THE SAMPLE SIZE NEEDED FOR THE CALCULATION OF A GLM TARIFF

BY

HANS SCHMITTER

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.

The article is a revised version of a paper presented at the XXXIst ASTIN Colloquium in Porto Cervo.

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, Y the Poisson-distributed number of claims and y an observation of Y, i.e. the observed number of claims. This means

$$\text{Prob}\{|Y/s - \lambda| \leq c\lambda\} \geq p, \tag{1}$$

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

$$\text{Prob}\left\{|Y - \lambda s| / \sqrt{\lambda s} \leq c\sqrt{\lambda s}\right\} \geq p$$

we have

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

ASTIN BULLETIN, Vol. 34, No. 1, 2004, pp. 249-262

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 y we thus get $y \geq 384.16$. 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. The result is presented in the form of an upper bound which can be calculated based on simple statistics. If the sample is larger than this upper bound, then the frequencies can be determined with the required accuracy. If, on the other hand, the sample is smaller than the upper bound, then the pre-defined accuracy is not guaranteed.

2. Notation

We use the following notation:

Y_i Poisson-distributed random number of claims of risk i ($i = 1, ..., n$). The Y_i are assumed to be independent.

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}$$

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

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.

$$\boldsymbol{b} = \begin{bmatrix} b_1 \\ \cdot \\ \cdot \\ \cdot \\ b_r \end{bmatrix}$$

y_i observed number of claims of risk i

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}$$

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

$L_n(\boldsymbol{b}, \boldsymbol{y}) = \sum_{i=1}^{n} \ln(f(y_i, \boldsymbol{b}))$ log-likelihood function of $\boldsymbol{b}$

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

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

If we require the partial derivatives of $L_n(\boldsymbol{b}, \boldsymbol{Y})$ to be equal to 0, then the resulting $b_1, \ldots, b_r$ are also random variables which we designate as $B_1, \ldots, B_r$ and, when arranged in vector form, as $\boldsymbol{B}$.

Because the Y_i are independent and because $\mathrm{Var}(Y_i) = \lambda_i$, the covariance of two elements of the score vector $\boldsymbol{U}_n(\boldsymbol{b})$, for instance the first and the second, is equal to

$$E\left[\sum x_{i1}(Y_i - \lambda_i) \sum x_{i2}(Y_i - \lambda_i)\right] = \sum x_{i1} x_{i2} \lambda_i. \tag{3}$$

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

$\boldsymbol{Q}^{-1} \to \mathrm{Cov}(\boldsymbol{B})$.

3. The case of class variables with two 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 $\boldsymbol{B}$ actually follows a joint normal distribution with covariance matrix $\boldsymbol{Q}^{-1}$ although this holds true only asymptotically. In this case, according to (2), the logarithm of the frequency estimate 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 $\sum x_{ij}B_j$. Writing $\boldsymbol{M}$ for $\boldsymbol{Q}^{-1}$ with elements m_{jk} we have

$$\operatorname{Var}\left(\sum_j x_{ij} \cdot B_j\right) = \sum_j \sum_k x_{ij} \cdot x_{ik} \cdot m_{jk}. \tag{4}$$

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 $\boldsymbol{Q}$ represent the expected numbers of claims of risks which simultaneously meet criteria j and k. The variance (4) is particularly simple if all $x_{ij} = 1$ for $j = 1, \ldots, r$. For such a risk (4) becomes

$$\operatorname{Var}\left(\sum_j B_j\right) = \sum_j \sum_k m_{jk}. \tag{5}$$

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

$$\sum_j \sum_k m_{jk} \leq \frac{1}{q_{11}} + \frac{1}{q_{22}} + \ldots\ldots + \frac{1}{q_{rr}}. \tag{6}$$

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.

