THE CREDIBILITY OF A SINGLE PRIVATE PASSENGER DRIVER

HOWARD C. MAHLER

Abstract

The credibility of the experience of an individual driver is determined by analyzing the accident records of private passenger drivers. For the particular data set analyzed, the risk parameters were found to be relatively stable over time, resulting in significant credibility being assigned to older years of data.

1. INTRODUCTION

In this paper, the accident records of private passenger drivers are analyzed using the methods developed by this author (in Mahler [2]) in order to estimate the credibility of the experience of an individual driver. The analysis is done using only the following classification variables: gender, state of licensing, and being licensed over an entire 14-year span.[1]

The use of additional years of experience (more than 10) is found to add significant information and is projected to do so for longer periods of time. For this particular data set, the risk parameters were found to be relatively stable.[2]

2. THE DATA SET

The data analyzed are for California private passenger drivers [1]. The data show the number of accidents annually in 1961–1963 and 1969–1974, for a sample of drivers licensed from 1961 to 1974. Thus, there

[1] Additional classification information was not available in the data set used.

[2] A larger data set, in terms of number of drivers, number of years of data, or classification information, may lead to a somewhat different conclusion.

are nine years of data for each driver, covering a 14-year period with a five year gap in the middle. The data are divided between male and female drivers. An extract from the data set is shown in Appendix A.

It should be noted that this data set allows an analysis only of accident frequency. No information is available on accident or claim severity.

3. CORRELATIONS

The correlations between years of data are shown in Exhibit 1. The key step in the analysis is to group together those pairs of years of data separated by the same number of years.[3] For example, there are five pairs separated by two years: 1961 and 1963, 1969 and 1971, 1970 and 1972, 1971 and 1973, and 1972 and 1974.[4]

The average correlations between pairs of years of data with different separations are shown in Exhibit 2. The correlations are all small, reflecting the low information-to-noise ratio. The correlations appear to decline gradually as the separation increases. This can be confirmed by fitting a linear regression to the average correlations or to the individual observed correlations.[5]

The results of fitting linear regressions to the individual observed correlations are:

$$\text{Male Drivers: correlation}(\lambda) = .03515 - .00064\lambda$$
$$\text{Female Drivers: correlation}(\lambda) = .03102 - .00126\lambda$$

Both regressions indicate a small, but significant, decline in the correlation as the separation between years λ increases.[6]

One can use these equations to approximate the covariance structure for separations of from one to 13 years. Also, it would not be unreasonable to use these equations to extrapolate the covariance structure for

[3] Mahler [2] makes the assumption that the correlation depends solely on the number of years of separation.

[4] There is a gap in the data from 1964 to 1968.

[5] One could also fit a weighted regression to the average correlations with weights equal to the number of observations underlying each average. The results of any of these three regressions are very similar.

[6] Both are significant at a 0.5% level. The t-statistics are −2.81 and −4.02, respectively, for 34 degrees of freedom.

EXHIBIT 1

Correlations (Male Drivers)

	1961	1962	1963	1969	1970	1971	1972	1973	1974
1961	1.0000	.0426	.0387	.0261	.0330	.0391	.0285	.0314	.0258
1962	.0426	1.0000	.0384	.0228	.0267	.0405	.0257	.0226	.0332
1963	.0387	.0384	1.0000	.0299	.0374	.0246	.0269	.0185	.0285
1969	.0261	.0228	.0299	1.0000	.0304	.0320	.0302	.0279	.0240
1970	.0330	.0267	.0374	.0304	1.0000	.0269	.0350	.0388	.0407
1971	.0391	.0405	.0246	.0320	.0269	1.0000	.0350	.0291	.0340
1972	.0285	.0257	.0269	.0302	.0350	.0350	1.0000	.0363	.0358
1973	.0314	.0226	.0185	.0279	.0388	.0291	.0363	1.0000	.0342
1974	.0258	.0332	.0285	.0240	.0407	.0340	.0358	.0342	1.0000

