Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

Judging Inference Adequacy in Logistic Regression

Dennis E. Jennings

To cite this article: Dennis E. Jennings (1986) Judging Inference Adequacy in Logistic Regression, Journal of the American Statistical Association, 81:394, 471-476

To link to this article: https://doi.org/10.1080/01621459.1986.10478292

Judging Inference Adequacy in Logistic Regression

DENNIS E. JENNINGS*

Inference for logistic regression based on the information matrix may be poor. This is noted in two examples in which confidence regions are examined. A measure to detect such inadequacies is presented; it judges the quadratic approximation to the likelihood surface, which justifies the usual procedure.

KEY WORDS: Confidence intervals; Binary data; Likelihood skewness; Information matrix.

1. INTRODUCTION

The logistic regression model is often used when the response is binary in nature. Parameters may be estimated by maximizing the likelihood, and confidence intervals are generally derived by inverting the observed (or expected) information matrix. This procedure can be justified by an asymptotic normality argument or as a quadratic approximation to the likelihood surface. However, the confidence intervals obtained in this manner may be poor, as illustrated in Section 2.

In the remaining sections a measure is developed to judge the adequacy of the inference obtained from the information matrix. The measure examines the cubic term of the Taylor expansion of the likelihood and thus is related to the skewness of the likelihood surface. Applications to several examples are given in Section 4. Finally, in Section 5 one possible approach for obtaining transformations to reduce the inference problem is discussed.

2. LOGISTIC REGRESSION

2.1 The Logistic Regression Model and Inference

The logistic regression model, or logistic model, may be employed when the data consist of a binary response and a set of explanatory variables. Specifically, for the tth case we observe $(y_t, x_{t1}, x_{t2}, \ldots, x_{tm})$, and n independent cases are sampled. The response y_t is binary and will be coded as 0 or 1 and generally referred to as failure or success. The vectors $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tm})'$ are the explanatory variables and may be continuous or discrete. The logistic model equates the logarithm of the conditional odds of success to failure to a linear function of the explanatory variables. Thus it proposes that

$$\text{logit}(p_t) = \log(p_t/(1 - p_t)) = x_{t1}\beta_1 + x_{t2}\beta_2 + \cdots + x_{tm}\beta_m.$$

A standard result in maximum likelihood estimation is that under certain regularity conditions $\hat{\boldsymbol{\beta}}$ is asymptotically normal with mean $\boldsymbol{\beta}$ and variance equal to the inverse of the observed or expected information matrix. This result is used in logistic regression to test hypotheses and form confidence regions. In this setting, the observed and expected information are identical and will be denoted by V.

Likelihood confidence regions can also be formed. These regions can be displayed as contours of constant likelihood in two dimensions and perhaps by cross-sections with three parameters. When more parameters are involved, a simplification of this surface is generally sought. One such simplification approximates the log-likelihood by a quadratic surface around the maximum likelihood estimate. Thus suppressing the dependence of the likelihood on X and $\mathbf{y}$, we have

$$\begin{aligned} l(\boldsymbol{\beta}) &\simeq l(\hat{\boldsymbol{\beta}}) + \tfrac{1}{2}\sum\sum (\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)\left.\frac{\partial^2 l}{\partial\beta_i\partial\beta_j}\right|_{\hat{\boldsymbol{\beta}}} \\ &= l(\hat{\boldsymbol{\beta}}) - \tfrac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'V(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ &\equiv l_q(\boldsymbol{\beta}). \end{aligned} \tag{2.1}$$

Using expression (2.1), the likelihood regions can be approximated by $\{\boldsymbol{\beta} \mid (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'V(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) < \tfrac{1}{2}c\}$. Now taking $c = \chi^2_m$ yields the same regions as would be attained from the usual assumption that $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, V^{-1})$. This connection results from the fact that normal distributions have quadratic log-likelihood surfaces.

Generally the likelihood-based regions are preferred to those based on the asymptotic normality of the sampling distribution [Cox (1970), Edwards (1972), Sprott (1973), and Hinkley (1978)]. A particular advantage is that the likelihood regions are invariant under 1–1 transformations of the parameters, which is not true for regions using the approximate normality of $\hat{\boldsymbol{\beta}}$. From a practical point of view it is desirable to use the approximation to the likelihood regions, with its ease in summarization, when the regions are close to likelihood regions. In this article, we develop methods to judge the adequacy of this approximation. Minkin (1983) suggested a component-wise evaluation of this approximation, which combines upper bounds for each observation to obtain an upper bound of the total error. Although this avoids a p-dimensional search, it may not be any easier to calculate for logistic regression data with large n and continuous explanatory variables. Also, with common sample sizes one might expect the evaluation based on the n components to be too large.

