Approximate sampling variances of maximum-likelihood probability estimates in a logit response function

J.E.O. Rege

International Livestock Research Institute, P.O. Box 5689, Addis Ababa, Ethiopia

Abstract

The maximum likelihood parameters estimated in logistic analysis are in terms of a transformation of the original response variable and, although inferences are easily made about the sources of variation in the linearized model using standard procedures applied in regression analysis, the variance-covariance structure of the predicted probabilities obtained following back-transformation of logits is complex and estimation of sampling variances normally require inversion of matrices and taking derivatives of the inverse of the link function evaluated at each prediction point. This paper presents a method for estimating sampling variances of such predicted probabilities without the need to invert any matrix or take derivatives of the link function. The method is based on the assumption that the exponent of a linear function of the logits is lognormal. It is demonstrated by way of a numerical example that this approximation is not different from the more complex methods applied by software such as SAS and GENSTAT.

1. Introduction

The difference between a logistic model and the conventional linear (regression) model is that the response variable in the former is binary or dichotomous. This difference is reflected both in the choice of parametric model and in the assumptions. These have been discussed in several texts (e.g. Fienberg, 1980; Hosmer and Lemeshow, 1989; McCullagh and Nelder, 1989) and are not the subject of this paper. It suffices to say that, in the logistic analysis, the parameters obtained are in terms of a transformation of the original response variable and, although inferences can be made about the sources of variation in the model using standard procedures applied in the regression models, the variance-covariance structure of the effects in the model following back-transformation to the original scale is illusive. Outputs from typical computer programs will include estimated coefficients ("parameter" estimates), standard errors of these coefficients, Z-values (ratio of coefficients to standard errors) and the likelihood ratio test comparing the fitted model to the one in which the coefficients of all the explanatory variables are zero. Some programs (e.g. GENSTAT, 1987) include the predicted probabilities, $\hat{\pi}$'s, of subclasses in the model and their standard errors in the output. Most programs, including CATMOD of SAS (1987), however, do not output subclass $\hat{\pi}$'s although these are presented at the "cell" level. While calculation of subclass $\hat{\pi}$'s by the inverse transformation of the logits from such programs should be straightforward (provided the parameterization used in obtaining the solutions is known), construction of their sampling variances needed for hypotheses test is not trivial. At whatever level the $\hat{\pi}$'s are calculated, the method applied involves taking inverses of matrices and derivatives of the inverse of the link function evaluated at each prediction (e.g. according to the method of Imrey et al., 1981). This paper presents a procedure for obtaining variances of the back-transformed "subclass means" obtained from a logistic analysis which requires neither matrix inversion nor taking of any derivatives. Although references are made to other computer software, the procedure is illustrated using the CATMOD procedure of SAS (1987) and the logistic regression of the GENSTAT (1987) softwares.

2. Overview

A transformation that is central to the logistic model is the logit transformation, which is defined, in terms of the probability parameter, π, as

$$Logit(\underline{\pi}) = \underline{l} = \mathbf{X}\underline{\beta}$$

where the i^{th} element of the vector $\underline{l}$, l_i, is the logit transformation of the i^{th} probability π_i calculated as $l_i = ln(\pi_i/1 - \pi_i)$, $\underline{\beta}$ is the vector of parameters and $\mathbf{X}$ is the incidence matrix. This tranforms a probability (π_i) to a real number, l_i, on a scale $-\infty$ to $+\infty$. The importance of this transformation is that has many of the desirable properties of a linear (regression) model: it is linear in parameters (β's) and is continuous.

