GOOD AND BAD DRIVERS—A MARKOV MODEL OF ACCIDENT PRONENESS

EMILIO VENEZIAN

Abstract

Existing models of the distribution of accidents among a population of drivers do not account for both the differences among individuals and those among age groups. This paper proposes a simple model to simultaneously explain these variations.

The model assumes that all drivers begin at some early age as "bad" drivers. Subsequently, drivers switch at random from the "bad" state, with high accident probabilities per mile driven, to a "good" state with low accident probabilities. The opposite transition, from "good" to "bad" states, also occurs at random. As the proportion of good drivers increases with age, the average frequency declines with age. The author develops in his paper the explicit mathematical equations of the model and a method of parameter estimation.

The model leads to three conclusions:

1. Classification efficiency, as measured by the SRI formula, can never achieve 100%. An upper bound of classification efficiency exists because the actual state of the driver at the inception of coverage is not known.

2. Underwriting and other risk assessment methods that tend to separate drivers in the good state from those in the bad state will offset some of the weaknesses in classification, increasing the efficiency of the risk assessment process as a whole.

3. Even with "perfect" risk assessment, that is, with complete separation of drivers in the good and bad states, efficiency will not reach 100% because subsequent switching during the policy period will create heterogeneity.

INTRODUCTION

Many authors have developed models which attempt to describe the statistical distribution of accidents. The simplest, in a sense, is the Poisson model, which assumes the probability of any individual having an accident in any given time period to be the same for all individuals and all time periods. Data on accidents do not often fit the predictions from this model.[1]

One way to account for the difference between predictions and data is to appeal to differences between individuals in their probabilities of having accidents, also called their accident proneness. The convenient way to develop this type of model is to assume that accident proneness fits a gamma distribution,[2] under which assumption the observed numbers of accidents have a negative binomial distribution. This distribution accounts for data somewhat more successfully than does a Poisson model.[3] Additional assumptions are needed, however, if one wishes to use the model to yield information about the relationship of accidents to age, or about the autoregressive structure of accidents.

A second way to explain the difference between data and the Poisson model is to suppose that accident proneness increases with every accident. The Polya model assumes that the likelihood of an individual having an accident in a time interval increases linearly with the number of accidents that the individual had prior to the beginning of the interval. Under this assumption, also, the observed numbers of accidents have a negative binomial distribution.[4]

Statistically, therefore, this model would describe the distribution of the numbers of accidents in a group just as successfully as would an assumed gamma distribution of accident proneness. Moreover, both models imply that the likelihood of having an accident increases linearly with the prior number of accidents; in the Polya model this is a behavioral assumption, whereas in the

---

[1] Hilary L. Seal, *Stochastic Theory of a Risk Business*, John Wiley & Sons, Inc., New York, 1969, pp. 12–29.

[2] Seal, *loc. cit.*; and Stanford Research Institute, "The Role of Risk Classifications in Property and Casualty Insurance: A Study of the Risk Assessment Process," Menlo Park, California, 1976.

[3] Seal, *loc. cit.*; Stanford Research Institute, *op. cit.*; and Donald C. Weber, "An Analysis of the California Driver Record Study in Context of a Classical Accident Model," *Accident Analysis and Prevention*, Vol. 44, 1972, pp. 109–116.

[4] William Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley & Sons Inc., New York, 1968, Vol. 1, 3rd Edition, pp. 121, 142, 143; and Seal, *loc. cit.*

gamma model it is a consequence of the information contained in the prior history.[5] The Polya model inherently leads to the prediction that accident frequency will increase with age, a prediction which is wrong for automobile accidents.[6]

A third way to account for the data is to assume that accident proneness varies over time; an extreme of this model, in which accident proneness was viewed as an all-or-none variable, has been studied.[7] A somewhat different approach is to assume that accident proneness is either "high" or "low," so that there are "good" and "bad" drivers;[8] this can be viewed as a polarization of either the heterogeneous proneness model or the episodic proneness model. Neither of these models has the ability to predict the variation of accident proneness with age; they both provide information on the autoregressive structure of observed accidents.

My interest in developing a model that could describe the statistical characteristics of accident distribution and simultaneously provide information on the age structure of accident rates first arose in early 1976, while I was reviewing an early draft of the Stanford Research Institute (SRI) report.[9] The draft of that report stated that the datum ". . . contradicts the simplistic view of a driver population made up of 'good' drivers and some 'bad' drivers," but did not contain a test of this "simplistic view." I performed a crude test of this model on eight age groups; I noted in a memorandum to SRI that the fit appeared to be adequate for each group, and added:

> It is also interesting that the accident likelihood for "good" drivers is much the same at the various ages, as it is for "bad" drivers. This suggests, among other things, a Markov model in which there are "good" and "bad" drivers but switching occurs from "good" to "bad" and vice versa. Again, the implications for merit rating could be important.

