TITLE: USES OF CLOSED CLAIM DATA FOR PRICING

AUTHOR: Mr. R. Michael Lamb

Mr. Lamb is Casualty Actuary for the Insurance Division of the State of Oregon. He received his FCAS designation in 1975 and holds masters degrees from Purdue and the University of Washington in mathematics and business administration. Michael is chairman of NAIC technical task forces on workers' compensation and on medical malpractice insurance and serves as a member of several others. He is on the CAS Editorial Committee.

REVIEWER: Mr. Richard S. Biondi

Mr. Biondi is Manager and Associate Actuary for Insurance Services Office in New York. At ISO, he is the staff representative on several committees, including the Commercial Automobile, General Liability and Professional Liability Actuarial Subcommittees. Dick received his FCAS in 1979 and holds an M.S. degree from the Polytechnic Institute of New York.

A fundamental problem of pricing insurance is: When all is known about claims from an accident-or policy-year, that year is too old to be relevant for next year's coverage. Thus, our ancestors began using aggregate historical patterns to estimate how incurred costs of recent periods would mature to full ultimate value.

The common accident-year model will be referred to as a representative of these development methods. The cost of claims from an accident year can be estimated at each of several points of time. The estimate at one time divided by the estimate at the previous time is an observed development ratio. Development stages are defined by a series of evenly-spaced time intervals measured from the beginning of each accident year. The latest observed ratios for each stage are usually averaged to estimate how a recent accident year yet to reach that stage will develop when it does pass through. The compound product of development ratios over all stages after a certain stage until the end of time - or to some prudent horizon - is a development factor for an accident year which has reached that stage.

The costs used in the process are usually the estimated incurred costs of claims reported to date.

The assumption that this statistic will follow historic patterns rests on a belief that claim personnel who establish reserves are both consistent and uneducable. Using paid costs instead of incurred costs is more objective, but disregards all information about open reported claims. There is a tradeoff of advantages to be considered.

Pricing insurance is like predicting adult traits of the next unborn generation of a species. Offspring are born and then grow teeth, hair, claws or fins and learn to walk, swim, hunt or fly and grow to adult size and strength. We can observe how youngsters of past generations have passed through stages to become adults and so can predict how todays children and adolescents will ripen. This is even called "development".

But in predicting the next and future generations, we must allow for evolution. In times of rapid evolution, many previous patterns of development into adults may not be accurate because the adults will be different. It becomes necessary to examine the very latest information about members of the species at every stage.

Evolution is called "trend" by an actuary. Trend factors are calculated across accident years much as

development factors are calculated across stages of maturity. So, the costs of a future accident year is estimated by essentially this formula:

This paper compares the common accident-year model with the uncommon closure-year model. Whereas an accident year includes all accidents or incidents occurring in a year, a closure year includes all claims reaching final disposition during a year regardless of when the incidents occurred.

Closed claim data offer the most recent objective information about final costs of insured risks. In times of uncertainty, this can be tremendously important-particularly when new methods of claim management or other aspects of claim disposition are significantly affecting costs independently from circumstances of true original incidents.

Closure-year models are uncommon because they do not represent an insurance product. Accident and policy years are more natural. Closed claim data is not temporally aligned with claims arising from, reported in, or covered by policies issued in a recent period.

Closure-year models are difficult to assemble. Ideally, you should have closure data for claims arising from all prior occurrence periods within a conservative horizon. Relying upon open claim reserves to represent early periods can cloud important distinctions between risks. Furthermore, unless the insured population is stable and your data source is universal, you must have exposure indices for each occurrence period. For application to a future coverage year, each occurrence period component of a closure year must be separately trended in the traditional algebraic model. Pure premium trends are the most natural, or they may be split into frequency and claim size portions. The use of external cost indices and trend residuals is not recommended without considerable study into how claim costs are determined by occurrence-period, closure-period, and intermediate-period influences. The simplest conception of a closure-year model for representing an accident year is

$$\text{Accident Year (T+1) cost per unit} = \sum_{J=1}^{M} \frac{\text{Accident Year (T-j+1) Claims Closed in Year T}}{\text{Accident Year (T-j+1) exposures}} \times \begin{pmatrix}\text{Trend Factor to Year T+1 from} \\ \text{(T-j+1)}\end{pmatrix}$$

where M is the number of years required for all claims to be closed. More interesting and useful models will be presented in later sections.

