ratio of Crash Severities Reportable to the MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. January 2007. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

3 Blower, D., and Matteson, A., Evaluation of Florida Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. December 2004. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

4 Blower, D., and Matteson, A., Evaluation of Iowa Crash Data Reported to the MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. August 2006. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

5 Blower, D., and Matteson, A., Evaluation of 2005 Louisiana Crash Data Reported to the MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. December 2006. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

6 Blower, D., and Matteson, A., Evaluation of Missouri Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. January 2004. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

7 Blower, D., and Matteson, A., Evaluation of 2005 Missouri Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. September 2006. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

8 Blower, D., and Matteson, A., Patterns of MCMIS Crash File Underreporting in Ohio. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. August 2003. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

9 Green, P.E., and Matteson, A., Evaluation of 2005 Ohio Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. November 2006. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

10 Green, P.E., and Matteson, A., Evaluation of 2006 Wisconsin Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann

Arbor, Michigan. March 2008. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

11 Green, P.E., and Matteson, A., Evaluation of 2005 Alabama Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. October 2007. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

12 Blower, D., and Matteson, A., Evaluation of 2005 Pennsylvania Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. September 2007. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

13 Green, P.E., and Matteson, A., Evaluation of 2005 Arizona Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. June 2007. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.

14 Green, P.E., and Matteson, A., Evaluation of 2006 South Carolina Crash Data Reported to MCMIS Crash File. University of Michigan Transportation Research Institute, Ann Arbor, Michigan. July 2008. Sponsor: Federal Motor Carrier Safety Administration, U.S. D.O.T.