Statistical Road Safety Modeling

Ezra Hauer

The hope is that statistical models fitted to historical data can be used to estimate the effect of road design elements on safety. Whether this can be done is not clear. A sign of trouble is that models based on diverse data sets tend not to yield similar results. Suggestions are made on how to increase the chance of success in this quest. Emphasis is on three questions: Which variables should serve in the model? What mathematical function should represent their influence? How does one check whether the representation of the influence of a variable is appropriate?

Statistical road safety modeling (SRSM) is the fitting of a statistical model to data. The data are about past accidents and traits for a set of road segments, intersections, or other infrastructure elements. The result of SRSM is an equation with the estimate of expected accident frequency on the left and a function of traits on the right. There are two uses for SRSM: to estimate the expected accident frequency of an infrastructure element based on its traits and to estimate the change in expected accident frequency caused by a change in a trait of an infrastructure element.

The computations associated with the two purposes are deceptively similar. For Purpose A, one plugs into the equation a set of trait values to obtain an estimate of expected accident frequency. For Purpose B, one does so twice, changing only the value of the trait whose effect on accident frequency is sought. This similarity in computation conceals a difference of essence; whereas Purpose A is largely unproblematic, Purpose B is fraught with difficulties.

To illustrate the difficulty, suppose that, holding all other values constant, a statistical model estimates X accidents a year with 10-ft lanes and Y accidents a year with 11-ft lanes. In the equation, the change of lane width has "caused" the change from X to Y. Does this mean that if two roads were built to be identical except that one has 10-ft lanes and the other has 11-ft lanes, one should expect the ratio of their accident frequencies to be Y/X? The correct answer is no, or, at most, "We do not know." In the data from which the model equation was built, roads with 10-ft and 11-ft lanes may be of differing vintage, may be located in diverse jurisdictions, and may differ in many traits that are either imperfectly represented in the model or absent from it altogether. Thus, the difference between X and Y reflects in part the complex influence of all the missing and imperfectly accounted-for causal factors and only in part the causal influence of lane width.

The essence of the problem is that lane width was not changed but that, in the data, some lanes were found to be 10 ft wide and others 11 ft wide. Finding is not the same as experimenting. Thus, one can measure the length of a metal rod, heat it, measure again, and, on this basis, conclude that the addition of heat caused the change in length of the rod. One may not conclude the same by measuring the length of two metal rods found to differ in temperature. Holland concluded, "No causation without manipulation" (*1*). Box et al. called data from historical records happenstance data and devoted 10 pages to the many obstacles that thwart attempts to interpret them by multivariate regression (*2*).

Still, the influence of causal factors that affect the probability of accident occurrence and their severity is surely reflected in data. The question is whether the tools of analysis can be honed sufficiently to reveal these cause–effect relationships. The importance of such tool honing derives from that fact that SRSM is the only practical way by which one can study the safety effect of many road design elements. Accordingly, the goal for this paper is to suggest a few ways with which to improve the SRSM process. When models produced by different researchers and based on diverse data sets produce similar results, this will be a sign that modeling is on the right track. Consistency of results is a necessary condition for making inferences about cause and effect. Until such consistency emerges, results of SRSM can be used for Purpose A but not Purpose B.

There are many books about multivariate statistical modeling, and sophisticated statistical software packages for this purpose are in common use. In this paper, emphasis is on some aspects of modeling that are perhaps less well covered in books—the questions of choosing, improving, and checking the functional form of the model.

The sought-after cause–effect relationships, although present in the data, are hidden by the randomness in accident counts, obscured by imprecise trait data, and covered by layers of interdependencies and missing information. This is why the process of model development is akin to the work of a short-sighted detective. Just like detective work, modeling is often tedious but cannot be routine; it requires intimate familiarity with the data, frequent exploratory sorties, much backtracking, and repeated model modifications. This task and the process are well served by an ordinary spreadsheet. The use of canned software, if it induces automation and routine, may not be an advantage. The spreadsheet keeps the data visible, makes exploration and visualization easy, facilitates the optimization needed for parameter estimation, and does not constrain the choice of functional form to be a generalized linear model.

MODEL EQUATION

The central element of SRSM is the model equation used to predict the number of accidents of some kind that may be expected to occur on an entity per unit of time, as a function of its traits (traffic, geometry, and environment). One should not be overly impressed by this mathematical expression. There is no theory to indicate how, for example, accident frequency should increase as traffic increases, or how it should be related to, say, the radius of horizontal curve. The SRSM process is basically that of curve fitting, in which the modeler chooses the function that is going to be fitted to the data. To make this choice, the modeler has no guidance from theory, not even from dimensional analysis. Nor is there much guidance in the data; a large number of mathematical functions could be chosen to fit the same

Apartment 1706, 35 Merton Street, Toronto, Ontario M4S 3G4, Canada.

Transportation Research Record: Journal of the Transportation Research Board, No. 1897, TRB, National Research Council, Washington, D.C., 2004, pp. 81–87.

data. To discriminate among those, one uses the notions of goodness of fit and simplicity. That is, of all the functions that can fit the data adequately, one usually chooses the function with the smallest number of parameters. However, there is no reason to think that underlying phenomenon indeed follows the chosen simple function, or that it follows any simple mathematical function. Thus, although the model equation is the centerpiece and final product of the SRSM process, it is basically chosen by the modeler on rather shaky grounds.

