STATISTICS IN MEDICINE, VOL. 16, 965–980 (1997)

A COMPARISON OF GOODNESS-OF-FIT TESTS FOR THE LOGISTIC REGRESSION MODEL

D. W. HOSMER,*[1] T. HOSMER,[2] S. LE CESSIE[3] AND S. LEMESHOW[1]

[1] *Department of Biostatistics and Epidemiology, University of Massachusetts, Arnold House, Box 30430, Amherst, MA 01004-0430, U.S.A.*

[2] *University Computer Centre, University of Massachusetts, Amherst, MA 01004, U.S.A.*

[3] *Department of Medical Statistics, University of Leiden, The Netherlands*

SUMMARY

Recent work has shown that there may be disadvantages in the use of the chi-square-like goodness-of-fit tests for the logistic regression model proposed by Hosmer and Lemeshow that use fixed groups of the estimated probabilities. A particular concern with these grouping strategies based on estimated probabilities, fitted values, is that groups may contain subjects with widely different values of the covariates. It is possible to demonstrate situations where one set of fixed groups shows the model fits while the test rejects fit using a different set of fixed groups. We compare the performance by simulation of these tests to tests based on smoothed residuals proposed by le Cessie and Van Houwelingen and Royston, a score test for an extended logistic regression model proposed by Stukel, the Pearson chi-square and the unweighted residual sum-of- squares. These simulations demonstrate that all but one of Royston's tests have the correct size. An examination of the performance of the tests when the correct model has a quadratic term but a model containing only the linear term has been fit shows that the Pearson chi-square, the unweighted sum-of-squares, the Hosmer–Lemeshow decile of risk, the smoothed residual sum-of-squares and Stukel's score test, have power exceeding 50 per cent to detect moderate departures from linearity when the sample size is 100 and have power over 90 per cent for these same alternatives for samples of size 500. All tests had no power when the correct model had an interaction between a dichotomous and continuous covariate but only the continuous covariate model was fit. Power to detect an incorrectly specified link was poor for samples of size 100. For samples of size 500 Stukel's score test had the best power but it only exceeded 50 per cent to detect an asymmetric link function. The power of the unweighted sum-of-squares test to detect an incorrectly specified link function was slightly less than Stukel's score test. We illustrate the tests within the context of a model for factors associated with low birth weight. © 1997 by John Wiley & Sons, Ltd. Stat. Med., Vol. 16, 965–980 (1997).

(No. of Figures: 0 No. of Tables: 7 No. of Refs: 24)

INTRODUCTION

The logistic regression model has become a widely used and accepted method of analysis of binary outcome variables. This popularity stems from the availability of easily used software in both mainframe and microcomputer packages and the ease of interpretation of the results of the

* Correspondence to: D. W. Hosmer

CCC 0277–6715/97/090965–16$17.50

Received May 1995
Revised May 1996

Table I. Estimated coefficients, standard errors and *p*-values for the fitted logistic model to the low birth weight data

Variable	Coefficient	Standard error	*p*-value
AGE	−0·022	0·034	0·511
LWT	−0.013	0·006	0·050
RACE_1	1·236	0·517	0·017
RACE_2	0·946	0·416	0·023
SMOKE	1·054	0·380	0·006
CONSTANT	0·332	1·118	0·764

fitted model, be it used for estimating probabilities and/or odds ratios. Commensurate with this increase in application has been an increase in statistical research on the model. One area of current research is the development of new methods to assess the adequacy of the fitted model. The need for this research has been motivated by limitations of currently available methods.

AN EXAMPLE OF THE PROBLEMS IN ASSESSING OVERALL GOODNESS-OF-FIT

To illustrate some of the problems with currently available methods of assessing overall goodness-of-fit[1] we present the results of the fit of a model using the low birth weight data in Appendix I of Hosmer and Lemeshow.[2] These data were collected at Baystate Medical Center in Springfield, Massachusetts, in 1986. The outcome variable was whether or not birth weight was less than 2500 grams. Data were collected on 189 births of which 59 were low birth weight and 130 were normal birth weight. The purpose of this example is to illustrate problems with assessing model fit rather than to provide a definitive analysis of these data. The independent variables used in this example are age of the mother (AGE), weight of the mother at the last menstrual period (LWT), race of the mother, (white, black, or other, coded into two design variables using white race as the referent group (RACE_1, RACE_2)) and whether or not the mother smoked, 1 = yes, 0 = no, (SMOKE)). To avoid differences between packages when ties are present in the estimated probabilities we added the value of an independent U(0, 1) variate to each AGE and LWT. The fitted model using the 'jittered' data agreed to three decimal places with the 'unjittered' data. Table I shows the results of fitting this logistic regression model, while Table II shows measures of overall goodness-of-fit obtained from six packages.

The fitted model shown in Table I contains variables known to be important risk factors for low birth weight. Mother's age, although not significant, was retained in the model because of its known biologic significance. All six packages mentioned in Table II obtained the same estimated coefficients and estimated standard errors.

The *p*-values for the goodness-of-fit statistics presented in Table II highlight current problems in trying to interpret summary tests of goodness-of-fit from packaged programs. First, the *p*-value for the Pearson chi-square statistic is, in this case, meaningless, as it is based on a contingency table whose expected cell frequencies are too small (all are less than one) to justify use of a chi-square distribution with 183 degrees-of-freedom. The statistic itself is a good measure of model adequacy; the problem with its current application lies in the way packages compute its *p*-value. We consider an alternative and easily implemented method that gives a correct *p*-value. Second, we obtain six different values of the Hosmer–Lemeshow goodness-of-fit statistic based on grouping subjects into deciles of risk. Four packages produce a statistic with a *p*-value $> 0{\cdot}1$, one with $0{\cdot}05 < p\text{-value} < 0{\cdot}1$ and one with a *p*-value $< 0{\cdot}05$. The problem is that the packages use

Table II. Value of the Pearson chi-square statistic, X^2, and values of the Hosmer–Lemeshow decile of risk staistic, $\hat{C}$, computed by six different packages

Statistic	Value	D.F.	p-value
Pearson-chi-square, X^2	180·8	183	0·532
BMDPLR's $\hat{C}$	18·11	8	0·020
LOGXACT's $\hat{C}$	13·08	8	0·109
SAS's $\hat{C}$	11·83	8	0·159
STATA's $\hat{C}$	12·59	8	0·127
STATISTIX's $\hat{C}$	12·11	8	0·147
SYSTAT's $\hat{C}$	14·70	8	0·065

different algorithms to select cutpoints that define the deciles. It is disconcerting to note that the statistic seems sensitive to choice of groups. All packages produce the same fitted model. However, depending on our choice of level of significance and particular package used, we might reach different conclusions on overall model fit.

CURRENTLY USED SUMMARY MEASURES OF FIT

The addition of goodness-of-fit tests and logistic regression diagnostic statistics to statistical software packages has made the once difficult task of using these methods to assess the adequacy of a fitted logistic regression model a routine step in the model building process. Any analysis should incorporate a thorough examination of logistic regression diagnostics, see Hosmer and Lemeshow,[2] Chapter 5, before reaching a final decision on model adequacy. We do not wish to understate the importance of the use of these statistics, but the focus of this paper is on overall goodness-of-fit tests.

