The sample size needed for the calculation of a GLM tariff

Hans Schmitter
Swiss Reinsurance Company

Mythenquai 50/60, CH-8022 Zurich
Tel. +41 1 285 4970 – Telefax +41 1 282 4970 – e mail Hans_Schmitter@swissre.com

Abstract

A simple upper bound for the variance of the frequency estimates in a multivariate tariff using class criteria is deduced. This upper bound is based exclusively on univariate statistics and can therefore be calculated before a GLM analysis is carried out. It can be used to estimate the number of claims that will be needed for a tariff calculation depending on the number of tariff criteria and the number of levels of each criterion.

Keywords

Poisson, frequency, multiplicative tariff, generalised linear model, sample size.

1. Introduction

When the estimate of the Poisson parameter for identical risks is required to lie close to the true value (e.g. within 10%) with high probability (e.g. 95%) the number of observed claims must exceed a certain minimum which can be determined in a straightforward way. Let λ be the Poisson parameter, s the number of risks, N the Poisson-distributed number of claims and n an observation of N, i.e. the observed number of claims. This means

$$(1)\ \text{Prob}\{|N/s-\lambda| \leq c\lambda\} \geq p$$

when we write c and p instead of 10% and 95%.

Using the normal approximation with expected value and variance equal to λs and rewriting (1) as

$$\text{Prob}\{|N-\lambda s| / \sqrt{(\lambda s)} \leq c\sqrt{(\lambda s)}\} \geq p$$

we have

$$0.1\sqrt{(\lambda s)} \geq 1.96$$

in our example with c=0.1 and p=0.95 and hence $\lambda s \geq 384.16$. This means the expected number of claims must exceed 384.16. Estimating the expected λs by the observed number n we thus get $n \geq 384.16$. Applying this result to the calculation of a tariff for identical risks one needs a sample with at least 385 claims (or any other minimum depending on appropriate values for c and p) in order to determine the claims frequency with the precision required.
To the author's knowledge no such rules guaranteeing sufficient precision are known in the case of tariffs using several rating criteria. It is intuitively clear that the minimum sample size will increase as the number of criteria increases but whether or not the available data is extensive enough is not known in advance. Often only after time-consuming analyses does

one discover that the statistical basis for the calculation of a sophisticated tariff was in fact too small.
The purpose of the present paper is to give simple rules for checking whether or not the available sample is large enough to allow the frequencies of a multivariate tariff to be calculated.

2. Notation

We use the following notation:

Y_i Poisson-distributed random number of claims of risk i (i = 1, ..., n)

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{bmatrix}$$

$$(2)\ \lambda_i = e^{\sum_{j=1}^{r} x_{ij} b_j}$$

λ_i is the Poisson parameter of risk i. (2) shows that we assume the dependence of the expected number of claims on the tariff criteria to be multiplicative. The x_{ij} are called covariates, the b_j parameters. In the following we assume $x_{i1}=1$ for all i. In this case the first parameter b_1 is called intercept.

$$\mathbf{b} = \begin{bmatrix} b_1 \\ . \\ . \\ . \\ b_r \end{bmatrix}$$

y_i observed number of claims of risk i

$$\mathbf{y} = \begin{bmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{bmatrix}$$

$f(y_i, \mathbf{b}) = e^{-\lambda_i} \lambda_i^{y_i} / y_i!$ probability for risk i to have y_i claims

$l_n(\mathbf{b}, \mathbf{y}) = \Sigma \log f(y_i, \mathbf{b})$ log-likelihood function of $\mathbf{b}$

For $l_n(\mathbf{b}, \mathbf{y})$ to reach a maximum, the r partial derivatives with respect to $b_1, \ldots\ldots, b_r$ must be equal to 0. If we replace the observations y_i by the random variables Y_i in $l_n(\mathbf{b}, \mathbf{y})$ the partial derivatives are also random variables. Let $\mathbf{U_n(b)}$ be the vector of the partial derivatives which is also called the score vector. Because the Y_i are Poisson-distributed this vector is

$$\mathbf{U}_n(\mathbf{b}) = \begin{bmatrix} \sum_{i=1}^{n} x_{i1}(Y_i - \lambda_i) \\ \vdots \\ \sum_{i=1}^{n} x_{ir}(Y_i - \lambda_i) \end{bmatrix}$$

If we require the partial derivatives of $l_n(\mathbf{b}, \mathbf{Y})$ to be equal to 0, then the resulting $b_1, \ldots\ldots, b_r$ are also random variables which we designate as $B_1, \ldots\ldots, B_r$ and, when arranged in vector form, as **B**.
Because the Y_i are independent and because $\mathrm{Var}(Y_i) = \lambda_i$, the covariance of two elements of the score vector $\mathbf{U_n(b)}$, for instance the first and the second, is equal to