The generic model equation is

$$Y = f(X_1, X_2, X_3, \ldots, X_n, \beta_1, \beta_2, \beta_3, \ldots, \beta_n) \quad (1)$$

where Y denotes the expected number of accidents per unit of time for entities with trait values $X_1, \ldots, X_n$ when the parameters are $\beta_1, \ldots, \beta_n$ and $f()$ denotes some function. Attention is usually on the task of estimating the β. However, the β depend entirely on what was chosen to serve as $f()$ and on which set of traits was chosen to serve as variables.

To illustrate, consider a researcher who seeks to determine how dimension Y depends on dimensions X_1 and X_2. The data consist of precise dimension measurements of 100 entities where 2 m $< X_1 <$ 10 m and 4.5 m $< X_2 <$ 5.5 m. When Y was plotted against X_1, Figure 1 was obtained.

By looking at the plot, the researcher decided that the linear model $Y = \beta_1 + \beta_2 X_1$ is a good choice. By using ordinary least squares, the estimate of β_1 was 4.97 and the estimate of β_2 was 0.75. The researcher did not know that X_1, X_2, and Y are the dimensions of various right-angle triangles with Y the hypotenuse. Had it been known, the model $Y = (X_1^{\beta_1} + X_2^{\beta_2})^{\beta_3}$ would have been used (and the estimates 2, 2, and 0.5 obtained for β_1, β_2, and β_3).

The moral of this story is manifold. First, as noted earlier, what is estimated to be the parameters (the β) depends entirely on what is chosen to be the model equation. The parameters β_1 and β_2 of the linear model (4.97 and 0.75) have nothing to do with the parameters β_1 and β_2 of the Pythagorean model (2 and 2). Whatever truth resides in a parameter is tied to the model equation and cannot be judged separately from it. Since the meaning and magnitude of the β depend on $f()$, not the other way around, the task of choosing and shaping the function $f()$ is of primary importance in modeling. The second lesson is that the true functional form of the model equation is nearly unfathomable. It is very unlikely that despite its relative simplicity, the use of the Pythagorean functional form would have occurred to most modelers (had they not already known of its applicability to the data, in which case no statistical modeling would have been necessary). Third, in this example, the linear form is clearly simpler

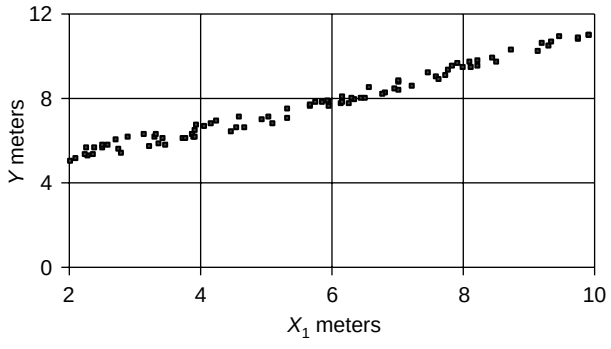

FIGURE 1 Measured values of Y when 2 m $< X_1 <$ 10 m and 4.5 m $< X_2 <$ 5.5 m.

that the Pythagorean form. Thus, the guidance of Occam's razor—prefer models with fewer parameters if they fit well—is not always good advice. In this example, and perhaps in most representations of the real world by a model, systematic preference for parsimony may be a hindrance in the quest for cause and effect.

Not all functions $f()$ can serve equally well. The reader will recognize such often used model equations as

$$(a) \quad Y = (\text{segment length}) \times (\beta_1 X_1 + \beta_2 X_2 + \cdots)$$

$$(b) \quad Y = (\text{segment length}) \times (\beta_0 X_1^{\beta_1} X_2^{\beta_2} \ldots) \quad (2)$$

$$(c) \quad Y = (\text{segment length}) \times e^{\beta_0} e^{\beta_1 X_1} e^{\beta_2 X_2} \ldots$$

In all these models, segment length is a logical variable. It captures the logical requirement that if N accidents are expected to occur on 1 mi of road, $2N$ accidents should be expected on an identical road that is 2 mi long.

Model *a* is additive. It corresponds to situations in which the presence of a trait adds a certain amount to Y/(unit length). This representation is natural for point hazards, such as a driveway or a narrow bridge, traits that influence accident occurrence on a small portion of a highway segment. The additive model is less well suited for representing the influence of such traits such as lane width or shoulder type that affect the probability of accident occurrence along significant portions of a highway segment. To explain, let X_1 in Model *a* represent traffic flow and let X_2 represent lane width. In Model *a*, the influence lane width is to add a certain number of accidents/unit length, whether average annual daily traffic (AADT) is 1,000 or 10,000. This is surely not true. A reflection of the same blemish is that the model will predict some accidents even if there is no traffic. One can obviate this problem by making each term in Model *a* also a function of traffic flow. Thus, instead of $\beta_2 \times$ (lane width), one could use, for example, $\beta_2 \times$ traffic flow $\times$ lane width. Similar expressions would have to represent the additive contributions of shoulder type, grade, and so forth. However, if traffic flow (or a function of it) is to multiply many summands in Model *a*, it is best to make it into a common multiplier. A similar argument could now be used for, say, shoulder type. That is, it could be argued that the effect of shoulder type should not be an addition that does not depend on the number of accidents caused by a certain traffic flow and lane width. A consistent argument along these lines leads to the use of a multiplicative model (such as *b* or *c*) for such variables as lane width and shoulder type. It appears therefore that the effect of variables that influence the probability of accident occurrence along significant portions of a highway segment is better represented by multiplicative than by additive factors.

