Journal of Econometrics 25 (1984) 241–262. North-Holland

CONVENIENT SPECIFICATION TESTS FOR LOGIT AND PROBIT MODELS

Russell DAVIDSON and James G. MacKINNON*

Queen's University, Kingston, Ont., Canada K7L 3N6

Received March 1983, final version received September 1983

We propose several Lagrange multiplier tests of logit and probit models, which may be inexpensively computed by means of artificial linear regressions. These may be used to test for various forms of model inadequacy, including the omission of specified variables and heteroskedasticity of known form. We perform a number of sampling experiments, in which we compare the small-sample properties of these tests and of likelihood ratio tests. One of the LM tests turns out to have better small-sample properties than any of the others. We then investigate the power of the tests against local alternatives, and conduct a further series of sampling experiments to compare the power of various tests.

1. Introduction

The logit and probit models, together with their multi-response and multivariate generalizations, are now widely used in applied econometric work. Such models are typically estimated by maximum likelihood methods which require the numerical maximization of a loglikelihood function. Since this is usually much more expensive than, say, calculating ordinary least squares estimates for a linear regression model, investigators often display a natural reluctance to test the specification of the model as thoroughly as would normally be done in the regression case. There is thus a clear need for specification tests of logit and probit models which are easy to understand and inexpensive to compute.

In this context it seems natural to investigate the use of Lagrange Multiplier, or score, tests, because they require only estimates under the null hypothesis, and can often be computed by means of artificial linear regressions. The literature on LM tests for logit and probit models is, however, remarkably limited. The recent survey of qualitative response models by Amemiya (1981) does not mention LM tests at all, and the survey of LM tests by Engle (1982) describes only one such test for logit and probit models, which appears to be new.

*We are greatly indebted to Alison Morgan for research assistance beyond the call of duty. We would also like to thank two referees, an Editor and an Associate Editor of this journal for helpful comments on an earlier version. This research was supported, in part, by grants from the Social Sciences and Humanities Research Council of Canada and the School of Graduate Studies and Research of Queen's University.

In this paper we discuss several varieties of LM test for logit and probit models, each of which may be computed by means of an artificial linear regression. For a given alternative hypothesis, there turn out to be two different artificial linear regressions which generate five different, but asymptotically equivalent, test statistics. These procedures may be used to test both for omitted variables, which was the case examined by Engle (1982), and for heteroskedasticity of known form. The latter is a serious problem in the case of logit and probit models, since it renders parameter estimates inconsistent.

We perform two sets of sampling experiments. In the first set, we examine the performance of six tests under the null: the five LM or pseudo-LM tests referred to above, and the Likelihood Ratio test. We find that one of the LM tests outperforms the other tests, in the sense that the small-sample distribution of the test statistic under the null more closely approximates the asymptotic distribution. Thus use of this test rather than any of the others will generally result in Type I error being known more accurately. We also find that different, asymptotically equivalent tests based on the *same* artificial regression may behave very differently in small samples.

In the second set of sampling experiments, we investigate the power of two LM tests and the LR test. We compare the power of the tests both at fixed critical values and at estimated critical values based on the previous experiments, so as to control the level of Type I error. We also examine how closely the distributions of the tests approximate their asymptotic distributions under local alternatives, which is all that asymptotic theory has to tell us about the power of these tests. The approximation turns out to be quite poor in many cases.

2. LM tests for logit and probit models

The tests we shall develop are applicable to a fairly wide class of binary choice models, of which the logit and probit models are by far the most commonly encountered varieties. In models of this class the dependent variable can take on only two values, which it is convenient to denote by 0 and 1. The probability that y_t, the tth observation on the dependent variable, is equal to 1 is given by $F(x_t(\beta))$. F is an increasing function of x_t which has the properties that $F(-\infty)=0$ and $F(\infty)=1$. x_t is a possibly nonlinear function, which depends on X_t, a row vector of exogenous variables, and β, a column vector of parameters to be estimated. In the commonly encountered linear case, $x_t(\beta)=X_t\beta$.

The only difference between the logit and probit models is that they employ different functions for F. In the case of the probit model,

$$F(x_t(\beta))=\Phi(x_t(\beta)), \tag{1}$$

where Φ denotes the cumulative distribution function of the standard normal variate. In the case of the logit model,

$$\begin{aligned} F(x_t(\beta)) &= \exp(x_t(\beta))/[1+\exp(x_t(\beta))] \\ &= 1/[1+\exp(-x_t(\beta))]. \end{aligned} \tag{2}$$

Note that for both (1) and (2), $F(-z)=1-F(z)$, a convenient property of which we shall make use below. Other binary choice models use other functions in place of (1) and (2) [see Amemiya (1981)]; provided these functions also have this symmetry property, everything we say below will apply to them as well.

We shall denote by $L_t(\beta; y_t)$ the contribution to the loglikelihood function made by the tth observation. It is easy to see that

$$\begin{aligned} L_t(\beta;1) &= \log F(x_t(\beta)), \\ L_t(\beta;0) &= \log F(-x_t(\beta)). \end{aligned} \tag{3}$$

Thus the loglikelihood function is

$$L(\beta; y) = \sum_{t=1}^{n} L_t(\beta; y_t). \tag{4}$$

In the linear case, this function is globally concave for both the logit and probit models, except in pathological cases where it is unbounded [see Amemiya (1981)]. Thus ML estimates may be found in a straightforward fashion by maximizing it.

We shall denote the gradient of (4) with respect to β by the row vector $g(\beta; y)$. Its ith component is

$$g_i(\beta; y) = \sum_{t=1}^{n} G_{ti}(\beta; y_t),$$

where

$$\begin{aligned} G_{ti}(\beta; y_t) = &\left[y_t F(x_t(\beta))^{-1} + (y_t - 1) F(-x_t(\beta))^{-1} \right] \\ &\times f(x_t(\beta)) X_{ti}(\beta). \end{aligned} \tag{5}$$

Here $X_{ti}(\beta)$ denotes the derivative of $x_t(\beta)$ with respect to β_i; in the linear case, this will simply be equal to X_{ti}. $f(z)$ denotes the first derivative of $F(z)$, and we have made use of the fact that $f(z)=f(-z)$. For the probit model, $f(z)=\phi(z)$, the standard normal density. For the logit model,

$$f(z) = \exp(-z)[1+\exp(-z)]^{-2}. \tag{6}$$

The ML estimates $\hat{\beta}$ must of course satisfy the first-order conditions

$$g(\hat{\beta}; y) = 0. \tag{7}$$

The variance–covariance matrix of $\hat{\beta}$ may be consistently estimated in at least three different ways. We define the information matrix, $\mathcal{I}(\beta)$, as the matrix whose ijth element is

$$\mathrm{E}_y\big[g_i(\beta; y) g_j(\beta; y)\big]. \tag{8}$$

This may of course be consistently estimated by minus the Hessian, evaluated at $\hat{\beta}$, but this estimator turns out to be inconvenient in this context.[1] A more convenient estimator is

$$\hat{\mathcal{I}}^1 = G(\hat{\beta})^T G(\hat{\beta}), \tag{9}$$

where $G(\hat{\beta})$ is the matrix with typical element $G_{ti}(\hat{\beta}; y_t)$. The use of $\hat{\mathcal{I}}^1$ in estimation and inference has been advocated by Berndt, Hall, Hall and Hausman (1974). The third way to estimate $\mathcal{I}(\beta)$ is simply to use $\mathcal{I}(\hat{\beta})$. It is easily derived that a typical element of $\mathcal{I}(\hat{\beta})$ is

$$\hat{\mathcal{I}}_{ij} = \sum_{t=1}^{n} \big[F(x_t(\hat{\beta}))F(-x_t(\hat{\beta}))\big]^{-1}\big[f(x_t(\hat{\beta}))\big]^2 X_{ti}(\hat{\beta}) X_{tj}(\hat{\beta}). \tag{10}$$

Notice that, by the definition of $\mathcal{I}$, (10) depends on y only through $\hat{\beta}$.

We are now ready to discuss LM tests. Suppose that we estimate the model subject to k distinct restrictions, where k is equal to or less than m, the dimension of β. For concreteness, the reader may wish to bear in mind the linear omitted variable case, in which

$$x_t(\beta) = X_{1t}\beta_1 + X_{2t}\beta_2, \tag{11}$$

and the restrictions are that $\beta_2 = 0$. However, everything that we shall say is valid much more generally. We shall denote the restricted ML estimates by $\tilde{\beta}$; in the above case, $\tilde{\beta}^T = (\tilde{\beta}_1^T, 0)$. It follows from standard results [Engle (1982)] that the restrictions may be tested using the LM statistic

$$g(\tilde{\beta})\tilde{\mathcal{I}}^{-1} g(\tilde{\beta})^T, \tag{12}$$

[1]Strictly speaking, it is of course $\mathcal{I} \equiv (1/n)\mathcal{I}$ which is consistently estimated by minus $1/n$ times the Hessian. Here and elsewhere we ignore this distinction when it is not important to the argument.

where $\tilde{\mathcal{I}}$ is any consistent estimate of $\mathcal{I}(\beta)$, which depends on $\tilde{\beta}$, and the dependence of g on y has been suppressed for notational convenience. Under the null hypothesis, this statistic will be asymptotically distributed as chi-squared with k degrees of freedom.