The inverse transformation,

$$\pi_i = \frac{\exp(l_i)}{[1 + \exp(l_i)]} \tag{1}$$

facilitates the expression of the probability (π_i) in terms of the parameters of the model. Estimates of model parameters, β's, can be obtained in several ways, the most common of which are the maximum-likelihood (ML)

and the weighted-least-squares (WLS) procedures. The maximum-likelihood method estimates the parameters of the linear model so as to maximize the value of the joint multinomial likelihood function of the responses. For example, the CATMOD procedure of SAS (1987) computes ML estimates for the standard log-linear model analysis according to Bishop et al. (1975). On the other hand, the WLS procedure (Grizzle et al., 1969; Koch et al., 1977) is implemented by minimizing the weighted residual sum of squares using, as weights, the elements contained in the inverse covariance matrix of the transformed vector of sample proportions. Most computer packages with logistic analysis capability (e.g. BMDP, SAS, GENSTAT, GLIM) use iterative algorithms to obtain the parameter estimates (McCullagh and Nelder, 1989). Hypothesis about linear combinations of the estimated parameters, including importance of a factor, are tested using various test statistics e.g. the Score Test (Cox and Hinkley, 1974; Dobson, 1983). For example, SAS tests for significance of effects in the model using the generalized Wald's (1943) statistic while the PLR program of BMDP uses various goodness of fit chi-square tests (e.g. the Brown test (1974)). GENSTAT, on the other hand, uses the deviance which results in likelihood ratio tests with asymptotic chi-square distribution. The method for estimating the variances and covariances of the estimated coefficients needed for these tests follows from well-developed theory of ML estimation (e.g. Rao, 1973). This theory states that the estimators are obtained from the matrix of second partial derivatives of the log-likelihood function. However, many computer packages including SAS, do not extend the analysis beyond "analysis of variance" and a presentation of the parameter estimates (β's) and standard errors. On the other hand, in order to make conclusions from an analysis, it is important to state the magnitude of the estimated effects on an easily understood scale. In the case of logistic analysis, the scale suitable for this purpose is different from the scale (or link function) used to achieve additivity of effects. Therefore it is necessary to back-transform the logit, l, to probability by substituting the estimated parameters, $\hat{\beta}$'s for β's in (1).

Thus

$$\text{Predicted probability} = \hat{\pi}_i = \frac{\exp(\mathbf{x}_i\underline{\hat{\beta}})}{\left[1+\exp(\mathbf{x}_i\underline{\hat{\beta}})\right]}$$

where $\underline{\hat{\beta}}$ is the estimated parameter vector and $\mathbf{x}_i$ is the i^{th} row of the design matrix.

Clearly, obtaining $\hat{\pi}_i$ should be straightforward whenever the parameter estimates are available. The problem is to estimate the variance-covariance matrix of the vector of probabilities in order to extend the hypotheses tests to $\underline{\pi}$. Imrey et al. (1981) has presented a method for obtaining the estimator of the asymptotic covariance matrix of the ML predicted probabilities. Nonetheless, tractable and easy-to-apply procedures for estimating sampling variances for predicted probabilities in logistic models remain generally unavailable. On the other hand, in virtually all cases in which logistic analyses are applied, the investigator is interested in the predicted probabilities and, usually, their standard errors. The following paragraphs present a method which can be used to obtain standard errors of predicted probabilities from any computer program so long as the variance-covariance matrix of the coefficients can be obtained and the parameterization used in calculating these coefficients is known. The method is illustrated using standard output from the CATMOD procedure of SAS and the logistic regression option of GENSTAT. It is demonstrated that this procedure can be extended to obtain estimates of standard errors for predicted probabilities in logistic analysis by any software provided that the software has an option to output the variance-covariance matrix of the coefficients.

3. Derivation

For a specified linear function of the parameters, the exponential function (1) can be used to estimate probability π_i given by

$$\pi_i = \frac{z_i}{(1+z_i)} \tag{2}$$

where $z_i = \exp(g_i)$ and g_i is a linear function of the parameters, β's.