The most serious conceptual problems with closure-year models relate to the passage of time over long horizons. We toil and spin in a multivariate world of infinite dimension in which relations between finite sets of factors do not remain constant. Significant changes over only a few years, however, mean important variables have not been included. The inability to recognize and usefully measure important influences is the true conceptual difficulty in any model. Other time-related problems will be discussed later.

Another criticism of closure-year models is that they ignore information offered by open claims. They resemble an extreme of payment development models. This criticism, however, leads us to see the value of closed claim data as fully-developed factual information about claims now reaching final disposition. During any times of changing claim management approaches or disposition methods which may affect costs, closed claim data should at least be used to supplement an accident-year model. The algebraic construction of closure-year models suggests closed claim trends correspond to accident-year development factors and certainly can explain and guide their selection.

Illustrations in this paper will mostly be drawn from the medical malpractice claims study of the National Association of Insurance Commissioners, which is the most extensive closed claim research effort in the public domain.

AN ALGEBRAIC MODEL

The model described hereafter relies on the work of Archer McWhorter (2), with some important variation. Let us define M to be the number of years required for all claims to be closed, or at least a reasonable horizon where remaining claims may be aggregated with little loss of precision, and

$N(t)$ = the ultimate number of claims for occurrence year t,

$n(t,u)$ = the number of claims from occurrence year t closed during closure-year u,

$g(j)$ = the fraction of occurrence-year claims closed in the j-th year, j=1 through M,

$r(t)$ = the claim frequency trend in year t, or $N(t)/N(t-1)$.

We can first use the number of claims closed in year T to estimate N (T+1) by a set of M equations:

$$n(T-j+1,T) = \frac{N\,(T+1) \cdot g(j)}{\prod_{k=1}^{j} r(T-k+2)}$$

or, if the claim frequency trend is reasonably constant,

$$n(T-j+1,T) = N(T+1) \cdot g(j) \cdot r^{-j}$$

The following table illustrates the estimation of N(1979) and the set of gs for various assumptions of a constant frequency trend. Claims closed in 1978 in the jth year from date of occurrence were equally divided between the jth and (j+1)th years preceding 1979. The sensitivity of the projected claim volume to the assumed frequency trend is readily apparent.

			Estimated closing pattern for 1979 occurrence year:			
j	Paid Claims Closed in 1978	r=	1.00	1.05	1.10	1.15
1	441		.059	.049	.040	.032
2	916		.123	.106	.091	.077
3	998		.134	.121	.109	.096
4	1,194		.160	.152	.143	.133
5	1,308		.175	.175	.172	.167
6	1,047		.140	.147	.152	.154
7	676		.090	.100	.108	.114
8	388		.052	.060	.068	.075
9	200		.027	.033	.039	.045
10	112		.015	.019	.024	.029
11	57		.008	.010	.013	.017
12	35		.005	.007	.009	.012
13	24		.003	.005	.007	.009
14	18		.002	.004	.006	.008
15+	61		.008	.013	.021	.032
Projected number of claims arising from 1979 occurrences			7475	9537	12,221	15,743

Source: NAIC Malpractice Claims, Vol. 2, No. 2 (1980).

Once the rs have been specified, the above M expressions give us M+1 unknowns. Since the sum of the gs equals 1.000, solutions can be found for N(T+1) and each g(j), j=1,2,...,M.

Now we divide the range of claim sizes into L intervals using a sequence d(0), d(1), d(2), ..., d(L) where d(0) = 0 and d(L) is a coverage limit or else d(L-1) is some practical bound and d(L) is infinity. then we define:

C(k) = average claim cost between d(k-1) and d(k),

P(k|j)=probability of a claim closed in the jth year having a cost between d(k-1) and d(k),

P(j)=probability of a claim closing in the jth year,

Y(t)=total claim costs for occurrence year t,

S(t)=claim size trend in year t.

A straight forward algebraic construction of Y(T+1) is

$$Y(T+1)= N(T+1) \sum_{k=1}^{L} C(k) \sum_{J=1}^{M} P(k|j) \cdot P(j).$$

Ordinarily, P(j) = g(j). Evaluating each P(k|j) and C(k) from closed claim data would begin by examining the distribution of claims closed in year T for each of the latest M occurrence years. This subset of claims for each occurrence year may be more homogeneous then the

whole and more likely to follow a theoretical pattern such as a log-normal distribution.