If in (4) some $x_{ij} = 0$ then the variance of $\sum x_{ij} \cdot B_j$ cannot be estimated immediately by (5) and (6). We first have to replace the parameters B_j by new parameters B_j^* for which

$$\sum_j x_{ij} \cdot B_j = \sum_j B_j^*$$

before we can apply (6). The relation between $\boldsymbol{B}$ and $\boldsymbol{B}^*$, $\boldsymbol{B} = \boldsymbol{A} \cdot \boldsymbol{B}^*$, is given by an r · r matrix $\boldsymbol{A}$ which we define as follows: Let $s - 1$ be the number of covariates which are equal to 0 ($2 \leq s \leq r$), so that, possibly after renumbering the parameters B_j,

$x_{i1} = 1$
$x_{i2} = ... = x_{is} = 0$
$x_{i,s+1} = ... = x_{ir} = 1.$

The elements of $\boldsymbol{A}$ are all equal to 0 except

$a_{jj} = 1$ if $x_{ij} = 1$
$a_{jj} = -1$ if $x_{ij} = 0$
$a_{1j} = 1$ if $x_{ij} = 0.$

Note that $\boldsymbol{A} = \boldsymbol{A}^{-1}$. Therefore, the solution $\boldsymbol{B}^*$ of $\boldsymbol{B} = \boldsymbol{A} \cdot \boldsymbol{B}^*$ can be easily calculated as

$$\boldsymbol{B}^* = \begin{bmatrix} B_1 + B_2 + \ldots + B_s \\ -B_2 \\ -B_3 \\ \vdots \\ -B_s \\ B_{s+1} \\ \vdots \\ B_r \end{bmatrix}$$

Let $\boldsymbol{X}$ be the n·r matrix of the covariates with respect to $\boldsymbol{B}$ and $\boldsymbol{Z}$ the n·r matrix of the covariates with respect to $\boldsymbol{B}^*$ defined by $\boldsymbol{Z} = \boldsymbol{X} \cdot \boldsymbol{A}$. As a result of the matrix multiplication $z_{i1} = z_{i2} = \ldots = z_{ir} = 1$.

Let $\boldsymbol{Q}^*$ be the matrix with elements q_{jk}^* defined as in (3) but with the covariates z_{ij} instead of x_{ij}. Then $\boldsymbol{Q}^{*-1}$ is the covariance matrix of $\boldsymbol{B}^*$ and its elements m_{jk}^* fulfil (6) when $q_{11}, \ldots, q_{rr}$ are replaced by $q_{11}^*, \ldots, q_{rr}^*$ to the right of the inequality sign.

The transition from $\boldsymbol{X}$ to $\boldsymbol{Z}$ should be interpreted in the following way: If e.g. $x_{i2} = 1$ means risk i is male, then $z_{i2} = 1$ means risk i is female etc.

An upper bound (6) can be calculated for every risk i. The highest of these upper bounds is the one with the lowest q_{jj}. It is found in the following way:

q_{11} is the total number of expected claims of the whole sample. q_{jj}, where $j > 1$, is the number of expected claims of those risks which meet criterion j. The number of expected claims of risks which do not meet criterion j is the difference $q_{11} - q_{jj}$. Define criterion j so that $q_{jj} \leq q_{11} - q_{jj}$.

Before proving (6) 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

By way of example let us look at the third segment, risks with large cars and age group 1, and estimate the variance of the logarithm of their frequency. In order to include an intercept term in the model we define $x_{i1} = 1$ for all i. Combining the car types small and medium into a new type "not large" we define $x_{i2} = 1$ if the car type is large and $x_{i2} = 0$ if it is not large; 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 $\boldsymbol{Q}$ by the observed numbers we get

$$\boldsymbol{Q} = \begin{pmatrix} \text{all} & \text{large} & \text{age 1} \\ \text{large} & \text{large} & \text{large and age 1} \\ \text{age 1} & \text{large and age 1} & \text{age 1} \end{pmatrix}$$

or, numerically instead of informally,

$$\boldsymbol{Q} = \begin{pmatrix} 268 & 15 & 80 \\ 15 & 15 & 1 \\ 80 & 1 & 80 \end{pmatrix}.$$

According to (6) the variance of the logarithm of the frequency estimate is at most equal to 1/268 + 1/15 + 1/80 = 0.08290. 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 0.08217.