We begin by setting the notation used to describe the model. Assume we are in the strictly binary case and observe n independent pairs $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i' = (x_{0i}, x_{1i}, \ldots, x_{pi})$, $x_{0i} = 1$, denotes a vector of $p + 1$ assumed fixed covariates for the ith subject and $y_i = 0, 1$ denotes an observation of the outcome random variable Y_i. Under the logistic regression model we assume that $P(Y_i = 1 | \mathbf{x}_i) = \pi(\mathbf{x}_i)$, where $\pi(\mathbf{x}_i) = e^{g(\mathbf{x}_i)}/(1 + e^{g(\mathbf{x}_i)})$, and $g(\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$. Parameter estimates are usually obtained by maximum likelihood and are denoted by $\hat{\boldsymbol{\beta}}' = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$. We denote the fitted values as $\hat{\pi}_i = \pi(\mathbf{x}_i, \hat{\beta})$.

Examining a model's goodness-of-fit involves determining whether the fitted model's residual variation is small, displays no systematic tendency and follows the variability postulated by the model. Evidence of lack-of-fit may come from a violation of one or more of these three characteristics. Thus, in the context of a logistic regression model, the essential components of fit are specified by the following assumptions:

(A1) the logit transformation is the correct function linking the covariates with the conditional mean, $\text{logit}[\pi(\mathbf{x})] = \mathbf{x}'\boldsymbol{\beta}$;
(A2) the linear predictor, $\mathbf{x}'\boldsymbol{\beta}$, is correct (we do not need to include additional variables, transformations of variables, or interactions of variables);
(A3) the variance is Bernoulli, $\text{var}(Y_i|\mathbf{x}_i) = \pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)]$.

Assessment of model fit may occur at a number of stages in the modelling process. We may use it as an aid in model development where our goal is to find violations primarily in (A2) and/or to

verify that a 'final' model does fit where the emphasis is more towards examining (A1) and (A3). In the case of a logistic regression model we are faced with the practical problem that assumptions (A1)–(A3) are not mutually exclusive. Specifically, assumption (A3) may be confounded with (A1) and/or (A2). If we violate (A2) and misspecify the linear predictor then the model-based estimate of the variance is also incorrect. Similarly, if we have the incorrect link function, with or without linear predictor misspecification, then the model-based estimate of the variance is also incorrect.

A useful conceptual framework for thinking about assessment of model fit is to consider the data as described by a $2 \times n$ contingency table. The two rows are defined by the values of the dichotomous outcome variable y and the n columns by the assumed number of possible distinct values taken on by the p non-constant covariates in the model. The replicated design occurs when there are fewer than n distinct values (patterns) of the covariates. The likelihood ratio D (deviance) and Pearson chi-square, X^2, statistics that compare observed values to those predicted by the fitted logistic regression model in the $2 \times n$ table are

$$D = -2\left\{\sum_{i=1}^{n} y_i \ln(y_i/\hat{\pi}_i) + (1-y_i)\ln[(1-y_i)/(1-\hat{\pi}_i)]\right\} \quad \text{and} \quad X^2 = \sum_{i=1}^{n} (y_i - \hat{\pi}_i)^2/\hat{\pi}_i(1-\hat{\pi}_i).$$

Evidence for model lack-of-fit occurs when the values of these statistics are large. Towards this end, many packages provide a p-value computed using the $\chi^2(n-p-1)$ distribution. For the situation considered in this paper, the strictly binary case, this p-value has no value. For the p-value to be a valid measure of model fit the number of columns in the table must be fixed and the sample size large enough that the estimated expected values in the table all exceed some minimum number such as five. Hosmer and Lemeshow[2] (Chapter 5) discuss two methods of grouping based on the ranked estimated logistic probabilities that form groups of equal numbers of subjects or use fixed cutpoints on the $[0, 1]$ interval. The statistic, based on 10 equal sized groups (called 'deciles of risk'), is denoted $\hat{C}$, and is currently computed in most statistical packages. We denote in this paper the statistic based on fixed cutpoints as $\hat{H}$ which is computed in few packages. Hosmer and Lemeshow[1] showed, via simulations, that when the logistic regression model is correct, assumptions (A1)–(A3) hold, and the estimated expected values are 'large' in all cells, the distributions of both $\hat{C}$ and $\hat{H}$ for g groups are well approximated by the chi-square distribution with $g-2$ degrees-of-freedom, $\chi^2(g-2)$.

SMOOTHED RESIDUAL BASED TESTS

The advantage of the Hosmer–Lemeshow type tests is that they are based on groupings of the estimated probabilities that are intuitively appealing and easily understood by subject matter scientists. The disadvantage is that the value of the statistic depends on the choice of cutpoints that define the groups. In addition, they may have low power for detecting certain types of lack-of-fit. le Cessie and van Houwelingen[3,4] note that because the Hosmer–Lemeshow tests are based on a grouping strategy in the 'y' space they lack power to detect departures from the model in regions of the '$\mathbf{x}$' space that yield the same estimated probabilities. For example, a model with a quadratic term may have widely different '$\mathbf{x}$' values with the same estimated probability.

Tsiatis[5] proposed an approach based on fixed groups in the '$\mathbf{x}$' space that yields a score test for fit. One can use this approach with models that contain quadratic terms or periodic functions since one can form groups to circumvent the difficulties noted by le Cessie and van Houwelingen.[3,4] However, by choosing fixed groups, the decision about whether or not the model fits still may depend on the particular choice of groups. le Cessie and van Houwelingen address these problems by proposing a class of tests based on smoothed residuals.

The motivation for the use of smoothed residuals comes from work on non-parametric regression. Copas[6] and Azzalini *et al.*[7] have used the idea of smoothing the values of the outcome variable to obtain a non-parametric estimate of the regression function. Copas uses the estimate mainly for plotting the observed outcome and the smoothed outcome versus predictor variables while Azzalini *et al.* use it to develop a pseudo-likelihood ratio test. Their basic idea is to compare a 'smoothed' value of the outcome variable for each subject (which is a weighted average of the 'y' values for other subjects 'near' the subject) to a similarly smoothed estimate of the logistic probabilities. We can define the idea of 'nearness' in terms of a distance measure in the '$\mathbf{x}$' space as suggested by le Cessie and van Houwelingen, or it can be in the 'y' space.

In this paper we consider weight functions defined using the uniform kernel for the '$\mathbf{x}$' space, as used by le Cessie and van Houwelingen, and a cubic weight in the 'y' space. The advantage of the cubic weight is that it is available as a smoothing option in a number of packages, thus lending itself readily to plotting.

