BULLETIN *of the*
MALAYSIAN
MATHEMATICAL
SCIENCES
SOCIETY

Bull. Malays. Math. Sci. Soc. (2) **27** (2004), 207–215

Significance Testing in Exact Logistic Multiple Regression

MEZBAHUR RAHMAN AND SHUVRO CHAKROBARTTY
Minnesota State University, Mankato, MN 56001, USA
e-mail: mezbahur.rahman@mnsu.edu and shuvro.chakrobartty@mnsu.edu

Abstract. Exact logistic regression is discussed for multiple regressors. The exact significance of a regressor is computed which can be used in simplifying the model and/or to compute the significance of a variable or a set of variables in the model. Bilingual education data is analyzed using the procedure mentioned in this paper.

2000 Mathematics Subject Classification: 62J02

1. Introduction

Following Cox ([1], Ch. 4), Tritchler [8] implemented an algorithm for the exact logistic regression analysis of a single regressor. Tritchler [8] used Fourier transformation algorithm given by [7] to compute the p value for testing the significance of the regression parameter. Mehta *et al.* [4] gave an efficient Monte Carlo method by networking the possible regressor values.

We assume that we have a set of independent variables (regressors) that are to be used for prediction in each of the following situations: (1) predicting whether a company's dealer will soon be mired in dire financial straits, (2) predicting if a person is likely to develop heart disease, (3) predicting whether a hospital patient will survive until being discharged, or (4) predicting whether a person has achieved a desirable competency level of a learning. In such scenarios, the response (dependent) variable is binary. Due to a wide range of applications, the binary response models are studied explicitly. For the latest developments in the area, the reader is referred to [6], [2] and the references therein.

Let $X_1, X_2, \cdots, X_p$ be p separate regressors and Y be a response (dependent) variable. Y can only take the values of '1' for 'success' and '0' for 'failure'. A random sample of n data points is taken from a phenomenon. A general binary model is assumed as

$$P(Y_i = 1) = \pi_i = E(Y_i \mid X_{1i}, X_{2i}, \cdots, X_{pi}), \quad i = 1,2,\ldots,n, \tag{1}$$

where $0 \le \pi_i \le 1$ and $P(\mathrm{Y}_i = 0) = 1 - \pi_i$. We define the logistic model as

$$\pi(\mathbf{x}_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right)} = \frac{\exp\left(\sum_{j=0}^{p} \beta_j x_{ji}\right)}{1 + \exp\left(\sum_{j=0}^{p} \beta_j x_{ji}\right)} \tag{2}$$

where $\beta_0, \beta_1, \cdots, \beta_{\mathrm{p}}$ are unknown constants and $\mathbf{x}_i$ is the row vector $(1\, x_{1i} \cdots x_{\mathrm{p}i})$. Notice that there is no error term on the right side of (2) because the left side is a function of $E(Y \mid X_1, X_2, \cdots, X_p)$, instead of *Y*, which serves to remove the error term.

If $y_1, y_2, \cdots, y_n$ is an observed binary sequence of size *i*, then

$$P\left(Y_1 = y_1, Y_2 = y_2, \cdots, Y_n = y_n \mid x_{11}, x_{12}, \cdots, x_{1n}, x_{21}, x_{22} \cdots, x_{2n} \cdots, x_{p1}, x_{p2}, \cdots, x_{pn}\right)$$
$$= \frac{\exp\left(\beta_0 s + \sum_{j=1}^{p} \beta_j t_j\right)}{\prod_{i=1}^{n}\left(1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ji}\right)\right)}, \tag{3}$$

where $s = \sum_{i=1}^{n} y_i$ and $t_j = \sum_{i=1}^{n} y_i\, x_{ji}$ for $j = 1, 2, \cdots, p$. Following Cox (1970), inference is based on the sufficient statistics $S = \sum_{i=1}^{n} Y_i$ and $T_j = \sum_{i=1}^{n} Y_i\, X_{ji}$ for $j = 1, 2, \cdots, p$, whose joint distribution is obtained by summing over all binary sequences generating each realization of $s, t_1, t_2 \cdots, t_{\mathrm{p}}$. Thus