LM statistics based on $\tilde{\mathcal{I}}^1$ and $\mathcal{I}(\tilde{\beta})$ are particularly attractive, because they can easily be computed using an OLS regression package. We consider the former case first. Using (9) and the definition of $g(\beta)$, the LM statistic (12) becomes

$$LM_1 = \iota^T G(\tilde{\beta})\big[G(\tilde{\beta})^T G(\tilde{\beta})\big]^{-1} G(\tilde{\beta})^T \iota, \tag{13}$$

where ι denotes an n-vector of ones. Expression (13) is clearly just the explained sum of squares from the artificial linear regression

$$\iota = G(\tilde{\beta})b + \text{errors}, \tag{14}$$

in which a vector of ones is regressed on the matrix $G(\tilde{\beta})$. This form of the LM statistic has been used by Godfrey and Wickens (1981) in a different context; it does not seem to have been suggested previously in the context of logit and probit models.

Regression (14) actually generates two valid test statistics. First there is the explained sum of squares, LM_1, which in this case is also equal to n times the *uncentred* R^2. Second, there is the statistic

$$F_1 = ((n - SSR)/k)/(SSR/(n - m)), \tag{15}$$

the distribution of which approaches $F(k, n - m)$ as n tends to infinity, and is calculated just like the ordinary F-statistic for linear regression models: the restricted sum of squared residuals is n, and the unrestricted sum of squared residuals from (14) is SSR. It is easily seen that F_1 is asymptotically equivalent to LM_1, since k times the numerator of F_1 is equal to LM_1, and the denominator tends to unity asymptotically under the null. This of course implies that when there is only one restriction, the t-statistic on the column of $G(\tilde{\beta})$ corresponding to the restriction will provide an asymptotically valid test statistic.

We now turn our attention to LM statistics based on $\mathcal{I}(\tilde{\beta})$. Let the matrix $R(\tilde{\beta})$ be constructed with typical element

$$R_{ti}(\tilde{\beta}) = \big[F(x_t(\tilde{\beta}))F(-x_t(\tilde{\beta}))\big]^{-\frac{1}{2}} f(x_t(\tilde{\beta}))X_{ti}(\tilde{\beta}), \tag{16}$$

and the vector $r(\tilde{\beta})$ be constructed with typical element

$$r_t(\tilde{\beta}; y_t) = y_t\big[F(-x_t(\tilde{\beta}))/F(x_t(\tilde{\beta}))\big]^{\frac{1}{2}} + (y_t - 1)\big[F(x_t(\tilde{\beta}))/F(-x_t(\tilde{\beta}))\big]^{\frac{1}{2}}. \tag{17}$$

Now consider the artificial linear regression

$$r(\tilde{\beta}) = R(\tilde{\beta})c + \text{errors}, \tag{18}$$

the explained sum of squares from which is

$$LM_2 = r(\tilde{\beta})^T R(\tilde{\beta})\left[R(\tilde{\beta})^T R(\tilde{\beta})\right]^{-1} R(\tilde{\beta})^T r(\tilde{\beta}). \tag{19}$$

It is easily verified that $r(\tilde{\beta})^T R(\tilde{\beta}) = g(\tilde{\beta})$, and that $R(\tilde{\beta})^T R(\tilde{\beta}) = \mathcal{I}(\tilde{\beta})$. Hence expression (19) is an LM statistic. The artificial regression (18) was suggested by Engle (1982) for the linear case.

Regression (18) actually generates three different test statistics. First, there is the explained sum of squares, LM_2. Second, there is n times the *uncentred* R^2 from the regression. Using the notation $\tilde{r} = r(\tilde{\beta})$ and $\tilde{R} = R(\tilde{\beta})$, this is

$$nR^2 = n\left[\left(\tilde{r}^T\tilde{R}(\tilde{R}^T\tilde{R})^{-1}\tilde{R}^T\tilde{r}\right)/(\tilde{r}^T\tilde{r})\right]. \tag{20}$$

As Engle (1982) points out, $\operatorname{plim}[(1/n)\tilde{r}^T\tilde{r}] = 1$, so that (20) is asymptotically equivalent to LM_2. Finally, there is the pseudo F-statistic

$$F_2 = \left((\tilde{r}^T\tilde{r} - SSR)/k\right)/(SSR/(n-m)), \tag{21}$$

which is analogous to F_1 in every respect: k times the numerator is equal to LM_2, and the denominator tends to unity asymptotically, under the null. Once again, the fact that F_2 is valid implies that when there is only one restriction, the t-statistic on the column of $\tilde{R}$ corresponding to the restriction will also be an asymptotically valid test statistic.

The specification of $x_t(\beta)$ as a non-linear function allows the tests just described to be used for a variety of purposes. The most obvious one is testing whether one or more additional variables should be added to a linear logit or probit model. A somewhat less obvious application is testing for heteroskedasticity. Since heteroskedasticity causes parameter estimates from logit and probit models to be inconsistent, it is a serious problem. Moreover, because these models are usually estimated using cross-section data, it is a problem which is likely to be encountered quite often.

Consider first the following latent variable model:

$$\begin{aligned} Y_t &= X_t\beta_1 + u_t, \qquad u_t \sim \mathrm{N}(0, \exp(2Z_t\beta_2)) \\ y_t &= 1 \quad \text{if} \quad Y_t > 0, \\ &= 0 \quad \text{if} \quad Y_t \le 0. \end{aligned} \tag{22}$$

Here Y_t is an unobserved (latent) variable, X_t and Z_t are row vectors of observations on exogenous variables, and β_1 and β_2 are vectors of unknown parameters. To ensure that both β_1 and β_2 are identifiable, we must specify that Z_t does not include a constant term. Clearly, when $\beta_2 = 0$, u_t will be N(0,1), and (22) will then yield the ordinary linear probit model. An LM test of the hypothesis that $\beta_2 = 0$ will test the ordinary probit model against the heteroskedastic alternative given by (22).

The above model implies that the probability that $y_t = 1$ is

$$\Phi[(X_t\beta_1)/\exp(Z_t\beta_2)] = F(x_t(\beta_1, \beta_2)), \tag{23}$$

where

$$x_t(\beta_1, \beta_2) = X_t\beta_1/\exp(Z_t\beta_2), \tag{24}$$

since for the probit model $F(z)$ is defined to be $\Phi(z)$. It is clear that we can specify x_t as in (24) for the logit model as well, which implies that

$$\log[P_t/(1 - P_t)] = X_t\beta_1/\exp(Z_t\beta_2) = x_t(\beta_1, \beta_2). \tag{25}$$

where P_t is the probability that $y_t = 1$. Since this logit specification involves no latent variable, expression (25) cannot properly be called a specification of heteroskedasticity. But it seems to us a reasonable specification nonetheless, and we shall for brevity use the term 'heteroskedasticity' to refer to it.

Given this specification of $x_t(\beta)$, it is now straightforward to use any of the LM or approximate LM tests discussed above to test for heteroskedasticity. For the benefit of practitioners, we note that

$$\begin{aligned} \partial x_t(\beta_1, \beta_2)/\partial\beta_1|_{\beta_1=\tilde{\beta}_1, \beta_2=0} &= X_t, \\ \partial x_t(\beta_1, \beta_2)/\partial\beta_2|_{\beta_1=\tilde{\beta}_1, \beta_2=0} &= -X_t\tilde{\beta}_1 Z_t, \end{aligned} \tag{26}$$

where these expressions are to be interpreted as row vectors. Using these in (5) or (16) and (17) allows one to generate the artificial regressions (14) or (18) very easily.