In order to prove (6) we use some results from section 6 of chapter III (Normal Densities and Distributions) of the second volume of Feller [1]:

a) A symmetric, positive definite r·r matrix $\boldsymbol{Q}$ defines an r-dimensional normal density centred at the origin (Feller's theorem 4). The form of this density is

$$\varphi(x) = \gamma^{-1} e^{-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}}$$

where γ is a constant and

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_r \end{bmatrix}$$

b) The vector $\boldsymbol{X}$ of the r normally distributed random variables $X_1, \ldots, X_r$ has expectation $E(\boldsymbol{X}) = \boldsymbol{0}$ and its covariance matrix $\boldsymbol{M} = \mathrm{Cov}(\boldsymbol{X})$ is the inverse of $\boldsymbol{Q}$ (Feller's theorem 3).

c) The variance of the marginal distribution of X_j is $\mathrm{Var}(X_j) = 1/q_{jj}$ (Feller's theorem 1 and relation 6.5).

d) Expected value and variance of the conditional variable $X_r \mid X_1, \ldots, X_{r-1}$ are $E(X_r \mid X_1, \ldots, X_{r-1}) = -q_{1r}/q_{rr} \cdot X_1 - \ldots - q_{r-1,r}/q_{rr} \cdot X_{r-1}$ and

$\mathrm{Var}(X_r \mid X_1, \ldots, X_{r-1}) = 1/q_{rr}$ (Feller's relation 6.13).

The definitions of r, $\boldsymbol{Q}$ and $\boldsymbol{M}$ 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 $\boldsymbol{Q}$ has the following properties:

$$q_{jj} > 0 \text{ for } j = 1, \ldots, r \tag{7}$$

$$0 \le q_{jk} \le q_{jj},\ q_{kk} \text{ for } j \ne k. \tag{8}$$

Moreover, $\boldsymbol{Q}$ is symmetric and positive definite. Therefore, according to a) it defines the density of an r-dimensional normal distribution.

Since Feller's variables X_j are our $B_j - E(B_j)$ proving relation (6) is the same as proving

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_r) \le 1/q_{11} + 1/q_{22} + \ldots + 1/q_{rr}. \tag{9}$$

We prove (9) by induction. For $r = 1$ (9) reduces to c) and is true. Assume it is true for $r - 1$. According to d) expected value and variance of the variable $X_r \mid X_1, \ldots, X_{r-1}$ are

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

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

Therefore

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

and

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

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

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

$\mathrm{Var}(X_1 + \ldots + X_r) = 1/q_{rr} + \mathrm{Var}(c_1 \cdot X_1 + \ldots + c_{r-1} \cdot X_{r-1})$. We look for the coefficients c_j which maximise this variance. Because of (7) and (8) we have $0 \le c_j \le 1$. Since for every j ($j = 1, \ldots, r - 1$) the second derivative

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

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

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

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

which proves (6).

4. Class variables with more than two 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 order to keep the notation simple, assume the covariates corresponding to these 2-level-variables are numbered $x_{i2}, \ldots, x_{ik}$. This means that they immediately follow the covariate $x_{i1} = 1$ for the intercept term. Designate the k levels by $l_1, ..., l_k$ and define the covariates $x_{i2}, \ldots, x_{ik}$ as

$x_{i2} = 1$ if $v = l_1$
0 otherwise

for $j = 3, \ldots, k\, x_{ij} = 1$ if $v \neq l_{j-1}$
0 otherwise

Consequently, a risk i for which the class variable v is equal to $v = l_1$ has the covariates $x_{i2} = \ldots = x_{ik} = 1$.

In this way it is possible to apply the procedure of the previous section also to the case of general class variables. The contribution of the class variable v to the upper bound (6) is $1/q_{22} + \ldots + 1/q_{kk}$. From (3) and the definition of $x_{i2}, \ldots, x_{ik}$ it can be seen that q_{22} is the expected number of claims of risks with $v = l_1$. For $j > 2$, q_{jj} is the expected number of claims of risks with $v \neq l_{j-1}$.