The '$\mathbf{x}$' space weight defining the distance between subject i and j is $w_{ij} = \prod_{k=1}^{p} u(x_{ik}, x_{jk})$ where $u(x_{ik}, x_{jk}) = 1$ if $|x_{ik} - x_{jk}|/s_k \leqslant c_u$ and is zero otherwise and s_k is the sample standard deviation of x_k. Choice of the weight constant, c_u, is important and was studied via simulation by le Cessie and van Houwelingen. They recommend setting c_u so that about $\sqrt{n}$ subjects have non-zero weights. The value of c_u used in the simulations reported in this paper is the value used in a SAS macro available from le Cessie and is $c_u = \frac{1}{2}(4/n^{1/(2p)})$. The cubic weights used define weights in the 'y' space and are given by the equation $w_{ij} = 1 - (|\hat{\pi}_i - \hat{\pi}_j|/c_{ci})^3$ if $|\hat{\pi}_i - \hat{\pi}_j| \leqslant c_{ci}$ and is zero if $|\hat{\pi}_i - \hat{\pi}_j| > c_{ci}$. Here the constant c_{ci} depends on i and is chosen such that $\sqrt{n}$ weights are non-zero for each subject. Fowlkes[8] has also used these weights in the context of a residual analysis in logistic regression.

The smoothed standardized residuals are $\hat{r}_{si} = \sum_{j=1}^{n} w_{ij}\hat{r}_j$ where $\hat{r}_j = (y_j - \hat{\pi}_j)/\sqrt{\{\pi_j(1-\hat{\pi}_j)\}}$ and the test statistic is $\hat{T}_r = \sum_{i=1}^{n} \hat{r}_{si}^2/\widehat{\mathrm{var}}(\hat{r}_{si}^2)$. Denote the statistics as $\hat{T}_{ru}$ when one uses the uniform kernel weights in the '$\mathbf{x}$-space' and $\hat{T}_{rc}$ when one uses the cubic weights in the 'y-space'. Computational details appear in the Appendix.

OTHER OVERALL TESTS OF GOODNESS-OF-FIT

Royston[9,10] proposed two procedures designed to detect departure from linearity in the logit that use partial sums of residuals. Royston did not specifically advocate the use of these tests for overall assessment of goodness-of-fit. However, since the tests are designed to be sensitive to departures in monotonicity in the logit or to detect a quadratic logit, then it seemed worthwhile to include them in the present study. The first test statistic is $P\hat{R}_1 = \max_{1 \leqslant l \leqslant n} |q_l|$ where $q_l = -\sum_{i=1}^{l} (y_{(i)} - \hat{\pi}_{(i)})$, $\hat{\pi}_{(i)}$ is the ith largest estimated logistic probability and $y_{(i)}$ is the associated value of the outcome variable. This test is called 'Royston monotone' in this paper as it was proposed to detect an overall departure from monotonicity in the logit. The second test is $P\hat{R}_2 = \max_{1 \leqslant l \leqslant n/2} |q_l - q_{n-l}|$ and is called 'Royston quadratic' as it was formulated to detect a quadratic departure in the model. Royston[2] provides easily computed transformations of the two test statistics that allow calculation of p-values from the standard normal distribution. Both tests are a special case of the test discussed in Beran and Miller.[11]

In the case of a single covariate the Royston monotone test is identical to a test proposed by Su and Wei.[12] Su and Wei proposed using a computationally intensive simulation to calculate the p-value. The computations for a model based on a sample of size n containing p main effect terms for continuous covariates are of order $n^p R$ where R is the number of simulations performed. The accuracy of the estimated p-value is a function of R, for example 500 simulations are needed to

estimate significance at the 5 per cent level to within 2 per cent with 95 per cent confidence. In preliminary simulations the performance of the Su and Wei method of obtaining the p-value was superior to Royston's analytic approach for models containing a single covariate. However, the overall performance, size and power, of the Su and Wei test was neither better nor worse than the other much less computationally intensive tests. Thus the Su and Wei test was not included in the detailed simulation study whose results are presented in this paper.

A generalized logistic model proposed by Stukel[13] provides a convenient model amenable to testing the adequacy of the fitted logistic model. The Stukel model uses a logit function with two additional parameters α_1 and α_2. These two parameters allow the generalized logistic model to be either symmetric or asymmetric with tails either lighter or heavier than is the case with the logistic model. The usual linear logistic model results when $\alpha_1 = 0$ and $\alpha_2 = 0$. A two degree-of-freedom test of the hypothesis that both parameters are equal to zero is obtained from the score test for the coefficients for the variables $z_1 = \frac{1}{2}\hat{g}^2 I(\hat{g} \geqslant 0)$ and $z_2 = -\frac{1}{2}\hat{g}^2 I(\hat{g} < 0)$, $\hat{g} = \mathbf{x}'\hat{\boldsymbol{\beta}}$, where $I(\cdot)$ is the indicator function returning the value one when the argument is true and zero otherwise. We denote this test statistic $S\hat{T}$.

Brown[14] developed a different two-parameter score test based on an extended logistic regression model proposed by Prentice.[15] A comparison of the Prentice model to the generalized logistic model by Stukel[13] showed that both offer the same level of flexibility in terms of generating alternative models but that Stukel's generalized logistic model is analytically easier to use since it does not require the integration needed with the Prentice model. Stukel[13] provides the expressions for the variables needed to carry out Brown's score test. We note that the Brown test is computed by program BMDPLR, but is not included in our simulations since the test is aimed at the same type of model departure as the Stukel test.

Copas[16] has suggested using the unweighted residual sum-of-squares, $\hat{S} = \sum(y_i - \hat{\pi}_i)^2$, to assess the model's adequacy. It is also a special case of the class of statistics considered by le Cessie and van Houwelingen[4] and we also consider this test statistic in the simulations.

SIMULATION RESULTS

We used simulations to study the properties of the overall goodness-of-fit tests. The goal was to assess the adequacy of the proposed null distribution of the statistics when the fitted logistic model was the correct model and to assess the power of the tests to detect a variety of departures from the logistic model.

Statistics used in the simulations are: the Pearson chi-square statistic, X^2; the unweighted sum-of-squares statistic, $\hat{S}$; the smoothed standardized residual-based test with weight functions based on a uniform kernel smooth, $\hat{T}_{ru}$; and the cubic distance smooth, $\hat{T}_{rc}$. In preliminary simulations the performance of the smoothed standardized residual based test and a similar statistic based on smoothed unstandardized residuals were comparable so we chose to use the standardized residual based test so that the results are more comparable to those in le Cessie and van Houwelingen.[3] Details on calculation of the estimated means, variances and the approach used to calculate the p-values for these four tests appear in the Appendix. Other statistics examined are the Hosmer–Lemeshow decile of risk statistic, $\hat{C}$, with 10 groups, and the Hosmer–Lemeshow fixed cutpoint statistic, $\hat{H}$, with up to 10 groups (Note: In some situations we used fewer than 10 groups when no estimated probabilities fell in certain intervals.) In all cases the degree-of-freedom was the number of groups minus two. We also included the two tests proposed by Royston, $P\hat{R}_1$ and $P\hat{R}_2$, and the score test, $S\hat{T}$, based on Stukel's extended logistic regression model. We performed all simulations on a DEC Alpha 2000 computer running the OSF operating system using a FORTRAN 5 program.

