ON THE STATISTICAL ESTIMATION OF COSTS OF CLAIMS

B. Ajne
Stockholm

1. For a number of reasons it is important for an insurance company to estimate the claims costs of a year within the different branches of non-life insurance as soon as possible after the end of the year. The claims cost of a year is hereby defined as the total cost, before taking reinsurance into account, of all claims generated by events that have occurred during the year. When the estimation has to be done, part of these claims will be reported and closed, others will be reported and still open, and the remaining ones will be incurred but not yet reported. The total cost of the claims is defined as the sum of all payments that have been made or will be made on account of the claims. Thus, in this definition no regard is paid to interest, i.e. no discount factors are applied to payments to be made in the future.

Instead of considering a year, we could consider an arbitrary period of twelve consecutive months. The estimation problem is the same, and estimates of the claims costs of consecutive twelve months periods will allow a closer following up of trends and yield predictions for the present year.

2. For estimates to be available quickly, it is necessary that the estimation procedure be founded on data that are available immediately at the end of the year or the latest twelve months period. This means a.o. that for the bulk of the open claims, individual estimates of reserves by claims adjusters are out of the question. In other words, the estimation procedure has to be basically of a statistical character. In addition, for continuous estimates to be produced it has to be well adapted to electronic data processing. Indata to the procedure have to be stored in the memories of the computer.

3. For simplicity of language we speak in the following only of years, it being silently understood that most of the reasoning is equally valid for arbitrary twelve months periods.

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press

Variables, the values of which could be available immediately after the end of a year, are e.g.

V_1 = Date of the year, e.g. 1972.
V_2 = Number of "small" claims reported during the year, irrespective of year of occurrence.
V_3 = Claims amount for "large" claims occurred and reported during the year, as estimated by claims adjusters.
V_4 = Claims payments during the year to claims occurred during the year.
V_5 = Number of "large" claims occurred and reported during the year.

In my company the claims adjuster has to contribute a judgement whether a reported claim is large or small. If large, he also has to give an estimated amount for the claim, with the exception of some types of liability claims. If major changes take place he is also required to modify his estimated amount. The lower limit for large claims is in general 50,000 sw. crowns (approx £ 5,000).

The idea behind the inclusion in V_2 of all small claims reported, whether occurred during the year or not, could be to balance the incurred but not reported small claims of the year by the reported small claims from earlier years. The corresponding effect for large claims could in most branches be achieved by keeping the files open for these claims during a short time interval after the end of the year.

4. In a system we have been working with in my company, variables V_1 through V_5 are recorded, actually on a monthly basis, for each branch and within branches for each type of claims (e.g. fire, burglary, water damage etc). The estimated cost for a certain type of claims is in most cases computed according to the equation

$$Y = a_2(V_1) \cdot V_2 + V_3 \tag{1}$$

and the estimated costs Y are summed over types of claims to branches, and over branches to those higher levels that may be interesting for management to look at.

The estimate (1) belongs to the class of estimates that are linear in variables V_2 through V_5. However the coefficient of V_2, $a_2(V_1)$, depends on V_1—the year under study. The coefficient can be regarded as a predicted average claims amount for small claims. The prediction is based on the average claims amounts during previous years for the type of claims in question. Before being fed into the system it is updated with the guessed effects of inflation and other circumstances that may influence the amount per claim.

5. The method with predicted average amounts for small claims naturally requires that the statistical variation of the average claims amount not be too large. As a rule of thumb one might take that the statistical variation be at most of the order of magnitude of a "normal" rate of inflation. For various types of claims within property insurance the coefficient of variation, i.e. standard deviation through mean value, for the claim distribution seems to be of the order of magnitude 1.5 to 2.5 (for motor insurance probably smaller), cf table in appendix 1. If n denotes the expected number of small claims per year, the coefficient of variation for the average claims amount per year is approximately $\sqrt{n}$ times smaller. Choosing e.g. the value $2/\sqrt{n}$ for this coefficient of variation and denoting by i the desired upper limit for the statistical variation (interpreted as the coefficient of variation), we get the equation for n

$$2/\sqrt{n} = i \tag{2}$$

E.g. $i = .10$ gives us $n = 400$ as the expected number of small claims per year, that is desired in order to apply the method to a certain type of claims. Correspondingly, $i = .05$ gives us a desired n-value of 1,600.