Note that while l is a linear function of the parameters, $\hat{\mathbf{g}} = \mathbf{Q}'\underline{\hat{\beta}}$ (for some matrix $\mathbf{Q}$) are linear functions of the estimated coefficients as typified by estimable contrasts in generalized linear models. For example, g_i may be INTERCEPT + β_{1i}, where β_{1i} is the parameter for the i^{th} level of the 1^{st} effect in the model. In this case π_i would be the predicted probability of this subclass.

Let

$$r_i = ln[\frac{z_i}{1+z_i}]$$

It can be shown (e.g. Mood et al., 1974; Rege, 1988) that the variance of r_i is approximately

$$var(r_i) = \{var(z_i)/[E(z_i)]^2\} + \{var(1+z_i)/[1+E(z_i)]^2\} - 2\{cov(z_i, 1+z_i)/E(z_i)E(1+z_i)\} \tag{3}$$

In the absence of true values of the mean and variance of z_i, $var(r_i)$ can be estimated by substituting estimates for parameters. The task is, therefore, to derive formulae for estimating both the mean and variance of z_i from the data.

It is to be recalled that the ML estimators, $\hat{\beta}$'s, are asymptotically normally distributed, with the true parameter values, β's, as means and asymptotic variance-covariance matrix, $\mathbf{\Sigma}$, equal to the inverse of the Fisher information matrix (McCullagh and Nelder, 1989). Indeed, it is the asymptotic normality of MLE's which lead to the asymptotic Chi-square distribution of the statistics for the goodness-of-fit (Bishop et al., 1975). Thus, $\hat{\mathbf{g}} = \mathbf{Q}'\underline{\hat{\beta}}$ is also normally distributed. Consequently, z_i is lognormal with mean, $E(z_i)$, and variance, $var(z_i)$, given by Lindgren (1976)

$$E(z_i) = exp[\mu_g + 1/2\sigma_g^2]$$

and

$$var(z_i) = \{exp(\sigma_g^2) - 1\}\{exp(2\mu_g + \sigma_g^2)\}$$

where μ_g and σ_g^2 are, respectively, the mean and variance of g_i.

Estimates of $E(z_i)$ and $var(z_i)$ are obtained by replacing $\hat{\mu}_g$ and $\hat{\sigma}_g^2$ for the parameters in the above formula. The logits, $\hat{g}$'s, are linear functions of the $\hat{\beta}$'s and hence their calculation and the calculation of their variances is straightforward. Once this is achieved, $var(r_i)$ is calculated from (3), noting that $var(1+z_i) = cov(z_i, 1+z_i) = var(z_i)$.

Noting that $\pi_i = \exp(r_i)$, the estimate of the variance of $\hat{\pi}_i$ is

$$\widehat{var(\hat{\pi}_i)} = var[exp(r_i)]$$

It can be shown (Rege, 1988) that

$$\widehat{var(\hat{\pi}_i)} \simeq \widehat{var(r_i)}(\hat{\pi}_i)^2$$

where $\hat{\pi}_i$ is the predicted probability computed from (2).

Finally, the standard errors of the predicted probabilities are calculated as

$$SE(\hat{\pi}_i) = [\widehat{var(r_i)}(\hat{\pi}_i)^2]^{1/2}$$

As has been stated, most packages have options for calculating elements of the matrix $\mathbf{\Sigma}$. With the exception of GENSTAT and GLIM, the deficiency of available packages is the fact that they do not provide sampling variances to test hypothesis on the back-transformed values (i.e. predicted probabilities) at the subclass levels. Even where this information is provided (e.g. in GENSTAT, 1987), the methodology used is complex and tedious and the documentation generally inadequate. Provided the ML estimates ($\hat{\beta}$'s) and the (asymptotic) variance-covariance matrix, $\mathbf{\Sigma}$, of these coefficients can be obtained, by whatever method, estimation of the sampling variances of predicted probabilities using the above procedure is straightforward.