Table III. Situations used to examine the null distribution of the test statistics

Covariate distribution	Logistic coefficients	Distributional charactertistics of the logistic probabilities ($n = 100$)				
		$\pi_{(1)}$	Q_1	Q_2	Q_3	$\pi_{(n)}$
U(−6, 6)	$\beta_0 = 0$, $\beta_1 = 0{\cdot}8$	0·009	0·087	0·5	0·913	0·991
U(−4·5, 4·5)	$\beta_0 = 0$, $\beta_1 = 0{\cdot}8$	0·029	0·144	0·5	0·856	0·971
U(−3, 3)	$\beta_0 = 0$, $\beta_1 = 0{\cdot}8$	0·087	0·231	0·5	0·769	0·913
U(−1, 1)	$\beta_0 = 0$, $\beta_1 = 0{\cdot}8$	0·313	0·400	0·5	0·600	0·687
N(0, 1·5)	$\beta_0 = 0$, $\beta_1 = 0{\cdot}8$	0·057	0·304	0·5	0·696	0·943
$\chi^2(4)$	$\beta_0 = -4{\cdot}9$, $\beta_1 = 0{\cdot}65$	0·009	0·025	0·062	0·202	0·965
3 Independent U(−6, 6)	$\beta_0 = 0$, $\beta_1 = \beta_2 = \beta_3 = 0{\cdot}8/3$	0·028	0·234	0·5	0·767	0·972
3 Independent N(0, 1·5)	$\beta_0 = 0$, $\beta_1 = \beta_2 = \beta_3 = 0{\cdot}8/3$	0·134	0·369	0·5	0·628	0·861
Independent U(−6, 6), N(0, 1·5) and $\chi^2(4)$	$\beta_0 = -1{\cdot}3$, $\beta_1 = \beta_2 = 0{\cdot}8/3$, $\beta_3 = 0{\cdot}65/3$	0·052	0·204	0·386	0·608	0·928

Null Distribution

We considered a number of different situations to examine the performance of the tests when the logistic model fit was the correct model. We chose the various distributions of the covariate to produce distributions of probabilities in the (0, 1) interval that one might encounter in practice. Table III describes the situations where the distribution of the covariate(s) is given, along with the true coefficients for the logistic model and expected values for the smallest, largest, and three quartiles for the resulting distribution of logistic probabilities for a sample of size 100 (that is, considering the distribution of $\pi(x)$ as a transformation of the distribution of x). The Uniform distribution on the (−6, 6) interval U(−6, 6), produces a symmetric distribution with mostly small or large probabilities, while the U(−1, 1) produces probabilities mostly in centre of the (0, 1) interval. A highly skewed right distribution results (mostly small but a few large probabilities), when the covariate has the $\chi^2(4)$ distribution. Other choices for the distribution produce a more uniform distribution of probabilities.

In all simulations we first generated a sample of size $n = 100$ or 500 values of the covariate(s) and then we generated the outcome variable by comparing an independently generated U(0, 1) variate, u, to the true logistic probability using the rule $y = 1$ if $u \leqslant \pi(x)$ and $y = 0$ otherwise. In all situations we used 500 replications. Table IV shows the per cent of time each of the statistics rejected the hypothesis of fit at the $\alpha = 0{\cdot}05$ level.

The results in Table IV indicate that in all but a few situations eight of the nine statistics reject at, or nearly at, the five per cent level when we used $\alpha = 0{\cdot}05$.

The Royston monotone test never rejected the hypothesis. These results suggest the need for further work on the analytical method proposed by Royston for calculating p-values.

We also see in Table IV that the cubic smoothed standardized residual statistic with a three variable model rejects too often. The reason for this is that the approximate variance estimator in (A4) underestimates the true sampling variance. This estimator also slightly overestimated the variance of the uniform kernel smooth in two of the multivariable situations leading to fewer rejections than expected.

Power

We examined the power of the tests of fit to detect three particular types of departure from the logistic model. The situations studied were the omission of a quadratic term in a continuous

Table IV. Simulatd per cent rejection at the $\alpha = 0{\cdot}05$ level using sample sizes of 100 and 500 with 500 replications. Confidence intervals are obtained using ± 2 per cent

Distribution/ sample size	Pearson-chi-square X^2		Unweighted sum-of-squares $\hat{S}$		Hosmer–Lemeshow $\hat{C}$		Hosmer–Lemeshow $\hat{H}$		Uniform kernel smooth $\hat{T}_{ru}$		Cubic weight smooth $\hat{T}_{rc}$		Royston monotone $P\hat{R}_1$		Royston quadratic $P\hat{R}_2$		Stukel score test $S\hat{T}$	
	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500	100	500
U(−6, 6)	5·2	3·4	4·4	4·2	6·1	3·8	3·6	4·0	6·0	3·6	7·0	5·2	0·0	0·0	0·2	0·0	6·5	4·0
U(−4·5, 4·5)	6·0	4·8	5·6	3·6	4·4	4·6	3·4	4·2	3·8	4·2	5·6	5·4	0·0	0·0	0·2	0·0	3·6	3·8
U(−3, 3)	7·8	4·8	3·4	6·2	3·4	5·4	4·2	5·6	2·4	5·0	3·6	6·0	0·0	0·0	3·8	2·0	3·1	7·2
U(−1, 1)	4·8	5·6	5·8	4·4	4·8	3·2	8·3	1·6	3·6	4·4	3·4	5·6	0·0	0·0	1·4	3·6	4·4	4·2
N(0, 1·5)	4·2	6·2	3·2	5·4	4·4	5·2	6·0	6·2	3·2	4·4	3·8	7·4	0·0	0·0	2·6	2·6	5·0	7·0
$\chi^2(4)$	3·4	4·1	4·4	5·3	5·8	6·2	6·5	5·7	3·6	7·3	6·1	9·3	0·0	0·0	0·0	0·0	3·7	3·6
3 Independent U(−6, 6)	6·4	6·2	4·4	4·2	5·8	6·4	7·0	5·6	6·2	4·6	13·6	12·6	0·0	0·0	1·0	0·1	6·0	4·2
3 Independent N(0, 1·5)	7·0	4·8	2·6	4·4	5·4	4·2	6·4	9·4	1·2	3·4	13·6	12·8	0·0	0·0	3·8	4·6	4·6	5·4
Independent U(−6, 6), N(0, 1·5) and $\chi^2(4)$	5·8	7·0	5·0	5·2	4·0	3·6	6·4	6·6	1·2	3·0	13·2	12·6	0·0	0·0	2·4	1·8	4·6	5·6

Table V. Coefficients for the generalized logistic model

Model	α_1	α_2
Probit	0·165	0·165
Complimentary log-log	0·620	−0·037
Long tails	−1·0	−1·0
Short tails	1·0	1·0
Asymmetric long-short tails	−1·0	1·0

variable, and the omission of the main effect for a dichotomous variable and its interaction with a continuous variable and an incorrectly specified link function. In all situations studied the distribution of the continuous covariate, x, was U(− 3, 3). The distribution of the dichotomous covariate, d, was Bernoulli (1/2) and was independent of the continuous covariate.