2.2 Inference Problems in Logistic Regression

Example 2.1. This example is similar to those given by Hauck and Donner (1977). The setting is an experiment with two groups, one receiving a treatment and the other a control. On completion each individual is labeled as a success or a failure. Suppose that our interest is whether an individual in one group is more likely to respond favorably, that is, whether

* Dennis E. Jennings is Assistant Professor, Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL 61801. This work is partly based on the author's Ph.D. dissertation, prepared at the University of Minnesota with fellowship support from the university. The author thanks his advisor, Kinley Larntz, for help and encouragement.

Table 1. Hypothetical Results From Two-Group Experiment

	Experiment 1		Experiment 2	
	Group 1	Group 2	Group 1	Group 2
Successes	100	198	100	199
Failures	100	2	100	1
$\hat{\beta}_1$	4.595		5.293	
$\hat{\beta}_1/se\hat{\beta}_1$	6.28		5.22	
99% C.I.	(2.73, 6.46)		(2.68, 7.91)	
Square root of lik. ratio test	12.43		12.73	

one group has a higher probability of success. This can be modeled in logistic regression by letting

$$x_i = 0 \text{ if the } i\text{th subject is in group one}$$

$$= 1 \text{ if the } i\text{th subject is in group two}$$

and

$$\log(P_i/(1 - P_i)) = \beta_0 + \beta_1 x_i.$$

If P_{G1} and P_{G2} are the probabilities of success for group one and group two, respectively, then $\log(P_{G2}/(1 - P_{G2})) - \log(P_{G1}/(1 - P_{G1})) = \beta_1$. That is, β_1 measures the difference in the probabilities of success on the logit scale.

Now imagine two experimenters collect data and get results given in Table 1. In both cases it appears clear that group two has a higher probability of success than group one. It is also clear that there is more evidence for this conclusion in experiment two. As expected, the estimates of β_1 in the logistic model show a larger difference between the two probabilities on the logit scale in experiment two (5.29 to 4.60).

Using the estimate of the standard error of $\hat{\beta}_1$ available from the inverse of the information matrix, $z = \hat{\beta}_1/se(\hat{\beta}_1)$ would be asymptotically $N(0, 1)$ under the null hypothesis that $\beta_1 = 0$. For these experiments z is 6.28 for experiment one and 5.22 for experiment two. Both yield strong evidence that $\beta_1 \neq 0$; however, it is surprising that the z-value for experiment two, which has stronger evidence for $P_{G2} > P_{G1}$ or equivalently $\beta_1 > 0$, is smaller than the z-value for experiment one. Hauck and Donner showed this phenomenon would occur as $\hat{P}_2 \to 1$ for any $\hat{P}_1$.

Alternatively, one might use the likelihood itself to test $\beta_1 = 0$. If $l(\boldsymbol{\beta}, X, y)$ is the logarithm of the likelihood of $\boldsymbol{\beta}$ and if $\boldsymbol{\beta}^*$ maximizes the likelihood over the set $\{\boldsymbol{\beta} \mid \beta_1 = 0\}$, then under $\beta_1 = 0$, $\Lambda = 2[l(\hat{\boldsymbol{\beta}}, X, \mathbf{y}) - l(\boldsymbol{\beta}^*, X, \mathbf{y})]$ asymptotically has a chi-squared distribution with one degree of freedom. Thus under $\beta_1 = 0$, the square root of Λ would be approximately $|N(0, 1)|$. In the two experiments the square roots of Λ obtained are 12.43 and 12.73. This is strong evidence in both cases for $\beta_1 \neq 0$, with more evidence expressed (larger Λ value) in experiment two as expected. It is interesting to note that although both $\hat{\beta}_1/se\hat{\beta}_1$ and the likelihood statistic are comparable to normal variables for a test of the same hypothesis, the magnitude of the latter is much greater. This hints at a problem with the inference based on the information matrix.

This problem becomes more apparent when 99% confidence intervals based on the information matrix are calculated. The intervals for β_1 are given in the table. Surprisingly, the interval for experiment two contains smaller values than does the interval for experiment one.

Since inference based on the information matrix can be viewed as approximating the likelihood surface, apparently the approximation is poor in this case. In this example there are only two parameters; thus it is not difficult to find the likelihood regions. The likelihood contours corresponding to nominal 50%, 75%, 95%, and 99% likelihood regions are plotted for the two experiments in Figure 1. The vertical lines represent the 99% confidence intervals for β_1 constructed from the information matrix. These do not agree well with the likelihood regions. One can note that these regions are not elliptical, as the regions obtained from the information matrix must be. In Figure 2, the 95% and 99% joint confidence regions obtained from the information matrix for experiment two are plotted with the corresponding likelihood regions they attempt to approximate. The elliptical regions poorly approximate the likelihood regions.