An analogous line of reasoning leads to the conclusion that multiplicative models (such as *b* and *c*) are not well suited for the representation of point hazards. To understand why, suppose that ℓ is segment length, A is the average number of accidents on a segment per unit length if the segment has no driveways, and B is the addition to accidents caused by one driveway. With n driveways, one may expect on this segment $\ell A + nB$ accidents. If one wishes to have a multiplicative model, the additive form can be rewritten as the product $\ell A[1 + (n/\ell)(B/A)]$. In this multiplicative expression, n/ℓ is driveway density and B/A is the number of accidents added by one driveway divided by the average number of nondriveway accidents per unit length of segment. Thus, what was done by nB in the additive model requires the more complex factor $[1 + (n/\ell)(B/A)]$ in the multiplicative model. The driveway factor requires (in addition to n and B) the presence A, which represents various conditions along the segment that appear to be extraneous to the representation of the

influence of driveways. Chances are that modeling of driveways will be more transparent with the simple additive rather than with the more complex multiplicative form.

Another consideration relates to the building blocks of which the models are commonly constituted. In Model *a*, the building block is of the form βX. Although there is no reason why $\beta g(X)$ could not be used [where $g()$ is some function], use of βX is common. Most phenomena related to safety are nonlinear. Thus, for example, the probability of a single-vehicle accident diminishes as traffic flow increases. The linear building block βX cannot easily capture this nonlinear reality (except by using a piecewise linear function with its many parameters). Similarly, in the multiplicative models, it is common to use factors such as X^{β} and $e^{\beta X}$. These, too, can be imperfect. First, as shown in Figures 2 and 3, they cannot represent a relationship that has a peak, a valley, or a point of inflection. Thus, for example, if it were true that accident frequency diminishes when lane width increases from 10 ft to 11 ft but then increases when lane width increases from 12 ft to 13 ft, this could not be found in the modeling process if such a building block were chosen for the model equation.

These considerations lead to the suggestion of a model form that is less subject to obvious shortcomings. First, the model equation should have both multiplicative and additive components. The role of the multiplicative component is to represent the influence of factors that naturally apply to a stretch of road (such as lane width or shoulder type), whereas the role of the additive component is to account for the influence of point hazards (such as driveways or narrow bridges). Second, the building blocks of the model equation should not be restricted to the commonly used expressions βX, X^{β}, and $e^{\beta X}$. In the absence of other guidance, the choice of the function for the building blocks should be such that it can replicate the form suggested by the data: peaks, valleys, and points of inflection. Thus, the generic form of the model equation for road segments could be

$$
\begin{aligned}
Y &= (\text{scale parameter}) \times [(\text{segment length for prediction}) \\
&\quad \times (\text{multiplicative portion}) + (\text{additive portion})] \\
\text{multiplicative portion} &= f(\text{traffic flow}) \\
&\quad \times g(\text{shoulder type}) \times \cdots \qquad (3) \\
\text{additive portion} &= h(\text{traffic flow, number of driveways}) \\
&\quad + i(\text{traffic flow, number} \\
&\quad \text{of short bridges}) + \cdots
\end{aligned}
$$

In Equation 3, Y is the accident frequency expected to occur on a road segment per year (or any unit of time) with the specified values for traffic flow, lane width, and so forth. Scale parameter may change from year to year, reflecting changes in weather, demography, accident reporting, and so forth—traits not accounted for in the model. Segment length for prediction is the aforementioned logical variable that is set to 1 for intersections at grade crossings, and $f()$, $g()$, $h()$, $i()$ denote building block functions. In principle, these functions could be of several variables. However, in practice, it is seldom possible to gain cues about multivariable functions from the data.

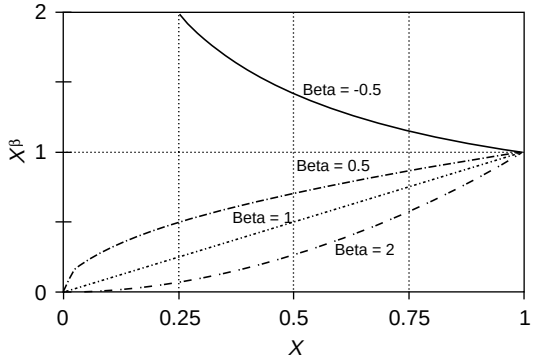

FIGURE 2 Possible shape of X^{β}.

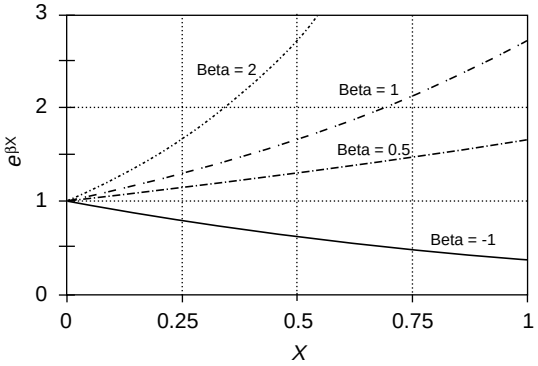

FIGURE 3 Possible shape of $e^{\beta X}$.