We used five different models to evaluate power with omission of a quadratic term from the model. We generated the outcome variable using a logistic model with logit $g(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ where we chose the values of the three coefficients such that $\pi(-1{\cdot}5) = 0{\cdot}05$, $\pi(3) = 0{\cdot}95$ and $\pi(-3) = J$ and $J = 0{\cdot}01$, 0·05, 0·1, 0·2 and 0·4. This generated models where the lack of linearity in the logit function became progressively more pronounced.

We used four different interaction models to study the power with omission of a dichotomous variable and its interaction term from the model. We generated the outcome variable from a model with logit $g(x, d) = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 xd$. We chose the four parameters such that $\pi(-3, 0) = 0{\cdot}1$, $\pi(-3, 1) = 0{\cdot}1$, $\pi(3, 0) = 0{\cdot}2$ and $\pi(3, 1) = 0{\cdot}2 + I$ where $I = 0{\cdot}1$, 0·3, 0·5 and 0·7. Thus the four models display progressively more interaction.

We examined five different models to assess the power to detect an incorrectly specified link function. We generated the values of the outcome variable from Stukel's generalized logistic model using the function $\eta(x) = 0{\cdot}8x$ as the linear predictor and values of the parameters α_1 and α_2 as specified in Table V. Stukel[13] noted that if $\alpha_1 = \alpha_2 = 0{\cdot}165$, then the resulting generalized logistic model has nearly the same shape as the probit model and when $\alpha_1 = 0{\cdot}62$ and $\alpha_2 = -0{\cdot}037$ has the same shape as the complimentary log-log model. We chose the remaining three situations to yield one model with both tails longer, one model with both tails shorter tails and an asymmetric model with one tail longer and one tail shorter than the logistic model.

The situations we used to examine the power of the tests were chosen to represent typical logistic regression models encountered in practice. We felt this approach was preferable to selecting biologically implausible parameter configurations where the tests would likely have high power for any sample size. The combination of two sample sizes, 100 and 500, and the various models examined yields results which provide an adequate picture of what types of departures from a linear logistic model the various tests can detect with moderate to high power.

Table VI shows the per cent of time each of the tests rejected the hypothesis of fit at the $\alpha = 0{\cdot}05$ level.

In Table VI(a) we see that the power is, as expected, poor when trying to detect models that are quite close to the logistic. As the departure from linearity in the logit increases, the power increases rapidly for all tests except the Royston monotone test. High power is attained for samples of size 100 in those situations where there are substantial differences over the entire [0, 1] interval between the true quadratic model and the fitted linear model. The power is near 90 per cent for samples of size 500 for even slight, $J = 0{\cdot}05$, departures from the linear logistic model.

The results in Table VI(b) show that all tests have low power for samples of size 100 and 500 to detect even a fairly extreme interaction. We performed additional simulations, whose

Table VI. Simulated per cent rejection of fit at the $\alpha = 0{\cdot}05$ using sample sizes of 100 and 500 with 500 replications, confidence intervals are obtained as ± 2 per cent

(*a*) *Quadratic models*

Statistic/sample size	Correct model									
	$J = 0{\cdot}01$		$J = 0{\cdot}05$		$J = 0{\cdot}1$		$J = 0{\cdot}2$		$J = 0{\cdot}4$	
	100	500	100	500	100	500	100	500	100	500
Pearson chi-square, X^2	7·2	9·6	35·2	90·0	59·4	98·6	84·5	100	98·2	100
Unweighted sum-of-squares, $\hat{S}$	6·8	6·2	35·4	90·4	60·4	99·2	84·9	100	98·7	100
Hosmer–Lemeshow, $\hat{C}$	7·6	6·4	29·1	80·0	52·6	97·0	77·2	100	94·9	100
Hosmer–Lemeshow, $\hat{H}$	3·8	3·4	8·6	40·0	7·2	77·0	25·0	98	62·8	100
Uniform kernel smooth, $\hat{T}_{ru}$	7·4	7·6	29·6	82·8	53·0	96·8	79·3	100	95·7	100
Cubic weight smooth, $\hat{T}_{rc}$	8·4	7·2	35·5	86·0	59·0	98·6	82·0	100	96·7	100
Royston monotone, $P\hat{R}_1$	0·0	0·0	0·0	0·0	0·6	30·0	5·6	86	27·5	100
Royston quadratic, $P\hat{R}_2$	0·4	0·2	5·6	39·8	20·0	83·8	50·3	100	87·1	100
Stukel score test, $S\hat{T}$	6·0	6·0	29·2	68·4	53·8	98·6	79·5	100	77·9	100

(*b*) *Interaction models*

Statistic/sample size	Correct model							
	$I = 0{\cdot}1$		$I = 0{\cdot}3$		$I = 0{\cdot}5$		$I = 0{\cdot}7$	
	100	500	100	500	100	500	100	500
Pearson chi-square, X^2	5·8	4·8	4·9	5·4	5·0	4·0	3·2	1·8
Unweighted sum-of-squares, $\hat{S}$	4·5	6·0	3·9	3·8	3·8	7·0	7·8	11·2
Hosmer–Lemeshow, $\hat{C}$	4·3	5·8	3·9	4·2	3·0	6·0	5·2	6·8
Hosmer–Lemeshow, $\hat{H}$	6·4	4·0	2·5	2·8	2·2	3·0	3·4	5·8
Uniform kernel smooth, $\hat{T}_{ru}$	3·4	4·2	2·3	4·0	2·2	4·6	4·8	6·6
Cubic weight smooth, $\hat{T}_{rc}$	3·8	5·4	4·9	5·6	4·0	6·0	6·4	9·4
Royston monotone, $P\hat{R}_1$	0·0	0·0	0·0	0·0	0·0	0·0	0·0	0·0
Royston quadratic, $P\hat{R}_2$	3·6	5·6	3·9	4·2	2·0	6·0	6·0	8·4
Stukel score test, $S\hat{T}$	3·8	5·6	6·4	3·4	3·4	7·0	4·2	10·2

(*c*) *Alternative link functions*

Statistic/sample size	Correct model									
	Probit		Complimentary log-log		Long tails		short tails		Asymmetric long-short tail	
	100	500	100	500	100	500	100	500	100	500
Pearson chi-square, X^2	5·0	14·4	2·2	20·8	8·2	24·0	0·0	27·8	10·0	29·8
Unweighted sum-of-squares, $\hat{S}$	5·8	31·8	5·4	33·4	8·2	35·8	11·2	37·8	21·8	41·0
Hosmer–Lemeshow, $\hat{C}$	2·2	21·8	4·8	24·4	4·8	23·4	2·2	27·2	16·6	29·8
Hosmer–Lemeshow, $\hat{H}$	3·6	32·4	6·0	27·8	7·2	28·6	5·6	31·6	22·8	34·6
Uniform kernel smooth, $\hat{T}_{ru}$	4·0	15·8	4·2	17·0	3·8	17·8	1·0	20·6	16·4	22·6
Cubic weight smooth, $\hat{T}_{rc}$	3·2	26·6	3·8	26·0	4·4	27·2	1·0	28·8	18·6	33·4
Royston monotone, $P\hat{R}_1$	0·0	0·8	0·0	1·0	0·0	1·0	0·0	1·4	0·0	1·6
Royston quadratic, $P\hat{R}_2$	0·6	19·8	2·2	18·6	2·6	22·8	0·0	25·8	19·4	27·6
Stukel score test, $S\hat{T}$	4·0	38·6	8·6	43·4	7·8	42·8	22·0	43·8	40·2	51·0

results are not shown, using the U(−6, 6) distribution for the continuous covariate that yielded power in the 40 per cent range when $I = 0{\cdot}7$ for samples of size 100. The reason for the increase in power is the U(−6, 6) distribution places proportionally more of the estimated probabilities in the upper tail where there is the greatest difference between the model fit and the correct model.