The last level, l_k, does not matter in the calculation of (6). If l_1 is given, we obtain the lowest value of (6) if we order the levels $l_1, \ldots, l_k$ so that l_k is the level with the highest number of expected claims. In the numerical example, l_3 is thus the level "small" with 143 expected claims.

Ordering $l_1, \ldots, l_k$ so that l_1 is the level with the lowest number of expected claims leads to the highest upper bound (6). In fact: let p_2 be the expected number of claims in level l_1, p_3 the expected number of claims in level l_2 etc. Then the contribution of the class variable to (6) is

$$\frac{1}{p_2} + \frac{1}{q_{11} - p_3} + \ldots + \frac{1}{q_{11} - p_k}.$$

Since $p_2 \leq p_j$ for $j > 2$, it follows that

$$\frac{1}{p_2} + \frac{1}{q_{11} - p_j} \geq \frac{1}{p_j} + \frac{1}{q_{11} - p_2}.$$

so that exchanging levels l_1 and l_{j-1} would not increase the value of (6).

We illustrate the handling of class variables with more than two levels again using the numerical example from [2] and estimating expected numbers of claims by observed numbers.

The car size with the lowest number of expected claims is "large" and in age group 1 there are less expected claims than in age group 2. Therefore, the risk with the highest upper bound (6) is given by car size "large" and age group 1. We define the covariates

x_{i1} = 1 for the intercept term

x_{i2} = 1 if car size = large
0 otherwise

x_{i3} = 1 if car size not medium
0 otherwise

x_{i4} = 1 if age group = 1
0 otherwise

Estimating the expected numbers of claims by inserting the observed numbers in the matrix $\boldsymbol{Q}$ we have

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

According to (6) the variance of the logarithm of the frequency estimate is at most equal to 1/268 + 1/15 + 1/158 + 1/80 = 0.08923 which is in fact higher than the value one obtains from the covariance matrix in appendix 2, namely 0.08224.

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

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 $e^{B_1 + B_2 + \ldots + B_r}$ should not be lower than $(1 - c) \cdot e^{b_1 + b_2 + \ldots + b_r}$ or higher than $(1 + c) \cdot e^{b_1 + b_2 + \ldots + b_r}$. Consequently the exponent $B_1 + B_2 + \ldots + B_r$ which follows a normal distribution should not

deviate from its expected value by more than $\ln(1-c)$. This defines the limit for the standard deviation of $B_1 + B_2 + \ldots + B_r$: let z_p be the value defined by $\text{Prob}\{|Z| \le z_p\} = p$, where Z follows the standard normal distribution (in the example with $p = 0.95$ and $c = 0.1$, $z_p = 1.96$).

We estimate the variance of $B_1 + B_2 + \ldots + B_r$ by the sum u of the reciprocal diagonal elements of $\boldsymbol{Q}$. If $u \le [\ln(1-c)]^2/z_p^2$ then our frequency estimate is sufficiently precise. Otherwise, assuming the composition of the sample remains the same, we determine a factor f by which all diagonal elements of $\boldsymbol{Q}$ are to be multiplied so that $u/f = [\ln(1-c)]^2/z_p^2$.

As a numerical example take again the motor insurance sample from [2] (see section 3). As has been shown in section 4 the segment of large cars and age group 1 has the highest upper bound of the variance, namely 0.08923. The factor f with which each q_{jj} is to be multiplied in order to get the sufficiently large sample is

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

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 $268 \cdot f = 8{,}276$.

As one anonymous referee points out the upper bound of the variance is heavily influenced by $q_{22} = 15$ corresponding to the segment of large vehicles. If we disregard the large vehicles and define x_{i2} to x_{i4} in another way so that e.g. for the car type with the second lowest number of claims (i.e. medium) and age group 1 all covariates are equal to 1, then the total number of claims needed in the sample is much smaller (2,715). In practice, we might do exactly that, i.e. tolerate that some rather insignificant tariff segments are less accurately rated provided that the accuracy of the important segments is sufficient.

Acknowledgement

I would like to thank both anonymous referees for their constructive remarks that led to a better presentation of the paper.