$$P\left(S = s, T_1 = t_1, T_2 = t_2, \cdots, T_p = t_p\right) = \frac{\left(\prod_{j=1}^{p} C_j\right)\exp\left(\beta_0 s + \sum_{j=1}^{p} \beta_j t_j\right)}{\prod_{i=1}^{n}\left(1 + \exp\left(\beta_0 + \sum_{j=1}^{\mathrm{p}} \beta_j x_{ji}\right)\right)}, \tag{4}$$

where C_j's are the numbers of distinct binary sequences yielding the values s and t_j's for the sufficient statistics. Exact inferences containing β_j's may be based on the conditional distribution $P(T_1 = t_1, T_2 = t_2, \cdots, T_p = t_p \mid S = s)$. Because of properties of the exponential family of distributions, the critical region defined by the upper tail values of t_j within the conditional reference set provides a uniformly most powerful unbiased test of $H_0 : \beta_j = \beta_{j0}$ versus $H_a : \beta_j > \beta_{j0}$ (Lehman 1959, p. 136). The conditional distribution used to test hypotheses concerning β_j is

$$P\left(T_j = t_j \mid S = s\right) = \frac{C_j \exp(\beta_{j0} t_j)}{\sum_{\forall l} C_{jl} \exp(\beta_{j0} t_{jl})}, \tag{5}$$

where l is an index ranging over all the values taken by T_j.

2. Tests of significance

In one-sided tests, such as $H_0 : \beta_j = \beta_{j0}$ versus $H_a : \beta_j > \beta_{j0}$, the p value can be computed as

$$P\left(T_j \geq t_j \mid S = s\right) = \frac{\sum_{i:T_j \geq t_j} C_i \exp(\beta_{j0} t_i)}{\sum_{\forall l} C_l \exp(\beta_{j0} t_l)} \tag{6}$$

which tests for significance of the particular variate in the model. Due to the pattern of products of frequencies and exponentiation of a linear function, the p values are easy to compute for the significance of a set of variates. For example, for testing, $H_0 : \beta_j = \beta_{j0}$ and $\beta_k = \beta_{k0}$ versus H_a: at least one of the β's not equal to the specified value can be tested by computing the p value as

$$\begin{aligned} P\left(T_j \in R_j \text{ or } T_k \in R_k \mid S = s\right) &= 1 - P\left(T_j \notin R_j \text{ and } T_k \notin R_k \mid S = s\right) \\ &= 1 - P\left(T_j \notin R_j\right) P\left(T_k \notin R_k\right) = 1 - \left(1 - 2P\left(T_j \geq t_j \mid S = s\right)\right) \\ &\quad \left(1 - 2P\left(T_k \geq t_k \mid S = s\right)\right) \end{aligned} \tag{7}$$

where R_j and R_k are the respective critical regions. The computations of such p values are demonstrated using a data set in Section 5. In equation (7) '$\geq$' will be replaced in one or both places by '$\leq$' depending on whether observed t_j and/or t_k lie in which tail of the distributions of T_j and T_k. Similar processes can be applied for testing a set of more than two parameters.

3. Inferences based on maximum likelihood

Maximum likelihood parameter estimation is studied extensively. Here we review the method as given in [5]. The log-likelihood function for the logit model (2) can be written as

$$\log L(\beta) = \sum_{i=1}^{n} \left\{ y_i \log \pi(X_i) + (1 - y_i) \log[1 - \pi(X_i)] \right\},$$

where β is vector valued. For methods with more than one parameter, the first-order conditions require that we simultaneously solve the $p + 1$ equations

$$u_j = \frac{\delta \log L(\beta)}{\delta \beta_j} = 0, \quad j = 0, 1, \cdots p$$

for which $$h_{jk} = \frac{\delta^2 \log L(\beta)}{\delta\beta_j \delta\beta_k} < 0, \; (j = 0,1,\cdots, p), (k = 0,1,\cdots, p).$$