Choosing Variables and Building Blocks

The building up of a model equation is best done by adding one building block after another and reestimating all parameters after each addition. The functional form of each building block is chosen so that it can replicate the peaks, valleys, and points of inflection indicated by the data. This can be done as follows.

Suppose that some variables were already introduced into the model and the corresponding parameters estimated; now, the introduction of variable V is contemplated. Should V be introduced into the model, and, if yes, what function represents its influence?

Begin by grouping the data entities into bins such that each bin contains entities with similar values of $V = V_1, V_2, \ldots$. Thus, for example, if V stands for lane width, there is a bin for segments with 10-ft lanes, another for segments with 11-ft lanes, and so forth; if V stands for percent of trucks, there may be bins for $V_1 = 0\%–2\%$, $V_2 = 3\%–4\%$, and so forth. For each entity in a bin, one has the recorded number of accidents and the number predicted by the current model that does not yet include the variable V. Summing over all entities in each bin, compute the sum of recorded accidents $N_{\text{recorded}}(V_i)$ and the sum of accidents predicted by the current model, $N_{\text{predicted}}(V_i)$. It will be shown that by examining the relationship between $N_{\text{recorded}}(V_i)$ and $N_{\text{predicted}}(V_i)$ for each bin, one can decide whether variable V should be included in the model and, if yes, what functional form can represent it. Because variables can be introduced into either the multiplicative or the additive part of the model, two cases need to be examined.

Case A

In Case A, the variable is to be represented in the multiplicative part of the model. For all bins $i = 1, 2, \ldots$, compute

$$
R(V_i) = \frac{N_{\text{recorded}}(V_i)}{N_{\text{predicted}}(V_i)} \qquad \text{and} \qquad \hat{\sigma}[R(V_i)] \cong \frac{\sqrt{N_{\text{recorded}}(V_i)}}{N_{\text{predicted}}(V_i)} \qquad (4)
$$

TABLE 1 *R* and Standard Error for Lane Width

1	2	3	4	5	6	7
V = Lane Width	Predicted	Recorded	*R*(*V*) = Recorded/Predicted	σ	*R*+σ	*R*–σ
10'	160.9	163	1.01	0.08	1.09	0.93
11'	719.4	698	0.97	0.04	1.01	0.93
12'	1251.2	1278	1.02	0.03	1.05	0.99
13'	308.7	307	0.99	0.06	1.05	0.94
14'	90.9	82	0.90	0.10	1.00	0.80
15'	3.0	6	2.02	0.83	2.85	1.20

Let *V* be lane width. The numbers in Table 1 are based on data for on-the-road injury accidents on undivided urban four lane roads (*3*). Thus, for example, on road segments with 10-ft lanes, 163 accidents were recorded, and the model, which at that point in the SRSM process already accounted for several variables, predicted for the same segments 160.9 accidents.

The first question can now be answered. A new variable is considered for introduction into the model equation when the estimates of *R*(*V*) (Column 4) form an orderly relationship with *V* (Column 1). Only then can the effect of *V* be captured by a function and made into a new building block. In Table 1 there is no evidence of such an orderly relationship. Therefore, in this case, lane width was not introduced into the model in the work by Hauer et al. (*3*).

A different conclusion was reached in the same project when the *R* for percent trucks were examined (Figure 4). For property-damage-only accidents, *R* are shown by full squares and their ±σ limits by full arrowheads; for injury accidents, the *R* and their ±σ limits are shown by empty symbols.

The progression of the *R*-ratios appears to be orderly, even if unexpected. To represent this asymmetric relation the functional form $\beta_1 \exp[\beta_2(\text{Percent Trucks}) + \beta_3(\text{Percent Trucks})]$ was selected (Figure 5).

Case B

In Case B, the variable is to be represented in the additive part of the model. For all bins $i = 1, 2, \ldots,$ compute

$$R(V_i) = N_{\text{recorded}}(V_i) - N_{\text{predicted}}(V_i) \quad \text{and}$$

$$\hat{\sigma}[R(V_i)] \cong \sqrt{N_{\text{recorded}}(V_i)} \tag{5}$$

When $R(V_i) > 0$, too few accidents are predicted for entities in bin *i;* when $R(V_i) < 0$, too many accidents are predicted. To render them equal, $R(V_i)$ needs to be added to the additive part of the model equation for $i = 1, 2, \ldots$. From here on, the considerations and procedures of Case A apply.

In summary, the twin questions A and B can be jointly answered by the examination of the *R*. Variable *V* is considered for inclusion in the model if its *R* form an orderly relationship with *V;* a function fitting *R* is then the building block for *V*. When those data are rich enough, the same approach could be used to estimate *R* that are functions of two or more variables.