(3) $E[\Sigma\, x_{i1}\,(Y_i - \lambda_i)\,\Sigma\, x_{i2}\,(Y_i - \lambda_i)] = \Sigma\, x_{i1}\, x_{i2}\, \lambda_i$

Let **Q** be the r·r-matrix with elements as in (3), i.e. $\mathbf{Q} = \mathrm{Cov}(\mathbf{U_n(b)})$. In maximum likelihood theory, it is shown that the distribution of the vector **B**, i.e. of the estimators of the parameters $b_1, \ldots\ldots, b_r$, is asymptotically normal (as $n \rightarrow \infty$), and that the inverse of **Q** tends to the covariance matrix of **B**:

$\mathbf{Q}^{-1} \rightarrow \mathrm{Cov}(\mathbf{B})$

3. The case of class variables with 2 levels

Following the example in the introduction the estimate for every frequency λ_i should be close, e.g. within $c\lambda_i$ (e.g. c=0.1) to the true value with high probability (e.g. 95%). For practical purposes we assume **B** actually follows a joint normal distribution with covariance matrix $\mathbf{Q}^{-1}$ although this holds true only asymptotically. In this case, according to (2), the logarithm of the frequency of risk i is the sum of r normally distributed variables $x_{ij}\, B_j$ and therefore also normally distributed. The probability of the estimate of λ_i to lie tolerably close to its expected value depends on the variance of $\Sigma\, x_{ij}\, B_j$. Writing **M** for $\mathbf{Q}^{-1}$ with elements m_{jk} we have

(4) $\mathrm{Var}(\sum_j x_{ij} \cdot B_j) = \sum_j \sum_k x_{ij} \cdot x_{ik} \cdot m_{jk}$.

When the tariff criteria take on only two values, e.g. the driver's sex which is male or female, the place of residence (rural or urban), the car size (big or small), the engine size (large or small) and so on, then the covariates x_{ij} have only two possible values for which it is convenient to choose 0 and 1. Thus $x_{ij} = 1$ when risk i meets criterion j and $x_{ij} = 0$ otherwise. As can be seen from (3), in this case the elements q_{jk} of **Q** represent the expected numbers of claims of risks which simultaneously meet criteria j and k. Now consider a particular risk i and assume, without loss of generality, $x_{ij} = 1$ for j=1, ..., r. This means (4) becomes

$$\mathrm{Var}(\sum_j B_j) = \sum_j \sum_k m_{jk} .$$

There exists an upper bound for this variance since, as we are going to show in the following

$$(5) \sum_j \sum_k m_{jk} \le 1/q_{11} + 1/q_{22} + \ldots\ldots\ldots + 1/q_{rr}$$

Note that the q_{jj} on the right side of the sign of inequality are the expected numbers of claims of risks which meet criterion j and can be estimated with simple univariate statistics.
Before proving (5) let us look at a numerical example which is known to all readers who have learnt the theory of generalised linear models using SAS. In the Technical Report P-243 [2] the following example is given:

risks	claims	car type	age group
500	42	small	1
1200	37	medium	1
100	1	large	1
400	101	small	2
500	73	medium	2
300	14	large	2

Suppose we are interested in the variance of the logarithm of the frequency estimate for risks with small cars and age group 1. In order to include an intercept term in the model we define $x_{i1} = 1$ for all i. Combining the car types medium and large into a new type „not small" we define $x_{i2} = 1$ if the car type is small and $x_{i2} = 0$ if it is not small; likewise $x_{i3} = 1$ if the age group is 1 and $x_{i3} = 0$ if it is 2. Estimating the expected numbers of claims in **Q** by the observed numbers we get

$$\mathbf{Q} = \begin{pmatrix} \text{all} & \text{small} & \text{age}_1 \\ \text{small} & \text{small} & \text{small\_and\_age}_1 \\ \text{age}_1 & \text{small\_and\_age}_1 & \text{age}_1 \end{pmatrix}$$

or, numerically instead of informally,

$$\mathbf{Q} = \begin{pmatrix} 268 & 143 & 80 \\ 143 & 143 & 42 \\ 80 & 42 & 80 \end{pmatrix}$$

According to (5) the variance of the logarithm of the frequency estimate is at most equal to 1/268+1/143+1/80 = 0.02322. A check with the covariance matrix in appendix 1 shows that a computer run does actually give a lower value for the estimated variance, namely 0.01645.