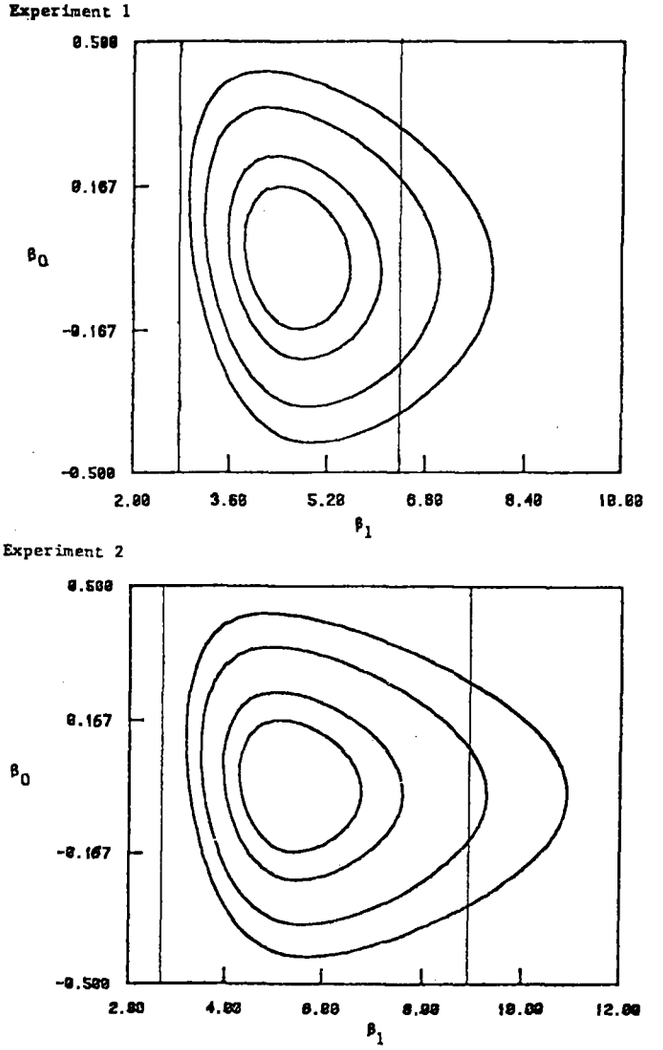

Figure 1. Likelihood Contours. Curves correspond to 50%, 75%, 95%, and 99% likelihood confidence regions. Horizontal lines indicate where 99% confidence intervals for β_1 fall. Such intervals would not agree well with likelihood intervals.

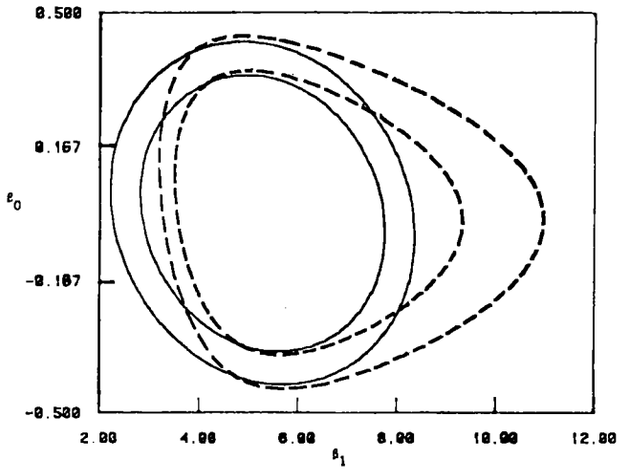

Figure 2. Confidence Regions Using the Likelihood and Information Matrix for Example 2.1, Experiment 2. Solid lines are 95% and 99% confidence regions based on asymptotic normality (information matrix). Dashed lines are 95% and 99% likelihood confidence regions. The shapes of regions are quite different.

Example 2.2. The previous example with a discrete X variable showed a lack of agreement between inference based on the information matrix and likelihood inferences. The following example shows that continuous X values can present similar problems and that coverage rates obtained from the usual inference procedure may be poor. Observations were generated from the logistic model $\text{logit}(p_i) = 2 + 4x_i$, where $x_i = -1(.02)1$. That is, Y_i, $i = 1, \ldots, 101$, were generated with $P(Y_i = 1) = 1/(1 + \exp(-(2 + 4x_i)))$ and $P(Y_i = 0) = 1 - P(Y_i = 1)$. This model was simulated 540 times, and 95% and 90% confidence intervals were constructed using the information matrix. Table 2 gives the number of times the confidence interval contained the true value, the number of times the interval was completely above the true value, and the number of times the interval was below the true value. Tests for whether $P(\text{contains true value}) = \alpha$ and $P(\text{below}) = P(\text{above}) = (1 - \alpha)/2$ for $\alpha = .95$ and .90 are rejected using the observed data. The coverage rates appear too high, about $97\frac{1}{2}\%$ for nominal 95% intervals and 96% for the 90% intervals. Also, when the interval does not cover the true value, it is usually above the true value. This suggests that the distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ is skewed.