Since variables are introduced into the model sequentially, the question is whether the order in which variables are considered can affect the variable inclusion decision and the choice of functional form. The answer is yes. The variables used in SRSM are usually correlated. One path of correlation is through traffic flow. Roads that are heavily traveled are built and maintained to different standards than lightly traveled roads. Therefore, one might expect that, for example, the values of *R*(Percent Trucks) might depend on whether the predicted values were produced by a model with only traffic flow in it or, say, with traffic flow and speed limit and lane width in the model. It follows that the decision on variable inclusion and functional form may need to be occasionally revisited. This is one of the reasons for which the process of model building is iterative, requires rechecking, backtracking, and revisions. It seems best to introduce the dominant variable traffic flow first. Following that, variables should be considered in the order in which they contribute to the increase in log-likelihood/parameter.

Variables Whose Value Changes Within a Segment

The data for SRSM are made up of records. When a data record pertains to a road segment, the goal is to make up segments that are homogeneous in variables such as AADT and lane width. Homogeneity is only partly achievable for two reasons. First, segments should not be shorter than the precision with which accident locations are recorded. Thus, for example, if accidents are located to the

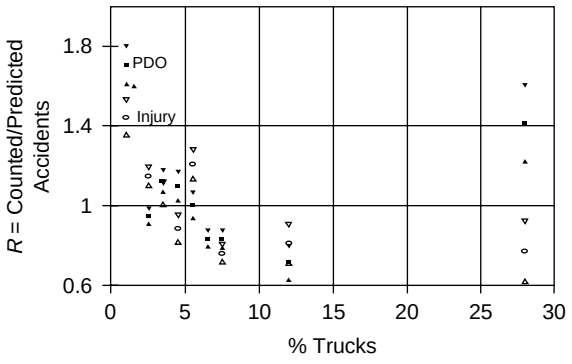

FIGURE 4 *R*-ratios versus variable percent trucks.

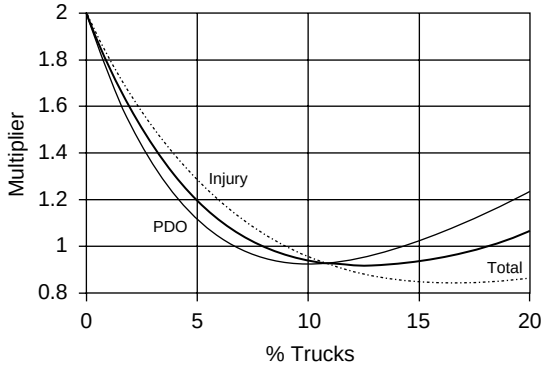

FIGURE 5 Building block for variable percent trucks.

nearest 0.1 mi, segments should not be shorter than that. However, even such short segments often are not homogeneous. Second, some variable values (e.g., grade or sight distance) may change continuously within a segment, no matter how short it is. For these reasons, it is important to write the model equation so that it can correctly accommodate variables whose value changes within a segment.

To show how this can be done, consider a segment of length L that is made up of two subsegments, one of length L_1 on which the grade is G_1, the other of length $L_2 = L - L_1$ on which the grade is G_2. On subsegments 1 and 2, one expects Y_1 and Y_2 accidents, whereas on the entire segment, by Equation 3, one expects

$$\begin{aligned} Y = Y_1 + Y_2 &= (\text{scale parameter}) \times [L_1 \times f(\text{traffic flow}) \\ &\quad \times g(\text{shoulder type}) \times \cdots \times p(\text{grade}_1) + L_2 \\ &\quad \times f(\text{traffic flow}) \times g(\text{shoulder type}) \\ &\quad \times \cdots \times p(\text{grade}_2) + (\text{additive portion})] \\ &= (\text{scale parameter}) \times \{L \times f(\text{traffic flow}) \\ &\quad \times g(\text{shoulder type}) \\ &\quad \times \cdots \times \left[\frac{L_1}{L}\, p(\text{grade}_1) + \frac{L_2}{L}\, p(\text{grade}_2)\right] \\ &\quad + (\text{additive portion})\} \end{aligned} \tag{6}$$

Thus, the function $p(\text{grade})$ takes the form $(L_1/L)p(\text{grade}_1) + (L_2/L)p(\text{grade}_2)$. This line of reasoning generalizes to the case when a segment is made of subsegments $i = 1, 2, \ldots, n$ of length L_i, on which variable V takes on values V_i. Let $p(V)$ denote the joint multiplicative contribution of $V_1, V_2, \ldots, V_i, \ldots, V_n$. Following the earlier reasoning,

$$P(V) = \sum_1^n \frac{L_i}{L}\, p(V_i) \tag{7}$$

This approach was introduced into SRSM by Miaou et al. (*4*).

NEGATIVE MULTINOMIAL LIKELIHOOD FUNCTION

The development of the model equation has been described as a process of introducing variables one after another such that all model parameters (β in Equation 1) are reestimated after a variable is added. Parameter estimation is always based on optimization of some kind. One may choose those parameter values that make the sum of weighed squared residuals smallest; alternatively, one may choose those parameter values that make the probability of the observed accident counts largest. The focus here is on the latter option, the so-called maximum likelihood estimation.