4. Computation details

It should be clear from the above presentation that only the vector of parameter estimates, $\underline{\hat{\beta}}$, and the matrix of estimated variances and covariances of these parameters, $\hat{\mathbf{\Sigma}}$, is required in order to calculate the standard errors of the predicted probabilities. The computations are quite simple. The output from the computer package can be saved on disc and the required information extracted into a file from where it can be read and the computations accomplished using a tailor-made program. Such a program has been developed for output from CATMOD of SAS (Rege, 1992). It is to be noted that, when the objective is to calculate $\underline{\hat{\pi}}$ and standard errors for levels of effects in a model, it is essential to know the parameterization used by the particular software. This point is illustrated in the example below using the CATMOD parameterization.

4.1. Parameter estimates

The coefficients presented by the CATMOD procedure are equivalent to "constant" estimates in the GLM procedure of SAS. That is, they are simply solutions to the linear equations. To obtain population marginal means which are equivalent to "least-squares means", further calculations are necessary. However, the nature of these calculations depend on the parameterization used to obtain the parameter estimates. For example, CATMOD constraints the sum of the coefficients of each factor in the model to sum to zero. The coefficient for the last level of the i^{th} factor is, however, not included in the output but can be calculated as $-S$, where S

is the sum of the $n_i - 1$ coefficients in the output; n_i being the number of levels in the i^{th} factor. Additionally, because the last level is not included in the variance-covariance matrix, its variance (and covariance with other estimates) must be constructed by noting that it is a linear function of the other $(n_i - 1)$ levels.
The advantage of this parameterization is that it affords the calculation of subclass parameters (population marginal means for subclasses) simply as the sum of the INTERCEPT (a) and the estimated subclass coefficient. That is, for each level j, of a factor, i, the corresponding logit, $\hat{g}_{ij}$ is calculated as $\hat{a} + \hat{\beta}_{ij}$.

Let $\mathbf{Q}'$ be a $p \times (p - k)$ matrix constructed such that

$$\hat{\underline{g}} = \mathbf{Q}'\hat{\underline{\beta}}$$

where p is the number of parameters in the model (i.e. all levels of effects including the intercept) and k is the number of classification (discrete) independent variables in the model.

The parameter vector, $\underline{\beta}$, is $(p - k) \times 1$. Because the $\hat{g}$'s are estimates of population marginal means (for subclasses), the elements of $\mathbf{Q}'$ depend on the parameterization used. It will be clear in examples given later that for the CATMOD procedure of SAS, the elements of the parameterization matrix, $\mathbf{Q}$ are 0's, 1's and -1's. On the other hand, the elements of $\mathbf{Q}$ for GENSTAT consist of 0's, 1's and $1/n_i$'s where n_i's are the number of levels of each factor. The difference arises from the fact that GENSTAT constraints the parameter of the first level of each class to zero while, as has been stated, SAS constraints the sum of the parameters to zero. In addition, GENSTAT does not output solution to the first level while SAS does not output solution to the last level. As an illustration, if the model has 2 effects with 2 and 3 levels, respectively, then the appropriate $\mathbf{Q}'$ to obtain subclass coefficients from SAS output is (for p=6, k=2):

$$\mathbf{Q}_1' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & -1 & -1 \end{bmatrix}$$

The corresponding $\mathbf{Q}'$ for GENSTAT would be

$$\mathbf{Q}_2' = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/3 \\ 1 & 0 & 1/3 & 1/3 \\ 1 & 1 & 1/3 & 1/3 \\ 1 & 1/2 & 0 & 0 \\ 1 & 1/2 & 1 & 0 \\ 1 & 1 & 1/2 & 0 \end{bmatrix}$$

The next step is to compute $\hat{\mathbf{g}} = \mathbf{Q}'\hat{\underline{\beta}}$ and the variance-covariance matrix of $\hat{\mathbf{g}}$, $\mathbf{V}_{\hat{\mathbf{g}}} = \mathbf{Q}'\hat{\mathbf{\Sigma}}\mathbf{Q}$. The required variances are on the diagonal of $\mathbf{V}_{\hat{\mathbf{g}}}$; the covariances in the off-diagonal are not of interest for the purposes of this presentation. The example below illustrates how the standard errors of $\hat{\pi}$'s are calculated from $\mathbf{V}_{\hat{\mathbf{g}}}$ using real data.