In order to prove (5) we use some results from section 6 of chapter III (Normal Densities and Distributions) of the second volume of Feller [1]. The definitions of r, **Q** and **M** used in this article are the same as in Feller, whereas the random variables $B_j - E(B_j)$ correspond to X_j in Feller's notation. From (3) it is seen that our covariance matrix **Q** has the following properties:

(6) $q_{jj} > 0$ for j=1,...., r

(7) $0 \leq q_{jk} \leq q_{jj}, q_{kk}$ for $j \neq k$.

According to Feller's 6.2 **Q** defines the density of an r-dimensional normal distribution $\varphi(\mathbf{x})$:

$$\varphi(\mathbf{x}) = \gamma^{-1} e^{-1/2\, \mathbf{x}^T \mathbf{Q}\, \mathbf{x}}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ . \\ . \\ . \\ x_r \end{bmatrix}$$

The vector of the r normally distributed random variables X_1,....., X_r,

$$\mathbf{X} = \begin{bmatrix} X_1 \\ . \\ . \\ . \\ X_r \end{bmatrix}$$

has expectation $E(\mathbf{X})=\mathbf{0}$ and $Cov(\mathbf{X})=\mathbf{Q}^{-1}$. The marginal distributions have

(8) $Var(X_j)=1/q_{jj}$.

Since Feller's variables X_j are our $B_j - E(B_j)$ proving relation (5) is the same as proving
(9) $\mathrm{Var}(X_1 + X_2 + \ldots\ldots\ldots\ldots + X_r) \leq 1/q_{11} + 1/q_{22} + \ldots\ldots\ldots + 1/q_{rr}$.

We prove (9) by induction. For r=1 (9) reduces to (8) and is true. Assume it is true for r-1. As Feller shows (6.13) for the conditional variable $X_r \mid X_1, \ldots\ldots, X_{r-1}$

$E(X_r \mid X_1, \ldots\ldots, X_{r-1}) = -q_{1r}/q_{rr} \cdot X_1 - \ldots\ldots\ldots\ldots - q_{r-1,\,r}/q_{rr} \cdot X_{r-1}$ and

$\mathrm{Var}(X_r \mid X_1, \ldots\ldots, X_{r-1}) = 1/q_{rr}$.

Therefore

$E(X_1 + \ldots\ldots + X_r \mid X_1, \ldots\ldots, X_{r-1}) = (q_{rr} - q_{1r})/q_{rr} \cdot X_1 + \ldots\ldots\ldots\ldots + (q_{rr} - q_{r-1,\,r})/q_{rr} \cdot X_{r-1}$
and

$\mathrm{Var}(X_1 + \ldots\ldots + X_r \mid X_1, \ldots\ldots, X_{r-1}) = 1/q_{rr}$.

Put for abbreviation $c_j = (q_{rr} - q_{jr})/q_{rr}$.
Since for arbitrary conditional random variables $X \mid Y$ the relation

$\mathrm{Var}(X) = E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[E(X \mid Y)]$ holds we have

$\mathrm{Var}(X_1 + \ldots\ldots + X_r) = 1/q_{rr} + \mathrm{Var}(c_1 \cdot X_1 + \ldots\ldots\ldots + c_{r-1} \cdot X_{r-1})$. We look for the coefficients c_j which maximise this variance. Because of (6) and (7) we have $0 \leq c_j \leq 1$. Since for every j (j=1,...., r-1) the second derivative

$$\frac{\partial^2}{\partial c_j^2}\mathrm{Var}(c_1X_1 + \ldots\ldots\ldots + c_{r-1}X_{r-1}) = 2\mathrm{Var}(X_j) > 0$$

the variance $\mathrm{Var}(c_1 \cdot X_1 + \ldots\ldots\ldots + c_{r-1} \cdot X_{r-1})$ is maximal either for $c_j = 0$ or $c_j = 1$. If the c_j are ordered appropriately then

$c_1 = \ldots\ldots\ldots = c_s = 1$, where $s \leq r-1$. The remaining $c_j = 0$ (for $j>s$). Thus

$$\begin{aligned}\mathrm{Var}(X_1 + \ldots\ldots + X_r) &\leq 1/q_{rr} + \mathrm{Var}(X_1 + \ldots\ldots\ldots + X_s)\\ &\leq 1/q_{rr} + 1/q_{11} + 1/q_{22} + \ldots\ldots + 1/q_{ss}\\ &\leq 1/q_{rr} + 1/q_{11} + 1/q_{22} + \ldots\ldots\ldots + 1/q_{r-1,\,r-1}\end{aligned}$$

which proves (8).