For the first-order conditions for the logit model (2), the likelihood expressions are written as

$$\frac{\delta \log L(\beta)}{\delta\beta} = U(\beta) = \sum_{i=1}^{n} \left[y_i - \pi(\mathbf{X}_i)\right]\mathbf{X}_i$$

and the negative of the second derivatives are

$$I(\beta) = -\frac{\delta^2 \log L(\beta)}{\delta\beta\delta\beta'} = \sum_{i=1}^{n} \pi(\mathbf{X}_i)[1 - \pi(\mathbf{X}_i)]\mathbf{X}_i' \, \mathbf{X}_i$$

where $U(\beta)$ is a $(p+1)\times 1$ vector and $I(\beta)$ is a $(p+1)\times(p+1)$ matrix. The matrix $I(\beta)$ plays a key role in the estimation procedure and yields the estimated variances and covariances of the estimates as by-product. The asymptotic variances and covariances of the logit estimates are obtained by inverting the Hessian (or expected Hessian) matrix or information matrix $I(\beta)$. Then the Newton-Raphson iterative solution of a system of equations can be used to obtain the solutions of β's. At the t^{th} iteration, estimates are obtained as

$$\hat{\beta}^{(t)} = \hat{\beta}^{(t-1)} + \left[I\left(\hat{\beta}^{(t-1)}\right)\right]^{-1} U\left(\hat{\beta}^{(t-1)}\right).$$

The least square estimates of β's are often used as the initial estimates. The quantity $\hat{\beta}_k \big/ \sqrt{D_{kk}}$ has asymptotic normal distribution where D_{kk} is the k^{th} diagonal element of $[I(\hat{\beta})]^{-1}$ and can be used in testing and in forming confidence intervals for a particular parameter β_k. Also, any subset of parameters can be tested using the following asymptotic χ^2 statistic.

$$\chi^2 = -[\log L_1 - \log L_0], \tag{9}$$

has approximate χ^2 distribution with $(n-q-1)-(n-p-1) = p-q$ degrees of freedom, where $q+1$ is the number of unknown parameters in the model under the null hypothesis, $\log L_1$ is the maximized log likelihood under the full model, and $\log L_0$ is the maximized log likelihood under the null hypothesis.

4. Motivation

Exact logistic regression is not a new phenomenon but in the wake of computational convenience is attracting more attention than before. Specialized software are not popular as they do not communicate effectively with the users. The algorithms are used in the software and are suggested by different authors are approximations, often through fourier transformations. Here we give algorithms to compute exact *p* values without any approximation. A data set is used to apply the exact logistic regression procedures. The exact *p* values will help to determine whether the particular factor is significant or not more accurately that using approximate *t* statistic for the maximum likelihood estimate. When particular factors are found to be significant then maximum likelihood method should be used in estimating or predicting the success probabilities. Often, the method of discriminant analysis (see [2], Section 1.5) gives higher rate of successful predictions but lacks properties like unbiasedness, consistency, and efficiency.

Table 1. Estimated Variance-Covariance Matrix (MLE)

Statistics	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
$\hat{\beta}_0$	0.8144	–0.0404	–0.1553	–0.0218
$\hat{\beta}_1$	–0.0404	0.0203	–0.0007	0.0025
$\hat{\beta}_2$	–0.1553	–0.0007	0.0395	–0.0007
$\hat{\beta}_3$	–0.0218	0.0025	–0.0007	0.0042

5. Application

The population studied is thirteen schools of the Salinas City Elementary School district in the County of Monterey in California. Participating students were fifth and sixth graders of limited English proficiency. This study was undertaken in 1996 to take an in depth look at the data gathered for the population of limited English proficient students and its role in redesignating students to fluent English proficiency status.