6. As already noted, the estimate (1) is linear in V_2 and V_3 with the coefficient of V_2 depending on time. If this dependence is chosen such that $a_2(V_1)$ follows some established price index, and if the amounts Y and V_3 are measured in fixed money-value according to this index, $a_2(V_1)$ will reduce to a constant and we will have a proper linear estimate in variables V_2 and V_3. This leads us to the idea of investigating the scope of linear estimates in variables V_2 through V_5 when all amounts are measured in fixed money-value according to some price index.

In appendix 2 are shown four sets of data. Each set consists of observed values of the "independent" variables V_2 through V_5 and the "dependent" variable Y during five consecutive years. A longer period of time had of course been desirable but for the

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press

moment being this was the longest period available with consistently defined data. The first two sets show fire and theft-and-burglary claims within branch B_1 while the last two comprise fire and water damage claims within branch B_2. All amounts are in fixed money-value, for branch B_1 according to one established price index and for branch B_2 according to a second such index. Each set of data was submitted to regression analysis. The results of the analyses together with some brief comments and explanations will be given in the next two paragraphs.

7. For the analysis standard programs for stepwise and multiple regression were used, to be found in *IBM*/360 Scientific Subroutines Package. The programs were used in a conversational form adapted to a direct access terminal.

The stepwise regression starts by choosing that variable among V_2, V_3, V_4 and V_5 which has the numerically largest correlation with Y. Equivalently, this is the variable V_i for which the residual sum of squares (Y_k, V_{ki} denote the observed values on Y, V_i for year k; $\bar{Y}$ and $\bar{V}_i$ denote the arithmetic means of the observations for the five-year period; b_i denotes the observed regression coefficient of Y on V_i)

$$\sum_k [Y_k - \bar{Y} - b_i(V_{ki} - \bar{V}_i)]^2$$

is as small as possible.

In each successive step that remaining variable is chosen which, together with the variables already chosen, yields the smallest residual sum of squares. If e.g. in step no. 1 variable V_2 was chosen, step no. 2 will pick that variable V_i; $i = 3, 4, 5$; for which

$$\sum_k [Y_k - \bar{Y} - b_2(V_{k2} - \bar{V}_2) - b_i(V_{ki} - \bar{V}_i)]^2$$

is as small as possible.

The residual sums of squares in the successive steps: $R_0 = \sum_k (Y_k - \bar{Y})^2$, R_1, R_2, ... will form a decreasing sequence. In each step a testvariable is computed which measures the significance of the reduction performed by the variable included. Assuming standard normal theory, this testvariable follows an F-distribution.

Below, the following additional terms and symbols are used in every regression situation:

$s(Y)$	= square root of $R_0/4$ = observed standard deviation of Y.
Variance reduction	= (R_0 — residual sum of squares of the variables included in the regression) as a percentage of R_0.
s red $)Y)$	= estimated standard deviation of (Y — the regression expression) = estimated standard error when using the regression expression to predict Y.

Finally, the low number of observations in the material means a low number of degrees of freedom when fitting one or more of the variables V_2 through V_5.

Correlations are thus a priori likely to be high, estimated standard deviations have large statistical errors and extrapolation into the future is hazardous. Anyhow, what follows is at least a piece of linear descriptive statistics.

8. Unless otherwise stated all amounts are expressed in 1,000 sw. crowns.

Branch B1, fire

$$\bar{Y} = 5{,}788, \; s(Y) = 1{,}213.$$

Stepwise regression:

Step no. 1. V_2 selected.
Variance reduction 84.3%, $s_{red}(Y) = 554$.
V_2 significant at the 5% level.
Regression: $Y = -1{,}108 + 3.480\, V_2$.

Step no. 2. V_3 selected.
Variance reduction 96.9%, $s_{red}(Y) = 304$.
V_3 not quite significant at the 10% level—but we include it'
Regression: $Y = -685 + 2.829\, V_2 + .6440\, V_3$.

