Bartlett Correction of Binomial Regression Analyses of Small Samples

Roy T. St. Laurent, Northern Arizona University, Flagstaff, AZ
Lisa A. Weissfeld, University of Pittsburgh, Pittsburgh, PA
Lawrence H. Moulton, The Johns Hopkins University, Baltimore, MD
(random order)

Many epidemiologic investigations are necessarily restricted in the number of subjects to be studied. Limited numbers of cases or exposed subjects may be available, or high data collection costs may limit the size of the study. Yet the need to analyze several confounding or effect modifying variables simultaneously is not necessarily diminished. Thus for dichotomous response variables, the use of binomial regression models (usually the logistic model) is often the analytic epproach of choice in such situations. In this paper we describe potential inferential pitfalls when analyzing data sets which are effectively "small," and we explain and discuss several possible courses of remedial action. The Bartlett correction to the likelihood ratio statistic is seen to be a particularly useful tool in this setting.

WHEN IS A SAMPLE SMALL?

The most commonly employed statistical testing procedures, the Wald, likelihood ratio, and score tests, are based on approximations justified on basis of their large-sample properties (1). The Wald test compares the maximum likelihood estimate (MLE) of the coefficient under test to its esymptotic standard error, while the likelihood ratio test (LRT) directly compares the maximized likelihood under the full model (that includes the variable under test) to the maximized likelihood under the reduced model (that excludes that variable). The score test is based on the derivative of the log likelihood function under the reduced model, and can be interpreted as a one-step approximation to the LRT. For a discussion of these methods see Cox and Snell (1989), Appendix 1.

We may call a sample "small" in relation to a given procedure when the approximation offered by the procedure is not sufficiently accurate. What may be considered small for a Wald test may not be small for a likelihood ratio test. This smallness derives from the combination of a variety of characteristics inherent in the data. First, as mentioned above, the sample may actually *be* small, perhaps consisting of only 10 to 20 observations. Second, there may be a large number of parameters to be estimated relative to the sample size: stable estimation of 10 parameters with 100 observations may be no easier than estimating 3 parameters with 30 observations. Third, the relative proportions of the values of the dichotomous outcome variable is important: there is far more information in a data set in which half the subjects have the response characteristic of interest than one in which only a handful have it. Hence, data sets in which the outcome is very rare (or very common) can be considered "small." Fourth, the functional relationship between the response and the explanatory variables affects the asymptotic approximation (2): the likelihood function for a risk difference (identity link) model is much flatter than that for an odds ratio (logit link) model, and therefore generally requires a larger sample size in order to obtain sufficiently precise inferences.

As an example, consider the data set of Finney (3) as reported by Pregibon (4). It consists of 39 observations on the binary response Y_i of occurrence (yes = 1; no = 0) of vasoconstriction in the fingers of subjects, with the volume (VOL) and average rate (RAT) of inspired air being the explanatory variables in the experiment. Finney proposed the model:

$$(RAT)^{\beta_1}\,(VOL)^{\beta_2} = \text{constant}$$

for curves of constant probability of occurrence. Assuming a logit link, this is equivalent to fitting the model:

$$\text{logit } Pr(Y_i=1) = \beta_0 + \beta_1 \log(RAT_i) + \beta_2 \log(VOL_i).$$

For comparison purposes, we also fit the identity (linear) link model:

$$Pr(Y_i=1) = \beta_0 + \beta_1 \log(RAT_i) + \beta_2 \log(VOL_i),\ i = 1,\ldots,39.$$

For the identity link model we have adopted the *ad hoc* procedure utilized by other workers (5-7) of bounding estimated probabilities between zero and unity. Estimated probabilities are assigned values of 10^{-6} or $1\text{-}10^{-6}$ during the fitting process when they become negative or greater than one, respectively. The standard test statistics for the null hypotheses that β_1 and β_2 are each zero are given in Table 1. Each test statistic has an asymptotic $\chi^2_{1d.f.}$ null sampling distribution. The greatest disparities are between the Wald procedure and the others, particularly in the case of the identity link function. For this data set, the identity link asymptotic standard errors are very small, due to a probability estimated to be negative and thus assigned the value 10^{-6}. When this occurs, referring the Wald statistic to the χ^2 distribution is not valid, although it remains so for the score and LR test statistics.