If a density function can be found to describe the size of claims closed in year T for occurrence year T-j+1, then $P(k|j)$ can be evaluated by the definite integral from a(k-1) to a(k), where

$$a(k) = \frac{d(k)}{\prod_{k=1}^{j} s(T-k+2)}$$

of, if the size trend is reasonably constant,

$$a(k) = d(k) \cdot s^{-j}$$

A first approximation for C(k) may be the average of closed claim amounts between d(k-1) and d(k). If L is large, then that may be sufficiently precise. Otherwise the effect on the average in each interval from a translation of the density functions could be determined using modern programmable calculators.

Those who think continuously may readily observe that if the density functions can be generalized to a joint function of both claim size and year j, then

$$Y(T+1)=N(T+1)\int_0^M \int_0^{d(L)} x\, f(XS^{-j}, j)\, dxdj$$

If the size of a claim is independent of the interval from incident to disposition, then

$f(x,j) = f(x|j)g(j) = f(x)j(j)$ and so

$$Y(T+1) = N(T+1) \int_0^M \int_0^{d(L)} x\, f(XS^{-j})g(j)\, dxdj.$$

A great many other algebraic models may be devised which rely on a knowledge of trends and claim size patterns.

Trending Methods

Some closed claim studies (1 and 6) have used trending methods to attempt to reduce the temporal alignment differences by "adjusting" the size of each claim to what might be expected for a common occurrence period. There have been two problems with these methods: (1) They are too simple - relying on elementary curves with only primative measures of significance or none at all, and (2) such techniques assume time passage affects claim costs totally independently of all other factors. The latter problem will be discussed in later sections. Some approaches to resolving the first problem appear here.

Closed claims have been commonly used to indicate claim size trends. Changes in the distribution of these claims among accident years, ranked by maturity, distort the patterns. Better understanding of both size and frequency trends can be gained by displaying closed claim data by closure period and maturity simultaneously. The NAIC publications (3, 4 and 5) give us good illustrations of how this may be done. Due to the obstinance of some insurers, however, we are unable

to observe reliable frequency patterns from this source.

The NAIC maturity and closure period table provides ratios of claim sizes from consecutive closure periods for each maturity range. The measures of variance also shown allows us to use one-way and two-way analysis of variance on the array of ratios to determine if significant patterns are present. Regression methods may be used across closure periods within maturity ranges on either ratios or dollar amounts but only on ratios across maturity ranges within closure periods.

Interpretation of the horizontal and vertical patterns requires some premise of whether the closure period influences costs independently of occurrence periods and maturity. In the simple trending methods, this distinction is overlooked because the time spans for which trend factors make adjustment have the same width whether measured between occurrence dates or between closure dates. If closure periods have an influence, then significant differences observed between maturity ranges within closure periods could mean trend factors should differ between maturities. Otherwise, such differences describe changes in trend ratios

across occurrence periods. A simple exponential trend is appropriate only of no significant differences are observed in the array of ratios. Otherwise, you must interpret the differences before you can select the series of trend factors to apply in an algebraic model such as previously described.

A reasonable way to determine whether costs are influenced by the closure period independently of the other factors is by reviewing correlations between claim costs and external indices for occurrence and closure periods. Again, the NAIC (5) has thoughtfully illustrated how this can be accomplished. Average paid claim costs are arrayed by closure and occurrence periods for various severity of injury ranges. Correlations can be tested for medical indexes, price indexes, and other economic indicators. One tenable theory is that temporary injuries or losses are compensated at actual costs in the period of occurrence while permanent disabilities are compensated with regard to prices and price changes at the time of disposition.

For a line of insurance like medical malpractice, precision may be gained by using trending methods to describe the residual claim cost changes remaining after adjustment using economic indexes and also by

separately reviewing trends by type or severity of loss.

Claim Size Distributions

From the wealth of published material about size distributions come two principal mathematical probability functions: the logrithmic normal distribution and the gamma distribution. The rationale behind using the log-normal is the central limit theorem from the depths of probability theory. This theorem says the sum of items taken from similarly distributed populations tends to be normal. The size of a claim is the product of a great many factors in this multivariate world, so the logarithm of claim size is the sum of many terms. Even if we do not know all the factors, we can still consider whether observed claim size patterns are log-normal. The gamma distribution is a more general one.