4.2. Example

The method is illustrated by an analysis of mortality records, coded as 1=alive and 0=dead, of 5261 calves produced by cows of several beef cattle genotypes in an on-station breed evaluation study. The original model proposed for the analysis included, as independent variables, dam genotype (17 levels), parous state of dam (2), sex of calf (2), management unit (2), year of birth of calf (8), date class of birth (8) and birth weight of calf (covariate). However, for purposes of this illustration, a simpler model which could easily run in both SAS and GENSTAT (for comparison) was fitted. The model included the effects of sex, management unit ("herd"), and date class of birth.
Results of ML analysis of variance (from SAS) are presented in Table 1. These results indicate that all effects in the model were significant. The likelihood ratio statistic (29.86) was not significant, indicating that the model provided a good fit to the data. The ML parameter estimates are presented in Table 2 with their standard errors and the chi-square statistics for testing the null hypothesis that the true parameters are equal to zero. The negative coefficient for SEX indicates that the first level (females) had lower probability of death (i.e. lower mortality rate) than the second level (male). On the other hand, the positive coefficient for HERD indicates that the first management unit was associated with higher mortality rate than the second. Patterns for date class of birth (DATECLS) can, in a similar way, be discerned from the coefficient estimates. However,

it is not possible to tell from these results whether, for a specific factor, the mortality in one subclass is significantly different from the figure in another subclass.

Source	df	Chi-square
Intercept	1	383.50(**)
Sex	1	11.08(**)
Herd	1	68.07(**)
Datecls	7	30.15(**)
Likelihood ratio	22	29.86(ns)

Table 1. Maximum likelihood analysis of variance for the example data from CATMOD of SAS (** = $p < .01$; ns=not significant)

Effect	Subclass	Estimate	SE	Chi-square
INTERCEPT		-1.976	.1009	383.50(**)
SEX	1	-0.143	.0429	11.08(**)
HERD	1	0.370	.0449	68.07(**)
DATECLS	1	0.816	.6001	1.85(ns)
	2	-0.541	.2174	6.19(*)
	3	-0.323	.1267	6.50(*)
	4	-0.075	.1225	0.38(ns)
	5	0.257	.1263	4.13(*)
	6	0.052	.1411	0.13(ns)
	7	0.103	.1636	0.40(ns)

Table 2. Maximum likelihood parameter estimates for the example data from CATMOD of SAS (** = $p < .01$; ns=not significant)

To make such a statement, one needs the standard errors, not of the coefficients, but of the inverse transformation (2).

For this model, $\mathbf{Q}'$ is a 13×10 matrix and is of the form

$$\mathbf{Q}_3' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \end{bmatrix}$$

The vector of coefficients, from Table 2, is

$$\underline{\beta}' = (-1.9763, -0.1428, .0.3704, 0.8159, -0.5409, -0.3230, -0.0753, 0.2568, 0.0515, 0.1030)$$

Subclass values ($\hat{g}$'s) needed to calculate subclass mortalities are obtained as

$$\hat{\mathbf{g}} = \mathbf{Q}_3'\hat{\underline{\beta}}$$

and are

$$\hat{\mathbf{g}}' = (-1.9763, -2.1191, -1.8335, -1.6059, -2.3467, \ldots, -2.2643)$$

Estimates of subclass mortalities are obtained from the inverse transformation of $\hat{\mathbf{g}}$ as

$$\underline{\hat{\pi}} = \frac{\exp(\hat{\mathbf{g}})}{[1 + \exp(\hat{\mathbf{g}})]}$$

These are

$$\underline{\hat{\pi}}' = (0.1217, 0.1073, 0.1378, 0.1672, 0.0873, ..., 0.0941)$$

From these we see, for example, that the overall mortality was 12.2% and that mortality of female calves was 10.7% while that for males was 13.8%. Similarly, mortality estimates for the two respective herds were 16.7 and 8.7%.