TABLE 1: Comparison of test statistics of the parameters in the Finney bioassay data for the logit and identity link functions

Variable	Test statistic		
	Wald	Score	LRT
Logit link function			
ln RAT	7.74	14.82	19.63
ln VOL	6.18	13.50	17.83
Identity link function			
ln RAT	118.66	12.19	13.93
ln VOL	3726.47	12.02	10.21

In practice, because many factors may contributa to the "smallness" of a particular data set, it will not always be clear as to when one should be wary of relying on large-sample properties. However, there are two warning signs that are easily recognizable. The first indication of suspect large sample properties is when a model parameter has an infinite MLE. This situation is often detected by good statistical software and accompanied by a printed warning message. Yet such automatic detection should not be relied upon, as many routines check only for divergence of the likelihood function and do not check

for coefficient divergence. Other program limitations may cause the program to terminate abnormally before the checks for convergence are performed. We recommend that the analyst verify convergence of both the value of the likelihood function and all parameter estimates by printing them at each iteration and inspecting them visually. Failing this, one should at least be alert to unusually large estimated coefficients or standerd errors; utilization of a forward stepwise selection approach facilitates finding such aberrations.

The second warning sign, which takes a little more effort to find, consists of comparing the LRT and Wald test statistics for the parameters of interest. Since these statistics are asymptotically equivalent, they should be close to one another in large sample situations; any appreciable difference between the two should be viewed as *a priori* evidence of a small sample situation. Some rather complex diagnostic techniques for comparing the degree of agreement of the Wald and LRT criteria have been developed (8, 9).

EXACT ANALYSES

The foregoing difficulties associated with using asymptotic inferential techniques on "small" data sets may be avoided by using the exact sampling distribution of the statistics of interest. In this context, "exact" means "without mathematical approximation within the framework of the assumed model" (1, p. 184). While this approach has been known for many years (10), the development and implementation of new algorithms has rendered it computationally feasible only recently (11, 12). In spite of these advances, important limitations on the use of exact methods remain. Currently, computer code is available only for handling logistic regression models; although there is no conceptual difficulty in adapting the method to other risk functions, the complexity of the required computer programming prohibits their in-house adaptation. Moreover, theoretical justification for the method in the case of non-logistic analyses of binomial data is lacking. Finally, for the currently available computer code: data sets are restricted to relatively small numbers of observations (<100), and to dichotomous covariates. As we hava discussed, even data sets with thousands of observations may still be considered small if the outcome is rare or many parameters need to be estimated. Handling continuous covariates is more computationally intensive, and the software may not be daveloped until faster computers are more widely in service.

When exact inference is feasible, there are two primary advantages to employing it. The first is in obtaining exact significance levels for tests of hypotheses; the second is in obtaining finite parameter estimates and confidence intervals, which make more biological sense than the corresponding infinite values from maximum likelihood analyses. However, the resulting inferences may be too conservative, as the principle employed is the same as that underlying Fisher's exact test, which is known to be too conservative (13). A major cause of this conservatism is the discrete nature of the response. To adjust for this, Hirji (14) recommends modifying exact tests (in the matched case-control setting) by evenly splitting the probability associated with the observed value of a test statistic, resulting in a method called a mid-*P* test.

As for the second advantage, although the estimates are finite, they are derived using Lehmann median unbiased estimation (15) instead of the customary maximum likelihood estimation procedure for obtaining logistic regression coefficients. The fundamental idea behind this approach is that an estimate b_L is chosen so that it is equally likely to be on either side of the true value β_L i.e. it must satisfy the conditions:

$$\Pr(b_L \leq \beta) \geq \tfrac{1}{2} \text{ and } \Pr(b_L \geq \beta) \geq \tfrac{1}{2}.$$

In the logistic regression setting, the estimation procedure is rather complicated (16). The resulting estimates do not, in general, have the same values as the MLEs. More importantly, they do not have the same *meaning* as MLEs; an investigator unfamiliar with them would need to give some thought as to their interpretation.

SMALL-SAMPLE ASYMPTOTIC METHODS

In addition to progress in exact methods of data analysis, the past decade has seen a great deal of development of other methods for improving upon the small-sample properties of standard inferential procedures. Most of these have been derived through calculating approximations via series expansions. An important approach that does not employ such calculations is that of resampling techniques such as the jackknife or the bootstrap (17). These computer-intensive methods, while conceptually simple, lie beyond the scope of the present article; their implementation may require e series of complex decisions that is highly situation dependent.

One aspect of maximum likelihood estimation which in some situations may be of concern is that its bias may be substantial in small samples. Although by definition the MLE is that value which is best supported by the data, for regulatory purposes it may be preferable to use a more nearly unbiased estimator. A relatively simple correction term may be added to the MLE for this purpose (18, 19). Calculation of the bias-correction term for logistic regression models is shown in the Appendix. For the identity link, the bias-correction term is zero. As an example, the MLEs of the coefficients of *ln* RAT and *ln* VOL in the logistic regression model for the Finney data set are 5.18 and 4.56, respectively, while the bias-corrected estimates are 4.07 and 3.60.