Density Function:	$\frac{e^{(\ln x - \mu)^2}}{\sqrt{2\pi}\,\sigma\,x}$	$\frac{c^\gamma}{\Gamma(\gamma)} x^{\gamma-1} e^{-cx}$
Mean:	$e^{\mu + \sigma^2/2}$	γ/c
Variance:	$e^{2\mu+\sigma^2}\left[e^{\sigma^2}-1\right]$	γ/c^2
Mode:	$e^{\mu-\sigma^2}$	$(\gamma-1)/c$
Skewness:	$(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$	$2/\sqrt{\gamma}$

The symbols μ and σ in the log-normal formulas represent the mean and standard deviation of the logarithms. The gamma formulas are determined by two constants, γ and c. For a gamma function to be of interest to us, usually γ must be greater than 1.000, which means that the mean must exceed the standard deviation. Unfortunately, this condition has ruled out gamma distributions for most types of insurance which have been critical problems recently because of the uncertainty of claim amounts as well as frequency.

A simple way to determine whether observed data might reasonably be described as log-normal is to see how well the mean and variance of the logarithms of observed claims fit the formulas above. The hypothetical mean and variance of the logarithms can be found by an iterative process since the sum of the log-mean and half the log-variance equals the logarithm of the object mean. The skewness and kurtosis of the logarithms should each be near zero. Skewness is a measure of asymetry. Kurtosis is a measure of non-normality.

Some computational formulas for skewness and kurtosis are:

$$\text{Skewness} = \frac{\sum(x^3) - 3\sum(x^2)\left(\sum x\right)/n + 2\left(\sum x\right)^3/n^2}{n\left[\frac{\sum(x^2) - \left(\sum x\right)^2/n}{n-1}\right]^{3/2}}$$

$$\text{Kurtosis}=\frac{\sum(x^4)-4\sum(x^3)(\sum x)/n+6\sum(x^2)(\sum x)^2/n^2-3(\sum x)^4/n^3}{n\left[\frac{\sum(x)^2-(\sum x)^2/n}{n-1}\right]^2}-3$$

If your data has fewer than 1000 paid claims, these computations may be more easily done on a programmable calculator than by convincing busy data processing people to give you sums, sums of squares, sums of cubes, and sums of fourth powers of both paid claim amounts and their logarithms with double precision.

Very likely, no theoretical distribution will fit observed insurance claim size data for several reasons such as these: 1) A popular premise is that small claims are overpaid and large claims underpaid. 2) Some groundless claims are paid for amounts less than probable defense costs(nuisance claims). 3) Many claims cluster about certain "target values" due to the need to approximate uncertain costs. 4) Economic factors operating over the occurrence, reporting, or closure period will "blur" the distributions. The latter effect may hinder analysis of accident-year claims as well as closure-year claims. Each is likely to be the sum of a continuum of log-normal distributions which will not be log-normal. (Products of log-normal distributions may be log-normal, but not sums.)

These departures from theoretical patterns can be

simulated on programmable calculators. The log-normal distribution, for instance, can easily be generated from standard random normal number generating routines. Such efforts can be tedious but necessary if the density function cannot be modified or successfully integrated to find theoretical means, variances, and skewnesses to be compared with descriptive statistics from actual data.

The Gini Index

The Gini Index of Concentration is another interesting statistic for comparing distributions. Named for its Italian inventor, the Gini index is a tool used in economics and demographics to measure inequality of distribution.

The associated Lorenz curve, L(x), in our application, represents the fraction of total claim costs which relates to claims closed for $x or less. The Gini index, G, is the ratio of the area between L(x) and an equal distribution curve (a 45-degree line when L is plotted against percentiles) to the total area beneath such curve. If the range of claim sizes is divided into k intervals by a sequence 0, d(1), d(2), ..., d(k) and p(t) is the percentage of paid claims at d(t) or less, then a standard method for calculating the Gini index is:

$$G = 1 - \sum_{t=0}^{k-1} (p(t+1) - p(t)) \left[L(d(t+1)) + L(d(t)) \right] .$$

Since the indexes are linearly related, very likely a great many dramatically different distributions could have the same Gini index (9). Nonetheless, the Gini index can be appropriately applied to a set of distributions understood to be substantially similar. Changes in the coefficient of variance, skewness, kurtosis, modality, or any feature in the shape of a distribution will affect the Gini index.