4.3. Estimation of standard errors

The asymptotic variance-covariance matrix, $\hat{\mathbf{\Sigma}}$, of the coefficients in Table 2 is a 10×10 symmetric matrix. The variance-covariance matrix of $\hat{\mathbf{g}}$, $\mathbf{V}_{\mathbf{g}}$, is a 13×13 symmetric matrix and can be calculated as $\mathbf{Q}'\hat{\mathbf{\Sigma}}\mathbf{Q}$. Leading diagonal elements of this matrix contain the variances of the elements in $\hat{\mathbf{g}}$. It is only these elements that are relevant for our purposes. Let $\mathbf{v}_{\mathbf{g}}$ be a vector of these elements arranged sequentially (from 1st row, 1st column to last row, last column).

Then

$$\underline{\hat{\mu}}_z = \exp(\hat{\mathbf{g}} + 0.5\mathbf{v}_{\mathbf{g}})$$

and

$$\hat{\mathbf{v}}_z = \{\exp(\mathbf{v}_{\mathbf{g}} - \mathbf{1})\} \# \exp\{2\hat{\mathbf{g}} + \mathbf{v}_{\mathbf{g}}\}$$

where $\underline{\hat{\mu}}_z$ and $\hat{\mathbf{v}}_z$ are the respective mean and variance vectors of $\mathbf{z} = \exp(\hat{\mathbf{g}})$, 1 denotes a vector of 1's and # denotes the direct product of the corresponding elements in the two vectors.

Effect	Subclass	$\hat{g}$	$\hat{\mu}_z$	$\hat{v}_z$	$\hat{\pi}$	SE of $\hat{\pi}$
Overall		-1.976	.1393	.000199	.122	.0108
Sex	Female	-2.119	.1209	.000182	.107	.0107
	Male	-1.834	.1608	.000303	.138	.0129
Herd	1	-1.606	.2018	.000446	.167	.0146
	2	-2.347	.0963	.000126	.087	.0093
Datecls	1	-1.160	.3958	.093341	.239	.1319
	2	-2517	.0827	.000348	.075	.0156
	3	-2.299	.1007	.000085	.091	.0076
	4	-2.052	.1290	.000117	.114	.0085
	5	-1.720	.1799	.000270	.152	.0118
	6	-1.925	.1469	.000294	.127	.0130
	7	-1.873	.1554	.000549	.133	.0174
	8	-2.264	.1072	.000746	.094	.0217

Table 3. Estimated means ($\hat{\mu}_z$) and variances ($\hat{v}_z$) of z=exp($\hat{g}$), predicted mortalities ($\hat{\pi}$) and standard errors (SE) for the example data using SAS ML estimates and variance-covariance matrix.

From the two vectors, $\underline{\hat{\mu}}_z$ and $\hat{\mathbf{v}}_z$, $var(r_i)$ for all i is calculated according to (3). Table 3 presents $\hat{\mathbf{g}}$, $\underline{\hat{\mu}}_z$, $\hat{\mathbf{v}}_z$, $\hat{\pi}$ and the standard errors $[(var(r_i))^{1/2}]$ of $\hat{\pi}_i$. The results obtained from GENSTAT using this procedure with an appropriate $\mathbf{Q}'$ were identical to those in Table 3. Results in the GENSTAT output (which includes both mortality and standard errors) were however slightly different. The difference was due to the manner in which GENSTAT computes subclass coefficients. The above model consists of 32 (2×2×8) cells or "samples". GENSTAT computes response functions, $\hat{g}$'s, for each of the 32 cells, obtains the inverse function of each then calculates predicted probabilities (mortalities in this case) of the marginals as an average of the appropriate cells. For example, the mean predicted probability for the first subclass of DATECLS from the above example would be calculated as the mean of the values of male and female calves born in both herds in the first DATECLS subclass.