The power to detect an alternative link function is poor for sample size 100 for all tests when the correct model is similar to the logistic model and it is less than about 40 per cent for sample size 500. This is not too surprising as the differences between the alternative links, probit and complementary log-log, and the logistic occur primarily in the tails which contain a small proportion of the estimated probabilities. The power of the Stukel score test is highest among all tests considered, but is, at best, only moderate even for sample size 500 and the asymmetric model that differs from the logistic over much of the [0, 1] interval.

The short tails models for sample size 100 presents an interesting situation. When both α_1 and α_2 are large and positive in the Stukel model, the probability function becomes quite steep and the fitted values tend to be either small or large. When this occurs, the test statistics, X^2 and $\hat{S}$-trace($\hat{\mathbf{V}}$), approach zero. However their estimated variances become quite large due to the range in the $\hat{\pi}$'s and thus in $\hat{c}$ and $\hat{d}$ (see Appendix). The normalized goodness-of-fit tests tend to be not significant since the numerator is small and the estimated variance is large. For example, the power of the Pearson chi-square statistic in Table VI(c) is less than the nominal alpha level when $\alpha_1 = \alpha_2 = 1$. Although not shown, there is a differential effect of the shape of the link on the Pearson chi-square and unweighted sum-of-squares statistic. However, as the two parameters, α_1 and α_2, become sufficiently large, both tests degenerate. With a sample of size 500 there are a sufficient number of estimated probabilities which are not near zero or one to allow the statistic to have a distribution which leads to power in the 25–30 per cent range.

The short tailed case affects the smoothed residual based tests with sample size 100. The actual sampling variance of the smoothed residual based test was much less than that estimated by (A4). The effect is that the test statistics tended to be systematically too small, and, as a result, the hypothesis of fit is not rejected sufficiently often. The reason for this is that with a steeply sloped probability function there is considerable agreement between the smoothed y and $\hat{\pi}$ when smoothing is in the y-space. Table VI(c) for the fitted model contains only one continuous covariate and the shape of the model affects both smoothed residual based tests. The tests perform much better with a sample size of 500 and have power in the 20–30 per cent range. In models with several covariates, smoothing in the '**x**-space' should provide greater power. These results combined with those in the multivariable null model simulation suggest the need for further research to improve the variance approximation given in (A4).

In summary, the results in Table VI show that overall the goodness-of-fit tests have, with the exception of the Royston monotone test, reasonable power for detecting a curvature type misspecification of the mean function and low power for interactions and an incorrect but still symmetric link function. The score test based on the Stukel model has moderate power to detect an asymmetric link function.

The overall performance of the Pearson chi-square statistic and unweighted sum-of-squares was superior to the other tests. The performance of the Hosmer–Lemeshow decile of risk statistic, the smoothed residual based tests and the Stukel score test were comparable. However, the work of le Cessie and van Houwelingen[3,4] suggest that with models more complicated than those used in our simulations the smoothed residual based tests will have greater power than the decile of risk test. It is not clear how the Stukel score test would perform with more complicated models and this merits study.

Table VII. Values of the goodness-of-fit statistics for the low birth weight model in Table I

Statistic	Value	Mean	Variance	*p*-value
Pearson chi-square, X^2	180·79	188·91	24·707	0·102
Unweighted sum-of-squares, $\hat{S}$	36·91	36·45	0·065	0·071
Hosmer–Lemeshow, $\hat{C}$	12·59	8	16	0·127
Hosmer–Lemeshow, $\hat{H}$	6·70	6	12	0·349
Uniform kernel smooth, $\hat{T}_{ru}$	0·514	0·579	0·028	0·641
Cubic weight smooth, $\hat{T}_{rc}$	1·469	0·801	0·018	$<0·001$
Royston monotone, $P\hat{R}_1$	5·88	*	*	0.300
Royston quadratic, $P\hat{R}_2$	8·05	*	*	0·049
Stukel score test, $S\hat{T}$	6·626	2	4	0·036

Consideration of the computational intensity, power and current availability in packages suggests that a practical strategy is to use the Pearson chi-square statistic and/or the unweighted sum-of-squares statistics in conjunction with the Hosmer–Lemeshow $\hat{C}$ statistic and the Stukel score test. The 2 by 10 table, of observed and estimated expected frequencies used to compute $\hat{C}$ provides a useful overall summary of the fit or lack-of-fit of the model and is easily understood by subject matter scientists. In addition it is good practice to examine plots of the residuals, perhaps smoothed, to support the visual inspection of the 2 by 10 table.

RETURN TO THE EXAMPLE

We return to an evaluation of the fit of the model for low birth weight shown in Table I. We used the scaled chi-square distribution to obtain *p*-values for the Pearson chi-square statistic and the smoothed residual based statistics and the standard normal distribution for the unweighted sum-of-squares statistic. Table VII gives results of the application of all nine tests.

The results in Table VII show that three of the nine statistics (the least powerful of the nine tests) have *p*-values $> 0·15$, three have *p*-values between 0·05 and 0·15, and three have a *p*-value $< 0·05$. The extremely small *p*-value for the smoothed residual based test using the cubic weight function may be due to the observed underestimation of the sampling variance of this test with multivariable models.

When we employ the recommended strategy of using the Pearson chi-square and/or unweighted sum-of-squares tests for power against overall non-linearity in the logit, the Hosmer–Lemeshow decile of risk statistic and 2 by 10 table for confirmatory evidence and the Stukel score test for power against a non-logit link we see that it suggests lack-of-fit of the model. Inspection of the 2 by 10 table showed a slight departure from model fit in the seventh decile of risk where there were more low birth weight babies observed than expected. Further modelling efforts suggested the inclusion of interactions between AGE and LWT and SMOKE and LWT. When we added these variables to the model, the fit improved to the point where the *p*-values were greater than 0·15. One requires a thorough examination of regression diagnostics as well as biologic plausibility, and in this case over fitting before one makes a decision on the final model. The recent paper by Harrell *et al.*[17] provides an excellent discussion with practical examples on issues surrounding the development of logistic regression models.