We stop here. Using the regression above we find (Y_{est} denotes values computed from the regression)

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press

Year	1	2	3	4	5
Y_{obs} (millions)	4.8	4.6	5.4	6.8	7.3
Y_{est} (millions)	4.7	4.9	5.2	6.6	7.5

The variables selected are V_2 and V_3, i.e. just those two variables that are used in the estimate (1). However, the regression comprises the constant term —685. One feels uneasy about having this negative constant in an equation between positive variables. It might even be argued that $V_2 = V_3 = 0$ should very likely imply $Y = 0$, i.e. no constant term at all should occur. The regression without constant term is

$$Y = 2.473\, V_2 + .6693\, V_3$$

which produces the series of estimated values

Year	1	2	3	4	5
Y_{est} (millions)	4.9	5.0	5.2	6.6	7.3

Variance reduction 95.7%; s_Y(red) = 291.

The estimate is practically as good as the one with constant term. In the constant money-value chosen, it assigns roughly 2,500 sw. crowns to each small claim reported, to which should be added 2/3 of the estimated large claims amount.

Branch B1, theft-and-burglary

$$\bar{Y} = 6{,}201, \quad s(Y) = 2{,}362.$$

Stepwise regression:

Step no. 1. V_4 selected.
Variance reduction 99.8%, $s_{red}(Y) = 129$.
V_4 significant at the 0.1% level.
Regression: $Y = 60 + 1.283\, V_4$.

Year	1	2	3	4	5
Y_{obs} (millions)	4.1	4.4	5.5	7.1	9.9
Y_{est} (millions)	4.2	4.3	5.5	7.2	9.8

A very good fit. If one dislikes the small constant term one could as well use

$$Y = 1.295\, V_4$$

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press

i.e. the ratio of final claims cost to claims paid during the year of occurrence is very stable at 1.3:1.

Branch B2, fire

$$\bar{Y} = 9{,}029.\ s(Y) = 1{,}606.$$

Stepwise regression:

Step no. 1. V_4 selected.
Variance reduction 95.2%, $s_{red}(Y) = 409$.
V_4 significant at the 1% level.
Regression: $Y = 3{,}711 + 1.132\ V_4$.

Step no. 2. V_3 selected.
Variance reduction 99.9%, $s_{red}(Y) = 57$.
V_3 significant at the 1% level.
Regression: $Y = 2{,}363 + .6430\ V_3 + .7004\ V_4$.

No difference between observed and estimated claims costs, as expressed in millions to one decimal place. However, the constant term is pretty large. Taking it away, results in the regression

$$Y = 1.257\ V_3 + .5031\ V_4$$

and a considerable decrease in variance reduction (to 91.8%) and increase in $s_{red}(Y)$ (to 530). For observed and estimated Y-values we get

Year	1	2	3	4	5
Y_{obs} (millions)	8.0	10.0	7.4	8.4	11.3
Y_{est} (millions)	7.6	10.3	7.3	7.8	11.8

The estimate is only slightly better than the corresponding estimate with V_4 replaced by V_2 which reads

$$Y = 2.232\ V_2 + 1.389\ V_3.$$

Branch B2, water damage

$$\bar{Y} = 11{,}263,\ s(Y) = 2{,}333.$$

Stepwise regression:

Step no. 1. V_2 selected.
Variance reduction 97.1%, $s_{red}(Y) = 459$.
V_2 significant at the 1% level.
$Y = -12{,}018 + 5.554\ V_2$.

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press

Year	1	2	3	4	5
Y_{obs} (millions)	9.1	9.4	10.5	13.0	14.4
Y_{est} (millions)	9.3	9.8	9.9	12.7	14.5

Next variable picked out—V_5—is not significant.

The estimate has a rather good fit but is of course useless because of the large negative constant. No satisfactory estimate without constant term was found. However, as seen from at least three of the four examples presented, claims costs have increased considerably more rapidly than the price index used to adjust them. The price indices used are of the type of consumer's price index and index of materials which, in Sweden as in many other countries, increase slower than e.g. index of wages. Average claims costs tend to increase at a rate somewhere in between the rates for these indices. This means that the coefficient $a_2(V_1)$ in estimate (1) is not neutralized in its dependence on V_1 by measuring Y and V_3 in relation to a consumer's prices type of index. It would rather still increase at a rate of, say, 3% a year. This means that the estimate (1) would be of the form

$$Y = a_2(1.03)^{V_1-5}\ V_2 + V_3\ (V_1 = 1, \ldots 5)$$

i.e. to get a time-independent regression the variable V_2 should be replaced by

$$V_2' = (1.03)^{V_1-5}\ V_2.$$

Using this variable instead in our regression analysis we get the following regression of Y on V_2' and V_3, without constant term

$$Y = 2.708\ V_2' + 3.807\ V_3$$

with variance reduction 96.5%, $s_{red}(Y) = 507$ and the following comparison between observed and estimated values