3. Small-sample properties of alternative tests

In this section we present the results of a number of sampling experiments designed to study the small-sample properties of the tests described above. The data generating process was either a logit or a probit model, with

$$x_t(\beta) = (\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \beta_3 X_{3t})/\exp(\beta_4 X_{3t}), \tag{27}$$

Table 1

Standard errors of alternative test statistics.[a]

Hyp.	Test	Exp.	Std. error	Exp.	Std. error	Exp.	Std. error	Exp.	Std. error
β_2	LR	50/L/3	1.092 (4.11)	100/L/3	1.051 (2.30)	200/L/3	1.064 (2.88)	100/L/6	1.080 (3.60)
	LM_1		1.124 (5.56)		1.068 (3.03)		1.068 (3.02)		1.116 (5.20)
	F_1		1.135 (6.05)		1.068 (3.06)		1.069 (3.08)		1.120 (5.37)
	LM_2		1.045 (2.02)		1.034 (1.50)		1.054 (2.43)		1.042 (1.88)
	F_2		1.239 (10.7)		1.106 (4.73)		1.101 (4.53)		1.473 (21.1)
	$2-nR^2$		1.213 (9.51)		1.104 (4.64)		1.099 (4.44)		1.441 (19.7)
β_3	LR	50/L/3	1.039 (1.75)	100/L/3	1.022 (1.00)	200/L/3	1.033 (1.49)	100/L/6	1.062 (2.76)
	LM_1		1.115 (5.14)		1.063 (2.81)		1.052 (2.35)		1.132 (5.89)
	F_1		1.123 (5.49)		1.065 (2.92)		1.053 (2.38)		1.138 (6.15)
	LM_2		1.015 (0.69)		1.012 (0.55)		1.029 (1.28)		1.043 (1.90)
	F_2		1.270 (12.1)		1.097 (4.36)		1.078 (3.48)		1.490 (21.9)
	$2-nR^2$		1.195 (8.73)		1.092 (4.13)		1.076 (3.42)		1.451 (20.2)
β_4	LR	50/L/3	1.164 (7.35)	100/L/3	1.102 (4.55)	200/L/3	1.031 (1.38)	100/L/6	1.115 (5.15)
	LM_1		1.484 (21.7)		1.336 (15.0)		1.163 (7.27)		1.546 (24.4)
	F_1		1.567 (25.3)		1.356 (15.9)		1.168 (7.53)		1.587 (26.2)
	LM_2		0.990 (0.46)		1.035 (1.55)		1.001 (0.05)		0.990 (0.46)
	F_2		1.083 (3.70)		1.087 (3.88)		1.038 (1.70)		1.280 (12.5)
	$2-nR^2$		1.080 (3.56)		1.086 (3.83)		1.038 (1.68)		1.270 (12.1)
β_2	LR	50/P/2	1.084 (3.74)	100/P/2	1.014 (0.61)	200/P/2	1.017 (0.77)	100/P/4	1.049 (2.20)
	LM_1		1.149 (6.65)		1.055 (2.48)		1.039 (1.74)		1.108 (4.82)
	F_1		1.157 (7.02)		1.057 (2.57)		1.040 (1.79)		1.112 (4.99)
	LM_2		1.031 (1.37)		0.992 (0.34)		1.006 (0.26)		1.002 (0.08)
	F_2		1.318 (14.2)		1.150 (6.72)		1.100 (4.49)		1.676 (30.2)
	$2-nR^2$		1.289 (12.9)		1.140 (6.27)		1.098 (4.38)		1.628 (28.1)
β_3	LR	50/P/2	1.064 (2.85)	100/P/2	1.049 (2.17)	200/P/2	1.030 (1.34)	100/P/4	1.029 (1.30)
	LM_1		1.178 (7.94)		1.127 (5.67)		1.070 (3.14)		1.138 (6.19)
	F_1		1.194 (8.69)		1.131 (5.86)		1.072 (3.20)		1.145 (6.47)
	LM_2		1.025 (1.14)		1.032 (1.44)		1.023 (1.01)		1.002 (0.11)
	F_2		1.326 (14.6)		1.179 (8.00)		1.120 (5.38)		1.696 (31.1)
	$2-nR^2$		1.287 (12.8)		1.172 (7.68)		1.118 (5.29)		1.647 (28.9)
β_4	LR	50/P/2	1.133 (5.95)	100/P/2	1.061 (2.73)	200/P/2	1.019 (0.84)	100/P/4	1.105 (4.70)
	LM_1		1.550 (24.6)		1.393 (17.6)		1.213 (9.52)		1.601 (26.9)
	F_1		1.632 (28.3)		1.415 (18.5)		1.220 (9.82)		1.640 (28.6)
	LM_2		0.915 (3.80)		0.957 (1.92)		0.971 (1.31)		0.925 (3.37)
	F_2		1.105 (4.68)		1.071 (3.18)		1.038 (1.72)		1.484 (21.7)
	$2-nR^2$		1.101 (4.54)		1.072 (3.20)		1.039 (1.72)		1.464 (20.8)

[a] Figures in parentheses after standard errors are asymptotic t-statistics for the hypothesis that the true standard deviation is one.

In $x/y/z$, x = number of observations, y = L if logit and P if probit, z = value of β_1.

where X_{1t} and X_{2t} were independent N(0, 1) variates (the same realizations being used in all experiments), and X_{3t} was a linear time trend. The sample size was always chosen as an integer multiple of 50, and the same sets of 50 observations on X_1, X_2 and X_3 were repeated the required number of times, to ensure that the matrix $(1/n)X^TX$ did not change as the sample size increased. The trend term X_3 was set equal to $0.10 + 0.01t$ for $t = 1$ to 50. We chose 50 as the smallest sample size to investigate because, in our experience, probit and logit models are rarely estimated using samples smaller than that (since smaller samples would rarely contain enough information to make estimation worthwhile). The largest sample size used in our experiments was 200. Because

computational cost is roughly proportional to sample size, to have investigated substantially larger sample sizes would have been prohibitively expensive.[2]

The null hypothesis was that $\beta_2 = \beta_3 = \beta_4 = 0$, so that only β_0 and β_1 were estimated under the null. For the experiments reported in this section, the null was always true, with $\beta_0 = 0$ and β_1 taking on various values. Basically, β_1 was chosen so that the model predicted y_t reasonably well, but not so well that X_1 was ever in danger of being a perfect classifier.

The test statistics examined in our experiments are LM_1, F_1, LM_2, F_2 and n times the R^2 from regression (18), as well as the likelihood ratio test statistic. We chose not to examine Wald-type test statistics because they are clearly unattractive in this context. Estimation of the null will here be easier than estimation of the alternative, and many investigators will wish to avoid the latter entirely by using one of the LM-type tests. If estimation of the alternative is undertaken, estimates under the null will normally already be available, so that calculation of the LR test will then be trivial, much easier than calculation of a Wald test. Moreover, just as there are several LM-type tests, so too are there several Wald-type tests, and attempting to deal with them all would have made reporting the experimental results quite difficult.

The results of eight experiments are presented in tables 1 through 3. The hypotheses that β_2, β_3 and β_4 were zero were each tested separately. Since the resulting test statistics have only one degree of freedom, we transformed them into test statistics which would be asymptotically N(0, 1) under the null. This was done by taking their square roots and multiplying by minus one if the coefficient of the test regressor in the artificial regression (14) or (18) is negative, or, in the case of the LR test, if the unrestricted coefficient estimate is negative.

All of the test statistics turned out to have means acceptably close to zero; thus we do not report this information. On the other hand, the standard errors of the various test statistics differed dramatically, and these are therefore presented in table 1. In this table, 'Hyp.' indicates which coefficient is being tested, and 'Exp.' indicates the sample size (50, 100 or 200), whether the logit or probit model was used, and the value of β_1. The numbers under 'Std. error' are the observed standard errors of the various test statistics in 1000 replications, and the numbers in brackets following these are asymptotic t-statistics for the hypothesis that the true standard deviation is unity.[3]

Several features of table 1 are striking. The only test statistic which ever has an estimated standard error of less than one, and the only test statistic for

[2] To perform the numerical maximizations, we used NAG subroutine E04EBF, which employs analytical second derivatives. In two of the 18,000 replications performed, this routine failed to find a maximum for the null hypothesis. These two replications were replaced.

[3] This t-statistic is $(s-1)(2N)^{\frac{1}{2}}$, where s is the standard error and N is the number of replications. Since N is always 1000, the normal approximation on which this statistic is based should be quite accurate.

Table 2

Performance of alternative test statistics – Logit.