			Sample composition					SE of $\hat{\pi}$		
Sample	N	N_d	Sex	Herd	Dateclass	g	$\hat{\pi}$	GENSTAT	SAS	NEW
1	5	0	1	1	1	-0.933	.282	.1386	.1387	.1456
2	83	5	1	1	2	-2.290	.092	.0191	.0191	.0193
3	370	49	1	1	3	-2.072	.112	.0106	.0106	.0106
4	381	46	1	1	4	-1.824	.139	.0118	.0118	.0119
5	249	47	1	1	5	-1.492	.184	.0157	.0157	.0158
6	152	21	1	1	6	-1.697	.155	.0170	.0170	.0170
7	80	9	1	1	7	-1.646	.162	.0220	.0219	.0220
8	47	3	1	1	8	-2.037	.115	.0262	.0263	.0266
9	1	0	1	2	1	-1.674	.158	.0914	.0915	.0992
10	108	5	1	2	2	-3.030	.046	.0102	.0102	.0103
11	340	15	1	2	3	-2.813	.057	.0063	.0063	.0063
12	305	31	1	2	4	-2.565	.071	.0075	.0075	.0075
13	205	21	1	2	5	-2.233	.097	.0105	.0104	.0105
14	137	11	1	2	6	-2.438	.080	.0103	.0103	.0104
15	87	9	1	2	7	-2.387	.084	.0129	.0129	.0130
16	28	5	1	2	8	-2.778	.059	.0144	.0145	.0147
17	3	2	2	1	1	-0.647	.344	.1544	.1545	.1598
18	45	7	2	1	2	-2.004	.119	.0242	.0242	.0244
19	345	53	2	1	3	-1.786	.144	.0127	.0127	.0127
20	379	69	2	1	4	-1.538	.177	.0139	.0139	.0138
21	276	67	2	1	5	-1.206	.230	.0180	.0180	.0180
22	186	37	2	1	6	-1.412	.196	.0196	.0196	.0196
23	100	20	2	1	7	-1.360	.204	.0256	.0256	0.0257
24	45	6	2	1	8	-1.751	.148	.0320	.0321	.0324
25	2	1	2	2	1	-1.388	.200	.1099	.1100	.1180
26	67	5	2	2	2	-2.745	.060	.0133	.0133	.0134
27	376	21	2	2	3	-2.527	.074	.0078	.0078	.0078
28	326	26	2	2	4	-2.279	.093	.0092	.0091	.0091
29	191	18	2	2	5	-1.947	.125	.0125	.0125	.01256
30	173	20	2	2	6	-2.152	.104	.0124	.0124	.0125
31	113	15	2	2	7	-2.101	.109	.0158	.0158	.0158
32	56	4	2	2	8	-2.492	.076	.0183	.0183	.0184

Table 4. Predicted probabilities and standard errors of the individual cells (samples) of the example model from GENSTAT, SAS and the NEW method.