---

[5] Seal, *loc. cit.*; and Stanford Research Institute, *op. cit.*

[6] R. C. Peck, R. S. McBride, and R. S. Coppin, "The Distribution and Prediction of Driver Accident Frequencies," *Accident Analysis and Prevention*, Vol. 2, 1971, pp. 243–299; and Stanford Research Institute, *op. cit.*

[7] Seal, *loc. cit.*

[8] Seal, *loc. cit.*

[9] Stanford Research Institute, *op. cit.*

That suggestion lay dormant until a recent discussion of merit rating and the efficiency of risk classifications at the Risk Theory Seminar held under the auspices of the American Risk and Insurance Association.[10] Most discussants agreed that the view of classification efficiency taken in the SRI report was inadequate. The most heated arguments centered on whether Richard G. Woll's study[11] went far enough in correcting the errors inherent in the SRI measure of efficiency. The discussion was largely hampered by failures to distinguish between the "expected value of accident proneness" taken over a set of individuals at a given time and that taken over time for a given individual. The distinction is important because variance in the time-averaged proneness among individuals can be reduced, in principle at least, by both classification and underwriting selectivity, whereas variance in accident proneness resulting from future random events cannot be reduced by anything short of clairvoyance. My interest in the Markov model was revived when I realized that the model I had suggested could clarify some of these issues. I retained the assumption of two states, "good" and "bad," not for historical reasons but rather because the available data do not permit much discrimination. The revived model now has been developed to the extent that it is useful. Full development of the quantitative aspects of merit rating is still needed. At the current level, however, the model is useful as a framework for considering issues of classification and underwriting.

A MARKOV MODEL

Consider an individual who can be in one of two states, "good" or "bad." Assume that drivers have an accident probability of θ_1 per mile driven when they are in the "good" state and that the analogous quantity for the "bad" state is θ_2. Assume that the expected number of miles driven per unit time does not depend on the state (in fact, that complication could be accommodated readily). Also assume that in any time interval dt, an individual has a probability adt of changing from the "good" state to the "bad" state and a probability bdt of changing from the "bad" state to the "good" state.

The probability, $p(t)$, of being in the "good" state at time t is governed by the differential equation

[10] The author is grateful to the institutions which support the Risk Theory Seminar and thereby create a forum for active exchange of views.

[11] Richard G. Woll, "A Study of Risk Assessment," *PCAS* LXVI, 1979.

$$\frac{dp(t)}{dt} = -ap(t) + b(1 - p(t)) \tag{1}$$

For an individual who is known to be in the "bad" state at some initial time t_0, the solution of this equation is

$$p(t) = \frac{b}{a + b}\left(1 - e^{-(a+b)(t-t_0)}\right) \tag{2}$$

Averaging Equation 2 over the time period from t_1 to $t_1 + \Delta t$ yields the probability that the individual, known to be in the "bad" state at t_0, is in the "good" state during this interval:

$$p(t_1, \Delta t) = \frac{b}{a + b}\left[1 - \frac{1 - e^{-(a+b)\Delta t}}{(a + b)\Delta t}\, e^{-(a+b)(t_1 - t_0)}\right] \tag{3}$$

If all we know is that the individual was in the "bad" state at t_0, then during the time interval from t_1 to $t_1 + \Delta t$ the expected value of the k^{th} moment of the accident proneness per mile driven is

$$E(\theta^k | t_1, \Delta t) = \theta_1^k p(t_1, \Delta t) + \theta_2^k[1 - p(t_1, \Delta t)] \tag{4}$$

This expression also is the expected value of θ^k, given t_1 and Δt, taken over individuals with the same θ_1, θ_2, a, b, and t_0.

The simplest assumption that can be made is that all parameters are the same for all individuals. This does not seem a realistic assumption *a priori*, but would be the most parsimonious one. A slightly more complex assumption is that t_0, the age at which people begin to switch from bad driving to good driving, relates to maturation so that t_0 for females may be somewhat lower than t_0 for males. Although this refinement still is very simplistic, it is of interest to develop the equations and test the ability of such a simple model to account for observations.

In most cases, data are available not by individual ages but only aggregated for all drivers within certain age spans, e.g., between ages t_1 and $t_1 + s$. Equation 3 can be modified to apply to the age span by averaging between the youngest and oldest ages included in the age span. If the age distribution within the span is uniform we obtain

$$p(t_1, \Delta t) = \frac{b}{a + b}\left[1 - \frac{1 - e^{-(a+b)\Delta t}}{(a + b)\Delta t}\,\frac{1 - e^{-(a+b)s}}{(a + b)s}\, e^{-(a+b)(t_1 - t_0)}\right] \tag{5}$$

and Equation 4 needs no modification.

The expected accident proneness between ages t_1 and $t_1 + s$ per unit time $\dot{m}$ is

$$E(\phi(t_1)) = \dot{m}E(\theta)$$

$$= \dot{m}\left[\theta_1 \frac{b}{a+b} + \theta_2 \frac{a}{a+b} + f\frac{b}{a+b}(\theta_2 - \theta_1)e^{-(a+b)(t_1-t_0)}\right] \quad (6)$$

$$\text{where } f = \frac{1 - e^{-(a+b)\Delta t}}{(a+b)\Delta t} = \frac{1 - e^{-(a+b)s}}{(a+b)s} \quad (7)$$

For very large t_1 the expected accident proneness approaches an asymptotic value $E(\phi_a)$ given by

$$E(\phi_a) = \lim_{t_1 \to \infty} E(\phi(t_1)) = \dot{m}\left(\theta_1 \frac{b}{a+b} + \theta_2 \frac{a}{a+b}\right) \quad (8)$$

Using this we can write Equation 6 in the form

$$\ln [E(\phi(t_1)) - E(\phi_a)] = \ln\left[\dot{m}f\frac{b}{a+b}(\theta_2 - \theta_1)\right]$$

$$+ (a+b)t_0 - (a+b)t_1 \quad (9)$$

The variance of the accident proneness per unit time is

$$V[\phi(t_1)] = \dot{m}[E(\theta^2) - E^2(\theta)]$$

$$= \dot{m}(\theta_2 - \theta_1)^2\left[\frac{ab}{(a+b)^2} + f\frac{b(b-a)}{(a+b)^2}e^{-(a+b)(t_1-t_0)}\right. \quad (10)$$

$$\left. -f^2\frac{b^2}{(a+b)^2}e^{-2(a+b)(t_1-t_0)}\right]$$

This variance also has an asymptotic value, $V(\phi_a)$:

$$V(\phi_a) = \lim_{t_1 \to \infty} V(\phi(t_1)) = \dot{m}(\theta_2 - \theta_1)^2\frac{ab}{(a+b)^2} \quad (11)$$