Correlations (Female Drivers)

	1961	1962	1963	1969	1970	1971	1972	1973	1974
1961	1.0000	.0285	.0290	.0159	.0145	.0188	.0337	.0043	.0134
1962	.0285	1.0000	.0284	.0241	.0279	.0217	.0202	.0063	.0123
1963	.0290	.0284	1.0000	.0322	.0200	.0236	.0247	.0171	.0180
1969	.0159	.0241	.0322	1.0000	.0412	.0195	.0380	.0188	.0205
1970	.0145	.0279	.0200	.0412	1.0000	.0225	.0154	.0337	.0164
1971	.0188	.0217	.0236	.0195	.0225	1.0000	.0270	.0217	.0249
1972	.0337	.0202	.0247	.0380	.0154	.0270	1.0000	.0308	.0374
1973	.0043	.0063	.0171	.0188	.0337	.0217	.0308	1.0000	.0412
1974	.0134	.0123	.0180	.0205	.0164	.0249	.0374	.0412	1.0000

Note the gap in information from 1964 through 1968.

EXHIBIT 2

Observed Average Correlations of Drivers' Experience Over Time

Difference Between Pairs of Years of Experience	Correlation		Number of Pairs of Years Observed
	Males	Females	
1	.0348	.0314	7
2	.0341	.0246	5
3	.0343	.0322	3
4	.0343	.0176	2
5	.0240	.0205	1
6	.0299	.0322	1
7	.0301	.0221	2
8	.0258	.0225	3
9	.0335	.0203	3
10	.0278	.0187	3
11	.0265	.0193	3
12	.0323	.0083	2
13	.0258	.0134	1

longer separations, provided one imposes the restriction that the correlations are not negative; i.e., that the correlations decline to zero and then remain there.[7] For the regression equation for male drivers, it takes 55 years for the correlation to decline to zero. For the female drivers, it takes 25 years for the correlation to decline to zero.

4. COVARIANCE STRUCTURE

In order to calculate the least squares credibilities, one has to estimate the covariance structure of the data. The required quantities are:

$$\begin{aligned} \tau^2 &= \text{between variance;} \\ C(\lambda) &= \text{covariance for data for the same risk, } \lambda \text{ years apart} \\ &= \text{``within covariance;''} \\ C(0) &= \text{within variance.} \end{aligned}$$

The within covariances will be estimated in terms of the correlations discussed in the previous section:

$$C(\lambda) = \text{correlation}(\lambda) \times C(0);$$

$$\text{correlation}(\lambda) = \begin{cases} \text{MAX}[0,\ .03515 - .00064\lambda] & \text{Male Drivers;} \\ \text{MAX}[0,\ .03102 - .00126\lambda] & \text{Female Drivers.} \end{cases}$$

The variances are estimated in Appendix B. The results are:

	Within Risk Variance	Between Risk Variance
Male Drivers	.0724	.0116
Female Drivers	.0377	.0057

In both cases the within variance is larger than the between variance.[8]

[7] It would be equally valid to extrapolate using an exponential regression fit to the correlations, as well as other methods of extrapolation. The use of a linear extrapolation is judged to be sufficient to illustrate the general technique.

[8] The Bühlmann credibility parameter K is the ratio of the within variance to the between variance. In these cases $K = 6.2$ and 6.6. The Bühlmann credibility for N years of data is given by $Z = N/(N + K)$.

