JSS 19 (3) (1971) 147-154

CLAIM FREQUENCY ANALYSIS IN MOTOR INSURANCE

by

T. GRIMES

Introduction

IT is a characteristic of motor insurance data that policies can be classified by many different factors, for example age of policyholder, No Claims Discount (N.C.D.), status, amount of voluntary excess and so on. The most 'complete' data on the experience of a portfolio would be a classification of each policy by the various rating factors, and by the number of claims which occur on that policy in the period of exposure. It will be shown that provided certain conditions are fulfilled, useful results can be obtained with far less information.

The method used

The model selected assumes that the claim frequency of a policy (that is, the average number of claims per year) is of the form

$$\mu+\alpha_i+\beta_j+\gamma_n+\ldots.$$

where α, β, γ, . . are rating factors (age of policyholder, N.C.D., status, etc.) and the subscripts range over the different values of the factor (e.g. if N.C.D. has five levels, the subscript of that factor can take five levels). The μ in the formula is a parameter whose use will be explained later.

It may be thought that this model is too simple to represent such data adequately. In fact, with a large number of factors a more complex model would be difficult to interpret, but it is possible to introduce compound factors (of the form $[\alpha\beta]_{ij}$) if necessary. The simple form of the model has been found to be adequate in practice.

The usual method of analysing data of the type described would be multiple regression, assuming that the observed claim frequency differed from the mean by a normally distributed random variable with zero mean and constant variance. This assumption is unlikely to be true in this case, as the 'observed values' of numbers of claims are all integers, but the resulting method of fitting still gives good results.

The problem has been considered by Feldstein (2), his results being reproduced here. First write the vector of parameters as

$$\boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \gamma_1 \\ \gamma_2 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

Each policy has a mean value of its claim frequency which is of the form $\mathbf{x}^T\boldsymbol{\beta}$ where $\mathbf{x}$ is a vector of zeros and ones, and the superscript T denotes transposition. The actual number of claims, y, is equal to $\mathbf{x}^T\boldsymbol{\beta}+u$ where u is $N(o,\sigma^2)$. The vector of observations, each observation being the number of claims on a policy, can thus be written

$$\mathbf{y} = X\boldsymbol{\beta}+\mathbf{u}$$

where $\mathbf{y}$ and $\mathbf{u}$ are vectors, and X is a matrix.

The least-squares solution of this equation for an estimate, is:

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^T\mathbf{y}$$

This can be written

$$\hat{\boldsymbol{\beta}} = N^{-1}\mathbf{C} \quad . \quad . \quad . (1)$$

where $N = X^T\mathbf{x}$, and $\mathbf{C} = X^T\mathbf{y}$. N is a matrix which has, as its p, qth element the number of policies which have a 1 in both the pth and qth positions of their $\mathbf{x}$-vectors. $\mathbf{C}$ is a vector which has, in its pth position, the total number of claims on all policies with a 1 in the pth position of their $\mathbf{x}$-vector. Thus, if $n_{ijk\ldots}$ is the total exposure in

'cell' $i,j,k,\ldots$ and $C_{ijk\ldots}$ is the total claims on policies in that cell, and:

$$N_{***\ldots} = \sum_{ijk} n_{ijk\ldots}, \quad C_{***\ldots} = \sum_{ijk} c_{ijk\ldots}$$
$$N_{i**\ldots} = \sum_{jk} n_{ijk\ldots}, \quad C_{i**\ldots} = \sum_{k} c_{ijk}$$
$$N_{ij*\ldots} = \sum_{k\ldots} n_{ijk} \quad .$$

etc.