For large values of t_1, the term $e^{-2(a+b)(t_1-t_0)}$ is much smaller than the other terms in Equation 10 so that we can rewrite this equation in the form

$$\ln[V(\phi(t_1) - V(\phi_a)] = \ln\left[\dot{m}(\theta_2 - \theta_1)^2 f \frac{b(b-a)}{(a+b)^2}\right] + (a+b)t_0 - (a+b)t_1 \tag{12}$$

Equations 9 and 12 indicate that semilogarithmic plots of the differences between the quantities of interest and their asymptotic values would be useful in identifying whether the model is adequate.

DISTRIBUTION OF THE NUMBER OF ACCIDENTS

If the probability of an individual's having an accident during a time interval is governed by the Poisson distribution, then the number of accidents experienced by that individual will be Poisson distributed with a parameter equal to the realization of the total proneness, even if the accident proneness parameter varies over time. Thus, for an individual whose average proneness over an interval Δt turns out to be ϕ, the probability of x accidents is

$$p(x|\phi,\Delta t) = e^{-\phi\Delta t}\frac{(\phi\Delta t)^x}{x!} \tag{13}$$

The k^{th} moment about the origin of the number of accidents for that individual is then

$$\begin{aligned} E(x^k|\phi,\Delta t) &= e^{-\phi\Delta t}\sum_{x=0}^{\infty}\frac{x^k(\phi\Delta t)^x}{x!} \\ &= \phi\Delta t e^{-\phi\Delta t}\frac{\partial}{\partial(\phi\Delta t)}\sum_{x=0}^{\infty}\frac{x^{k-1}(\phi\Delta t)^x}{x!} \\ &= (\phi\Delta t)\left[E(x^{k-1}|\phi\Delta t) + \frac{\partial}{\partial(\phi\Delta t)}E(x^{k-1}|\phi\Delta t)\right] \end{aligned} \tag{14}$$

From this recursion equation we obtain

$$\begin{aligned} E(x^0|\phi,\Delta t) &= 1 \\ E(x^1|\phi,\Delta t) &= \phi\Delta t \\ E(x^2|\phi,\Delta t) &= (\phi\Delta t)^2 + (\phi\Delta t) \\ E(x^3|\phi,\Delta t) &= (\phi\Delta t)^3 + 3(\phi\Delta t)^2 + (\phi\Delta t) \\ E(x^4|\phi,\Delta t) &= (\phi\Delta t)^4 + 6(\phi\Delta t)^3 + 7(\phi\Delta t)^2 + (\phi\Delta t) \end{aligned} \tag{15}$$

From these expressions we find the moments of the distribution of the numbers of accidents by the equation

$$E(x^k) = \int_0^\infty E(x^k|\phi, \Delta t)g(\phi\Delta t)d(\phi\Delta t) \tag{16}$$

where $g(\phi\Delta t)$ is the probability density function of $\phi\Delta t$ taken over the same set of individuals as the number of accidents experienced in Δt.

Thus,

$$\left.\begin{aligned} E(x) &= \Delta t E(\phi) \\ E(x^2) &= (\Delta t)^2 E(\phi^2) + \Delta t E(\phi) \\ E(x^3) &= (\Delta t)^3 E(\phi^3) + 3(\Delta t)^2 E(\phi^2) + \Delta t E(\phi) \\ E(x^4) &= (\Delta t)^4 E(\phi^4) + 6(\Delta t)^3 E(\phi^3) + 7(\Delta t)^2 E(\phi^2) + \Delta t E(\phi) \end{aligned}\right\} \tag{17}$$

It follows from these equations that

$$E(x) = \Delta t E(\phi), \text{ and} \tag{18}$$

$$V(x) - E(x) = E(x^2) - E^2(x) - E(x) = (\Delta t)^2 V(\phi) \tag{19}$$

Equations 18 and 19 demonstrate that the model given in the preceding section specifies both the expected value of the number of accidents, $E(x)$, and the "excess variance," $V(x) - E(x)$, as a function of age.

Let $M(x)$ denote the mean number of accidents observed among N individuals, and $S(x)$ denote the calculated value of the excess variance; we hypothesize that these quantities follow the model described above. In order to test this hypothesis, we need estimates of the sampling variability of these quantities. For $M(x)$, the calculation is standard; the variance of $M(x)$ is simply $V(x) \div N$. For $S(x)$, the calculation is not as familiar since it must account not only for the variance of the sample estimates of the variance and the mean, but also for the covariance between these. The basic results can be obtained from most good books in statistics.[12] Neglecting terms of order N^{-2} we obtain

$$V(S(x)) = \frac{\mu_4(x) - \mu_2^2(x) - 2\mu_3(x) + \mu_2(x)}{N} \tag{20}$$

where $\mu_j(x)$ is the j^{th} central moment of x, for $j \geq 2$.

[12] Harald Cramer, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N.J., 1964, pp. 347–348.