The resulting covariance structure is:

$$\tau^2 = \begin{cases} .0116 & \text{Male Drivers} \\ .0057 & \text{Female Drivers} \end{cases}$$

$$C(0) = \begin{cases} .0724 & \text{Male Drivers} \\ .0377 & \text{Female Drivers} \end{cases}$$

For $\lambda \geqq 1$:

$$C(\lambda) = \begin{cases} .0724 \times \text{MAX}[0,\ .03515 - .00064\lambda] & \text{Male} \\ .0377 \times \text{MAX}[0,\ .03102 - .00126\lambda] & \text{Female} \end{cases}$$

In addition, the following example, with much more quickly shifting risk parameters over time, will be provided for illustrative purposes of contrast. The assumed covariance structure is:

$$\tau^2 = .01;$$

$$C(0) = .07;$$

For $\lambda \geqq 1$:

$$C(\lambda) = .07 \times \text{MAX}[0,\ .5 - .05\lambda].$$

5. CREDIBILITIES

In the case of using the latest N years of data, with the complement of credibility given to the overall mean, Mahler[9] develops the following N linear equations in N unknowns which can be solved for the least squares (Bühlmann/Bayesian) credibilities:

$$\sum_{j=1}^{N} Z_j\,(\tau^2 + C(|i - j|)) = \tau^2 + C(N + \Delta - i) \qquad i = 1,2, \ldots N$$

where:

Z_j = the credibility assigned to year j, with $j = N$ the most recent year of data;

τ^2 = between variance;

[9] Equation 11.3 in Mahler [2].

$C(\lambda)$ = covariance for data for the same risk, λ years apart = "within covariance;"

$C(0)$ = within variance;

Δ = the length of time between the latest year of data used and the year being estimated.

Using the covariance structure from the previous section, these equations produce the credibilities shown in Exhibit 3. Given the relatively small amount of data used, the estimated credibilities are subject to a fair amount of uncertainty.[10]

For both the male and female drivers, the credibilities calculated for older years are relatively close to those for more recent years. The sum of the credibilities as shown in Exhibit 4 increases as the number of years of data increases in a manner that is not unexpected. For male drivers the total credibility is approximately $N/(N + 5)$. For female drivers the total credibility is approximately $N/(N + 6)$.

In the example for contrast, the most recent year gets much more weight than older years, since the correlations quickly decrease to zero. The sum of the credibilities is much higher for the use of between one and five years of data than is the case for the California data, since the correlations are higher in this example for contrast.

6. SQUARED ERRORS

Mahler[11] gives the following equation for the expected squared error between the observation and prediction:

[10] The values shown for the use of more than 15 years of data are subject to even more uncertainty, since they are based on an extrapolation of the covariance structure beyond that estimated from the data set.

[11] Equation 11.2 in Mahler [2].

$$V(Z) = \sum_{i=1}^{N} \sum_{j=1}^{N} Z_i Z_j(\tau^2 + C(|i - j|))$$
$$- 2 \sum_{i=1}^{N} Z_i(\tau^2 + C(N + \Delta - i))$$
$$+ \tau^2 + C(0),$$

where all the symbols are defined as before and Z_i is the credibility assigned to year i and the complement of credibility is given to the overall mean.

Exhibit 5 displays the squared errors corresponding to the use of the least squares credibilities calculated in the previous section. For both the male and female drivers, the squared errors decline slowly and at a gradually declining rate as more years of data are added. In the example for contrast, the squared error declines significantly with the use of a single year of data, then declines somewhat with the use of a few additional years, and then levels off more quickly than for the California driver data.

7. CONCLUSIONS

The data set analyzed in this paper was one of two analyzed in a paper by Emilio Venezian [3]. In this paper, the data is analyzed in a more detailed manner using the methods developed in Mahler [2]. This analysis leads to the conclusion that the risk parameters are shifting at a relatively slow rate, which explains why Dr. Venezian, for this data set, was not able to reject the hypothesis that relative accident rates are stable.