Year	1	2	3	4	5
Y_{obs} (millions)	9.1	9.4	10.5	13.0	14.4
Y_{est} (millions)	9.2	10.0	10.1	13.2	14.0

9. The findings of the foregoing paragraph may be briefly summarized as follows.

Variable V_2—number of small claims reported during the year of

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press

occurrence—has its strongest position for the water damage claims. This is quite natural, as large claims are of relatively little importance for this type of claims. Furthermore, claims payments during the year of occurrence (V_4), has at least in my company a bad reputation as predictor of the final claims cost. The distribution of water damage claims over the year is fairly dependent on weather conditions (cold or warm winter, cold or warm autumn) which has a certain influence on V_2 but a still stronger one on V_4.

Variable V_3—estimated amount of large claims—quite naturally has its strongest position for fire damage cost, strongly dependent as this is on the large claims result. This variable was picked out in both the stepwise regressions for fire damage claims.

Variable V_4—claims paid during year of occurrence—fits the burglary claims especially well. These have fairly short claim settlement durations and the proportion of incurred but not reported claims is not very high. Also, variable V_2 comes to some disadvantage during the period studied because of the rapidly rising average claims cost, cf the discussion for water damage in the foregoing paragraph.

Variable V_5, finally, i.e. number of large claims reported during the year of occurrence, has not had much success in the material presented. Its chief use in my company is for third party liability personal injury claims, where it is a little more meaningful to speak of an average amount for large claims than it is for fire claims, and where the great difficulty of giving quick estimates of individual claims makes variable V_3 very hard to use.

APPENDIX I

Observed values for the coefficient of variation of claims less than 50.000 sw. crowns.

Branch	Type of claims	Year(s) of experience	Coeff. of variation
Liability insurance for industry and enterprises	liability	1968	2.1
Industrial fire excl loss of profits	fire	1968	1.4
Burglary insurance for industry and enterprises	burglary	1968	1.9
		1969	1.7
Combined shop insurance	mixed	1968	2.1
		1969	1.8
Water damage insurance for industry and enterprises	water damage	1968	1.6
		1969	1.6
Comprehensive home-owner's	fire	1968	2.4
	burglary	1968	1.9
Combined small houses	fire	1965-66	2.4
	water-damage	1965-66	1.6
Combined mansions	fire	1966-68	1.7
	water-damage	1966-68	1.4

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press

APPENDIX 2

Claims statistics for a five-year period. Amounts in 1,000 sw crowns, adjusted with respect to insex.

Branch and type claims	Year	No of small claims reported	Estimated amount for large claims	Claims paid during year of occurrence	No of large claims reported	Total claims cost
		(V_3)	(V_3)	(V_4)	(V_5)	(Y)
B_1, fire	1	1,635	1,227	3,359	9	4,788
	2	1,738	1,043	3,427	9	4,617
	3	2,020	335	4,235	4	5,418
	4	2,062	2,207	3,782	20	6,783
	5	2,453	1,916	5,419	19	7,332
B_1, burglary and theft	1	3,072	52	3,209	1	4,081
	2	3,132	301	3,314	2	4,446
	3	3,462	417	4,241	4	5,532
	4	4,190	134	5,658	2	7,067
	5	5,865	503	7,609	4	9,880
B_2, fire	1	912	4,503	3,899	25	8,048
	2	752	5,936	5,550	21	10,029
	3	615	4,612	2,928	24	7,359
	4	680	4,375	4,640	22	8,403
	5	898	6,814	6,478	29	11,305
B_2, water damage	1	3,839	51	3,734	1	9,050
	2	3,921	153	4,060	2	9,380
	3	3,943	85	3,872	1	10,491
	4	4,480	464	5,922	5	13,034
	5	4,777	366	6,224	6	14,360

https://doi.org/10.1017/S0515036100006024 Published online by Cambridge University Press