After additional algebra to relate the moments of x to those of $\phi\Delta t$ we obtain

$$V(S(x)) = \frac{1}{N}[\mu_4(\phi\Delta t) - \mu_2^2(\phi\Delta t) + 4\mu_3(\phi\Delta t) + 2\mu_2(\phi\Delta t) + 4\mu_1(\phi\Delta t)\mu_2(\phi\Delta t) + 2\mu_1^2(\phi\Delta t)] \tag{21}$$

where $\mu_1(\phi\Delta t) = E(\phi\Delta t) = \Delta t E(\phi) = \mu_1(x)$.

In the special case of a completely homogeneous population, one in which the realization of ϕ is identical across all individuals, $S(x)$ reduces to

$$E_1(S(x)) = 0 \tag{22}$$

$$V_1(S(x)) = \frac{2}{N}\mu_1^2(\phi\Delta t) \simeq \frac{2}{N}\mu_1^2(x) \tag{23}$$

Since $S(x)$ is asymptotically normally distributed, the value

$$Z_1 = \frac{S(x)}{M(x)}\sqrt{\frac{N}{2}} \tag{24}$$

is, asymptotically, a unit variance normal deviate, and provides a test of significance for the excess variance against the null hypothesis of zero excess variance that corresponds to a Poisson process with no heterogeneity.

COMPARISON OF THE MODEL WITH DATA

In order to test the adequacy of the model we must compare the model predictions to data. A convenient set of data is that drawn from licensed drivers in California in 1961–1963.[13] The published data include the mean numbers of accidents by year of age for ages 17 through 30 and by five year age groups for ages 21 through 76. The data are available for males and females separately, and sufficient information is provided to allow the calculation of the excess variance for each sex in age groups spanning five years. The relevant data are shown in Tables I and II. It is of some interest that the excess variance greatly exceeds its standard deviation, as indicated by the large values of Z_1 found for most age groups in both sexes. This indicates that the excess variance does not arise from sampling variability.

[13] Peck, McBride, and Coppin, *loc. cit.*

TABLE I

AVERAGE NUMBER OF ACCIDENTS BY AGE AND SEX FOR AGES 17–30

Age Group	Males		Females	
	Number*	Average Accidents	Number*	Average Accidents
17	11**	0.727***	4**	0.250
18	1114	0.532	763	0.213
19	1399	0.476	955	0.219
20	1683	0.419	1146	0.198
21	1521	0.396	1182	0.149
22	1600	0.355	1182	0.163
23	1678	0.308	1182	0.129
24	1757	0.311	1182	0.126
25	1836	0.301	1182	0.123
26	1721	0.298	1315	0.113
27	1794	0.288	1315	0.124
28	1867	0.310	1315	0.094
29	1940	0.279	1315	0.132
30	2014	0.277	1315	0.129

* Estimated from totals for the three-year age spans (see text) by assuming numbers are linear with age and requiring that the mean average accidents for an age span be equal to the weighted mean of individual years; totals for age groups may differ from those in the original article because of rounding to the nearest integer.

** Smallest integer consistent with data given in original article.

*** Table 12 of Peck, McBride, and Coppin gives 0.737 for this value, which is inconsistent with the data for single and married males given separately in the same table. The value given here is consistent with the disaggregated data.

TABLE II

AVERAGE NUMBER OF ACCIDENTS AND EXCESS VARIANCE, BY SEX, FOR AGE GROUPS

Age Group	Males				Females			
	N	$M(x)$	$S(x)$*	$Z_1(x)$	N	$M(x)$	$S(x)$**	$Z_1(x)$
18–20	4196	0.468	0.062	6.1	2863	0.209	0.017	3.1
21–25	8392	0.332	0.054	10.5	5910	0.138	0.018	7.1
26–30	9336	0.290	0.047	11.1	6574	0.118	0.013	6.3
31–35	10200	0.256	0.058	16.2	7534	0.119	0.017	8.8
36–40	10573	0.250	0.039	11.3	8612	0.122	0.012	6.5
41–45	10127	0.231	0.041	12.6	8113	0.122	0.012	6.3
46–50	9041	0.234	0.031	8.9	6671	0.126	0.009	4.1
51–55	7466	0.226	0.034	9.2	5253	0.108	0.013	6.2
56–60	5949	0.224	0.023	5.6	3807	0.124	0.006	2.1
61–65	4608	0.226	0.038	8.1	2706	0.118	0.015	4.7
66–70	3419	0.193	0.030	6.4	1822	0.112	0.011	3.0
71–75	2027	0.179	0.010	1.8	952	0.136	0.007	1.1
$\geq$ 76	1372	0.200	0.038	5.0	452	0.142	0.025	2.6

* Calculated from the distribution of male licenses by age and number of reported accidents given in Table 10 of Peck, McBride, and Coppin.

** Calculated from the distribution of female licenses by age and number of reported accidents given in Table 11 of Peck, McBride, and Coppin.

The values of the model parameters could be established by statistical fitting techniques. In view of the complexity of the task, however, we used a much simpler procedure, as described in the Appendix to this paper. The resulting parameter values are shown in Table III. For ease of comparison to the paper by Peck, McBride, and Coppin,[14] the parameters for males and females are shown in terms of the age recorded in that paper, which corresponds to two years more than the age t_1 used in our equations. Since the relevant variable is the difference $t_1 - t_0$, we can then use the age, as recorded in the article by Peck, McBride, and Coppin, with the value of t_0 from Table III to compute the relevant quantities.