Given the relatively limited information available on each driver in this data set, additional years of each driver's past accident record provide useful information for predicting his or her future relative accident frequency. Therefore, accident records from 10 or 15 years ago would be given significant credibility. However, it is important to keep in mind that credibility is a relative concept. The 10-year-old accident information is being given significant weight, but only relative to the weight given

EXHIBIT 3, PART 1

MALE DRIVERS

Credibility (based on assumed covariance structure, $\Delta = 1$)

Years Between Data and Estimate	Number of Years of Data Used							
	1	2	3	4	5	10	15	20
1 (Most Recent)	16.8%	14.4%	12.6%	11.2%	10.1%	7.0%	5.5%	4.7%
2		14.3	12.5	11.1	10.0	6.9	5.4	4.6
3			12.5	11.1	10.0	6.8	5.3	4.5
4				11.0	9.9	6.7	5.2	4.4
5					9.9	6.6	5.1	4.3
6						6.6	5.0	4.2
7						6.5	5.0	4.1
8						6.4	4.9	4.0
9						6.4	4.8	4.0
10						6.4	4.8	3.9
11							4.7	3.8
12							4.7	3.8
13							4.6	3.7
14							4.6	3.7
15							4.6	3.6
16								3.6
17								3.6
18								3.5
19								3.5
20								3.5
Total Credibility	16.8%	28.7%	37.6%	44.4%	49.9%	66.3%	74.2%	79.0%

EXHIBIT 3, PART 2

FEMALE DRIVERS

Credibility (based on assumed covariance structure, $\Delta = 1$)

Years Between	Number of Years of Data Used							
Data and Estimate	1	2	3	4	5	10	15	20
1 (Most Recent)	15.7%	13.6%	12.0%	10.8%	9.8%	7.0%	5.8%	5.2%
2		13.5	11.9	10.6	9.7	6.9	5.6	5.0
3			11.8	10.5	9.5	6.7	5.4	4.8
4				10.4	9.4	6.5	5.2	4.6
5					9.3	6.4	5.1	4.4
6						6.2	4.9	4.3
7						6.1	4.8	4.1
8						6.0	4.7	4.0
9						6.0	4.5	3.8
10						5.9	4.4	3.7
11							4.3	3.6
12							4.3	3.5
13							4.2	3.4
14							4.1	3.3
15							4.1	3.2
16								3.1
17								3.1
18								3.0
19								3.0
20								2.9
Total Credibility	15.7%	27.1%	35.7%	42.3%	47.7%	63.7%	71.4%	76.0%

EXHIBIT 3, PART 3

Example for Contrast

Credibility (based on assumed covariance structure, $\Delta = 1$)

Years Between Data and Estimate	Number of Years of Data Used							
	1	2	3	4	5	10	15	20
1 (Most Recent)	51.9%	37.3%	32.7%	31.3%	31.0%	30.7%	30.2%	30.1%
2		28.2	22.2	20.2	19.7	19.6	19.6	19.5
3			16.1	13.2	12.4	12.5	12.8	12.7
4				8.8	7.5	7.8	8.2	8.1
5					4.1	4.7	5.0	4.9
6						2.6	2.5	2.6
7						.9	.5	.6
8						−.5	−1.4	−1.3
9						−2.1	−3.6	−3.6
10						−4.1	−6.6	−6.6
11							−.8	−.9
12							1.8	1.5
13							2.7	2.3
14							3.0	2.3
15							3.0	1.9
16								1.4
17								.9
18								.5
19								.4
20								.7
Total Credibility	51.9%	65.5%	71.1%	73.5%	74.7%	72.1%	76.9%	78.0%

EXHIBIT 4

SUM OF CREDIBILITIES OF THE INDIVIDUAL YEARS OF DATA

Number of Years of Data Used	Male Drivers	Female Drivers	Example For Contrast
1	16.8%	15.7%	51.9%
2	28.7	27.1	65.5
3	37.6	35.7	71.1
4	44.4	42.3	73.5
5	49.9	47.7	74.7
10	66.3	63.7	72.1
15	74.2	71.4	76.9
20	79.0	76.0	78.0
30	84.0	81.1	81.3
40	86.8	84.7	83.2
50	87.8	87.1	85.0
60	89.1	88.5	86.7

EXHIBIT 5

SQUARED ERRORS*

Number of Years of Data Used	Male Drivers	Female Drivers	Example For Contrast
0**	.0840	.0434	.0800
1	.0816	.0423	.0585
2	.0800	.0416	.0538
3	.0787	.0410	.0524
4	.0778	.0405	.0520
5	.0770	.0402	.0519
6	.0764	.0399	.0519
7	.0759	.0397	.0519
8	.0755	.0395	.0519
9	.0751	.0393	.0519
10	.0748	.0392	.0518
15	.0738	.0387	.0514
20	.0732	.0385	.0514
30	.0727	.0383	.0513
40	.0725	.0382	.0512
50	.0724	.0381	.0511
60	.0723	.0380	.0511

* Expected squared error between the observation and prediction, where the prediction employs the least squares credibilities.