Rep.	Nobs.	Slope	Hyp.	Test	Std. error	0.01 tail	0.05 tail	0.10 tail	0.05 crit.
1000	50	3	β_2	LR	1.092 (4.11)	0.020 (3.18)	0.076 (3.77)	0.138 (4.01)	2.16 (2.05–2.30)
				LM_1	1.124 (5.56)	0.022 (3.81)	0.085 (5.08)	0.148 (5.06)	2.23 (2.06–2.39)
				LM_2	1.045 (2.02)	0.012 (0.64)	0.059 (1.31)	0.124 (2.53)	2.04 (1.94–2.14)
			β_3	LR	1.039 (1.75)	0.012 (0.64)	0.052 (0.29)	0.100 (0.00)	1.97 (1.86–2.13)
				LM_1	1.115 (5.14)	0.016 (1.91)	0.077 (3.92)	0.137 (3.90)	2.19 (2.07–2.32)
				LM_2	1.015 (0.69)	0.010 (0.00)	0.044 (0.87)	0.095 (0.53)	1.92 (1.81–2.04)
			β_4	LR	1.164 (7.35)	0.032 (6.99)	0.086 (5.22)	0.153 (5.59)	2.34 (2.18–2.45)
				LM_1	1.485 (21.7)	0.092 (26.1)	0.182 (19.2)	0.249 (15.7)	3.03 (2.88–3.29)
				LM_2	0.990 (0.46)	0.006 (1.27)	0.050 (0.00)	0.101 (0.11)	1.97 (1.82–2.08)
1000	100	3	β_2	LR	1.051 (2.30)	0.010 (0.00)	0.062 (1.74)	0.117 (1.79)	2.04 (1.95–2.14)
				LM_1	1.068 (3.03)	0.013 (0.95)	0.067 (2.47)	0.122 (2.32)	2.04 (1.97–2.18)
				LM_2	1.034 (1.50)	0.006 (1.27)	0.057 (1.02)	0.109 (0.95)	2.00 (1.91–2.09)
			β_3	LR	1.022 (1.00)	0.009 (0.32)	0.053 (0.44)	0.102 (0.21)	2.01 (1.85–2.12)
				LM_1	1.063 (2.81)	0.015 (1.59)	0.057 (1.02)	0.128 (2.95)	2.05 (1.72–2.18)
				LM_2	1.012 (0.55)	0.007 (0.95)	0.052 (0.29)	0.096 (0.42)	1.98 (1.84–2.09)
			β_4	LR	1.102 (4.55)	0.011 (0.32)	0.082 (4.64)	0.141 (4.32)	2.19 (2.05–2.33)
				LM_1	1.336 (15.0)	0.060 (15.9)	0.141 (13.2)	0.211 (11.7)	2.76 (2.53–2.94)
				LM_2	1.035 (1.55)	0.006 (1.27)	0.061 (1.60)	0.115 (1.58)	2.00 (1.87–2.29)
1000	200	3	β_2	LR	1.064 (2.88)	0.016 (1.91)	0.067 (2.47)	0.125 (2.64)	2.11 (1.98–2.27)
				LM_1	1.068 (3.02)	0.017 (2.22)	0.062 (1.74)	0.133 (3.48)	2.09 (1.94–2.24)
				LM_2	1.054 (2.43)	0.012 (0.64)	0.065 (2.18)	0.120 (2.11)	2.08 (1.97–2.22)
			β_3	LR	1.033 (1.49)	0.009 (0.32)	0.057 (1.02)	0.108 (0.84)	1.99 (1.89–2.18)
				LM_1	1.052 (2.35)	0.014 (1.27)	0.062 (1.74)	0.115 (1.58)	2.04 (1.91–2.21)
				LM_2	1.029 (1.28)	0.009 (0.32)	0.056 (0.87)	0.107 (0.74)	1.98 (1.89–2.16)
			β_4	LR	1.031 (1.38)	0.012 (0.64)	0.058 (1.16)	0.112 (1.26)	2.02 (1.89–2.15)
				LM_1	1.163 (7.27)	0.031 (6.67)	0.101 (7.40)	0.145 (4.74)	2.31 (2.20–2.50)
				LM_2	1.001 (0.05)	0.010 (0.00)	0.045 (0.73)	0.103 (0.32)	1.91 (1.82–2.02)
1000	100	6	β_2	LR	1.080 (3.60)	0.021 (3.50)	0.075 (3.63)	0.127 (2.85)	2.12 (2.01–2.33)
				LM_1	1.116 (5.20)	0.022 (3.81)	0.077 (3.92)	0.153 (5.59)	2.21 (2.07–2.35)
				LM_2	1.042 (1.88)	0.009 (0.32)	0.063 (1.89)	0.117 (1.79)	2.06 (1.95–2.19)
			β_3	LR	1.062 (2.76)	0.018 (2.54)	0.069 (2.76)	0.118 (1.90)	2.13 (1.98–2.32)
				LM_1	1.132 (5.89)	0.029 (6.04)	0.093 (6.24)	0.146 (4.85)	2.26 (2.11–2.49)
				LM_2	1.043 (1.90)	0.014 (1.27)	0.063 (1.89)	0.114 (1.48)	2.09 (1.94–2.26)
			β_4	LR	1.115 (5.15)	0.019 (2.86)	0.080 (4.35)	0.134 (3.58)	2.19 (2.05–2.33)
				LM_1	1.546 (24.4)	0.116 (33.7)	0.204 (22.3)	0.284 (19.4)	3.06 (2.89–3.30)
				LM_2	0.990 (0.46)	0.005 (1.59)	0.044 (0.87)	0.095 (0.53)	1.91 (1.83–2.08)

which the hypothesis that the standard deviation is unity cannot be rejected most of the time, is LM_2. Of particular interest is the fact that F_2 and $2 - nR^2$ [n times the R^2 from regression (18)], although based on exactly the same artificial regression as LM_2, always have larger standard deviations than LM_2, often so large that they will clearly yield highly unreliable inferences. There seems to be an explanation for this. Note that

$$nR^2 = \left(n/\tilde{r}^T\tilde{r}\right) LM_2, \tag{28}$$

the first factor being a random variable with a plim of unity. Unless there is substantial negative covariance between this factor and LM_2, nR^2 will have greater variance than LM_2, which is exactly what we find in table 1. Similarly, F_2 is related to LM_2 by

$$kF_2 = \left[(n-m)/\left(\tilde{r}^T\tilde{r} - LM_2\right)\right] LM_2. \tag{29}$$

Since $\tilde{r}^T\tilde{r}$ is $O(n)$ while LM_2 is O(1), under the null, the first factor in (29) will tend to be very similar to the first factor in (28), so that kF_2 and nR^2 can be expected to be very close. Indeed, table 1 shows that the standard errors of F_2

Table 3

Performance of alternative test statistics – Probit.

Rep.	Nobs.	Slope	Hyp.	Test	Std. error	0.01 tail	0.05 tail	0.10 tail	0.05 crit.
1000	50	2	β_2	LR	1.084 (3.74)	0.014 (1.27)	0.069 (2.76)	0.134 (3.58)	2.07 (1.98–2.22)
				LM_1	1.149 (6.65)	0.016 (1.91)	0.082 (4.64)	0.155 (5.80)	2.18 (2.08–2.38)
				LM_2	1.031 (1.37)	0.006 (1.27)	0.054 (0.58)	0.117 (1.79)	1.98 (1.88–2.08)
			β_3	LR	1.064 (2.85)	0.017 (2.22)	0.061 (1.60)	0.120 (2.11)	2.08 (1.94–2.20)
				LM_1	1.178 (7.94)	0.029 (6.04)	0.094 (6.38)	0.162 (6.54)	2.30 (2.19–2.47)
				LM_2	1.025 (1.14)	0.006 (1.27)	0.052 (0.29)	0.111 (1.16)	2.00 (1.88–2.10)
			β_4	LR	1.113 (5.95)	0.022 (3.81)	0.089 (5.66)	0.148 (5.06)	2.26 (2.16–2.36)
				LM_1	1.550 (24.6)	0.113 (32.7)	0.211 (23.4)	0.287 (19.7)	3.10 (2.96–3.30)
				LM_2	0.915 (3.80)	0.003 (2.22)	0.025 (3.63)	0.071 (3.06)	1.73 (1.67–1.81)
1000	100	2	β_2	LR	1.014 (0.61)	0.010 (0.00)	0.051 (0.15)	0.103 (0.32)	1.98 (1.84–2.12)
				LM_1	1.055 (2.48)	0.012 (0.64)	0.059 (1.31)	0.131 (3.27)	2.06 (1.91–2.21)
				LM_2	0.992 (0.34)	0.009 (0.32)	0.044 (0.87)	0.096 (0.42)	1.93 (1.80–2.05)
			β_3	LR	1.049 (2.17)	0.012 (0.64)	0.063 (1.89)	0.116 (1.69)	2.03 (1.94–2.17)
				LM_1	1.127 (5.67)	0.021 (3.50)	0.087 (5.37)	0.139 (4.11)	2.23 (2.10–2.31)
				LM_2	1.032 (1.45)	0.011 (0.32)	0.057 (1.02)	0.112 (1.26)	1.98 (1.92–2.13)
			β_4	LR	1.061 (2.73)	0.013 (0.95)	0.061 (1.60)	0.122 (2.32)	2.05 (1.91–2.15)
				LM_1	1.393 (17.6)	0.062 (16.5)	0.164 (16.5)	0.237 (14.4)	2.80 (2.55–3.04)
				LM_2	0.957 (1.92)	0.002 (2.54)	0.032 (2.61)	0.086 (1.48)	1.79 (1.73–1.95)
1000	200	2	β_2	LR	1.017 (0.77)	0.012 (0.64)	0.056 (0.87)	0.094 (0.63)	2.01 (1.88–2.17)
				LM_1	1.039 (1.74)	0.013 (0.95)	0.061 (1.60)	0.107 (0.74)	2.05 (1.94–2.19)
				LM_2	1.006 (0.26)	0.008 (0.64)	0.051 (0.15)	0.091 (0.95)	1.98 (1.85–2.12)
			β_3	LR	1.030 (1.34)	0.008 (0.64)	0.053 (0.44)	0.113 (1.37)	1.99 (1.89–2.10)
				LM_1	1.070 (3.14)	0.011 (0.32)	0.070 (2.90)	0.125 (2.64)	2.07 (1.98–2.21)
				LM_2	1.023 (1.01)	0.007 (0.95)	0.052 (0.29)	0.110 (1.05)	2.00 (1.87–2.09)
			β_4	LR	1.019 (0.84)	0.007 (0.95)	0.056 (0.87)	0.106 (0.63)	2.05 (1.92–2.20)
				LM_1	1.213 (9.52)	0.044 (10.8)	0.106 (8.13)	0.170 (7.38)	2.39 (2.19–2.76)
				LM_2	0.971 (1.31)	0.003 (2.22)	0.049 (0.15)	0.095 (0.53)	1.92 (1.84–2.08)
1000	100	4	β_2	LR	1.049 (2.20)	0.017 (2.22)	0.066 (2.32)	0.109 (0.95)	2.09 (1.97–2.22)
				LM_1	1.108 (4.82)	0.018 (2.54)	0.080 (4.35)	0.136 (3.79)	2.19 (2.06–2.34)
				LM_2	1.002 (0.08)	0.005 (1.59)	0.052 (0.29)	0.099 (0.11)	1.98 (1.85–2.07)
			β_3	LR	1.029 (1.30)	0.012 (0.64)	0.061 (1.60)	0.107 (0.74)	2.03 (1.93–2.15)
				LM_1	1.138 (6.19)	0.023 (4.13)	0.087 (5.37)	0.135 (3.69)	2.27 (2.15–2.40)
				LM_2	1.002 (0.11)	0.007 (0.95)	0.050 (0.00)	0.097 (0.32)	1.96 (1.87–2.05)
			β_4	LR	1.105 (4.70)	0.015 (1.59)	0.081 (4.50)	0.142 (4.43)	2.15 (2.04–2.25)
				LM_1	1.601 (26.9)	0.112 (32.4)	0.230 (26.1)	0.311 (22.2)	3.17 (3.04–3.43)
				LM_2	0.925 (3.37)	0.003 (2.22)	0.029 (3.05)	0.072 (2.95)	1.76 (1.68–1.90)