then:

$$N = \begin{bmatrix} N_{***\ldots} & N_{1**\ldots} & N_{2**\ldots\ldots} & N_{*1*\ldots\ldots} \\ N_{1**\ldots} & N_{1**} & 0 & N_{1\,1*\ldots\ldots} \\ N_{2**\ldots} & 0 & N_{2**} & N_{2\,1*\ \ldots} \\ \vdots & & & \\ N_{*1*} & N_{1\,1*\ldots} & & N_{*1*\ldots}\ 0 \\ N_{*2*\ldots} & N_{1\,2*} & & 0\ N_{*2*} \\ \vdots & & & \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} C_{***} \\ C_{1\,2*} \\ C_{2**} \\ \vdots \\ C_{*1*} \\ C_{*2*} \\ \vdots \end{bmatrix}$$

This implies that the only information required to solve the regression equations is the total claims for each level of each factor, and the total exposure for each level of each pair of factors. A complete breakdown of the data by individual policies is not required, and this considerably simplifies data collection (and handling).

The above theory is derived on the assumption that each policy has equal exposure. If this is not true, the distributional assumptions may go astray, but it is unlikely that too much bias will be introduced.

A numerical example

These data are for the policies of a single office exposed during the October–December Quarter of 1967. The data relate to all policies with comprehensive cover with no voluntary excess, insuring for social, domestic and pleasure use, a small vehicle registered in 1965 or later and garaged in a particular rating area. The data are classified by age of policyholder and N.C.D., all other possible factors being ignored. The exposure of 3575 policy-years is one quarter of the average of two censuses, taken at 31.9.67 and 31.12.67. The exposure and claims are:

		Exposure					
	N.C.D.	*Age of policyholder*					
	years	17–22	23–26	27–65	66–90	Total	*Claims*
Exposure	0	122	50	293	10	475	115
	1	79	48	340	8	475	87
	2	46	36	347	11	440	67
	3	23	30	254	5	312	48
	4 or more	21	77	1680	95	1873	202
	total	291	241	2914	129	3575	519
Claims		79	45	379	16	519	

The equation for $\hat{\boldsymbol{\beta}}$ is thus:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 3575 & 475 & 475 & 440 & 312 & 1873 & 291 & 241 & 2914 & 129 \\ 475 & 475 & 0 & 0 & 0 & 0 & 122 & 50 & 293 & 10 \\ 475 & 0 & 475 & 0 & 0 & 0 & 79 & 48 & 340 & 8 \\ 440 & 0 & 0 & 440 & 0 & 0 & 46 & 36 & 347 & 11 \\ 312 & 0 & 0 & 0 & 312 & 0 & 23 & 30 & 254 & 5 \\ 1873 & 0 & 0 & 0 & 0 & 1873 & 21 & 77 & 1680 & 95 \\ 291 & 122 & 79 & 46 & 23 & 21 & 291 & 0 & 0 & 0 \\ 241 & 50 & 48 & 36 & 30 & 77 & 0 & 241 & 0 & 0 \\ 2914 & 293 & 340 & 347 & 254 & 1680 & 0 & 0 & 2914 & 0 \\ 129 & 10 & 8 & 11 & 5 & 95 & 0 & 0 & 0 & 129 \end{bmatrix}^{-1} \begin{bmatrix} 519 \\ 115 \\ 87 \\ 67 \\ 48 \\ 202 \\ 79 \\ 45 \\ 379 \\ 16 \end{bmatrix}$$

This (formal) equation cannot be solved immediately because the matrix is singular. This is generally true of this type of problem—the matrix has nullity equal to the number of factors (2 in this example) because the sums of the rows corresponding to any one factor are

equal to the first row (rows 2–6 and 7–10 here). Feldstein (2) suggests striking out the rows and columns corresponding to α_1, β_1, γ_1, etc. (implicitly setting them to zero) and solving the resulting equations. He then calculates 'adjusted deviations', in effect requiring that:

$$\left.\begin{array}{l} \sum_i \alpha_i N_{i**\ldots} = 0 \\ \sum_j \beta_j N_{*j*\ldots} = 0 \\ \qquad \text{etc.} \end{array}\right\} \qquad \ldots \qquad (2)$$