Information on 257 participating students were recorded and displayed in Table 3. In Tables 3, “E” represents English score, “S” represents Spanish score, “Y” represents the number of years in the program and “B” represents the redesignation in the bilingual status. In the variable “B”, ‘1’ indicates success of the participant in the program and ‘0’ indicates failure. In the variables “E” and “S”, higher the score means higher the proficiency.

Redesignation in the program is done by the evaluator after considering the three variables “E”, “S” and “Y”, and the personal judgement of the evaluator. Here we will model using Logistic regression models. Goal is to give a rule by which one can be redesignated. The assumption is that the redesignation of the present data is done by an expert and future redesignation is possible by an auto-mated rule with the help of the present analysis.

The Logistic model given in (2) is estimated using the maximum likelihood method (MLE) given in Powers ([5], Section 3.3.3) as described in Section 3 as

$$\hat{\pi} = \frac{\exp(1.0999 + 0.0127\,\text{Eng} - 0.1816\,\text{Spa} - 0.0474\,\text{Yrs})}{1 + \exp(1.0999 + 0.0127\,\text{Eng} - 0.1816\,\text{Spa} - 0.0474\,\text{Yrs})}$$

The variance-covariance matrix for the estimates of the parameters is computed using the delta method as $[I(\hat{\beta})]^{-1}$ in Section 3 and displayed in Table 1. The maximum likelihood estimates and the corresponding *p* values are displayed in Table 2. In Table 2, the likelihood ratio statistic as in (9) is represented as χ^2, Est. represents the MLE estimates of the parameters, *Z* is the studentized statistic, p_z is the *p* value for *Z* statistic, p_{χ^2} is the *p* value using the χ^2 statistic, Exact *p* is the *p* value using the exact method as described in Section 2. In testing $H_0 : \beta_1 = 0$ and $\beta_2 = 0$

Table 2. Estimates and *p* values

	Est.	Z	p_z	χ^2	p_{χ^2}	Exact p
$\hat{\beta}_0$	1.0999	1.2188	0.2229			
$\hat{\beta}_1$	0.0127	0.0891	0.9290	0.0078	0.9296	$2P(T_1 \geq 210 \mid S = 138) = 0.8204$
$\hat{\beta}_2$	–0.1816	–0.9137	0.3609	0.8410	0.3591	$2P(T_2 \geq 553 \mid S = 138) = 0.3889$
$\hat{\beta}_3$	–0.0474	0.7314	0.4645	0.5410	0.4620	$2P(T_3 \geq 673 \mid S = 138) = 0.4177$

versus H_a : at least one of β_1 and β_2 non-zero. p value $= 1 - (1 - 2P(T_1 \geq 210) \mid S = 138))\ (1 - 2P(T_2 \leq 553 \mid S = 138)) = 0.8902$.

Similarly, the *p* values for testing significance of the other two subsets of parameters (β_1, β_3) and (β_2, β_3) are respectively, 0.8954 and 0.6442. Asymptotic method such as likelihood ratio χ^2 gives the following three *p* values for the respective pairs of parameters as 0.6551, 0.7286, and 0.4789. The *p* value for testing significance of all three parameters β_1, β_2 and β_3 using likelihood ratio test is 0.6680, and using the exact method is $= 1 - (1 - 2P(T_1 \geq 210) \mid S = 138))\ (1 - 2P(T_2 \leq 553 \mid S = 138))\ (1 - 2P(T_3 \leq 673 \mid S = 138)) = 0.9361$.

6. Conclusion

In Section 5, we notice that the exact p values for the individual parameters are comparable with the asymptotic p values except for β_1 where the difference is noticeable. But for a set of two or three parameters the differences in p values are high. In asymptotic computations, the Chi-square inferences are based on independence of parameter estimates but in reality they are not as can be seen in Table 1. The differences in p values are clear even though the data is large and the covariances among the estimates are very small.