The Gini index comes from a class of statistics with asymptotically normal distributions, known as "U-statistics". Tests for significance of differences exist beyond the scope and allowed length of this paper (8). Multiplying a distribution by a constant does not change the Gini index, but the index is sensitive to the number and selection of data points used in its construction (11). Our application concerns substantially similar distributions with the same means, so the use of a static set of data points should not distort comparisons of Gini indexes.

The following exhibit illustrates a comparison of claim distributions according to descriptive statistics and Gini indexes. The pattern shown first is for 421 paid claims closed in the second half of 1977 arising from occurrences in the first half of 1974 as reported

COMPARISON OF CLAIM DISTRIBUTIONS

Size of Claim	NAIC Malpractice Claims from first half of 1974 closed during the second half of 1977		Random generation from a log-normal distribution		Random generation from log-normal with small claims overpaid and large claims underpaid	
	Number of claims	Amount of claims	Number of claims	Amount of claims	Number of claims	Amount of claims
$ 1 to $1,999	80	$ 74,080	1,571	$ 1,654,718	226	$ 211,673
2,000 to 4,999	101	308,656	1,804	6,071,217	191	626,524
5,000 to 9,999	64	427,712	1,785	12,976,943	170	1,230,519
10,000 to 19,999	55	716,925	1,698	24,339,768	157	2,212,814
20,000 to 49,999	59	1,671,057	1,738	55,477,202	141	4,461,099
50,000 to 99,999	32	2,029,632	818	57,184,521	63	4,481,277
100,000 to 199,999	19	2,634,312	378	52,642,581	29	4,199,820
200,000 to 499,999	7	1,690,535	169	49,257,642	12	3,063,956
500,000 to 999,999	3	1,713,312	26	17,331,377	10	7,315,137
1,000,000 and over	1	1,400,000	13	22,441,297	1	1,251,388
	421	$12,666,221	10,000	$299,377,266	1000	$29,054,207
Mean		$30,086		$29,938		$29,054
Standard Deviation		92,263		87,727		91,541
Log-Mean		8.89		9.15		8.84
Log-Standard Deviation		1.64		1.54		1.67
Gini Index		.763		.704		.765

by the NAIC (5). A log-normal distribution with the same mean and standard deviation should have a log-mean of 9.14 and a log-standard deviation of 1.53. a random generation of 10,000 log-normal numbers with approximately the same sample mean and standard deviation is shown for comparison. Several smaller simulations strongly suggest the differences in these statistics and in the Gini indexes is more than random.

The higher Gini index for the NAIC data suggests a greater peakedness which might plausibly be explained by the hypothesis that small claims are overpaid and large ones are underpaid. A random generation of 1000 claims from a population with log-mean 8.83 and log-standard deviation 1.70 produced the third pattern shown in the exhibit after claim amounts less than $1000 were amplified by a factor which increased from 1.00 to 2.00 as the amount decreased from $1000 to $0 and the excess portion of claims over $500,000 was multiplied by .75. The differences between this third pattern and that of the NAIC data are well within the bounds of random variation.

DIFFERENCES BETWEEN RISKS

The greatest value of closed claim data is as factual information on the costs of insuring certain

risks. If our fears about temporal alignments can be overcome, we should be anxious to use closed claim data for determining the attributes or classification schemes which distinguish individual risks. Recent accident-year data refined by class or attribute may not disclose differences, or may display false differences, since development factors are typically based on aggregated data.

One way to examine the claims costs effect of a certain attribute, such as smoking for drivers or board certification for physicians, would be to construct separate algebraic models for each data group. But that would be subject to the weaknesses of all the trending methods and would be a laborious task, especially if the attribute has several values.

Multivariate statistical methods can be straight forward and enable the researcher to either control or manipulate several variables at once. Hence, if temporal alignment is feared to be influencing comparative observations from closed claim data, then a sensible remedy should be to include time values in a multivariate analysis. Such methods are able to recognize the interactions of factors, so the earlier criticism of simple

trending methods which assume time affects costs independently of other factors becomes less worrysome.

Multivariate statistical methods have an advantage of precision. The significance of differences observed for any any variable is measured by comparing those differences to measures of "error", "unexplained", or "random" variance. When other factors are present which "explain" additional portions of the variance, then the error variance is reduced. Seen this way, multivariate methods are indispensible for reviewing closed claim data.

In the remaining pages attention is given to analysis of variance and multiple regression. Thorough and understandable discussions of these methods can be found in the references (12 and 14). Brief mention will be made of more advanced methods and their application.