To maximize likelihood, one must have an equation that expresses the probability of the observed accident counts as a function of the known entity traits and the unknown parameters that need to be estimated. To write such an expression requires an assumption about the nature of the random process by which accident counts arise. One can assume, for example, that the count of accidents is Poisson distributed with a mean = Y (Equation 1). This assumption implies that entities with the same measured traits have the same mean. Experience with data shows that this is usually not a good assumption. Alternatively, one may assume that the count of accidents on the entities in the data is Poisson distributed with a mean = $Y\theta$, where Y is determined by Equation 1, and θ for each entity is drawn at random from a (Gamma) distribution. The multiplier θ represents the extent to which the mean of an entity in the data differs from the mean of an average entity with the same measured traits. With these assumptions, the count of accidents obeys the negative binomial distribution. This is adequate when the data are for a period that is short enough that the traits, including traffic flow, can be assumed to remain approximately the same. When θ is a multiplier that characterizes a specific entity (as under the negative binomial assumption), and some of the measured traits (e.g., traffic flow or speed limit) change over time, the negative binomial distribution generalizes into the negative multinomial distribution. The following derivation of the negative multinomial likelihood function is based on the work of Guo (*5*). It is given here partly for its adaptation to SRSM and partly to serve those who may wish to use the result for modeling on a spreadsheet. The main results are Equations 14 and 15.

Let μ_{ij} denote the mean accident frequency for entity i in period j such that

$$\mu_{ij} = Y_{ij}\theta i \tag{8}$$

Y_{ij} is the average of the μ_{ij} (in period j) for an imagined population of entities all with the same measured traits as entity i but differing in many other traits. It is assumed that the distribution of θ_i in that imagined population is gamma with the mean = 1 and variance = $1/\varphi$, the probability density function of which is

$$f(\theta_i) = \frac{\theta_i^{\varphi-1} e^{-\varphi\theta_i} \varphi^{\varphi}}{\Gamma(\varphi)} \tag{9}$$

For entity i, accident counts $a_{i1}, a_{i2}, \ldots, a_{ij}, \ldots, a_{in_i}$ were recorded in periods $1, 2, \ldots, j, \ldots, n_i$. These are assumed to be Poisson distributed with means $Y_{ij}\theta_i$. If so, for θ_i given, and independent accident counts,

$$P(a_{i1}, a_{i2}, \ldots, a_{in_i} | \theta_i) = \prod_{j=1}^{n_i} \frac{(Y_{ij}\theta_i)^{a_{ij}} e^{-Y_{ij}\theta_i}}{a_{ij}!} \tag{10}$$

However, when θ_i is not given, only its distribution is known (Equation 9)

$$P(a_{i1}, a_{i2}, \ldots, a_{in_i}) = \int_0^{\infty} \prod_{j=1}^{n_i} \frac{(Y_{ij}\theta_i)^{a_{ij}} e^{-Y_{ij}\theta_i}}{a_{ij}!} \frac{\theta_i^{\varphi-1} e^{-\varphi\theta_i} \varphi^{\varphi}}{\Gamma(\varphi)} d\theta_i \tag{11}$$

For notational brevity use

$$a_i = \sum_{j=1}^{n_i} a_{ij} \qquad Y_i = \sum_{j=1}^{n_i} Y_{ij} \qquad \prod_{j=1}^{n_j} \theta_i^{a_{ij}} = \theta_i^{\sum_{j=1}^{n_i} a_{ij}} = \theta_i^{a_i} \tag{12}$$

With this notation, Equation 11 yields

$$\begin{aligned} P(a_{i1}, a_{i2}, \ldots, a_{in_i}) &= \int_0^{\infty} \prod_{j=1}^{n_i} \frac{(Y_{ij}\theta_i)^{a_{ij}} e^{-Y_{ij}\theta_i}}{a_{ij}!} \frac{\theta_i^{\varphi-1} e^{-\varphi\theta_i} \varphi^{\varphi}}{\Gamma(\varphi)} d\theta_i \\ &= \frac{\varphi^{\varphi} \prod_{j=1}^{n_i} Y_{ij}^{a_{ij}}}{\left(\prod_{j=1}^{n_i} a_{ij}!\right)\Gamma(\varphi)} \int_0^{\infty} \theta_i^{a_i+\varphi-1} e^{-\theta_i(Y_i+\varphi)} d\theta_i \\ &= \frac{\varphi^{\varphi}\left(\prod_{j=1}^{n_i} Y_{ij}^{a_{ij}}\right)\Gamma(a_i + \varphi)}{\left(\prod_{j=1}^{n_i} a_{ij}!\right)\Gamma(\varphi)(Y_i + \varphi)^{a_i+\varphi}} \end{aligned} \tag{13}$$

This is the negative multinomial distribution of accident counts. Replacing Y_{ij} by the model equation makes the result in Equation 13 a likelihood function that depends on observed traits, some of which may change from one period to another, the parameter vector $\boldsymbol{\beta}$ and the additional parameter φ. For reasons explained in Hauer (*6*), instead of using φ that is common to all segments, one should use φ_i = (segment length) × φ.

Those parameter values are sought that maximize likelihood or, equivalently, its logarithm. For this purpose, the constants containing $a_{ij}!$ can be omitted. Thus, the contribution to the log-likelihood of entity i over periods 1 to n_i is

$$\ln(\ell_i) = \varphi_i \ln(\varphi_i) + \left[\sum_{j=1}^{n_i} a_{ij} \ln(Y_{ij})\right] + \ln[\Gamma(a_i + \varphi_i)] - \ln[\Gamma(\varphi_i)] - (a_i + \varphi_i)\ln(Y_i + \varphi_i) \quad (14)$$

Maximum likelihood parameter estimates are those that maximize

$$\ln(\ell) = \sum_{\forall i} \ln(\ell_i) \quad (15)$$

This formulation of the likelihood function has one principal advantage. It allows for the correct representation of changes in traffic flow and other variables that occur on each entity over time. Even changes in unmeasured variables (weather, reporting limits, etc.) can now be appropriately accommodated. The practical consequence is that the data sets used for SRSM do not have to be limited to periods over which traffic and other changes are negligible. One can use data for all past years and thereby richer data sets that produce better models.