Regions using the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and the information matrix are easier to use than likelihood regions and easily yield marginal regions if our approximations are appropriate. Our goal is to find a measure of the adequacy of such approximations to likelihood regions. The idea is to use such approximations when appropriate and make some adjustments when they are not. Adjustments might include transforming parameters or actually plotting likelihoods.

3. JUDGING INFERENCE ADEQUACY

3.1 The Quadratic Approximation

The basic assumption that yields inference based on the inverse of the observed information matrix is that the log-likelihood is well approximated by a quadratic expansion about $\hat{\boldsymbol{\beta}}$ as given by expression (2.1). If the third and all higher derivatives were zero, expression (2.1) would be exact. Asymptotically, these higher-order terms are dominated by the quadratic term. However, this does not imply that the cubic terms are negligible in any finite sample. Therefore, for a first-order evaluation of the quadratic approximation, we might examine the third derivatives of the likelihood. Large values will suggest potential problems.

3.2 Evaluating the Quadratic Approximation

The procedure outlined here compares the approximate region, obtained using the quadratic approximation, in terms of its agreement with the likelihood region. Specifically we examine the likelihood along the boundary of the approximate region. Lemma 1 gives the general form of these boundary points, and Lemma 2 gives an approximation to the relative error at these points.

Lemma 1. Boundary points of the quadratic region are of the form

$$\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}} + (\chi^2_{m,\alpha}/\mathbf{h}'V\mathbf{h})^{1/2}\mathbf{h} \equiv \hat{\boldsymbol{\beta}} + \mathbf{h}c_q(\mathbf{h}).$$

This follows directly from (2.1) and the form of the confidence region.

Table 2. Coverage Rates

	Confidence Interval		
	$\beta_0 = 2$	$\beta_1 = 4$	Expected
	95% Confidence Intervals		
Above true value	10 (.019)	9 (.017)	13.5 (.025)
Contains true value	527 (.976)	528 (.978)	513 (.95)
Below true value	3 (.006)	3 (.006)	13.5 (.025)
G^2	9.3	10.86	
	90% Confidence Intervals		
Above true value	18 (.033)	13 (.024)	27 (.05)
Contains true value	518 (.959)	520 (.963)	486 (.90)
Below true value	4 (.007)	7 (.013)	27 (.05)
G^2	24.6	23.2	

NOTE: G^2 is the likelihood ratio test statistic to test whether observed values are consistent with expected values. G^2 is approximately distributed as X^2_2 if expected rates hold. $X^2_2(.05) = 5.99$.

Lemma 2.

$$\Delta_{\mathbf{h}} = \frac{\sqrt{\chi^2_{m,\alpha}}}{3} \frac{\sum_i \sum_j \sum_k h_i h_j h_k v_{ijk}}{|\mathbf{h}'V\mathbf{h}|^{3/2}}$$

estimates the relative error in approximating the likelihood by the quadratic approximation at $\boldsymbol{\beta}_0$, the boundary point in the direction $\mathbf{h}$ from $\hat{\boldsymbol{\beta}}$. Here,

$$v_{ijk} = \partial^3 l(\beta)/\partial\beta_i\partial\beta_j\partial\beta_k$$

evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$.

For a fixed direction $\mathbf{h}$, we might approximate $|l(\boldsymbol{\beta}_0) - l_q(\boldsymbol{\beta}_0)|$ by $|l_c(\boldsymbol{\beta}_0) - l_q(\boldsymbol{\beta}_0)|$, where l_c is the cubic approximation to the likelihood and l_q is the quadratic approximation of (2.1). This is the error on the likelihood scale if we use $\boldsymbol{\beta}_0$ as a boundary point. The size of this error can be judged relative to $\frac{1}{2}\chi^2_{m,\alpha}$, the desired difference in likelihood at the boundary point and maximum likelihood estimate. As a relative error, this is

$$\Delta_{\mathbf{h}} = \frac{|\{l_q(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\} - \{l_c(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})\}|}{|l_q(\boldsymbol{\beta}_0) - l(\hat{\boldsymbol{\beta}})|} = \frac{|l_q(\boldsymbol{\beta}_0) - l_c(\boldsymbol{\beta}_0)|}{\frac{1}{2}\chi^2_{m,\alpha}}.$$

Since

$$l_c(\boldsymbol{\beta}) - l_q(\boldsymbol{\beta}) = \tfrac{1}{6}\sum_i \sum_j \sum_k (\beta_i - \hat{\beta}_i)(\beta_j - \hat{\beta}_j)(\beta_k - \hat{\beta}_k)v_{ijk},$$

$\Delta_{\mathbf{h}}$ can be written as given in the lemma.