Analysis of Variance

The NAIC studies (4 and 5) have a great many illustrations of analysis of variance. The basic concept is very elementary. Variation between group or cell means is compared against residual or random variation "within" groups. For several factors and several groups, it is analogous to tests using the standard t-statistic for two group means.

The F-test for significance assumes the populations are normally distributed. The computation process assumes

homogeneity of the variances within groups and that the dependent variable has continuous measure with equal intervals. The latter assumption is not of concern in our applications. Non-homogeneity of variance increases the residual variance and makes the F-test more conservative. Simulation models with pure premium distributions - products of non-normal frequency and amount distributions - on a programmable calculator have found non-normality to also reduce F-ratios. Nevertheless, the NAIC analysis (4 and 5) have found several large F-ratios. The conclusion is that analysis of variance with standard F-tests is very robust. Nonparametric analysis of variance methods based on rankings may be used to verify results, but are less powerful for detecting false hypotheses.

The following tables and calculations illustrate the concept of analysis of variance with two independent variables. The illustrated analysis seeks to determine whether the average cost of physicians' malpractice claims differs by type of practice and uses the year of claim disposition as a "control" variable. Occurrence year may be a more natural control or possibly both occurrence year and time required for disposition could

PHYSICIANS AND SURGEONS
MALPRACTICE CLAIMS BY TYPE OF PRACTICE AND YEAR OF CLAIM DISPOSITION

Type of Practice		Closure Year 1975	1976	1977	1978	Total
Institutional	Claims	88	108	86	89	371
	Indemnity	1,420,734	1,313,063	1,607,970	2,695,430	7,037,197
Prof. Corp. or Ptnship	Claims	873	1,370	1,333	1,585	5,161
	Indemnity	18,646,181	27,808,980	33,798,124	54,212,985	134,466,270
Self-Employed	Claims	1,775	2,674	2,457	2,730	9,636
	Indemnity	37,207,032	60,277,675	57,968,550	89,203,413	244,656,670
Employed	Claims	150	201	293	219	863
	Indemnity	2,705,521	3,188,934	7,433,108	5,423,285	18,750,848
Resident	Claims	3	10	8	13	34
	Indemnity	194,333	128,705	116,875	100,411	540,324
Total	Claims	2,889	4,363	4,177	4,636	16,065
	Indemnity	60,173,801	92,717,357	100,924,627	151,635,524	405,451,309

Total sum of squares of raw amounts: 68,739,329,480,645

Source: NAIC Malpractice Claims, Vol. 2, Number 2 (1980).

Computations for Analysis of Variance

Correction from raw amounts to deviations from the mean $= \frac{(405,451,309)^2}{16,065} = 10,232,851,796,051$

Total Sum of Squares = 68,739,329,480,645
-10,232,851,796,051
58,506,477,684,594

Between All Groups Sum of Squares $= \sum_{\text{types}} \sum_{\text{years}} \frac{(\text{indemnity})^2}{\text{claims}} - C$

= 10,684,688,432,954
-10,232,851,796,051
451,836,636,903

Between Types of Practice Sum of Squares $= \sum_{\text{types}} \frac{(\sum \text{indemnity})^2}{\sum \text{claims}} - C$

= 10,264,702,339,317
-10,232,851,796,051
31,850,543,266

Between closure years Sum of Squares $= \sum_{\text{years}} \frac{(\sum \text{indemnity})^2}{\sum \text{claims}} - C$

= 10,621,930,858,946
-10,232,851,796,051
389,079,062,895

Interaction Sum of Squares = 451,836,636,903
- 31,850,543,366
-389,079,062,895
30,907,030,742

Residual Sum of Squares = 58,506,477,684,594
- 451,836,636,903
58,054,641,047,691

FINAL ANALYSIS OF VARIANCE TABLE

	Degrees of Freedom	Mean Squares	F-ratio	level of significance
Between types of practice	4	7,962,635,817	2.203	.066
Between Years	3	129,693,020,965	35.887	.000
Interaction	12	2,575,585,895	.713	.740
Residual	16,045	3,613,959,229		
TOTAL	16,064			

be used in a three-way analysis of variance. Closure year was selected for ease of illustration.