Equations (1) and (2) together imply that

$$\mu = \frac{C_{***\ldots}}{N_{***\ldots}} \text{ (the overall claim frequency)}$$

These two operations can be performed in one step by removing rows corresponding to α_1, β_1, etc. and adding rows corresponding to equations (2). Our example becomes:

$$\hat{\beta} = \begin{bmatrix} 3575 & 475 & 475 & 440 & 312 & 1873 & 291 & 241 & 2914 & 129 \\ 475 & 0 & 475 & 0 & 0 & 0 & 79 & 48 & 340 & 8 \\ 440 & 0 & 0 & 440 & 0 & 0 & 46 & 36 & 347 & 11 \\ 312 & 0 & 0 & 0 & 312 & 0 & 23 & 30 & 254 & 5 \\ 1873 & 0 & 0 & 0 & 0 & 1873 & 21 & 77 & 1680 & 95 \\ 241 & 50 & 48 & 36 & 30 & 77 & 0 & 241 & 0 & 0 \\ 2914 & 293 & 340 & 347 & 254 & 1680 & 0 & 0 & 2914 & 0 \\ 129 & 10 & 8 & 11 & 5 & 95 & 0 & 0 & 0 & 129 \\ 0 & 475 & 475 & 440 & 312 & 1873 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 291 & 241 & 2914 & 129 \end{bmatrix}^{-1} \begin{bmatrix} 519 \\ 87 \\ 67 \\ 48 \\ 202 \\ 45 \\ 379 \\ 16 \\ 0 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \cdot 145175 \\ \cdot 078525 \\ \cdot 028526 \\ \cdot 004337 \\ \cdot 008380 \\ -\cdot 029563 \\ \cdot 086423 \\ \cdot 027329 \\ -\cdot 010540 \\ -\cdot 007923 \end{bmatrix} = \begin{array}{ll} \mu & \\ 0 & \\ 1 & \\ 2 & \text{N.C.D.} \\ 3 & \\ 4 & \\ 17\text{–}22 & \\ 23\text{–}26 & \text{Age} \\ 27\text{–}65 & \\ 66\text{–}90 & \end{array}$$

For comparison, the 'parameters' obtained by taking only the marginal totals

$$\left[\frac{C_{i**..}}{N_{i**..}} - \frac{C_{***..}}{N_{***..}}\right]$$

etc., are:

$$\begin{bmatrix} \cdot145175 \\ \cdot096930 \\ \cdot037983 \\ \cdot007098 \\ \cdot008671 \\ -\cdot037327 \\ \cdot126303 \\ \cdot041547 \\ -\cdot015113 \\ -\cdot021144 \end{bmatrix} = \mu \begin{cases} \left.\begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix}\right\} \text{N.C.D.} \\ \left.\begin{matrix} 17\text{–}22 \\ 23\text{–}26 \\ 27\text{–}65 \\ 66\text{–}90 \end{matrix}\right\} \text{Age} \end{cases}$$

These parameters have the property (2) required, but are obviously more 'extreme' than the solutions of the regression equations. This is because age and N.C.D. are related (young policyholders having low N.C.D., and so on), and the marginal total parameters are overallowing for the variation. The multiple regression method corrects for such associations in the exposure.

Actual claims and predicted claims (*m.r.* parameters) are shown below:

Predicted claims

		Age 17–22	23–26	27–65	66–90	*Total*
	0	37·8	12·6	62·5	2·2	115·0
	1	20·5	9·6	55·5	1·3	87·0
N.C.D.	2	10·9	6·4	48·2	1·6	67·0
	3	5·5	5·4	36·3	·7	48·0
	4 or more	4·2	11·0	176·5	10·2	202·0
	total	79·0	45·0	379·0	16·0	519·0

Actual claims

		Age 17–22	23–26	27–65	66–90	*Total*
	0	45	9	59	2	115
	1	18	16	53	0	87
N.C.D.	2	8	8	48	3	67
	3	6	3	39	0	48
	4 or more	2	9	180	11	202
	total	79	45	379	16	519

(The method of fitting always gives exact equality of total actual and predicted claims for each factor.)