References

1. D.R. Cox, *Analysis of Binary Data*, London, UK: Methuen, 1970.
2. D.W. Hosmer Jr. and S. Lemeshow, *Applied Logistic Regression*, Second edition. New York: Wiley, 2000.
3. E.L. Lehmann, *Testing Statistical Hypothesis*, New York: John Wiley, 1995.
4. R.C. Mehta, N.R. Patel, and P. Senchaudhuri, Efficient Monte Carlo methods for conditional logistic regression, *Journal of the American Statistical Association* **95 (449)** (2000), 99−108.
5. D.A. Powers, *Statistical Methods for Categorical Data Analysis*, London, UK: Academic Press, 2000.
6. T.P. Ryan, *Modern Regression Methods*, New York: Wiley, 1997.
7. R.C. Singleton, An algorithm for computing the mixed-radix fast Fourier transformation, *IEEE Transactions on Audio and Electroacoustics* **17** (1969), 93−103.
8. D. Tritchler, An algorithm for exact logistic regression, *Journal of the American Statistical Association* **79** (1984), 709−711.

Keywords: Binary response model; Discriminant analysis; Goodness-of-fit; Least square estimate; Maximum likelihood estimate.

Table 3. Bilingual Data

ID#	E	S	Y	B	ID#	E	S	Y	B	ID#	E	S	Y	B	ID#	E	S	Y	B
1	1	4	7	1	66	1	4	7	1	131	3	6	4	0	196	1	5	3	0
2	2	4	4	1	67	3	4	5	1	132	3	5	5	1	197	2	3	4	0
3	3	4	6	1	68	3	4	2	1	133	1	4	7	0	198	1	3	1	0
4	1	4	5	1	69	1	5	5	0	134	1	4	7	1	199	3	2	0	0
5	1	3	6	0	70	1	4	7	0	135	1	4	7	1	200	3	4	3	0
6	1	3	5	1	71	1	4	4	1	136	1	5	4	1	201	1	4	0	0
7	1	3	7	1	72	3	4	3	0	137	3	4	7	1	202	1	4	7	0
8	1	3	7	0	73	2	4	4	1	138	1	3	7	1	203	1	4	3	0
9	1	4	7	1	74	1	5	5	0	139	3	4	7	1	204	1	4	7	0
10	1	5	7	0	75	1	4	7	0	140	1	4	7	1	205	1	4	5	0
11	1	5	7	0	76	1	4	4	1	141	1	4	7	1	206	1	4	5	1
12	1	4	7	1	77	3	4	1	0	142	1	4	7	1	207	3	4	5	0
13	1	5	7	1	78	2	4	5	0	143	1	5	7	1	208	1	4	7	0
14	1	4	7	0	79	1	5	5	0	144	1	4	7	0	209	3	4	6	0
15	1	4	6	1	80	1	4	7	0	145	1	4	7	1	210	1	4	5	0
16	1	4	7	0	81	1	4	5	1	146	1	4	7	0	211	1	5	7	0
17	1	4	7	0	82	3	4	2	0	147	1	4	7	1	212	1	4	7	0
18	1	4	7	0	83	2	4	6	0	148	1	3	7	1	213	1	4	6	0
19	3	3	7	0	84	1	4	7	0	149	1	4	5	0	214	1	4	6	0
20	1	4	7	0	85	2	2	4	0	150	1	4	7	1	215	1	5	5	0
21	1	4	6	1	86	4	4	3	0	151	3	4	2	1	216	2	4	6	0
22	1	3	6	1	87	1	4	6	1	152	1	3	7	1	217	1	4	7	0
23	1	4	7	0	88	2	4	5	1	153	1	4	5	1	218	1	4	6	0
24	1	3	4	1	89	1	3	6	1	154	1	4	6	1	219	1	4	4	0
25	1	4	6	0	90	1	4	5	1	155	2	3	6	0	220	1	4	5	0
26	1	4	7	1	91	1	4	5	1	156	2	4	6	0	221	1	3	4	1
27	1	4	7	0	92	4	4	6	0	157	1	4	6	1	222	2	4	3	1
28	1	4	7	0	93	1	5	6	1	158	1	4	6	0	223	2	4	4	1
29	1	4	7	0	94	1	5	5	0	159	3	4	6	1	224	1	5	5	1
30	1	4	7	0	95	1	4	6	1	160	2	5	7	0	225	1	3	2	0
31	1	4	2	1	96	1	4	6	1	161	3	5	4	0	226	1	4	0	1
32	1	4	7	0	97	1	4	7	0	162	1	4	5	1	227	1	4	0	1
33	1	4	6	0	98	1	4	6	1	163	1	4	5	0	228	3	4	0	1
34	1	4	7	1	99	1	4	5	1	164	1	5	4	0	229	1	4	2	1
35	1	4	7	1	100	1	5	5	1	165	1	5	2	0	230	3	4	2	0
36	1	4	6	0	101	2	5	3	1	166	1	5	1	0	231	1	4	0	1
37	1	4	7	1	102	5	4	1	1	167	1	4	6	1	232	1	4	2	0
38	1	4	7	0	103	1	4	4	1	168	1	5	7	0	233	3	3	0	0
39	1	4	5	1	104	1	3	4	1	169	2	5	6	0	234	5	4	5	0
40	1	4	7	0	105	3	4	2	1	170	5	5	6	0	235	1	4	5	0
41	3	4	5	1	106	2	4	4	1	171	1	5	4	0	236	1	3	5	1
42	1	4	6	1	107	1	3	6	1	172	1	5	1	0	237	3	3	0	1
43	1	4	7	0	108	2	3	5	1	173	1	5	5	0	238	3	4	2	0
44	1	4	7	1	109	1	3	7	0	174	1	5	6	0	239	2	5	7	1
45	1	4	7	0	110	1	5	5	1	175	1	5	4	0	240	3	5	4	1
46	1	3	7	0	111	1	4	6	1	176	3	4	6	0	241	1	4	5	1
47	1	5	7	0	112	1	4	6	1	177	1	4	7	0	242	1	5	4	0
48	1	4	6	1	113	1	4	0	0	178	4	5	7	1	243	1	5	2	1
49	1	4	5	1	114	1	3	2	0	179	1	5	7	0	244	1	4	5	1
50	1	3	6	1	115	1	4	6	1	180	2	5	6	0	245	1	4	6	1