SUMMARY

The use of overall summary measures of goodness-of-fit of logistic regression models has become an important and easily performed step in model building. Decisions on model fit using tests based on cutpoints may depend more on choice of cutpoints than on fit or lack-of-fit. Tests not dependent on cutpoints proposed by le Cessie and van Houwelingen[3,4] and Royston,[9,10] a score test for an incorrectly specified link function proposed by Stukel[13] the cutpoint based tests of Hosmer and Lemeshow,[2] the Pearson chi-square statistic and the unweighted sum-of-squares test have been studied via simulation under both null and alternative scenarios. The simulation results showed that all but the Royston monotone maintained the correct size. The Pearson chi-square and unweighted sum-of-squares statistics had the highest power for omission of a quadratic term. All tests had low power to detect a continuous dichotomous variable interaction. All tests had more power to detect lack-of-fit due to model misspecification when the logit was non-monotone increasing (decreasing) under the alternative than when it was monotone under both null and alternative models. None of the tests studied had high power to detect an incorrectly specified link function with sample size 100. The Stukel score test had moderate for sample size 500.

Because of the superior power of the Pearson chi-square/unweighted sum-of-squares statistics and Stukel score test in the simulations, we recommend their use and the use of cutpoint and/or smoothed residual based tests for confirmation of model fit or lack-of-fit. Assessment of significance using the Pearson chi-square statistic should use the conditional mean and the variance estimate and one should compute p-values using the scaled chi-square distribution. When using the unweighted sum-of-squares statistic one should compute a standardized statistic with p-values obtained from the standard normal distribution. In all cases one must keep in mind the lack of power with small sample sizes to detect subtle deviations from the logistic model. Thus the choice of both the logistic regression model and its covariates should have a strong biological or clinical basis.

APPENDIX I: EXPRESSING THE SQUARED RESIDUALS TESTS AS QUADRATIC FORMS IN THE RESIDUALS

Application of the sum-of-squared residual-based test statistics requires computation of estimates of their respective means and variances under the assumption that the logistic model fit is correct. Derivation of means and variances is simplified if we express the test statistics as quadratic forms in the residuals.

Let $\mathbf{W}$ represent an n by n matrix of weights whose ith row, $\mathbf{w}_i$, contains the weights for the 'nearness' or distance of subject i to subjects 1 to n. In vector form the estimated standardized residuals are $\hat{\mathbf{r}} = \hat{\mathbf{V}}^{-1/2}\hat{\mathbf{e}}$, where $\hat{\mathbf{V}} = \text{diag}[\hat{v}_i = \hat{\pi}_i(1-\hat{\pi}_i)]$ is an n by n diagonal matrix. Thus the smoothed standardized residuals are $\hat{\mathbf{r}}_s = \mathbf{W}\hat{\mathbf{r}} = \mathbf{W}\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{e}}$. We can express the smoothed standardized residual based test statistics in matrix form as

$$\begin{aligned}\hat{T}_r &= \hat{\mathbf{r}}'\mathbf{W}'\mathbf{D}_r^{-1}\mathbf{W}\hat{\mathbf{r}} \\ &= \hat{\mathbf{e}}'\mathbf{V}^{-1/2}(\mathbf{W}'\mathbf{D}_r^{-1}\mathbf{W})\hat{\mathbf{V}}^{-1/2}\hat{\mathbf{e}} \qquad (1)\end{aligned}$$

where $\mathbf{D}_r$ is an n by n diagonal matrix that contains the diagonal elements of the matrix $\mathbf{WW}'$.

The expression in (1) shows that the test statistics only use the main diagonal elements of the respective full covariance matrices of the smoothed residuals. If we used the full covariance matrices to define the quadratic forms, then the statistics simplify to the Pearson chi-square statistic. Thus we can think of the Pearson chi-square statistic as a full covariance matrix version

of the smoothed residual-based tests. We also obtain the Pearson chi-square statistic if we do no smoothing (that is, use $\mathbf{W} = \mathbf{I}_{n\times n}$, the identity matrix).

APPENDIX II: MOMENTS AND ASYMPTOTIC DISTRIBUTION OF THE GOODNESS-OF-FIT STATISTICS

We begin by summarizing results that provide expressions for calculating large sample approximations for the mean and variance of the Pearson chi-square statistic. McCullagh[18–20] has derived moments conditional on $\hat{\boldsymbol{\beta}}$, work summarized in McCullagh and Nelder.[21] Recently, Osius and Rojek[22] considered this problem and showed that, in the strictly binary case, the conditional and unconditional moments are asymptotically equivalent. Here we summarize the results of Osius and Rojek. By substracting n the statistic simplifies to

$$X^2 - n = (\mathbf{1} - 2\hat{\boldsymbol{\pi}})'\hat{\mathbf{V}}^{-1}\hat{\mathbf{e}}.$$

First-order approximations presented in le Cessie and van Houwelingen[3] show $\hat{\boldsymbol{\pi}} \cong \boldsymbol{\pi} + \mathbf{Me}$, $\hat{\mathbf{e}} \cong (\mathbf{I} - \mathbf{M})\mathbf{e}$ and $\hat{\mathbf{V}} \cong \mathbf{V}$ where $\mathbf{M} = \mathbf{VX}(\mathbf{X}'\mathbf{VX})^{-1}\mathbf{X}'$ is the logistic regression version of the hat matrix. Using these approximations for purposes of calculating moments we obtain $X^2 - n \cong (\mathbf{1} - 2\boldsymbol{\pi})'\mathbf{V}^{-1}(\mathbf{I} - \mathbf{M})\mathbf{e}$ which yields that the asymptotic mean is zero and the variance is $\mathrm{var}(X^2 - n) \cong \mathbf{c}'(\mathbf{I} - \mathbf{M})\mathbf{Vc}$ where $\mathbf{c}' = (\mathbf{1} - 2\boldsymbol{\pi})'\mathbf{V}^{-1}$. Hence we may obtain an estimate of the variance as the residual sum-of-squares, $\hat{R}$, from the regression of $\hat{\mathbf{c}}$ on $\mathbf{X}$ with weights $\hat{\mathbf{V}}$. Osius and Rojek show that when one has fit the correct model the asymptotic distribution of $(X^2 - n)/\sqrt{R}$ is Normal(0, 1). Equivalent results appear in Windmeijer.[23] Our simulations suggest that, for small samples, we obtain better distributional results if we use an estimate of the conditional mean and variance obtained by McCullagh[16], namely

$$\hat{E}(X^2 \mid \boldsymbol{\beta}) = (n - p - 1) - \frac{1}{2}\sum_{i=1}^{n}(1 - 6\hat{v}_i)\hat{m}_{ii}/\hat{v}_i + \frac{1}{2}\sum_{i=1}^{n}\tilde{\hat{c}}_i\hat{c}_i\hat{m}_{ii}\hat{v}_i$$

where $\hat{m}_{ii}$ is the ith diagonal of the hat matrix $\hat{\mathbf{M}}$, $\tilde{\hat{c}}_i$ is the fitted value obtained from the weighted regression used to compute the unconditional variance. We obtain the conditional variance by multiplying $\hat{R}$ by $(n - p)/n$.