MODEL EVOLUTION

An extract from the table of parameters developed in the course of the project by Hauer et al. (*3*) is given in Table 2. The ellipses (. . .) indicate omitted columns (variables) and rows (models).

The addition of a new variable is manifest in Table 2 by a new row. The first variable to be introduced was AADT, represented by $(\text{AADT}/10{,}000)^{2.16}e^{-0.31(\text{AADT}/10{,}000)}$. The introduction of AADT increased $\ln(\ell)$ by 272. The next variable entering the model, and also represented by a two-parameter function, was % trucks. It increased $\ln(\ell)$ by only 15.7, or 7.9 per parameter. In general, the larger the increase in $\ln(\ell)$/parameter, the earlier the variable should be introduced. However, the affinity of variables (e.g., AADT and percent trucks or degree of curve and curve length) may also influence the sequence of variable introduction.

Note that when the % trucks variable was added to the model, the parameters of the AADT building block changed. The change of previously estimated parameters with the introduction of a new variable is caused partly by the association between the new variable and the variables introduced earlier and partly by the precision with which parameters can be estimated. Scanning a column of Table 2 from top to bottom allows one to form an opinion about the magnitude of this joint effect. If a previously estimated parameter tends to change as new variables are added, one may expect that it would continue to do so. Conversely, when a parameter changes little as new variables are added, one may assume that it would remain stable. Fluctuations that are of a size that is consistent with the statistical precision of parameter estimation are a sign that the attainable stability has been reached. Larger fluctuations or a consistent trend in a parameter are a sign that parameter stability has not been reached and a causal interpretation is not possible.

EXAMINING GOODNESS OF FIT

When a model is used for prediction (Purpose A earlier), it is important that it fit well throughout the range of each variable. The cumulative residuals (CURE) method is a useful tool for checking and adjusting the fit (*7*). The residual—the difference between the number of recorded and predicted accidents—is the basic element by which to judge fit. Standard procedure is to examine the plot residuals against the variable of interest. In SRSM this may not be sufficiently revealing, as is illustrated in Figure 6.

However, when the same residuals are cumulated, as shown in Figure 7, the nature of the fit is revealed at once.

It is now clear that for 0 < AADT < 2,000, the count of accidents tends to be larger than the predicted values, and the opposite is true when 4,000 < AADT. Therefore, the functional form used in this case ($Y = \alpha\text{AADT}^{\beta}$) is not appropriate and needs to be replaced by a function that can rise more steeply near the origin and will have lower values for larger AADT. In general, a good CURE plot is one that oscillates around 0. A bad CURE plot is one that is entirely above or below 0 (except at the edges). If there are ranges of a variable where the CURE plot is consistently increasing or consistently decreasing, the fit in that range is poor and may be improved by a suitable modification of the functional form. A vertical jump in the CURE plot indicates an outlier. This should trigger a detailed investigation of that data point.

In the language of probability, a CURE plot is a random walk. For a random walk of this kind one can find limits beyond which the plot should go only rarely. Only the procedure for computing the limits is presented here; their derivation is given elsewhere (*6*). After sorting the N residuals in the increasing order of the variable of interest, number them consecutively 1, 2, . . . , n, . . . , N. Compute the squared residual for each n. Let $\hat{\sigma}^2(n)$ denote the sum of these squared residuals from 1 to n. Compute

TABLE 2 Sequence of Models and Parameters (*3*)

Model	AADT		% Trucks		D°		Curve L	Speed Limit	…	ln(ℓ)
										-6250.8
1	2.16	-0.31								-5979.2
2	1.93	-0.23	-0.18	0.06						-5963.5
3	1.90	-0.21	-0.16	0.06	-0.20	0.07				-5958.8
4	1.91	-0.22	-0.16	0.06	-0.15	0.06	-0.58			-5958.0
5	2.01	-0.26	-0.10	0.06	-0.16	0.06	Elim.	4.64		-5932.7
…										
10	2.25	-0.29	-0.10	0.06	-0.16	0.06	Elim.	3.97		5907.7
11	2.15	-0.26	-0.09	0.06	-0.17	0.06	Elim.	3.99	…	5901.5

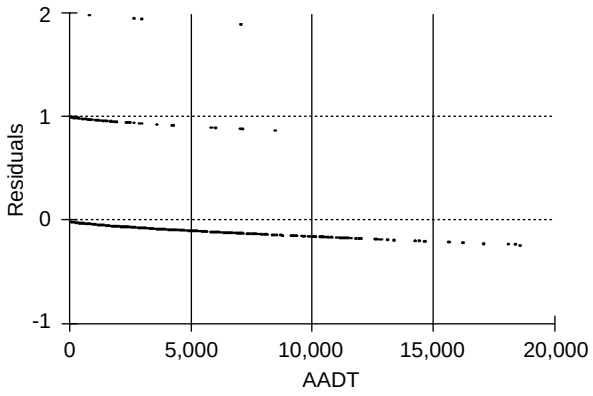

FIGURE 6 Plot of residuals against AADT (1,039 rural two-lane road segments in Maine).