and nR^2 are always extremely similar, much more so than those of any of the other test statistics.

It seems clear from table 1 that we would always wish to use LM_2 rather than the pseudo-LM statistics F_2 and nR^2 based on the same artificial regression. The choice between LM_1 and F_1 is not so clearcut, since the standard errors of those two statistics tend to be very similar. However, close examination of table 1 shows that the standard error of F_1 always exceeds that of LM_1, which is in turn always greater than one, so that LM_1 clearly dominates F_1. This then leaves three test statistics which are worth examining more closely: LR, LM_1 and LM_2. Detailed results for these three are presented in tables 2 and 3. Here 'Rep.' and 'Nobs.' indicate the number of replications and the sample size, respectively. Besides the standard errors, we here report the proportion of the time that the test statistics exceeded 2.576, 1.960 and 1.645 (the 0.01, 0.05 and 0.10 critical values of the standard normal distribution) under '0.01 tail', '0.05 tail' and '0.10 tail', respectively. These are followed by estimated asymptotic absolute t-statistics for the hypotheses that these proportions are 0.01, 0.05 and 0.10.[4] Finally, under '0.05 crit.' we report estimated critical values for tests with a size of 0.05, together with estimated 95% confidence intervals on these estimates.[5]

Tables 2 and 3 are largely self-explanatory. It is evident that LM_2 is much the best behaved of the three test statistics in almost all cases, followed at some distance by LR, and at a long distance by LM_1. This last always rejects the null hypothesis more often than it should, sometimes rejecting more than 10% of the time at a nominal 1% level. The performance of all the test statistics tends to improve as the sample size increases, and tends to worsen when β_1 is increased. Tests of $\beta_2 = 0$ and $\beta_3 = 0$ tend to be better behaved than tests of $\beta_4 = 0$, perhaps because of the greater non-linearity involved in the heteroskedastic alternative.

The poor performance of LM_1 relative to LM_2 is not entirely unexpected. As we show in the appendix, the random variable towards which all of the LM test statistics tend asymptotically, LM_0, depends on y only through the gradient; the information matrix does not depend on y at all. LM_2 differs from LM_0 only because the former uses $\mathcal{I}(\tilde{\beta})$ while the latter uses $\mathcal{I}(\beta^0)$, where β^0 is the true parameter vector. Note that since $\tilde{\beta}$ is an ML estimate, $\mathcal{I}(\tilde{\beta})$ is asymptotically efficient for $\mathcal{I}(\beta^0)$. In contrast, LM_1 uses $\tilde{\mathcal{I}}^1$, which *does* depend directly on y, and must therefore be a less efficient estimator than $\mathcal{I}(\tilde{\beta})$. Since the asymptotic distribution of LM_0 depends on $\mathcal{I}(\beta^0)$ being non-random, we would expect that LM_2, which uses a relatively efficient

[4] This t-statistic is equal to $(\hat{p} - p)/[p(1-p)/N]^{\frac{1}{2}}$, where $\hat{p}$ is the observed proportion and p is the expected proportion if the test statistic were really N(0,1). Use of the normal approximation to the binomial is justified by the fact that N is always 1000.

[5] Note that these confidence bounds, which are based on non-parametric inference, are not symmetric around the estimate. For details on their calculation, see Mood and Graybill (1963, pp. 406–409).

estimate of $\mathcal{I}(\beta^0)$, should more closely approximate the distribution of LM_0 than LM_1. For a similar argument, see Davidson and MacKinnon (1983).

The results of these experiments are thus quite definite. You are least likely to make a Type I error if you use LM_2. Indeed, except for tests of $\beta_4 = 0$, where it tends to reject the null less often than it should, LM_2 seems to have a small-sample distribution which is remarkably close to its asymptotic one. The likelihood ratio test is less reliable than LM_2, but still reasonably well behaved. However, LM_1, F_1, F_2 and $2 - nR^2$ are often very badly behaved, and may reject a true null hypothesis much too often.

4. Power of the tests

In this section we investigate the power of the tests dealt with previously. In order to do so, we first investigate the asymptotic distribution of the LM test statistic when the null hypothesis is false, but the data generating process, or DGP, is assumed to be 'close' to the null. On the basis of this asymptotic distribution, which is of course the same for all the LM and pseudo-LM tests and for the LR test, we know what the power of the tests should be asymptotically. We can then see how this compares with the actual power of the tests in small samples.

The parameter vector β may be partitioned into two column vectors, β_1 and β_2, of lengths $(m-k)$ and k, respectively. The 'true' value of β_1 is β_1^0, and the 'true' value of β_2 is 0, where the meaning of 'true' should become clear in a moment. Thus β^0 is the vector whose first $m-k$ elements are β_1^0 and whose last k elements are 0. The information matrix from a sample of size n will be defined by

$$\mathcal{I} \equiv n\imath \equiv \mathrm{E}_0\big[g(\beta^0; y)^T g(\beta^0; y)\big], \tag{30}$$

where the expectation in (30) is taken assuming that $\beta = \beta^0$, and $\imath$ represents the average information contained in one observation.

The DGP is characterized by the loglikelihood function

$$\sum_{t=1}^{n} L_t(a_t, \beta^0; y_t), \tag{31}$$

where

$$\begin{aligned} L_t(a_t, \beta^0; 1) &= \log\big[F(x_t(\beta^0)) + a_t\big], \\ L_t(a_t, \beta^0; 0) &= \log\big[F(-x_t(\beta^0)) - a_t\big]. \end{aligned} \tag{32}$$

The numbers a_t have the properties that

$$0 \le F(x_t(\beta^0)) + a_t \le 1, \tag{33}$$

and

$$n^{-\frac{1}{2}} \sum_{t=1}^{n} a_t = \mathrm{O}(1). \tag{34}$$

Each individual a_t therefore becomes small like $n^{-\frac{1}{2}}$ as n becomes large, so that the DGP approaches the null hypothesis in large samples.