One might look for the direction where this relative error is largest. This yields a measure of inference adequacy based on the cubic approximation to the likelihood surface, which will be denoted as $\Delta = \max_{\mathbf{h}} \Delta_{\mathbf{h}}$. If this relative error is large, inference based on the observed information is likely to be poor. A small value implies that third derivatives do not greatly affect our inference statements. There is no claim here that the cubic approximation is adequate. It is hoped that when the quadratic approximation is poor the cubic approximation will be significantly different. This can be anticipated by noting that an upper bound for the error at $\hat{\boldsymbol{\beta}} + c\mathbf{h}$ is

$$c^3 \left| \sum_i \sum_j \sum_k \left[\frac{\partial^3 l(\boldsymbol{\beta})}{\partial\beta_i\partial\beta_j\partial\beta_k} \right]_{\hat{\boldsymbol{\beta}} + \varepsilon\mathbf{h}} \right| \Big/ 6,$$

where $0 < \varepsilon < c$ for any c. Thus $|l_c(\mathbf{h}) - l_q(\mathbf{h})|$ approximates this upper bound by evaluating the third derivatives at $\hat{\boldsymbol{\beta}}$ instead of $\hat{\boldsymbol{\beta}} + \varepsilon\mathbf{h}$.

For problems with a single parameter, this corresponds to a measure examined by Sprott (1973). There it can be viewed as local measure of skewness of the likelihood function. In the multiparameter problem, we are looking at this measure on all slices of the likelihood surfaces through $\hat{\boldsymbol{\beta}}$.

3.3 Calculation of the Cubic Measure of Inference Inadequacy

In this section an iterative procedure is developed to calculate Δ. The procedure succeeds by noting that Δ is not affected by a linear transformation of the parameters or the norm of $\mathbf{h}$. Thus a transformation is made so that the information matrix is the identity. This implies that if $\|\mathbf{h}\| = 1$, we need only maximize the numerator. An algorithm to accomplish this is presented.

The following theorem implies that linear transformations do not affect Δ.

Theorem 3.1. Let A be an $m \times m$ full rank linear transformation. Then for $\boldsymbol{\phi} = A^{-1}\boldsymbol{\beta}$, $V_{\boldsymbol{\phi}} = A'V_{\boldsymbol{\beta}}A$ and $\Delta_{\boldsymbol{\beta}} = \Delta_{\boldsymbol{\phi}}$.

This follows by evaluating $\partial l/\partial\phi_j$ by the chain rule and differentiating to get second and third derivatives.

Theorem 3.1 allows us to solve a simpler maximization problem by assuming $V = I$. If $V \neq I$ then we transform to new coordinates $\boldsymbol{\phi}$ by $\boldsymbol{\phi} = V^{1/2}\boldsymbol{\beta}$, where $V^{1/2}$ is $m \times m$ symmetric such that $V^{1/2}V^{1/2} = V$. Letting $a = \sqrt{\chi^2_{m,\alpha}}/3$ by Theorem 3.1,

$$V_{\boldsymbol{\phi}} = V^{-1/2}VV^{-1/2} = I$$

and

$$\Delta = \Delta_{\boldsymbol{\phi}} = \max_{\mathbf{h}\in R^m} \frac{a \cdot \left| \sum_i \sum_j \sum_k h_i h_j h_k v^{\boldsymbol{\phi}}_{ijk} \right|}{|\mathbf{h}'\mathbf{h}|^{3/2}} \equiv \max_{\mathbf{h}\in R^m} a\Gamma(\mathbf{h}).$$

Now $\Gamma(c\mathbf{h}) = \Gamma(\mathbf{h})$; thus

$$\Delta_{\boldsymbol{\phi}} = a \cdot \max_{\|\mathbf{h}\|=1} \Gamma(\mathbf{h}) = a \cdot \max_{\|\mathbf{h}\|=1} \left| \sum_i \sum_j \sum_k h_i h_j h_k v^{\boldsymbol{\phi}}_{ijk} \right|.$$

An algorithm for maximizing

$$\Phi(\mathbf{h}) = \left| \sum_i \sum_j \sum_k h_i h_j h_k v^{\boldsymbol{\phi}}_{ijk} \right|$$

is developed that relies on the fact that for this problem if $\mathbf{h}^*$ locates a local maximum, the gradient at $\mathbf{h}^*$ will be in the same direction. This mimics the algorithm used by Bates and Watts (1980), and further details can be found in Jennings (1982). Because of the symmetry in the v_{ijk} values here, the gradient has rth component

$$g_r = 3\sum_j \sum_k h_j h_k v^{\boldsymbol{\phi}}_{rjk}.$$

The procedure has the following steps:

1. Choose an initial guess $\mathbf{h}^0$;
2. Calculate the gradient $\mathbf{g}^i$ at $\mathbf{h}^i$, normalize $\tilde{\mathbf{g}}^i = \mathbf{g}^i/\|\mathbf{g}^i\|$;
3. If $(\tilde{\mathbf{g}}^i)'\mathbf{h}^i > 1 - \varepsilon$, take $\mathbf{h}^i$ as solution; otherwise take $\mathbf{h}^{i+1} = \tilde{\mathbf{g}}^i$ and repeat from step 2; and
4. Calculate Φ at solution.