5. Conclusion

As has been alluded to, SAS does not compute subclass predicted probabilities (and their standard errors). However, like GENSTAT, SAS computes ML coefficients and corresponding probabilities for the response functions, $\hat{g}$'s, at the "sample" or cell level. Additionally, the two programs use identical procedures in the calculation of the statistics at the cell level. To facilitate a comparison of the "new" method with the methods used by these two programs, cell probabilities and their standard errors were calculated from the two outputs using the procedures outlined above. Table 4 presents predicted probabilities and standard errors extracted from the outputs of SAS and GENSTAT along with those obtained by employing the "NEW" method using coefficients ($\hat{\beta}$'s) and variance-covariance matrix ($\hat{\Sigma}$) from these packages. Results for the NEW method were the same regardless of which output $\hat{\beta}$ and $\hat{\Sigma}$ were obtained from, provided the appropriate $\mathbf{Q}'$ was used. Estimates of standard errors by SAS, GENSTAT and the NEW method were strikingly similar. Indeed, the only samples with what appeared to be real differences in SE's were those with 5 observations or less (i.e. samples 1, 9, 17 and 25 in Table 4), especially for cells with a single outcome (either all dead or all alive). In such cases, all the three methods differed in estimated SE's. It is to be noted that the NEW method was applied to figures from outputs. Thus rounding-off errors are likely to have had a contribution to the differences in estimated SE's between this method and the two packages. In conclusion, the NEW method does not differ in terms of estimated SE's from the more tedious and computationally demanding methods which involve inversion of matrices and taking of derivatives of the inverse of the link function. Moreover, the logic behind the method is simple and its implementation should be straightforward even for relatively unsophisticated users of logistic analysis. In addition to estimation of standard errors, specific statistical tests (e.g. the t-test) can be conducted and confidence intervals constructed using the variance-covariance matrix of the estimated probabilities.

Acknowledgements

We would like to thank Miss Roman Tirfie for word-processing the manuscript. The sample data was obtained from the Matopos Research Station of the Department of Research and Specialist Services, Ministry of Lands, Agriculture and Rural Resettlement, Zimbabwe. We wish to thank Mrs Siboniso Moyo and the present and past staff of the station.

References

[1] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (1975) *Discrete Multivariate Analysis: Theory and Practice*. The MIT Press, Cambridge, MA.

[2] Brown, M.B. (1974) The identification of sources of significance in two-way contingency tables. *Applied Statistics*, **23**, 405-413.

[3] Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman and Hall, London.

[4] Dobson, A. (1983). *An introduction to Statistical Modelling*. Chapman and Hall, New York.

[5] Fienberg, S.E. (1980) *The Analysis of Cross-Classified Categorical Data*. 2nd Ed. The MIT Press, Cambridge, MA.

[6] GENSTAT (1987). *The GENSTAT 5 Reference Manual*. Oxford Univ. Press, Oxford.

[7] Grizzle, J., Starmer, F. and Koch, G. (1969) Analysis of categorical data by Linear Models. *Biometrics*, **25**, 489-504.

[8] Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression*. John Wiley and Sons, New York.

[9] Imrey, P.B., Koch, G.G. and Stokes, M.E. (1981) Categorical data analysis: Some reflections on the log-linear model and logistic regression. Part I: Historical and Methodological Overview. *International Statistical Review*, **49**, 265-283.

[10] Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. and Lehnen, R.G. (1977) A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics*, **33**, 133-158.

[11] Lindgren, B.W. (1976) *Statistical Theory*. 3rd Ed. Macmillan Pub. Co., New York.

[12] McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*. 2nd Ed. Chapman and Hall, New York.

[13] Mood, A.M., Graybill, F.A. and Boes, D.C. (1974) *Introduction to the Theory of Statistics*. 3rd Ed. McGraw Hill Pub., New York.

[14] Rao, C.R. (1973) *Linear Statistical Inference and its Application*. 2nd Ed. John Wiley and Sons, New York.

[15] Rege, J.E.O. (1988) Estimation of sampling variances of moment-type estimates of genetic parameters. *Bulletin Animal Health and Production in Africa*, **36**, 112-119.

[16] Rege, J.E.O. (1992) *LOGMLVAR: A computer programme for estimating sampling variances of predicted probabilities from maximum likelihood estimates in a logit response function*. International Livestock Centre for Africa (ILCA), Addis Ababa, Ethiopia.

[17] SAS (1987) *SAS/STAT Guide for Personal Computers*, Version 6. Cary, NC: SAS Institute Inc.

[18] Wald, A. (1943) Tests of statistical hypotheses concerning general parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426-482.