The asymptotic moments of $\hat{S}$ are $\mathrm{E}[\hat{S} - \mathrm{trace}(\mathbf{V})] \cong 0$ and $\mathrm{var}[\hat{S} - \mathrm{trace}(\mathbf{V})] \cong \mathbf{d}'(\mathbf{I} - \mathbf{M})\mathbf{Vd}$ where $\mathbf{d}$ is the vector with general element $d_i = (1 - 2\pi_i)$. Thus we can obtain an estimate of the variance as the residual sum-of-squares from the regression of $\hat{\mathbf{d}}$ on $\mathbf{X}$ with weights $\hat{\mathbf{V}}$. The expressions for the mean and variance given here are simpler than those obtained by Copas.[16] We obtained our results by simplifying $\hat{S}$ under the strictly binary assumption and then using first-order approximations of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\pi}}$. We used the standardized statistic $[\hat{S} - \mathrm{trace}(\hat{\mathbf{V}})]/\sqrt{\{\widehat{\mathrm{var}}[\hat{S} - \mathrm{trace}(\hat{\mathbf{V}})]\}}$ to assess significance using the standard normal distribution.

We obtain moments for the smoothed standardized residual based test statistics using first-order approximations $\hat{T}_r \cong \mathbf{e}'\mathbf{A}_r\mathbf{e}$ where $\mathbf{A}_r = (\mathbf{I} - \mathbf{M})'\mathbf{Q}_r(\mathbf{I} - \mathbf{M})$ and $\mathbf{Q}_r = \mathbf{V}^{-1/2}(\mathbf{W}'\mathbf{D}_r^{-1}\mathbf{W})\mathbf{V}^{-1/2}$. Well-known results for moments of quadratic forms, Seber,[24] yield that

$$\mathrm{E}(\hat{T}_r) = \mathrm{trace}(\mathbf{A}_r\mathbf{V}) \tag{2}$$

and

$$\mathrm{var}(\hat{T}_r) = \sum_{i=1}^{n} a_{rii}^2 v_i(1 - 6v_i) + 2\ \mathrm{trace}(\mathbf{A}_r\mathbf{VA}_r\mathbf{V}). \tag{3}$$

We obtain estimates of the moments by using $\hat{\mathbf{V}}$ in all expressions.

Early in the simulations we encountered time problems when implementing the variance estimator in (3). The difficulty is that the computations are of order n^4 to evaluate the matrix $\mathbf{A}_r$. As a result we used the approximation given in the Appendix of le Cessie and van Houwelingen[3]

$$\widehat{\text{var}}(\hat{T}_r) \cong 2\left(\frac{2}{3}\right)^p \frac{\text{trace}(\mathbf{W}\mathbf{W}')}{n^2} \tag{4}$$

for the $\widehat{\text{var}}(\hat{T}_r)$. This approximation reduced the order of computation to n^2.

Our simulations show that for small samples we may better approximate the distributions of the Pearson chi-square statistic, X^2, and le Cessie and van Houwelingen's smoothed residual test $\hat{T}_r$, by scaling the statistic and using a chi-square distribution whose degrees-of-freedom depend on the estimated mean and variance. The details of this method may be found in le Cessie and van Houwelingen[3] and are summarized here for convenience. For a particular goodness-of-fit statistic, denoted simply as *Stat*, we calculate the p-value as $\Pr[\chi^2(v) \geqslant bStat]$ where $b = 2\hat{E}/\widehat{\text{var}}$, $v = 2\hat{E}^2/\widehat{\text{var}}$ and $\hat{E}$ and $\widehat{\text{var}}$ are the estimated mean and variance of the statistic *Stat*.

ACKNOWLEDGEMENT

S. le Cessie wishes to acknowledge gratefully the support of a NATO science fellowship.

REFERENCES

1. Hosmer, D. W. and Lemeshow, S. 'A goodness-of-fit test for the multiple logistic regression model', *Communications in Statistics*, **A10**, 1043–1069 (1980).
2. Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression*, Wiley, New York, 1989.
3. le Cessie, S. and van Houwelingen, J. C. 'A goodness-of-fit test for binary data based on smoothing residuals', *Biometrics*, **47**, 1267–1282 (1991).
4. le Cessie, S. and van Houwelingen, J. C. 'Testing the fit of a regression model via score tests in random effects models', *Biometrics*, **51**, 600–614 (1995).
5. Tsiatis, A. A. 'A note on a goodness-of-fit test for the logistic regression model', *Biometrika*, **67**, 250–251 (1980).
6. Copas, J. B. 'Plotting p against x', *Applied Statistics*, **32**, 25–31 (1980).
7. Azzalini, A., Bowman, A. W. and Hardle, W. 'On the use of nonparametric regression for model checking', *Biometrika*, **76**, 1–11 (1989).
8. Fowlkes, E. B. 'Some diagnostics for binary regression via smoothing', *Biometrika*, **74**, 503–515 (1987).
9. Royston, P. 'The use of cusums and other techniques in modeling continuous covariates in logistic regression', *Statistics in Medicine*, **11**, 1115–1129 (1992).
10. Royston, P. 'Cusum plots and tests for binary variables', *STATA Technical Bulletin*, **STB-12** 16–17 (1993).
11. Beran, R. J. and Millar, P. W. 'Tests of fit for logistic models', *in* Mardia, K. V. (ed), *The Art of Statistical Science*: *A Tribute to G. S. Watson*, Wiley, New York, Chapter 12 (1992).
12. Su, J. Q. and Wei, L. J. 'A lack-of-fit test for the mean function in a generalized linear model', *Journal of the American Statistical Association*, **86**, 420–426 (1991).
13. Stukel, T. A. 'Generalized logistic models', *Journal of the American Statistical Association*, **83**, 426–431 (1988).
14. Brown, C. C. 'On a goodness-of-fit test for the logistic model based on score statistics', *Communications in Statistics*, **11,** 1087–1105 (1982).
15. Prentice, R. L. 'A generalization of the probit and logit methods for dose response curves', *Biometrics*, **32**, 761–768 (1976).
16. Copas, J. B. 'Unweighted sum of squares test for proportions', *Applied Statistics*, **38**, 71–80 (1989).
17. Harrell, F. E., Lee, K. L. and Mark, D. B. 'Tutorial in Biostatistics, Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicine*, **15**, 361–387 (1996).
18. McCullagh, P. 'On the asymptotic distribution of Pearson's statistics in linear exponential family models', *International Statistical Review*, **53**, 61–67 (1985).

19. McCullagh, P. 'Sparse data and conditional tests', *Bulletin of the International Statistical Institute, Proceedings of the 45th Session of the ISI* (Amsterdam), Invited Paper 28, **3**, 1–10 (1985).
20. McCullagh, P. 'The conditional distribution of goodness-of-fit statistics for discrete data', *Journal of the American Statistical Association*, **81**, 104–107 (1986).
21. McCullagh, P. and Nelder, J. *Generalized Linear Models*, 2nd edn, Chapman Hall, New York, 1989.
22. Osius, G. and Rojek, D. 'Normal goodness-of-fit tests for multinomial models with large degrees of freedom', *Journal of the American Statistical Association*, **87**, 1145–1152 (1992).
23. Windmeijer, F. A. G. 'Goodness-of-fit in linear and qualitative-choice models', Tinbergen Institute Research Series No. 29, Tinbergen Institute, Faculty of Economics and Econometrics, University of Amsterdam, Amsterdam, The Netherlands, 1992.
24. Seber, G. A. F. *Linear Regression Analysis*, Wiley, New York, (1977).