The convergence criterion in step 3 can be varied to get desired accuracy in Δ. The use of $\varepsilon = .00001$ generally achieved at least three-decimal accuracy in Δ. A starting value might be chosen from the set of basis vectors $\mathbf{e}_i$. Since $\Phi(\mathbf{e}_i) = v^{\boldsymbol{\phi}}_{iii}$, $\max_i \Phi(\mathbf{e}_i) = \Phi(\mathbf{e}_j)$ if $\max_i v^{\boldsymbol{\phi}}_{iii} = v^{\boldsymbol{\phi}}_{jjj}$. Thus choosing $\mathbf{e}_j$ corresponding to the largest diagonal element of the array of third derivatives is appropriate. Convergence of the algorithm in the examples considered appears quite satisfactory. Two-digit accuracy generally occurs in four or fewer iterations. Thus far, various starting values have always yielded the same maximum. Once we have the third derivative matrix, each iteration takes on the

order of m^2 calculations. Thus minimal computer time is used per iteration, and the number of iterations needed may increase with m and the desired accuracy.

4. APPLICATION IN LOGISTIC REGRESSION

The procedure discussed is easily applied in logistic regression. The second derivative matrix V is generally available from any logistic regression program, since it is needed for inference. In addition, the third derivatives of the log-likelihood must be calculated. A complete discussion of the calculation, assuming we have found the maximum likelihood estimate, will be given. The method will then be applied to several data sets.

4.1 Calculation of Δ

Given X the $n \times m$ design matrix and response $\mathbf{y}$ an $n \times 1$ vector, we find the MLE $\hat{\boldsymbol{\beta}}$. After obtaining $\hat{\boldsymbol{\beta}}$, we calculate $\hat{p}_t = [1/(1 + \exp(-\hat{\boldsymbol{\beta}}'\mathbf{x}_t))]$, the estimated probability of success for the tth individual. Here $V = X'WX$, where W is an $n \times n$ diagonal matrix with elements $p_t(1 - p_t)$. The third derivatives evaluated at $\hat{\boldsymbol{\beta}}$ are

$$v_{ijk} = \sum_{t=1}^{n} x_{ti}x_{tj}x_{tk}\hat{p}_t(1 - \hat{p}_t)(2\hat{p}_t - 1).$$

Now we can find $V^{-1/2}$ (a LINPACK routine was used) and calculate the third derivatives in the transformed parameterization $\boldsymbol{\phi} = V^{1/2}\boldsymbol{\beta}$. If $B = V^{-1/2}$ and has entries b_{ij}, then

$$v^{\phi}_{ijk} = \sum_r \sum_s \sum_t v_{rst}b_{ri}b_{sj}b_{tk}.$$

Next, we find $\mathbf{h}^*$, which maximizes

$$\Phi(\mathbf{h}) = \left|\sum \sum \sum h_i h_j h_k v^{\phi}_{ijk}\right|$$

for $\mathbf{h}$ on the unit circle by iterative procedure described in Section 3.3. Then $\Delta = (\sqrt{\chi^2_{m,\alpha}}/3)\Phi(h^*)$.

4.2 Applications

This inference adequacy measure was calculated for several logistic regression data sets. Table 3 gives the results when one is considering the effect on 95% confidence intervals.

In the two group experiments (Example 2.1) the relative errors are .57 and .81 for experiment one and experiment two, respectively. This is a clear indication of potential inference problems, as it suggests in experiment one that the true likelihood at a boundary point could be off by as much as $.57\chi^2_{2,.05} = 3.41$ when the desired difference from the MLE is $\chi^2_{2,.05} = 5.99$. Thus the boundary defined may contain points that are actually boundaries for 75% or 99% likelihood confidence regions ($\chi^2_{2,.75} = 2.77$ and $\chi^2_{2,.99} = 9.21$).