In the final Analysis of Variance Table, the sums of squared deviations from the means ("sum of squares") are divided by the statistical degrees of freedom to achieve "mean squares", which are the estimates of variance used in this process. Each of these is compared with the residual mean squares to determine the level of significance. Note that if closure years had not been included in the analysis, the residual mean squares would have been greater, the F-ratio for types of practice would have been lower, and the level of significance would have been greater. (The level of significance is the probability observed differences could occur randomly.)

Multiple Regression

Multiple regression estimates the magnitude of relations between factors and has more general capability then analysis of variance. Control variables can be included more naturally. If the number of observations in the groups or cells are unequal, multiple regression is preferred. There is an R^2 statistic to describe the portion of total variance "explained" by the set of independent variables and an F-ratio for significance.

However, the calculations are much more extensive. Desk calculators are impractical beyond four or five independent variables.

The desired expression is of the form

$$Y=A+B(1)\ X\ (1)+B(2)\ X\ (2)+\ldots+B(N)\ X\ (N)$$

where Y is the dependent variable, the set of X's is the set of independent variables, and the coefficients A and B(1) to B(N) are found to minimize the squared deviations from the predicted values. The process requires solution of a set of equations:

$$r(1,1)\ b(1)+\ldots+r(1,N)b(N)=r(y,1)$$

$$\vdots \qquad\qquad \vdots$$

$$r(N,1)b(1)+\ldots+r(N,N)b(N)=r(y,N)$$

where r(i,j) is the correlation between X(i) and X(j), r(y,i) is the correlation between Y and X(i), and b(i) is the standardized regression coefficient.

The importance of a single factor X(k) is usually evaluated by the significance of the contribution it makes to R^2. Multiple regression strategies are a modern art form, admirably discussed by Cohen and Cohen (12). For our applications, the preferred strategy apparently is to first include the necessary control variables such as time of occurrence, determine an R^2

for this limited set of independent variables, then add the particular variable of interest such as age, gender, or marital status, and redetermine R^2.

Because of random fluctuations, any variable added to the set of independent variables will always increase R^2. Most researchers prefer to use a corrected or "shrunken" R^2 which is a better estimate of the population R^2. With k independent variables and sample size n, the corrected value is:

$$R_c^2 = 1 - (1-R^2)\ \frac{n-1}{n-k-1}.$$

In the simple case of two independent variables, we can define the simipartial correlation, sr, of Y and X(2) to be the correlation between Y and X(2) not related to X(1). Then,

$$R^2=\frac{r^2(y,1)+r^2(y,2)-2r(y,1)r(y,2)r(1,2))}{1-r^2(1,2)}$$

$$sr^2=(r(y,2)-r(y,1)r(1,2))^2/(1-r^2(1,2))$$

$$F(X(2))=sr^2(n-3)/(1-R^2).$$

Independent variables must be discrete or nominal for analysis of variance, but may be continuous for multiple regression. Continuous variables usually provide the greatest information value.

Discriminant Analysis

A set of independent variables may be used to estimate group membership as the dependent variable.

Most applications discriminate between two groups, but discriminant analysis can be adapted to three or more groups. Some natural applications for insurance are classification of risks and answering claim management questions: Which claims will be paid? Which claims will include law suits? What will be the outcome of arbitration? Which claims will reopen?

Discriminant analysis is becoming recognized as a highly sophisticated risk management tool. As soon as any untoward incident occurs, the particulars may be fed into a discriminant function at a computer terminal and the likelihood of a compensable event is rapidly determined. The risk manager can promptly act to contain the costs. This technology is being introduced at hospitals in various parts of the country. The basic data is from closed claims. The exclusion of incidents which have not produced claims may not seriously reduce the predictive accuracy in many instances. Even if extensive incident data is available, insurance claim costs are clearly necessary in corresponding detail.

The selection of predictive variables is another modern art form. Separate discriminant functions should be constructed for nominal variables such as

gender, marital status, and medical specialty. Astonishingly, negative or highly positive correlations between independent variables increases discriminatory power (13).

After discrimant analysis has been used to predict which claims will result in payment, a natural step is to use multiple regression to estimate the amount of payment for each - a loss reserving method.

A genuine time problem may result if such techniques are based on internal data sources only. For instance, if a hospital constructed and periodically revised its discriminant function for compensable events based on its own data, then its own success at decreasing costs would also decrease the predictive accuracy of the predictive variables. Costs would then increase again until predictive power is reestablished. The insurance industry has cycles like that.