TABLE III

VALUES OF COMPUTED PARAMETERS

Parameter	Value	Units
a	0.03	Per Year
b	0.17	Per Year
θ_1	4.20×10^{-6}	Per Mile
θ_2	18.76×10^{-6}	Per Mile
t_0, male	18.37	Years
t_0, female	16.02	Years

Figures 1a and 1b show the average numbers of accidents during a three-year period ($\Delta t = 3$) involving drivers in the age range 17 through 30 years, displayed by year of age ($s = 1$). The data for males are shown in Figure 1a; the data for females are shown in Figure 1b. In each case the asymptotic value has been subtracted from the observation and the range of plus and minus one standard deviation is shown. The lines shown in these figures represent Equations 6 and 16, with the relevant parameter values from Table III. The fit is generally adequate, though not outstanding.

Figures 2a and 2b display the corresponding data over the entire age range ($s = 3$ for 18 through 20 years, $s = 5$ for other ages). The line for males is in general agreement with the observations at all age groups up to the 61–65 year age group; beyond that, there may be some departure. In the case of females, however, the line follows the data only up to the 31–35 year age group, with what appear to be progressively larger departures after that age. More sophisticated fitting of the parameters would not improve the fit of the line to the data, since the data do not appear to be log-linear as implied by Equation 9.

[14] Peck, McBride, and Coppin, *loc. cit.*

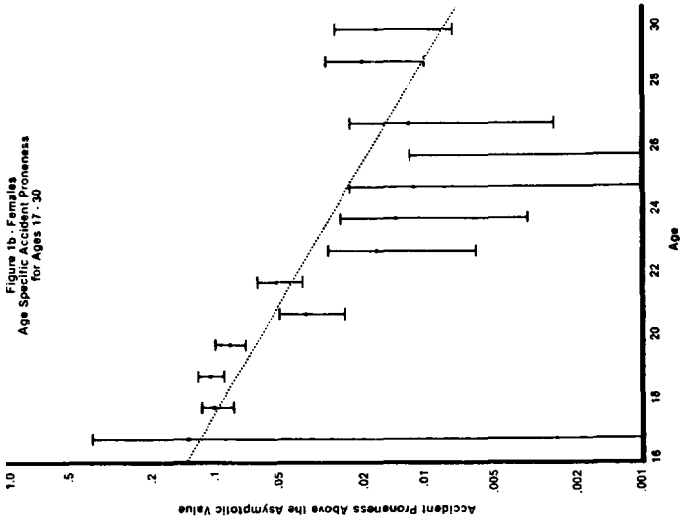


Figure 1b - Females
Age Specific Accident Proneness
for Ages 17 - 30
Accident Proneness Above the Asymptotic Value
1.0
.5
.2
.1
.05
.02
.01
.005
.002
.001
16
18
20
22
24
26
28
30
Age

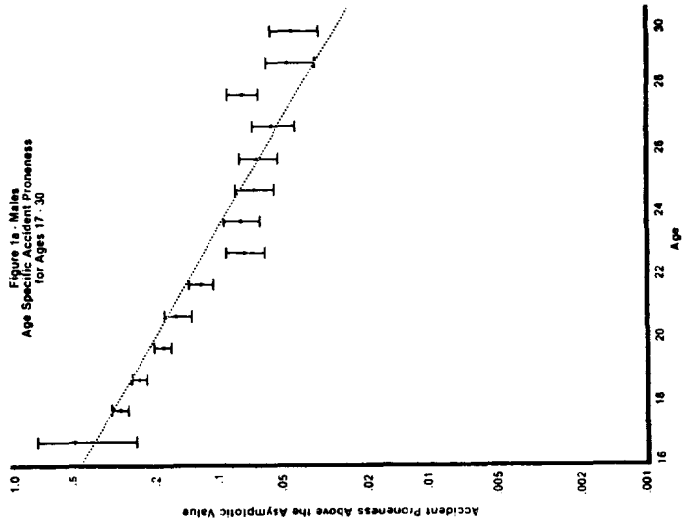


Figure 1a - Males
Age Specific Accident Proneness
for Ages 17 - 30
Accident Proneness Above the Asymptotic Value
1.0
.5
.2
.1
.05
.02
.01
.005
.002
.001
16
18
20
22
24
26
28
30
Age

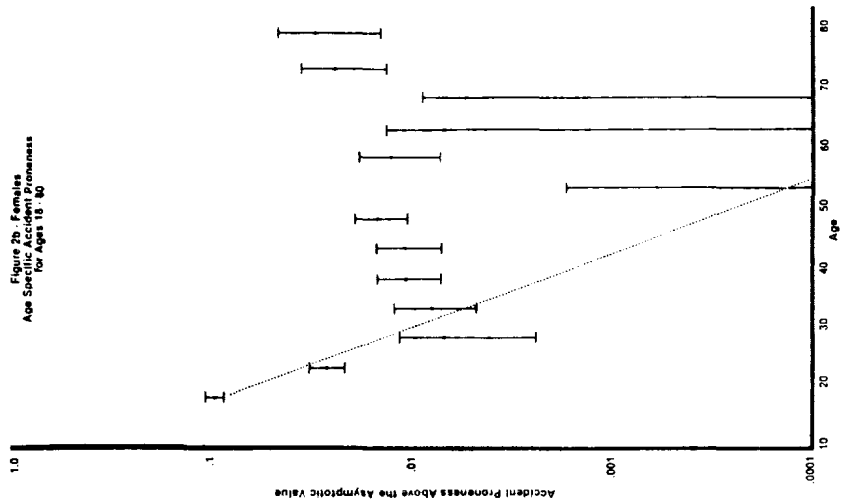


Figure 2b. Females
Age Specific Accident Proneness
for Ages 18 - 80
Accident Proneness Above the Asymptotic Value
1.0
.1
.01
.001
.0001
10
20
30
40
50
60
70
80
Age