This characterization of the DGP is unrelated to the alternative hypothesis against which the LM test is constructed. The standard case where the DGP is a special case of the alternative hypothesis is easily accommodated, however. Suppose that, for some scalar α of order $n^{-\frac{1}{2}}$, the probability that $y_t = 1$ is $F(x_t(\beta_1^0, \alpha\beta_2^0))$, which is nested in the parametric family $F(x_t(\beta_1, \beta_2))$. This parametric family clearly includes the null at $\beta_1 = \beta_1^0$, $\alpha = 0$. Then the a_t, being of order $n^{-\frac{1}{2}}$, can to that order be approximated by

$$\alpha \sum_{j=1}^{k} \beta_{2j}^0 f\big(x_t(\beta_1^0, 0)\big) X_{tj}(\beta_1^0, 0), \tag{35}$$

where $X_{tj}(\beta_1, \beta_2)$ denotes the derivative of x_t with respect to the jth element of β_2. If the a_t were defined by (35), then the results of our analysis below would correspond with standard results for the case where the DGP is embedded within the alternative hypothesis.

Now define the $1 \times m$ vector

$$\Lambda = n^{-\frac{1}{2}} \sum_{t=1}^{n} a_t \big[G_t(\beta^0; 1) - G_t(\beta^0; 0)\big], \tag{36}$$

where, as before, $G_t(\beta^0; y_t)$ is the contribution to $g(\beta^0; y)$ from the tth observation. From (5) we see that

$$\begin{aligned} G_t(\beta^0; 1) &= f\big(x_t(\beta^0)\big) X_t(\beta^0) / F\big(x_t(\beta^0)\big), \\ G_t(\beta^0; 0) &= -f\big(x_t(\beta^0)\big) X_t(\beta^0) / F\big(-x_t(\beta^0)\big), \end{aligned} \tag{37}$$

where $X_t(\beta^0)$ is the gradient of x_t (a row vector of length m) evaluated at β^0.

Λ and $\mathscr{I}$ may be partitioned according to the distinction between β_1 and β_2 as follows:

$$\Lambda = [\Lambda_1 \quad \Lambda_2], \qquad \mathscr{I} = \begin{bmatrix} \mathscr{I}_{11} & \mathscr{I}_{12} \\ \mathscr{I}_{21} & \mathscr{I}_{22} \end{bmatrix}.$$

It is shown in the appendix that the asymptotic distribution of all the LM statistics is non-central chi-squared, with non-centrality parameter

$$n\left[\Lambda_2 - \Lambda_1 \mathcal{I}_{11}^{-1}\mathcal{I}_{12}\right](\mathcal{I}^{-1})_{22}\left[-\mathcal{I}_{21}\mathcal{I}_{11}^{-1}\Lambda_1^T + \Lambda_2^T\right]. \tag{38}$$

Here $\mathcal{I}_{11}^{-1}$ denotes the inverse of $\mathcal{I}_{11}$, while $(\mathcal{I}^{-1})_{22}$ denotes the 22-block of $\mathcal{I}^{-1}$. It is easily derived that

$$(\mathcal{I}^{-1})_{22} = \left[\mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}\right]^{-1}. \tag{39}$$

Now define the matrix R as the matrix with typical element $R_{ti}(\beta^0)$, where the latter was defined in (16), and partition it as $[R_1\ R_2]$. In addition, define the $n \times 1$ column vector r_0 by the equation

$$r_{0t} = a_t\left[F(x_t(\beta^0))F(-x_t(\beta^0))\right]^{-\frac{1}{2}}. \tag{40}$$

Clearly $\mathcal{I} = R^TR$ and $\Lambda = n^{\frac{1}{2}}r_0^TR$. In addition, define

$$M_1 = I - R_1\left(R_1^TR_1\right)^{-1}R_1^T. \tag{41}$$

Then it is evident that

$$\mathcal{I}_{ij} = R_i^TR_j \quad \text{for} \quad i, j = 1,2, \tag{42}$$

and that

$$(\mathcal{I}^{-1})_{22} = \left(R_2^TM_1R_2\right)^{-1}; \tag{43}$$

this result follows immediately from (39). Making use of these results, we can reduce expression (38) for the non-centrality parameter to

$$r_0^TM_1R_2\left(R_2^TM_1R_2\right)^{-1}R_2^TM_1r_0. \tag{44}$$

This expression may readily be computed by means of artificial linear regressions.

The results we have just derived are strictly valid only in the limit, as the sample size tends to infinity and the DGP tends to the null hypothesis. If they are to be useful, these results should be approximately valid when the sample size is moderate and the DGP is some distance from the null. To see whether this is in fact the case, and to compare the power of alternative tests, we conducted a further series of sampling experiments, similar to the ones reported on in section 3; the results are presented in tables 4 and 5.

Table 4

Power of LM and LR tests – Logit.[a]

Rep.	Nobs.	Alt.	Hyp.	R.N.C.P.	Test	Mean	Std. error	Power-1.96	Power-est.	Est.
500	50	$\beta_2 = 0.704$	β_2	1.96	LR	1.79 (3.80)	0.974 (0.82)	0.418	0.356	2.16
					LM_1	1.79 (4.17)	0.906 (2.97)	0.440	0.330	2.23
					LM_2	1.66 (7.87)	0.841 (5.03)	0.384	0.344	2.04
			β_3	0.11	LR	0.17 (1.47)	0.969 (0.97)	0.038	0.034	1.97
					LM_1	0.20 (1.80)	1.059 (1.87)	0.062	0.048	2.19
					LM_2	0.17 (1.47)	0.946 (1.71)	0.032	0.036	1.92
			β_4	−0.42	LR	−0.44 (0.32)	1.086 (2.73)	0.094	0.050	2.34
					LM_1	−0.59 (2.74)	1.395 (12.5)	0.192	0.028	3.30
					LM_2	−0.36 (1.60)	0.923 (2.42)	0.038	0.038	1.97
500	100	$\beta_2 = 0.498$	β_2	1.96	LR	1.88 (1.77)	0.978 (0.69)	0.450	0.422	2.04
					LM_1	1.90 (1.37)	0.963 (1.17)	0.488	0.436	2.04
					LM_2	1.80 (3.91)	0.900 (3.17)	0.428	0.420	2.00
500	200	$\beta_2 = 0.352$	β_2	1.96	LR	2.01 (1.03)	0.996 (0.11)	0.506	0.464	2.11
					LM_1	2.01 (1.12)	0.981 (0.61)	0.514	0.476	2.09
					LM_2	1.96 (0.04)	0.949 (1.62)	0.500	0.464	2.08
500	50	$\beta_3 = 5.33$	β_3	1.96	LR	1.87 (1.98)	1.025 (0.81)	0.464	0.460	1.97
					LM_1	2.13 (3.34)	1.163 (5.15)	0.568	0.496	2.19
					LM_2	1.74 (5.40)	0.909 (2.88)	0.438	0.446	1.92
500	100	$\beta_3 = 3.77$	β_3	1.96	LR	1.87 (2.00)	0.973 (0.86)	0.478	0.448	2.01
					LM_1	2.04 (1.71)	1.102 (3.22)	0.540	0.512	2.05
					LM_2	1.81 (3.56)	0.917 (2.62)	0.452	0.444	1.98
500	200	$\beta_3 = 2.67$	β_3	1.96	LR	1.98 (0.45)	0.990 (0.33)	0.508	0.496	1.99
					LM_1	2.07 (2.25)	1.065 (2.05)	0.538	0.506	2.04
					LM_2	1.94 (0.32)	0.957 (1.36)	0.498	0.494	1.98
500	50	$\beta_4 = 4.00$	β_4	1.96	LR	1.16 (14.2)	1.261 (8.25)	0.290	0.180	2.34
					LM_1	1.05 (18.6)	1.090 (2.84)	0.230	0.014	3.30
					LM_2	0.90 (25.1)	0.942 (1.83)	0.118	0.118	1.97
500	100	$\beta_4 = 2.83$	β_4	1.96	LR	1.58 (8.38)	1.017 (0.53)	0.352	0.284	2.19
					LM_1	1.63 (6.88)	1.063 (2.00)	0.380	0.140	2.76
					LM_2	1.36 (15.5)	0.866 (4.25)	0.244	0.226	2.00
500	100	$\beta_4 = 1.41$	β_4	0.98	LR	0.90 (1.71)	1.066 (2.08)	0.172	0.118	2.19
					LM_1	1.01 (0.55)	1.204 (6.46)	0.236	0.082	2.76
					LM_2	0.81 (4.08)	0.954 (1.46)	0.130	0.114	2.00
500	200	$\beta_4 = 2.00$	β_4	1.96	LR	1.64 (7.09)	1.009 (0.30)	0.382	0.354	2.02
					LM_1	1.78 (3.46)	1.126 (3.98)	0.464	0.330	2.31
					LM_2	1.52 (10.7)	0.917 (2.61)	0.320	0.348	1.91

[a] Figures in parentheses after means and standard errors are t-statistics for the hypotheses that the mean is equal to R.N.C.P. and the standard error is equal to one.