4. Class variables with more than 2 levels

Class variables may assume more than two levels. For example the variable „car size" in [2] can have one of the three levels „small", „medium" or „large". A class variable v with k

levels (k>2) can be replaced by k −1 variables each having only 2 levels in the following way: call the k levels l_1, ..., l_k and define the covariates for k-1 2-level-variables as

x_1=1 if v=l_1
0 otherwise

x_2=1 if v ϵ {l_1, l_2}
0 otherwise
.....
x_{k-1}=1 if v ϵ {l_1, l_2,....., l_{k-1}}
0 otherwise

This one-to-one relation of a k-level-variable and k-1 2-level-variables makes it possible to apply the procedure of the previous section in the estimation of an upper bound for the variance of a claim frequency also to the case of general class variables. We illustrate this again using the numerical example from [2].

Suppose we are again interested in the variance of the logarithm of the frequency estimate for risks with small cars and age group 1. This time, however, we do not combine the car types medium and large into a new type but keep them separate. We define

x_1=1 for the intercept term

x_2=1 if car size = small
0 otherwise

x_3=1 if car size ϵ {small, medium}
0 otherwise

x_4=1 if age group=1
0 otherwise

Estimating the expected numbers of claims by inserting the observed numbers in the matrix **Q** we have

$$\mathbf{Q} = \begin{pmatrix} 268 & 143 & 253 & 80 \\ 143 & 143 & 143 & 42 \\ 253 & 143 & 253 & 79 \\ 80 & 42 & 79 & 80 \end{pmatrix}$$

According to (5) the variance of the logarithm of the frequency estimate is at most equal to 1/268+1/143+1/253+1/80 = 0.2718 which is higher than the value one gets from the covariance matrix in appendix 2, namely 0.01737.

5. An upper bound for the minimum number of observed claims needed in a sample

Suppose a tariff calculation is based on a sample which contains q claims and consider a class variable with k levels. The upper bound (5) of the variance will be the same for all tariff segments if the same number of claims have been observed for each level, i. e. q/k. Now let us replace the class variable by k-1 2-level-variables as in section 4. The numbers of observed claims corresponding to these 2-level-variables are q/k, 2q/k, 3q/k,.....,(k-1)q/k. They will appear in the main diagonal of the matrix **Q** when we use (5). This applies to all r class variables. Let k_j be the number of levels of variable number j. Bearing in mind that $k_1=1$ and $q_{11}=q$ because of the intercept we obtain for the right hand side of (5)

(10) $1/q\cdot[1+k_2\cdot(1+1/2+1/3+.....+1/(k_2-1))+.....+k_r\cdot(1+1/2+1/3+.....+1/(k_r-1))]$.

We now return to the problem stated in the introduction: the estimate of λ_i should lie with high probability p (e.g. 95%) close to its expected value (e.g. within 10%). This means, writing c for 10%, the estimate

$e^{B_1+B_2+.....+B_r}$ should not be lower than $(1-c)\cdot e^{b_1+b_2+.....+b_r}$ or higher than $(1+c)\cdot e^{b_1+b_2+.....+b_r}$. Consequently the exponent $B_1+B_2+....+B_r$ which follows a normal distribution should not deviate from its expected value by more than log(1-c). This defines the limit for the standard deviation of $B_1+B_2+....+B_r$: let z_p be the value defined by $Prob\{|Z|\leq z_p\}=p$, where Z follows the standard normal distribution (in the example with p=0.95 and c=0.1, $z_p=1.96$). Then we get from (10) if we call the expression in brackets u,

$u=[1+k_2\cdot(1+1/2+1/3+.....+1/(k_2-1))+.....+k_r\cdot(1+1/2+1/3+.....+1/(k_r-1))]$,

(11) $q = z_p^2\cdot u/\log[(1-c)]^2$

As a numerical example take again the motor insurance sample from [2] (see section 3). Suppose c=0.1 and $z_p=1.96$. There are 3 car size levels and 2 age group levels, so $k_2=3$ and $k_3=2$. Hence from (10) u=7.5 so that (11) yields q=2,595.