Table 3. Relative Errors for 95% Confidence Intervals for Two-Group Experiments (successes/failures)

Group 1	Group 2	Δ
100/100	198/2	.568
100/100	199/1	.810
50/50	60/40	.033
50/50	80/20	.122
50/50	98/2	.580
50/50	99/1	.804
25/75	10/90	.218
25/75	5/95	.337
25/72	3/97	.450
25/75	1/99	.804
A data set from Example 3.2		.429
Vasoconstriction data		.719

Table 4. Vasoconstriction Data (Finney 1947)

Vol.	Rate	Resp.	Vol.	Rate	Resp.
3.7	.825	1	.4	2.0	0
3.5	1.09	1	.95	1.36	0
1.25	2.5	1	1.35	1.35	0
.75	1.5	1	1.5	1.36	0
.8	3.2	1	1.6	1.78	1
.7	3.5	1	.6	1.5	0
.6	.75	0	1.8	1.5	1
1.1	1.7	0	.95	1.9	0
.9	.75	0	1.9	.95	1
.9	.45	0	1.6	.4	0
.8	.57	0	2.7	.75	1
.55	2.75	0	2.35	.03	0
.6	3.0	0	1.1	1.83	0
1.4	2.33	1	1.1	2.2	1
.75	3.75	1	1.2	2.0	1
2.3	1.64	1	.8	3.33	1
3.2	1.6	1	.95	1.9	0
.85	1.415	1	.75	1.9	0
1.7	1.06	0	1.3	1.625	1
1.8	1.8	1			

NOTE: Vol. = volume and Resp. = response. The response y indicates the occurrence (1) or nonoccurrence (0) of vasoconstriction in the skin of the digits.

Several other possible outcomes for the experiment with two groups are considered. A clear increase is noted as one of the probabilities of success approaches zero or one.

In Example 2.2, a simulation was performed in which confidence intervals based on the information matrix proved unreliable. A typical data set was selected from those generated. Although for this data set the true value is contained in the 95% confidence region, $\Delta = .761$. This indicates potential problems with approximate inference under this model.

The problem may occur in actual data, as can be noted in the vasoconstriction data examined by Finney (1947). The response is the occurrence (1) or nonoccurrence (0) of vasoconstriction in the skin of the digits. The explanatory variables are volume and rate of air inspired. The data are given in Table 4. In actuality, only three subjects were tested, one 9 times, the second 8 times, and the third 22 times. However, the information matching results to subjects is not available. For purposes of example only, we will assume independent observations. The model suggested by Pregibon (1979) is $\text{logit}(p) = \beta_0 + \beta_1\log(\text{VOLUME}) + \beta_2\log(\text{RATE})$. Using this model and considering 95% confidence intervals, $\Delta = .719$, which suggests major inference problems.

One property of Δ that has not been noted is that it goes to zero at the rate of $1/\sqrt{n}$. Thus, in a given experiment, doubling the observations will approximately reduce Δ by a factor of $\sqrt{2}$. This effect is noted in the table (by observing that the line 50/50 99/1 .804 gives a value for Δ which is a factor of $\sqrt{2}$ larger than for the first line 100/100 198/2 .568).

5. PARAMETER TRANSFORMATION

If inference inadequacy is found, alternatives must be sought. One approach is to try to find a new parameterization for which

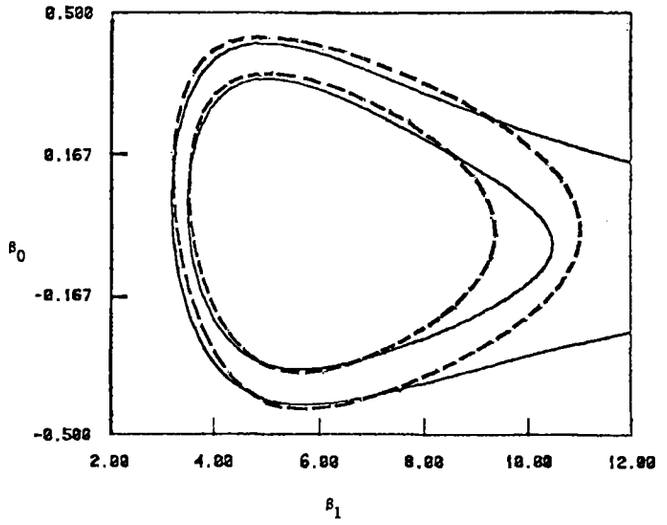

Figure 3. Confidence Regions Using the Likelihood and Transformed Parameters for Example 2.1. Solid lines are 95% and 99% confidence regions using asymptotic normality on transformed parameters. Dashed lines are 95% and 99% likelihood confidence regions.

the approximation is more acceptable. The inference adequacy measure could be used to compare suggested transformations. More ambitiously, one might try to find a parameterization with relative error equal to zero. Generally this will be unattainable, for $\Delta = 0$ implies all mixed third derivatives must be zero. Instead, the following discussion considers a restricted set of parameterizations and focuses on reducing only the $\partial^3 l/\partial\beta_i^3$ terms. Kass (1984) gives a geometrical view of the parameterization problem.