Table 3 (*cont'd*)

51	1	3	5	1	116	1	4	4	1	181	5	5	2	1	246	2	3	6	1
52	1	5	6	1	117	1	3	7	0	182	1	5	5	1	247	2	4	6	1
53	1	5	6	1	118	1	3	7	0	183	1	5	0	0	248	1	4	6	1
54	1	4	6	1	119	1	5	7	1	184	1	5	5	1	249	1	4	6	1
55	1	5	6	1	120	3	4	2	1	185	1	5	7	0	250	3	4	6	1
56	3	3	6	1	121	1	4	2	1	186	1	5	4	0	251	3	4	4	1
57	1	4	1	1	122	3	4	4	1	187	1	4	7	0	252	1	4	2	1
58	1	4	5	0	123	1	3	2	0	188	4	5	3	1	253	1	4	2	1
59	1	3	1	1	124	1	4	2	1	189	1	5	3	0	254	1	4	7	1
60	1	4	7	0	125	1	4	2	1	190	2	3	6	0	255	2	4	4	1
61	1	4	6	1	126	3	4	2	0	191	1	4	2	1	256	3	4	5	1
62	3	4	5	1	127	1	3	4	1	192	1	3	1	0	257	1	4	5	1
63	1	4	7	0	128	3	4	2	1	193	3	2	0	0					
64	1	4	7	1	129	3	4	4	1	194	2	4	3	0					
65	1	4	6	1	130	1	5	5	1	195	1	4	0	1					