References

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

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

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

APPENDIX 1

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

;
```

The GENMOD Procedure

Model Information

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	large notlarge
age	2	1 2

Parameter Information

Parameter	Effect	car	age
Prm1	Intercept		
Prm2	car	large	
Prm3	car	notlarge	
Prm4	age		1
Prm5	age		2

Estimated Covariance Matrix

	Prm1	Prm2	Prm4
Prm1	0.005710	−0.005293	−0.005637
Prm2	−0.005293	0.07164	0.004298
Prm4	−0.005637	0.004298	0.01808

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	– 1.6427	0.0756	– 1.7908	– 1.4946	472.57	<. 0001
car	large	1	– 1.4300	0.2677	– 1.9546	– 0.9053	28.54	<. 0001
car	notlarge	0	0.0000	0.0000	0.0000	0.0000	.	.
age	1	1	– 1.4276	0.1345	– 1.6912	– 1.1641	112.75	<. 0001
age	2	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

The table Analysis of Parameter Estimates shows that the Genmod procedure sets the values for "car notlarge" and "AGE 2" to 0. The logarithm of the frequency of large cars and age 1 is thus estimated by the sum of the estimates for Intercept, car large and age 1, – 1.6427 – 1.4300 – 1.4276 = – 4.5003.

Its variance is Var(Intercept) + Var(car large) + Var(age 1) + 2 · Cov(Intercept, car large) + 2 · Cov(Intercept, age 1) + 2 · Cov(car large, age 1). The table Parameter Information on the previous page shows the correspondence of Prm1 to Prm5 and the terms Intercept, car large, car notlarge, age 1 and age 2. Therefore

$$
\begin{aligned}
\text{Var(intercept + car large + age 1)} &= \text{Var(Intercept)} + \text{Var(car large)} + \text{Var(age 1)} \\
&\quad + 2 \cdot \text{Cov(Intercept, car large)} + 2 \cdot \text{Cov(Intercept, age 1)} + 2 \cdot \text{Cov(car large, age 1)} \\
&= \text{Var (Prm1)} + \text{Var(Prm2)} + \text{Var(Prm4)} \\
&\quad + 2 \cdot \text{Cov(Prm1, Prm2)} + 2 \cdot \text{Cov(Prm1, Prm4)} + 2 \cdot \text{Cov(Prm2, Prm4)} \\
&= 0.005710 + 0.07164 + 0.01808 - 2 \cdot 0.005293 - 2 \cdot 0.005637 + 2 \cdot 0.004298 \\
&= 0.08217.
\end{aligned}
$$

APPENDIX 2

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

;
```

The GENMOD Procedure

Model Information

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

Estimated Covariance Matrix

	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

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	−1.3168	0.0903	−1.4937	−1.1398	212.73	<.0001
car	large	1	−1.7643	0.2724	−2.2981	−1.2304	41.96	<.0001
car	medium	1	−0.6928	0.1282	−0.9441	−0.4414	29.18	<.0001
car	small	0	0.0000	0.0000	0.0000	0.0000	.	.
age	1	1	−1.3199	0.1359	−1.5863	−1.0536	94.34	<.0001
age	2	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

$$
\begin{aligned}
\text{Var(intercept + car large + age 1)} &= \text{Var(Intercept)} + \text{Var(car large)} + \text{Var(age 1)} \\
&\quad + 2 \cdot \text{Cov(Intercept, car large)} + 2 \cdot \text{Cov(Intercept, age 1)} + 2 \cdot \text{Cov(car large, age 1)} \\
&= \text{Var(Prm1)} + \text{Var(Prm2)} + \text{Var(Prm5)} \\
&\quad + 2 \cdot \text{Cov(Prm1, Prm2)} + 2 \cdot \text{Cov(Prm1, Prm5)} + 2 \cdot \text{Cov(Prm2, Prm5)} \\
&= 0.008150 + 0.07418 + 0.01847 - 2 \cdot 0.007772 - 2 \cdot 0.004623 + 2 \cdot 0.003113 \\
&= 0.08224.
\end{aligned}
$$