A simple set of transformations are those taking a single β_i into a single new parameter. We call such a transformation a univariate transformation. Thus $\boldsymbol{\phi} = g(\boldsymbol{\beta})$, such that $\phi_i = g_i(\beta_i)$. An advantage of such transformations is that marginal confidence intervals can be related to the original parameters by transforming the endpoints of intervals obtained in the new parameterization (provided the g_i are monotonic).

We would like to choose a univariate transformation so as to reduce the relative error of the quadratic approximation. This can be viewed as trying to get small third derivatives. In particular, the transformation $\phi_i = \exp((-v_{iii}/3v_{ii})\beta_i)$ for all i implies that $v_{iii}^{\boldsymbol{\phi}} = 0$ for all i. This univariate transformation produces local symmetry for the log-likelihood surface along the axes, that is, in directions $\mathbf{h}$ with $h_i = 1$ for some i and $h_j = 0$ if $i \neq j$. It does not guarantee improvement along any other direction. In fact, Δ may be increased or decreased. The method tends to be successful when the $\{v_{iii}\}$ are the dominant terms in the array of third derivatives.

Obviously, we expect improvement in one-parameter situations. Generally, one-parameter situations do not arise in logistic regression; however, one might use the logistic parameterization to estimate the probability of success in a single population and very clear improvements can be noted [see Jennings (1982)].

Substantial improvement is also found in the two-group data of Example 2.1. Here if we apply $\phi_i = \exp((-v_{iii}/3v_{ii})\beta_i)$, the curvature measure Δ is reduced from .57 to .08 in experiment one and from .81 to .08 in experiment two. Confidence intervals formed on the transformed scale and transformed back to intervals for β are (3.13, 7.48) and (3.41, 11.29), respectively, which agree more closely with the likelihood contours exhibited in Figure 1. Also the interval for experiment two does not contain smaller values than those contained in the interval from experiment one, as was the case under the usual inference method. Perhaps the strongest argument for the validity of this transformation is the agreement of 95% and 99% likelihood confidence regions with those regions obtained from the transformation (Figure 3). The agreement is much better than the regions obtained in the original scale using the information matrix (Figure 2).

This parameterization is affected by linear transformation of the x-variables and equivalently of the parameters. If a linear transformation is applied to $\boldsymbol{\beta}$ to obtain new parameterization $\boldsymbol{\gamma} = T\boldsymbol{\beta}$, where T is $m \times m$ and of full rank, then $\phi_i = \exp((-v_{iii}/3v_{ii})\gamma_i)$ is a different parameterization than when this transformation is applied to $\boldsymbol{\beta}$. This suggests we might try to find a linear transformation such that the diagonal elements of the third derivative array are dominant.

The linear transformation such that $V = I$ might be considered. In this case the quadratic expansion suggests that the parameters are approximately independent. If that is the case the mixed third derivatives would be small. Of course a large Δ implies this quadratic expansion is not very good, so the mixed derivatives need not be small. The resulting transformation is not a univariate transformation of $\boldsymbol{\beta}$, but might be considered.

In the vasoconstriction data, direct application of the univariate transformation discussed earlier reduces Δ from .72 to .49 when the x-variables are scaled to have mean zero. Rotating the parameters to obtain a diagonal information matrix and then applying the univariate transformation produces $\Delta = .42$.

Other procedures for finding useful transformations are needed. Useful inference in logistic regression depends on recognizing when such difficulties arise and developing procedures to handle them.

[*Received November 1983. Revised August 1985.*]

REFERENCES

Bates, D. M., and Watts, D. G. (1980), "Relative Curvature Measures of Nonlinearity" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 40, 1–25.

Cox, D. R. (1970), *The Analysis of Binary Data*, London: Methuen.

Edwards, A. W. F. (1972), *Likelihood*, Cambridge, England: University Press.

Finney, D. J. (1947), "The Estimation From Individual Records of the Relationship Between Dose and Quantal Response," *Biometrika*, 34, 320–334.

Hauck, W. W., and Donner, A. (1977), "Wald's Test as Applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association*, 72, 851–853.

Hinkley, D. V. (1978), "Likelihood Inference About Location and Scale Parameters," *Biometrika*, 65, 253–261.

Jennings, D. E. (1982), *Inference and Diagnostics for Logistic Regression*, Ph.D. dissertation, University of Minnesota, School of Statistics.

Kass, R. (1984), "Canonical Parameterizations and Zero Parameter Effects Curvature," *Journal of the Royal Statistical Society*, Ser. B, 46, 86–92.

Minkin, S. (1983), "Assessing the Quadratic Approximation to the Log Likelihood Function in Nonnormal Linear Models," *Biometrika*, 70, 367–372.

Pregibon, D. (1979), *Data Analytic Methods for Generalized Linear Models*, Ph.D. dissertation, University of Toronto, Department of Statistics.

Sprott, D. A. (1973), "Normal Likelihoods and Their Relation to Large Sample Theory of Estimation," *Biometrika*, 60, 457–465.