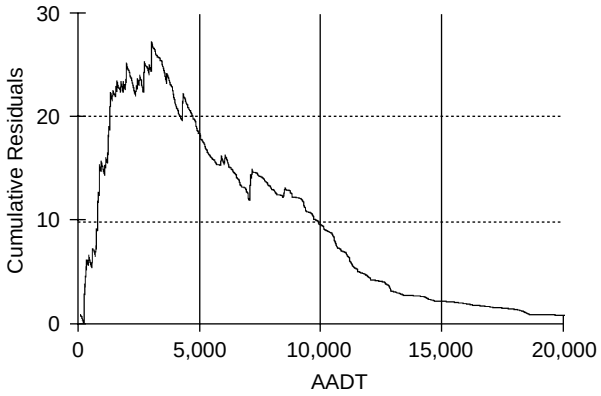

FIGURE 7 Cumulative residuals for data in Figure 6.

$$\sigma^* = \hat{\sigma}(n)\sqrt{1 - \frac{\hat{\sigma}^2(n)}{\hat{\sigma}^2(N)}} \tag{16}$$

and add $\pm 2\sigma^*(n)$ limits to the CURE plot.

The standard plot of residuals for the model and data from the work by Hauer et al. (*3*) and used in earlier examples is shown in Figure 8. It is not easy to tell from here whether the chosen function $(\text{AADT}/10{,}000)^{\beta_1} e^{\beta_2 \text{AADT}/10{,}000}$ makes for a good fit.

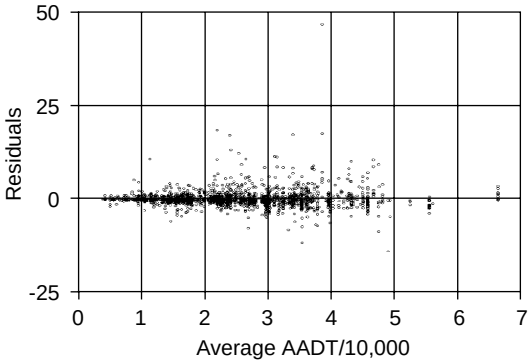

FIGURE 8 Residuals versus AADT/10,000.

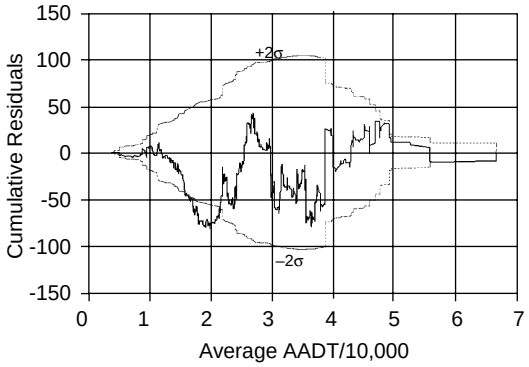

FIGURE 9 Cumulative residuals with $\pm 2\sigma^*$ bands.

The CURE plot for the same data is shown in Figure 9. The CURE plot oscillates around 0, as required. However, it encroaches on the lower limit and is therefore only marginally acceptable. The encroachment is partly due to an outlier (at AADT ≈ 38,000), which was kept in the data.

CONCLUSIONS

Learning about cause and effect from happenstance data is difficult, perhaps impossible. For road safety, it is important to try because, in many cases, other research approaches are not available. In this paper, an attempt was made to focus on what is important and neglected: how to build up the model equation, how to choose the appropriate functional forms, how to decide whether to include a variable in the model, and how to examine whether the chosen functional form fits the data. The approach advocates the gradual building up of the model equation, exploration, and backtracking—an approach that shuns routine and automation.

REFERENCES

1. Holland, P. Statistics and Causal Inference. *Journal of the American Statistical Association*, Vol. 81, 1986, pp. 945–960.
2. Box, E. P., W. G. Hunter, and J. S. Hunter. *Statistics for Experimenters*. John Wiley and Sons, New York, 1978.
3. Hauer, E., F. M. Council, and Y. Mohammedshah. Safety Models for Urban Four-Lane Undivided Road Segments. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1897*, TRB, National Research Council, Washington, D.C., 2004, pp. 96–105.
4. Miaou, S.-P., P. S. Hu, T. Wright, A. K. Rathi, and S. C. Davis. Relationship Between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach. In *Transportation Research Record 1376*, TRB, National Research Council, 1992, pp. 10–18.
5. Guo, G. Negative Multinomial Regression Models for Clustered Event Counts. *Sociological Methodology*, Vol. 26, 1996, pp. 113–132.
6. Hauer, E. Overdispersion in Modeling Accidents on Road Sections and in Empirical Bayes Estimation. *Accident Analysis and Prevention*, Vol. 33, 2001, pp. 799–808.
7. Hauer, E., and J. Bamfo. Two Tools for Finding What Function Links the Dependent Variable to Explanatory Variables. *Proc., ICTCT 97* (International Cooperation on Theories and Concepts in Traffic Safety), Lund, Sweden, 1997, pp. 1–7.

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