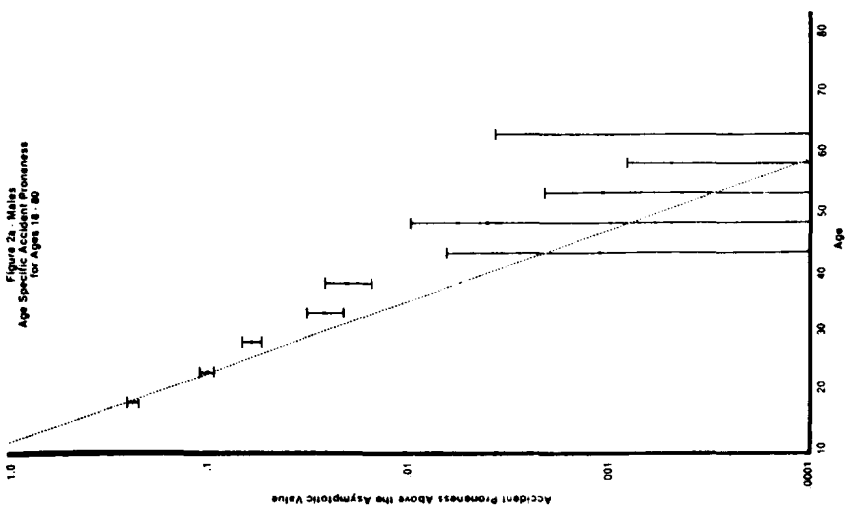


Figure 2a. Males
Age Specific Accident Proneness
for Ages 18 - 80
Accident Proneness Above the Asymptotic Value
1.0
.1
.01
.001
.0001
10
20
30
40
50
60
70
80
Age

Figures 3a and 3b show the data for the excess variance by age. The lines are calculated from Equations 8 and 17. The model agrees well with data for males but not particularly well with data for females. It is possible that the fit could be improved by relaxing the assumption that the mileage driven each year is *independent of age*. Using age-specific intensities of exposure would change the fitted values of all parameters, and might or might not lead to an improved fit. It would be feasible to examine this issue if questionnaire responses on mileage driven by a sample of the drivers studied were available. The exercise would be especially meaningful if the variance of mileage driven, as well as the mean, could be established for each age group and the equations were modified to allow for this variability.

DISCUSSION

The model presented here goes beyond the highly simplistic view of "good" and "bad" drivers by creating a model of transitions from one state to the other. The assumption that all drivers are "bad" at some suitably low age is then sufficient to account for differences in mean accident rates between age groups and heterogeneity within age groups. In this paper we have assumed that most parameters do not vary between individuals. More realistically, one might expect that mileage driven would be correlated not only with sex, as assumed here, but also with age, vehicle driven, and the characteristics of the territory in which most driving is done. Moreover, there is probably a quality weighting of the miles driven because of varying road and traffic characteristics. The values of the accident proneness parameters θ_1 and θ_2 may also vary across individuals and across driving environments. The characteristic age at maturation t_0 and the transition rates a and b also could be assumed to vary between individuals. A generalization of the model would include the specification of a joint distribution function for all these variables.

Keeping in mind the fact that the model is simplistic, it explains surprisingly well, within a simple theoretical structure, heterogeneity within and between ages and sexes. Because of this success, it is interesting to examine some implications of the model.

To begin with, the model is based on the assumption that an individual's accidents are generated by a Poisson process with time dependent parameters, yet the distribution of the numbers of accidents taken across individuals is not Poisson. Similarly, the numbers of accidents for a given individual taken over time subintervals will not exhibit a Poisson distribution around the mean for the interval as a whole, but will show clustering for subintervals during which the

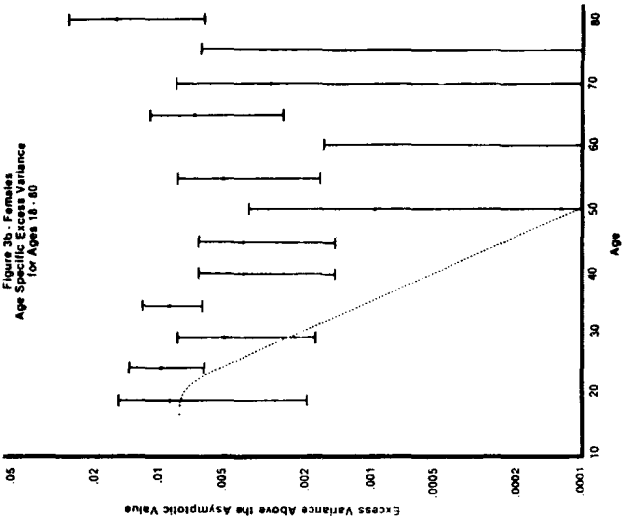


Figure 3b - Females
Age Specific Excess Variance
for Ages 10 - 80
Excess Variance Above the Asymptotic Value
.05
.02
.01
.005
.002
.001
.0005
.0002
.0001
10
20
30
40
50
60
70
80
Age