Note that the necessary number of claims could be higher than the value q given in (11) if the number of claims corresponding to the various class levels is not the same for each level of the same class. In this case, the sample size needed can be determined assuming the composition of the sample remains the same. In the example taken from [2] the segment of large cars and age group 1 has the highest upper bound of the variance because it contains the lowest number of claims. Using the method of section 4 we define

$x_2=1$ if car size = large
0 otherwise

$x_3=1$ if car size ϵ {small, large}
0 otherwise

and leave x_1 and x_4 unchanged. Then **Q** becomes

$$\mathbf{Q} = \begin{pmatrix} 268 & 15 & 158 & 80 \\ 15 & 15 & 15 & 1 \\ 158 & 15 & 158 & 43 \\ 80 & 1 & 43 & 80 \end{pmatrix}$$

The upper bound of the variance, i.e. the sum of the reciprocal diagonal elements, is 0.08923. Let us call this sum v and the factor with which each q_{jj} is to be multiplied in order to get the sufficiently large sample f. Similarly to (11) we have in this case

$$f = z_p^2 \cdot v / \log[(1-c)]^2$$

or in our numerical example f = 30.88. The sample size needed is thus 30.88 times larger than the given sample with a total number of claims of 268 f = 8,276.

6. References

[1] Feller, W. (1971) An Introduction to Probability Theory and Its Applications, volume 2, 2nd edition. Wiley.

[2] SAS Technical Report P-243, SAS/STAT Software: The GENMOD Procedure, Release 6.09, Cary, NC: SAS Institute Inc., 1993. 88 pp.

Appendix 1

```
data insure;
  input n c car$ age;
  ln=log(n);
  cards;
  500    42  small    1
  1200   37  notsmall 1
  100     1  notsmall 1
  400   101  small    2
  500    73  notsmall 2
  300    14  notsmall 2
```

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.INSURE
Distribution	POISSON
Link Function	LOG
Dependent Variable	C
Offset Variable	LN
Observations Used	6

Class Level Information

Class	Levels	Values
CAR	2	notsmall small
AGE	2	1 2

Parameter Information

Parameter	Effect	CAR	AGE
PRM1	INTERCEPT		
PRM2	CAR	notsmall	
PRM3	CAR	small	
PRM4	AGE		1
PRM5	AGE		2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3	21.7570	7.2523
Scaled Deviance	3	21.7570	7.2523
Pearson Chi-Square	3	19.5238	6.5079
Scaled Pearson X2	3	19.5238	6.5079
Log Likelihood	.	827.9851	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-1.3472	0.0912	218.0410	0.0001
CAR	notsmall	1	-0.9064	0.1227	54.5936	0.0001
CAR	small	0	0.0000	0.0000	.	.
AGE	1	1	-1.2034	0.1337	80.9743	0.0001
AGE	2	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

Estimated Covariance Matrix

Parameter Number	PRM1	PRM2	PRM4
PRM1	0.008324	-0.006725	-0.004879
PRM2	-0.006725	0.01505	-0.000984
PRM4	-0.004879	-0.000984	0.01788

$$\text{Var(intercept + age1)} = 0.008324 - 2 \cdot 0.004879 + 0.01788$$
$$= 0.016446$$

Appendix 2

```
data insure;
  input n c car$ age;
  ln=log(n);
  cards;
  500    42  small   1
  1200   37  medium  1
  100     1  large   1
  400   101  small   2
  500    73  medium  2
  300    14  large   2
```

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.INSURE
Distribution	POISSON
Link Function	LOG
Dependent Variable	C
Offset Variable	LN
Observations Used	6

Class Level Information

Class	Levels	Values
CAR	3	large medium small
AGE	2	1 2

Parameter Information

Parameter	Effect	CAR	AGE
PRM1	INTERCEPT		
PRM2	CAR	large	
PRM3	CAR	medium	
PRM4	CAR	small	
PRM5	AGE		1
PRM6	AGE		2

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	2.8207	1.4103
Scaled Deviance	2	2.8207	1.4103
Pearson Chi-Square	2	2.8416	1.4208
Scaled Pearson X2	2	2.8416	1.4208
Log Likelihood	.	837.4533	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-1.3168	0.0903	212.7321	0.0001
CAR	large	1	-1.7643	0.2724	41.9587	0.0001
CAR	medium	1	-0.6928	0.1282	29.1800	0.0001
CAR	small	0	0.0000	0.0000	.	.
AGE	1	1	-1.3199	0.1359	94.3388	0.0001
AGE	2	0	0.0000	0.0000	.	.

Estimated Covariance Matrix

Parameter Number	PRM1	PRM2	PRM3	PRM5
PRM1	0.008150	-0.007772	-0.006344	-0.004623
PRM2	-0.007772	0.07418	0.006556	0.003113
PRM3	-0.006344	0.006556	0.01645	-0.002592
PRM5	-0.004623	0.003113	-0.002592	0.01847

$$\text{Var(intercept + age1)} = 0.008150 - 2 \cdot 0.004623 + 0.01847$$
$$= 0.017374$$