** Relying solely on the overall mean, the expected squared error is the between variance plus the within variance.

to the other data that is available. The credibility depends on the value of the information contained in the overall mean, which is given the complement of credibility. This depends, in turn, on the classification information available. If, for example, data on the principal place of garaging of the car being driven or the age of the driver were available and incorporated in the analysis, then the credibility assigned to older accident data would differ.

This general method of analysis should be useful when applied to other sets of data.

REFERENCES

[1] K.W. Kwong, J. Kuan, and R.C. Peck, *Longitudinal Study of California Driver Accident Frequencies I: An Exploratory Multivariate Analysis,* Department of Motor Vehicles, State of California, Sacramento, California, 1976.

[2] H.C. Mahler, "An Example of Credibility and Shifting Risk Parameters," *PCAS* LXXVII, 1990, p. 225.

[3] E.C. Venezian, "The Distribution of Automobile Accidents . . . Are Relativities Stable over Time?" *PCAS* LXXVII, 1990, p. 309.

APPENDIX A

SAMPLE EXTRACT OF CALIFORNIA DRIVER ACCIDENT DATA*

Nine Year Total Number of Accidents	Single Year Accidents**	Male Drivers	Female Drivers
4	000010111	0	1
4	000011020	1	1
4	000012100	0	1
4	000020002	1	0
4	000020110	3	0

* Taken from Appendix I of Longitudinal Study of California Driver Accident Frequencies [1]. The various combinations of single year accidents that occurred for the 54,165 drivers in the sample are shown. The nine year total number of accidents observed ranged from 0 to 9.

** Columns 1 through 9 represent single year accident totals for years 1961, 1962, 1963, 1969, 1970, 1971, 1972, 1973, and 1974 respectively.

APPENDIX B

ANALYSIS OF VARIANCE

The total sum of squares of deviations from the grand mean for the data is given by:

$$\text{Total Sum of Squared Deviations} = \sum_i \sum_t X_{it}^2 - X^2/N,$$

$$\text{where: } X = \sum_i \sum_t X_{it},$$

$$N = \sum_i \sum_t 1.$$

$$\text{Within Risk Sum of Squared Deviations} = \sum_i \sum_t X_{it}^2 - \sum_i X_i^2/n_i$$

$$\text{where: } X_i = \sum_t X_{it},$$

$$n_i = \sum_t 1.$$

$$\text{Between Risk Sum of Squared Deviations} = \sum_i n_i\left(\frac{X_i}{n_i} - \frac{X}{N}\right)^2,$$

$$= \sum_i X_i^2/n_i - X^2/N.$$

Total Sum of Squares = Within Sum of Squares
+ Between Sum of Squares.

To get the variances, one divides each sum of squares by the product of (number of years of data − 1) × (number of drivers − 1). For the data sets examined here, the number years of data is nine. The number of male drivers is 30,293 and the number of female drivers is 23,872.

It should be noted that for the credibility analysis only the relative size of the variances is used. Therefore, as long as the sums of squared deviations are each divided by the same number, the result of the credibility analysis will be the same.

The sum of squared deviations are:

	Within	Between	Total
Male Drivers	17,555	2,815	20,370
Female Drivers	7,193	1,092	8,285

The resulting estimated variances are:

	Within Variance	Between Variance	Total Variance
Male Drivers	.0724	.0116	.0841
Female Drivers	.0377	.0057	.0434