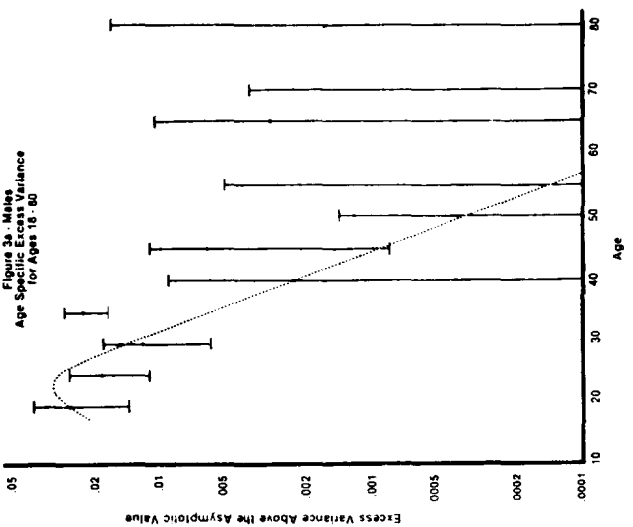


Figure 3a - Males
Age Specific Excess Variance
for Ages 10 - 80
Excess Variance Above the Asymptotic Value
.05
.02
.01
.005
.002
.001
.0005
.0002
.0001
10
20
30
40
50
60
70
80
Age

individual is in the "bad" driver state. An alternative statement of this last comment is that the distribution of times between accidents will not exhibit the exponential distribution that would be expected for a Poisson process of constant rate.

The model also illustrates the importance of maintaining clarity as to what is meant by "expected value." The expected value of the accident proneness per unit time for individuals of age t_1 is

$$E(\phi) = \dot{m}\left[\theta_1 \frac{b}{a+b} + \theta_2 \frac{a}{a+b} + \frac{b}{a+b}(\theta_2 - \theta_1)e^{-(a+b)(t_1 - t_0)}\right] \tag{25}$$

when the averaging process is over individuals about whom no other information is available. For individuals known to be in the "good" state at age t_1, the expected proneness per unit time over the following T years is

$$E(\phi|\text{good}) = \dot{m}\left[\theta_1 \frac{b}{a+b} + \theta_2 \frac{a}{a+b} - \frac{a}{a+b}(\theta_2 - \theta_1)\frac{1 - e^{-(a+b)T}}{(a+b)T}\right] \tag{26}$$

Similarly

$$E(\phi|\text{bad}) = \dot{m}\left[\theta_1 \frac{b}{a+b} + \theta_2 \frac{a}{a+b} + \frac{b}{a+b}(\theta_2 - \theta_1)\frac{1 - e^{-(a+b)T}}{(a+b)T}\right] \tag{27}$$

Thus the expected value, taken over time, for a given individual (who must be in either one or the other state at age t_1) is not the same as the expected value taken over individuals, except in two special cases:

(1) $t_1 = t_0$ and $T = 0$, and

(2) $t_1 = T = \infty$

The variances will differ correspondingly. A group selected for identical ages and initial states will develop heterogeneity just because of the random changes of individuals within that group.

The model developed in this paper has interesting implications relative to the continuing controversy regarding classifications, homogeneity, and under-

writing freedom.[15] It has become almost commonplace to say that the role of classification is not to predict the number of accidents that an individual will have during a time interval, but rather to predict the likelihood of that individual's having an accident.[16] By that criterion, 100% homogeneity, in the sense of no excess variance,[17] requires the identity of *realization* of proneness, not just identity in the *expected value* of proneness. This is unachievable.

The model has some important implications relative to merit rating. Individuals whose ages are close to t_0 will be in the "bad" driver state in almost every instance. Therefore, their prior accident records will contain additional information about their likely accident experience only to the extent that individuals differ with respect to mileage driven or other parameters assumed constant in this paper. For mature individuals, the situation is quite different. The age of mature individuals is not a good predictor of initial state, since at advanced ages very nearly 15% are in the "bad" state and 85% are in the "good" state. Among individuals who have just had an accident, nearly 45% will be in the "bad" state and only 55% will be in the "good" state. There is substantial persistence in a state; nearly 85% of the individuals in the "bad" state and 97% of those in the "good" state at any instant will remain in the state for at least one full year. Thus, a mature individual's prior accident record has substantial predictive value.

The model suggests that merit rating relativities will *increase* with age, and rapidly so at ages close to t_0. Though we have no data for drivers at ages close to t_0, the data from North Carolina,[18] shown in Figure 4, suggest that this prediction is correct. Further, the model suggests that the mileage driven by young people with accidents should be quite different from the mileage driven by young people without accidents; the difference should decrease with age.

---

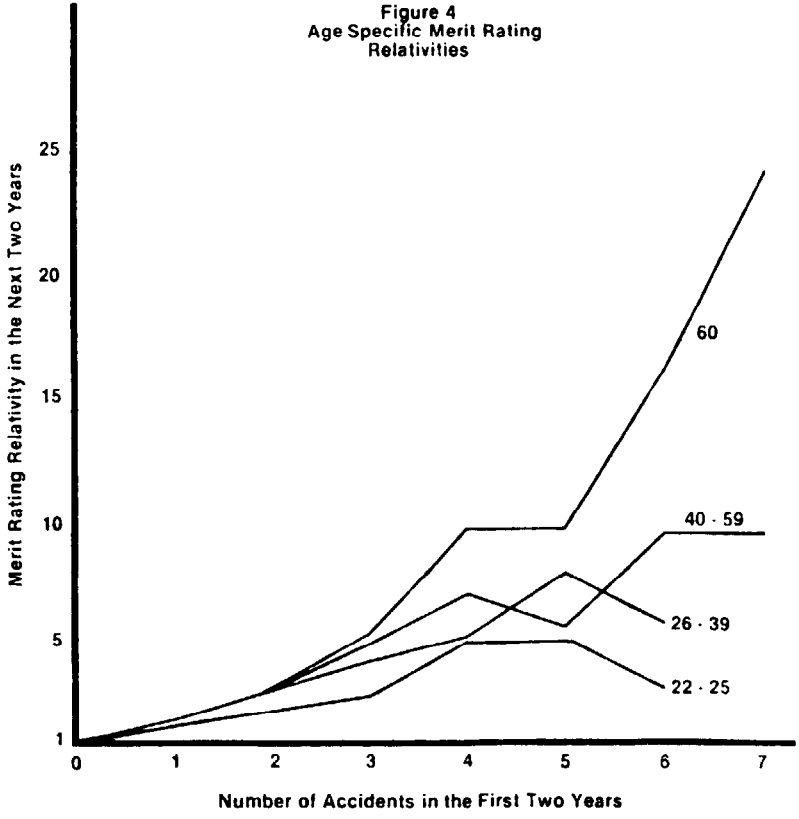
[15] Robert A. Bailey and LeRoy J. Simon, "Two Studies in Automobile Insurance Ratemaking," *The Astin Bulletin*, Vol. I, 1961, pp. 192–217.