When considering these results, it should be remembered that this is a two-factor example of a method intended for multi-factor problems, and that the two factors chosen are probably those most associated on exposure.

Variances of the estimates

The question naturally arises: if we can solve for the parameters without obtaining data as complete as we usually require for multiple regression, what information are we losing? The answer is that we cannot, without making further assumptions, estimate the variances of the estimated paramaters $\hat{\boldsymbol{\beta}}$.

If the random errors are assumed to be $N(o,\sigma^2)$, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 N^{-1}$$

From this, we can see that variance var($\hat{\boldsymbol{\beta}}$) is proportional to N^{-1}.

It would be possible to estimate σ^3 by

$$(\mathbf{y}^T\mathbf{y} - \hat{\boldsymbol{\beta}}^T X^T\mathbf{y})/K$$

(where K is the number of degrees of freedom of the estimate), if we had sufficient information to calculate it. The assumption of normality is so suspect that it is probably not worth considering this case further.

It can be shown that if the claims on a policy are a Poisson variable, we have

$$\text{var}(\hat{\boldsymbol{\beta}}) = N^{-1}CN^{-1}$$

where C is a matrix of the same form as N, but with the expected claims in place of the exposure. This can be estimated by using actual, rather than expected (which are unknown) claims. Even this estimate requires a breakdown of actual claims by pairs of factors.

Minimum χ^2 methods

If

$$p_{ijk\ldots} = \mu + \alpha_i + \beta_j + \gamma_k + \ldots,$$

it is possible to define a χ^2 function

$$\chi^2 = \sum_{ijk} \frac{(n_{ijk\ldots} p_{ijk\ldots} - c_{ijk\ldots})}{\text{var}(c_{ijk\ldots})}$$

where the form of var($c_{ijk\ldots}$) depends on the distribution of claims per policy. For example

Distribution	var $(c_{ijk\ldots})$
Binomial	$n_{ijk\ldots}\, p_{ijk\ldots}\,(1-p_{ijk\ldots})$
Poisson	$n_{ijk\ldots}\, p_{ijk\ldots}$
Negative Binomial	$n_{ijk\ldots}\, p_{ijk\ldots}\,(1+kp_{ijk\ldots})$

The multiple regression equations used can be obtained using a χ^2 function with var($c_{ijk\ldots}$) proportional to $n_{ijk\ldots}$ only.

The Poisson Distribution was assumed in the papers by Bailey and Simon (1) and Mehring (4), who also used a multiplicative model

$$p_{ijk\ldots} = \mu\alpha_i\beta_j\gamma_k\ldots$$

Conclusion

The method described here has been applied to more complicated cases (see Johnson (3) for an example), and has produced reasonable results. The data required for the application of the method are simple in character, and do not present any formidable difficulties for a computer-aided statistician. It is obviously possible to apply the method to analyse costs per claim or per policy-year.

REFERENCES

1. Bailey, R. A. and Simon, L. (1960). Two studies in Automobile Insurance Ratemaking, *ASTIN Bull. I*, 192.
2. Feldstein, M. S. (1966) A binary variable multiple regression method of analysing factors affecting peri-natal mortality and other outcomes of pregnancy. *J.R.S.S.* (A) **129**, 61.
3. Johnson, P. D. (1969) The Analysis of Motor Insurance Statistics, O.E.C.D. road research programme—symposium on statistical methods in the analysis of road accidents, 14–16 April 1969.
4. Mehring, J. (1964) Ein mathematisches Hilfsmittel für Statistik—und Tariffragen in der Kraftsfahrtversicherung, *Blätter der deutschen Gesellschaft für Versicherungsmathematik, VII*, 111.