Factor Analysis

Factor analysis is an extremely complex computational methodology for discovering natural dimensions behind a number of simple quantitative measures. Psychological tests, for example, measure qualities by asking a great many questions. Most often researchers are not aware of the fundamental dimensions and must seek to learn these from many simple measures.

This analytic technique may eventually be used to find comparatively few important complex dimensions represented by the several hundred variables in closed claim data collection instruments of recent studies.

DESIGNING CLOSED CLAIM DATA

Underwriting, pricing, loss reserving, claim management, and loss prevention are only separated by brief steps and perspectives. The fact that claim files typically contain only sufficient information to establish coverage, establish defenses, and compute payments should not validly prevent us from seeking information important for other functions.

The data should be designed to answer important questions or test important theories. If the task is so well defined, then questions can be easily imagined relevant to the hypothesis and sample sizes determined from formulas in the statistics books.

Unfortunately, the examples of closed claim studies in the public domain have arisen from crises in various kinds of liability insurance where there have been low frequencies, phenomenal variances, myriads of socio-underwriting theories pointing in all directions and often conflicting, no deadlines for new theories, unresponsive rating systems, and simplistic ratemaking methods.

Because of these situations, few assumptions could be made about expected patterns or variances from them and no specific lists of hypothesis could be prescribed in advance. The first purpose of the closed claim studies has been to provide an understanding of the statistical dimensions and to evaluate the importance of hypotheses as they become expressed. Classical research designs and statistical analysis have had to come second.

The closed claim data collection instruments have had to be comprehensive. With no reliable knowledge of what factors may be importantly related to claim occurrences or costs, or of the nature of those relations, or of the variances from such, no sampling techniques could be intelligently chosen and no data item could be dismissed. Hence, the forms have been designed to describe as completely as possible the insured, the claimant, the relations of the insured and the claimant, the incident, the relations of the insured and the claimant to the incident, other persons and factors related to the incident, the loss endured by the claimant, the paths taken to final disposition, and the resulting indemnities and expenses. Then hopes have been expressed that the forms were not so formal as to preclude other significant factors from being discovered.

Actuaries bent on substituting fact for impression should understand this background and learn from it how closed claim data can be an imaginative source for designing responsive rating systems, observing trends, and answering important questions before crisis situations occur.

REFERENCES

CLOSED CLAIM STUDIES

1. All-Industry Committee, Special Malpractice Review: 1974 Closed Claim Survey - Technical Analysis of Survey Results, Insurance Services Office, New York, November 1976.
2. McWhorter, Archer, Jr., "Drawing Inferences from Medical Malpractice Closed Claim Studies", Journal of Risk and Insurance, March 1978, pp. 79-94.
3. NAIC Malpractice Claims, National Association of Insurance Commissioners, Vol. 1, Number 4, May 1977.
4. NAIC Malpractice Claims, National Association of Insurance Commissioners, Vol. 2, Number 1, December 1978.
5. NAIC Malpractice Claims, National Association of Insurance Commissioners, Vol. 2, Number 2, 1980.
6. Product Liability Closed Claim Survey: A Technical Analysis of Survey Results, Insurance Services Office, New York 1977.

GINI INDEX

7. Gastwirth, Joseph L., "The Estimation of the Lorenz Curve and Gini Index", The Review of Economics and Statistics, August 1972, pp. 306-316.
8. Hoeffding, Wassily, "A Class of Statistics with Asymptotically Normal Distribution", The Annals of Mathematical Statistics, Vol. XIX, 1948, pp. 293-325.
9. Morrow, James S., "Toward A More Normative Assessment of Maldistribution: The Gini Index", Inquiry, Vol. XIV, September 1977, pp. 278-292.
10. Paglin, Morton, "The Measurement and Trend of Inequality: A Basic Revision", The American Economics Review, Vol. LXV, Number 4 (September 1975), pp. 598-609.
11. Petersen, Hans-Georg, "Effects of Growing Incomes on Classified Income Distributions, the Derived Lorenz Curves, and Gini Indices", Econometrica, Vol. 47, Number 1 (January 1979).

ANALYTICAL METHODS

12. Cohen, Jacob and Patricia Cohen, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, John Wiley & Sons, New York, 1975.

13. Eisenbeis, Robert A., "Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics", The Journal of Finance, Vol. XXXII, Number 3 (June 1977), pp. 875-900.
14. Kerlinger, Fred N., Foundations of Behavioral Research (Second Edition), Holt, Rinehart, and Winston, New York, 1973.