[16] Michael A. Walters, "Risk Classification Standards," *PCAS* LXVIII, 1981 and Richard G. Woll, *op. cit.*

[17] Stanford Research Institute, *op. cit.*

[18] J. Richard Stewart and B. J. Campbell, "The Statistical Association Between Past and Future Accidents and Violations," The University of North Carolina, Highway Safety Research Center, Chapel Hill, N.C., 1972.

Figure 4
Age Specific Merit Rating
Relativities



In the model, the age relative to t_0 is the best objective *classification* predictor of an individual's state as "good" or "bad" driver. Prior accident record helps, but a complete development of the conditional probabilities would be needed to evaluate quantitatively the contribution of this variable to reduced heterogeneity. "Subjective" or "underwriting" judgments also could be used to determine whether an individual is in the "good" or "bad" state. Such judgments, if less than 100% efficient in separating drivers in the "good" state from those in the "bad" state, would have more impact if applied to mature drivers than if applied to young drivers. This contrasts with the usual perception of industry practices.

The applicability of models such as the one presented in this paper is often limited. In the case of automobile insurance, a major limitation is created by the fact that the coverage extends to an automobile and is not limited to a driver. Even if the model were an accurate representation of reality, its direct quantitative application to automobile insurance might not be warranted. The model does provide some interesting insights into the data that is needed to evaluate the model's validity. More sophisticated fitting of parameters seems much less important than assessing the interpersonal variation in all variables modelled and the impact of other variables, such as actual mileage driven.

APPENDIX
METHOD OF PARAMETER ESTIMATION

In order to check the fit of the model presented in this paper, the model's parameters must be determined. One parameter, mileage driven per year, is not readily accessible. This parameter is not of major importance, since it is merely a scaling factor; it is very convenient, however, since knowing the ratio of mileage driven by males to miles driven by females reduces by one the number of parameters to be estimated. I have used data based on a 1969–1970 survey by the Federal Highway Administration,[19] which indicates that, per person, per year, females drive approximately 48% of the mileage driven by males. The fitting was therefore performed on the basis that males drive 12,000 miles and females drive 5,800 miles per year.

The initial stage of the fitting relied on the fact that the mean numbers of accidents, minus the asymptotic value at high ages of the number of accidents,

[19] Motor Vehicle Manufacturers Association of the United States, Inc., "MVMA Motor Vehicle Facts & Figures 1978," Detroit, Michigan, p. 49. The information is based on unpublished data from the National Personal Transportation Survey conducted by the Bureau of the Census for the Federal Highway Administration, 1969–1970.

must be linear in a semilogarithmic plot. Since the asymptotic value is

$$\phi_a \Delta t = \dot{m}\Delta t \left(\theta_1 \frac{b}{a+b} + \theta_2 \frac{a}{a+b}\right), \tag{A1}$$

the value for females and males must be in the same ratio as the mileages. A few trials using the data for ages 17 through 30 gave $\phi_a \Delta t$ values of 0.23 for males and 0.11 for females. The slope of the semilogarithmic plots is $-(a + b)$, and the fact that both sets of data could be accommodated by $a + b = 0.20$ gave a preliminary indication that the model was promising and that the assumption of common parameters was tenable. This procedure also provided constraints on the parameters, giving two equations in four unknowns.

In order to determine all four of these unknowns, plus the values of t_0 for males and females, additional relationships were needed.

The asymptotic value of the excess variance provided one such relation:

$$S_a = S[x(\infty)] = (\dot{m}\Delta t)^2(\theta_2 - \theta_1)^2 \frac{ab}{(a+b)^2} \tag{A2}$$

Solving this with Equation A1 we obtain

$$\theta_2 = \frac{1}{\dot{m}\Delta t}\left[\phi_a \Delta t + \sqrt{\frac{b}{a}}\, S_a\right] \tag{A3}$$

Equation A3 allows solving for θ_2 if a value of b is assumed. Finally, we determined t_0 using the mean number of accidents at a recorded age[20] of 17 as estimated from the value at other ages:

$$\bar{x}_{17} - \phi_a = \dot{m}\Delta t \frac{b}{a+b}(\theta_2 - \theta_1) f e^{-(a+b)(t_1 - t_0)} \tag{A4}$$

This procedure was tried at various values of b until a reasonable fit, based on the excess variance at young ages, was obtained. A few trials sufficed. The selected parameters are shown below.

Parameter	Value	Units
a	0.03	Per Year
b	0.17	Per Year
θ_1	4.20×10^{-6}	Per Mile
θ_2	18.76×10^{-6}	Per Mile
t_0, male	18.37	Years
t_0, female	16.02	Years

[20] Recorded age is at the midpoint of the study for people in the middle of the age bracket. It therefore corresponds to $t_1 + 2$.