Standard errors for estimates under 'Power-1.96' may be calculated as $[p(1-p)/500]^{\frac{1}{2}}$, where p is the estimated power.

In these experiments $x_t(\beta)$ was given by (27), and one of β_2, β_3 or β_4 was always non-zero; which one, and its value, are indicated under 'Alt.' in the tables. With two exceptions, the value of the non-zero parameter was chosen so that when the hypothesis being tested (indicated under 'Hyp.') was the alternative which generated the data, λ, the square root of the noncentrality parameter, would 1.96. The value of λ is shown under 'R.N.C.P.' in the tables.

Table 5

Power of LM and LR tests – Probit.[a]

Rep.	Nobs.	Alt.	Hyp.	R.N.C.P.	Test	Mean	Std. error	Power-1.96	Power-est.	Est.
500	50	$\beta_2 = 0.428$	β_2	1.96	LR	1.86 (2.24)	0.989 (0.34)	0.448	0.382	2.07
					LM_1	1.90 (1.37)	0.959 (1.29)	0.474	0.374	2.18
					LM_2	1.71 (6.56)	0.848 (4.82)	0.384	0.374	1.98
500	100	$\beta_2 = 0.303$	β_2	1.96	LR	1.94 (0.37)	0.991 (0.29)	0.494	0.486	1.98
					LM_1	2.00 (0.91)	1.007 (0.21)	0.520	0.482	2.06
					LM_2	1.85 (2.67)	0.901 (3.14)	0.458	0.480	1.93
500	200	$\beta_2 = 0.214$	β_2	1.96	LR	1.96 (0.01)	0.958 (1.33)	0.496	0.480	2.01
					LM_1	2.00 (0.93)	0.974 (0.81)	0.516	0.474	2.05
					LM_2	1.91 (1.15)	0.915 (2.69)	0.488	0.482	1.98
			β_3	0.12	LR	0.19 (1.49)	1.000 (0.00)	0.058	0.054	1.99
					LM_1	0.20 (1.68)	1.047 (1.49)	0.062	0.054	2.07
					LM_2	0.19 (1.43)	0.992 (0.25)	0.058	0.052	2.00
			β_4	−0.46	LR	−0.49 (0.58)	1.064 (2.03)	0.098	0.074	2.05
					LM_1	−0.60 (2.56)	1.285 (9.01)	0.160	0.090	2.39
					LM_2	−0.47 (0.33)	1.031 (0.97)	0.068	0.080	1.92
500	50	$\beta_3 = 3.25$	β_3	1.96	LR	2.01 (1.15)	0.977 (0.73)	0.522	0.454	2.08
					LM_1	2.36 (7.95)	1.112 (3.55)	0.652	0.522	2.30
					LM_2	1.83 (3.61)	0.825 (5.53)	0.450	0.430	2.00
500	50	$\beta_3 = 4.91$	β_3	2.96	LR	2.60 (8.11)	0.985 (0.47)	0.748	0.704	2.08
					LM_1	3.03 (1.56)	1.027 (0.85)	0.844	0.772	2.30
					LM_2	2.27 (20.5)	0.758 (7.64)	0.678	0.650	2.00
500	100	$\beta_3 = 2.30$	β_3	1.96	LR	1.94 (0.52)	1.007 (0.23)	0.506	0.480	2.03
					LM_1	2.18 (4.15)	1.158 (5.00)	0.576	0.480	2.23
					LM_2	1.84 (2.82)	0.929 (2.25)	0.482	0.470	1.98
500	200	$\beta_3 = 1.63$	β_3	1.96	LR	1.94 (0.58)	0.933 (2.13)	0.480	0.468	1.99
					LM_1	2.08 (2.56)	1.036 (1.14)	0.532	0.492	2.07
					LM_2	1.89 (1.89)	0.888 (3.55)	0.466	0.446	2.00
500	50	$\beta_4 = 3.64$	β_4	1.96	LR	1.31 (12.5)	1.153 (4.85)	0.316	0.230	2.26
					LM_1	1.20 (16.2)	1.049 (1.55)	0.246	0.028	3.10
					LM_2	0.99 (24.8)	0.877 (3.90)	0.140	0.210	1.73
500	100	$\beta_4 = 2.57$	β_4	1.96	LR	1.53 (9.59)	0.997 (0.10)	0.336	0.302	2.05
					LM_1	1.67 (5.99)	1.088 (2.79)	0.388	0.140	2.80
					LM_2	1.33 (16.4)	0.850 (4.74)	0.228	0.296	1.79
500	200	$\beta_4 = 1.82$	β_4	1.96	LR	1.70 (6.05)	0.942 (1.85)	0.398	0.368	2.05
					LM_1	1.92 (0.70)	1.089 (2.82)	0.490	0.340	2.39
					LM_2	1.55 (11.0)	0.833 (5.29)	0.334	0.350	1.92

[a]See notes to table 4.

Setting $\lambda = 1.96$ ensures that the power of a test at the 0.05 level should be approximately 0.5, since if the test statistic were distributed as $N(\lambda, 1)$, it should be greater than 1.96 half the time (in addition, of course, it would be less than -1.96 a very small fraction of the time). The design of our experiments was such that λ was always very small when the alternative which generated the data was not also the hypothesis under test; for the most part, therefore, we report results only for the case where they were the same.

According to the asymptotic theory, each of the test statistics should be $N(\lambda, 1)$. We therefore report the mean and standard error of the statistics, together with t-statistics (asymptotic ones in the latter case) for the hypotheses that the mean is λ and the standard error is one. It is evident from the tables that the asymptotic approximation is somewhat deficient. The mean is often significantly too small for LM_2, and is sometimes significantly different from λ (but not always too small) for LR and LM_1. The standard deviation is often significantly different from unity, tending to be too small for LM_2. When the mean is less than λ, power suffers. In consequence, LM_2 always has less power using a critical value of 1.96 than either of the other test statistics. Such a comparison is not fair, however, because LM_2 was also less likely to reject the null when it was true. A more reasonable comparison is to use different critical values for each of the test statistics, based on the estimated 0.05 critical values from the experiments reported on in section 3. Test powers at these estimated critical values are shown under 'Power-est.', and the estimated critical values under 'Est.'. The performance of all three tests is now quite similar, with LM_1 and LR tending to do slightly better than LM_2 in most cases. However, LM_1 has quite low power when testing the hypothesis that $\beta_4 = 0$. The reason for this appears to be that the standard deviation of LM_1, though greater than one, is substantially less than it was when the null was true, so that the estimated critical values are too conservative.

By and large, the asymptotic results do not appear to be seriously misleading when the alternative which generated the data is $\beta_2 \neq 0$ or $\beta_3 \neq 0$, but when $\beta_4 \neq 0$ they do. In almost all cases when $\beta_4 \neq 0$, the actual mean of the test statistic is far less than λ. This discrepancy diminishes as the sample size is increased, and diminishes even more strikingly when the value of λ is halved to 0.98 (see table 4). The effect of each of these changes is to reduce the value of β_4, so that the DGP is closer to the null hypothesis. This suggests that the poor performance of the asymptotic theory in this case is due to the nonlinearity of $x_t(\beta)$ with respect to β_4.

The results of this section do not provide any clearcut guidance on the choice of tests. The fine performance of LM_2 under the null does not carry over to the alternative, although on a size-corrected basis it is rarely much less powerful than the other tests. The LR test generally performs quite well, and would seem to be preferable to LM_1 because of its better performance against $\beta_4 \neq 0$ and under the null. Unfortunately, the LR test is expensive. In our experiments, calculating the three LR test statistics took more than ten times as much computer time as calculating both LM_1 and LM_2 for all three alternatives.

5. Conclusion

In this paper we have discussed two forms of the LM test for logit and probit models. These were derived in the context of general non-linear binary

choice models, and may be used to test for both omitted variables and heteroskedasticity, among other things. In a series of sampling experiments, we compared these two forms of the LM test, along with several pseudo-LM tests based on the same artificial regressions, and the likelihood ratio test. We found that LM_2 tends to be the most reliable test under the null, but not the most powerful. We also found that pseudo-LM tests may behave very differently from genuine LM tests, even though based on the same artificial regression.

Appendix

In this appendix we show that, under local alternatives as specified by (31) to (34) in the text, both LM_1 and LM_2 tend asymptotically to the same random variable, LM_0. We then work out the asymptotic distribution of LM_0. Our specification of local alternatives is more general, and our results more explicit, than the treatments in standard references such as Engle (1982).

We first derive an expression for $\tilde{\beta} = (\tilde{\beta}_1^T, 0^T)^T$, the vector of constrained parameter estimates. $\tilde{\beta}_1$ is defined by the likelihood equations

$$g_1(\tilde{\beta}_1, 0; y) = \sum_{t=1}^{n} G_{1t}(\tilde{\beta}_1, 0; y_t) = 0, \tag{A.1}$$

where g_1 and G_{1t} denote gradients of L and L_t with respect to β_1 only. A Taylor expansion of these likelihood equations yields

$$0 = n^{-\frac{1}{2}} \sum_{t=1}^{n} G_{1t}(\beta^0; y_t) + n^{\frac{1}{2}}(\tilde{\beta}_1 - \beta_1^0)^T \left[n^{-1} \sum_{t=1}^{n} H_{11t}(\beta^0; y_t) \right] + O(n^{-\frac{1}{2}}), \tag{A.2}$$

where H_t is the contribution from the tth observation to the Hessian of the loglikelihood, and H_{11t} is the block corresponding to β_1.

We must show that the Taylor expansion in (A.2) is valid. First of all, we note that $E(G_{1t}(\beta^0; y_t)) = O(n^{-\frac{1}{2}})$. This follows from the facts that $E_0(G_{1t}(\beta^0; y_t)) = 0$, and that the difference between the likelihoods used in calculating these two expectations (a_t) is $O(n^{-\frac{1}{2}})$. The law of large numbers allows us to conclude that

$$n^{-1} \sum_{t=1}^{n} G_{1t}(\beta^0; y_t) = O(n^{-\frac{1}{2}}).$$

Similarly,

$$E(H_t(\beta^0; y_t)) = E_0(H_t(\beta^0; y_t)) + O(n^{-\frac{1}{2}}). \tag{A.3}$$

Thus from the central limit theorem and the standard result that $E_0(-H_t) = E_0(G_t^T G_t)$, we may deduce that

$$n^{-1}\sum_{t=1}^{n} H_t(\beta^0; y_t) = n^{-1}E_0\left[\sum_{t=1}^{n} H_t(\beta^0; y_t)\right] + O(n^{-\frac{1}{2}})$$

$$= -\imath + O(n^{-\frac{1}{2}}). \tag{A.4}$$

Consequently,

$$n^{-1}\sum_{t=1}^{n} H_{11t}(\beta^0; y_t) = -\imath_{11} + O(n^{-\frac{1}{2}}),$$

and is therefore not only O(1) but also bounded away from zero, since asymptotic identifiability requires that $\imath$ be strictly positive definite. From (A.2) than, $\tilde{\beta}_1 - \beta_1^0 = O(n^{-\frac{1}{2}})$, and the Taylor expansion is justified. In fact,

$$n^{\frac{1}{2}}(\tilde{\beta}_1 - \beta_1^0) = \imath_{11}^{-1}\left[n^{-\frac{1}{2}}\sum_{t=1}^{n} G_{1t}^T(\beta^0; y_t)\right] + O(n^{-\frac{1}{2}}). \tag{A.5}$$

The LM statistic in any of its various forms may be written as

$$g_2(\tilde{\beta}; y)(\mathcal{I}^{*-1})_{22} g_2(\tilde{\beta}; y)^T, \tag{A.6}$$

where $n\mathcal{I}^{*-1}$ is some estimator, consistent if $\beta = \beta^0$, of n times the inverse of the information matrix. The choice of this estimator does not affect the fact that, under local alternatives, $n\mathcal{I}^{*-1}$ is asymptotically non-stochastic and equal to $\imath^{-1}$, because the likelihood differences a_t introduce only terms of order $n^{-\frac{1}{2}}$.

By (A.4) and (A.5),

$$\begin{aligned} n^{-\frac{1}{2}} g_2(\tilde{\beta}; y) &= n^{-\frac{1}{2}}\sum_{t=1}^{n} G_{2t}(\tilde{\beta}; y_t) \\ &= n^{-\frac{1}{2}}\sum_{t=1}^{n} G_{2t}(\beta^0; y_t) \\ &\quad + n^{\frac{1}{2}}(\tilde{\beta}_1 - \beta_1^0)^T\left[n^{-1}\sum_{t=1}^{n} H_{12t}(\beta^0; y_t)\right] + O(n^{-\frac{1}{2}}) \\ &= n^{-\frac{1}{2}} g(\beta^0; y)\begin{bmatrix} -\imath_{11}^{-1}\imath_{12} \\ I_k \end{bmatrix} + O(n^{-\frac{1}{2}}), \end{aligned} \tag{A.7}$$

where I_k is the $k \times k$ identity matrix. Thus the LM statistic (A.6) is equal to

$$n^{-1} g(\beta^0; y) \begin{bmatrix} -\imath_{11}^{-1}\imath_{12} \\ I_k \end{bmatrix} (\imath^{-1})_{22} \left[-\imath_{21}\ \imath_{11}^{-1} I_k \right] g(\beta^0; y)^T + O(n^{-\frac{1}{2}}). \tag{A.8}$$

The asymptotic term in (A.8) is the random variable called LM_0 in the text; it is evident that LM_0 depends on y only through the gradient $g(\beta^0; y)$.

The expectation of $n^{-\frac{1}{2}} g(\beta^0; y)$ is easily calculated. It is

$$\mathrm{E}\left(n^{-\frac{1}{2}} g(\beta^0; y) \right) = n^{-\frac{1}{2}} \sum_{t=1}^{n} a_t \left[G_t(\beta^0; 1) - G_t(\beta^0; 0) \right] = \Lambda. \tag{A.9}$$

The equality here follows from the definition of Λ [see (36)], and from the fact that $\mathrm{E}_0(g(\beta^0; y)) = 0$. By the argument used to derive (A.4), it is evident that

$$\begin{aligned} \mathrm{var}\left[n^{-\frac{1}{2}} g(\beta^0; y) \right] &= \mathrm{E}\left[n^{-1} g(\beta^0; y)^T \left(g(\beta^0; y) - \Lambda \right) \right] \\ &= \imath + O(n^{-\frac{1}{2}}). \end{aligned} \tag{A.10}$$

It is now straightforward to compute that

$$\mathrm{var}\left[n^{-\frac{1}{2}} g(\beta^0; y) \begin{bmatrix} -\imath_{11}^{-1}\imath_{12} \\ I_k \end{bmatrix} \right] = \left[(\imath^{-1})_{22} \right]^{-1} + O(n^{-\frac{1}{2}}). \tag{A.11}$$

Let the random vector on the left-hand side of (A.11) be denoted by x^T. By a central limit theorem applied to $n^{-\frac{1}{2}} g(\beta^0; y)$, x is asymptotically normal. Its mean is

$$\mu = \left[-\imath_{21}\imath_{11}^{-1}\ I_k \right] \Lambda^T,$$

and its covariance matrix is $[(\imath^{-1})_{22}]^{-1}$, which we shall call A. Then it is immediate from (A.8) that, if one ignores terms not of leading order, LM_0 is equal to $x^T A^{-1} x$.

It is a standard result [see, for example, Hogg and Craig (1978, p. 413)] that if x is distributed as $\mathrm{N}(\mu, A)$, then the statistic $x^T A^{-1} x$ has the noncentral chi-squared distribution, with number of degrees of freedom equal to the rank of A and non-centrality parameter $\mu^T A^{-1} \mu$. In this case the rank of A is k (the length of β_2), and the non-centrality parameter is

$$\left[\Lambda_2 - \Lambda_1 \imath_{11}^{-1}\imath_{12} \right] (\imath^{-1})_{22} \left[-\imath_{21}\imath_{11}^{-1}\Lambda_1^T + \Lambda_2^T \right], \tag{A.12}$$

which is equivalent to expression (38) in the text.

Although this calculation is strictly for the LM statistic, a similar calculation shows that (A.12) gives the non-centrality parameter for the LR statistic as well.

References

Amemiya, Takeshi, 1981, Qualitative response models: A survey, Journal of Economic Literature 19, 1483–1536.

Berndt, Ernst R., Bronwyn H. Hall, Robert E. Hall and Jerry A. Hausman, 1974, Estimation and inference in nonlinear structural models, Annals of Economic and Social Measurement 3, 653–665.

Davidson, Russell and James G. MacKinnon, 1983, Small sample properties of alternative forms of the Lagrange multiplier test, Economics Letters 12, 269–275.

Engle, Robert F., 1982, Wald, likelihood ratio and Lagrange multiplier tests in econometrics, in: Z. Griliches and M. Intriligator, eds., Handbook of econometrics (North-Holland, Amsterdam).

Godfrey, Leslie G. and Michael R. Wickens, 1981, Testing linear and log-linear regressions for functional form, Review of Economic Studies 48, 487–496.

Hogg, Robert V. and Allen T. Craig, 1978, Introduction to mathematical statistics, 4th ed. (Macmillan, New York).

Mood, Alexander M. and Franklin A. Graybill, 1963, Introduction to the theory of statistics, 2nd ed. (McGraw-Hill, New York).