A focus of recent statisticel activity has been the development of saddlepoint approximations, so-called because they result from approximating Fourier representations of probability densities evaluated at points of relative flatness (20). A comprehensive review of these may be found in Reid (21). When a canonical link function is used, as in logistic regression, double-saddlepoint methods can be used to adjust tests of added parameters (22). However, they will not work when a non-canonical link is used, e.g. when a risk difference model is posited. Nor can they be applied when an MLE in the full model is infinite.

A method of more general applicability is that of calculating a Bartlett correction factor to adjust the likelihood ratio test (19, 23, 24). A Bartlett correction is a scaling factor applied to the LRT statistic so that the moments of the resulting statistic more closely match those of the approximating chi-squared distribution. This is a relatively simple calculation for any generalized linear model, although more effort is required for other models, e.g., the conditional logistic model for the analysis of matched data (25, 26). Suppose we wish to evaluate the statistical significance of q variables when added to a model already containing p terms. The Bartlett corrected LRT for any generalized linear model (24) is obtained by dividing the usual LR statistic by: $[1 + (\varepsilon_{p+q} - \varepsilon_p)/q\ ]$, where

$$\varepsilon_p = \tfrac{1}{4}\mathrm{tr}(\mathbf{H}\mathbf{Z}_d^2) - \tfrac{1}{3}\mathbf{1}_N^T\mathbf{G}\mathbf{Z}^{(3)}(\mathbf{F}+\mathbf{G})\mathbf{1}_N$$

$$+ \tfrac{1}{12}\mathbf{1}_N^T\mathbf{F}(2\mathbf{Z}^{(3)} + 3\mathbf{Z}_d\mathbf{Z}\mathbf{Z}_d)\mathbf{F}\mathbf{1}_N \ .$$

Here, $\mathbf{Z} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T$, $\mathbf{Z}_d = \text{diag}\{z_{11} \ldots z_{NN}\}$, $\mathbf{Z}^{(3)} = \{z_{ii}^3\}$, and $\mathbf{X}$ is the covariate matrix. For logit link models, $\mathbf{W} = \mathbf{V}$, $\mathbf{H} = 6\mathbf{V}^2 - \mathbf{V}$, $\mathbf{G} = \mathbf{0}$, and $\mathbf{F} = \mathbf{V}\,\text{diag}\{1-2\mu_i\}$ while for identity link models,

$$\mathbf{W} = \mathbf{V}^{-1},\ \mathbf{H} = 2\mathbf{V}^{-3} - 6\mathbf{V}^{-2},\ \mathbf{G} = \mathbf{V}^{-2}\text{diag}\{2\mu_i - 1\},\ \text{and}\ \mathbf{F} = \mathbf{0},$$

where for the i-th response Y_i, $E(Y_i) = \mu_i$, and $\mathbf{V} = \text{diag}\{\mu_i(1 - \mu_i)\}$. In the next section we present examples of the use of Bartlett correction factors; in the Appendix is a SAS/IML® routine for the calculation of these factors.

Adjustment of confidence intervals to increase the agreement between actual and nominal coverage rates is more complicated. One approach is to improve the usual Wald interval by applying a variance stabilizing transformation suggested by differential geometric considerations (27). However, this requires the simultaneous solution of a set of complicated differential equations, which may be accomplished safely only by someone trained in numerical analysis. A more straightforward approach would be to invert the aforementioned saddlepoint tests; this would involve a double limit search to find the endpoints of the desired interval, still not a computationally attractive solution. For this reason, here we have concentrated on improving the accuracy of tail probability estimation.

EXAMPLES OF USE OF BARTLETT CORRECTION FACTORS

Returning to the bioassay data of Finney, in Table 2 we present the results of Bartlett correction of the likelihood ratio tests shown in Table 1. The LRT statistics, which are too liberal (the actual Type I error is higher than the nominal value (25)), are substantially modified by this correction, although all of the corresponding p-values are quite small.

TABLE 2: Comparison of uncorrected and Bartlett corrected likelihood ratio tests of parameters in the Finney bioassay data for the logit and identity link functions

	Variable			
	ln RAT		*ln* VOL	
	Test statistic		Test statistic	
	LRT	BC-LRT*	LRT	BC-LRT*
Logit link function				
χ^2	19.6	17.9	17.8	15.2
p-value	9.4×10^{-6}	2.3×10^{-5}	2.4×10^{-5}	9.7×10^{-5}
(Ratio)†	(2.4)		(4.0)	
Identity link function				
χ^2	13.9	13.1	10.2	6.7
p-value	1.9×10^{-4}	2.9×10^{-4}	1.4×10^{-3}	9.6×10^{-3}
(Ratio)†	(1.5)		(6.9)	

* BC-LRT, Bartlett corrected likelihood ratio test.
† LRT p-value / BC-LRT p -value

For another numeric example illustrating the use of some of these small sample methods, we reanalyze the data reported in Brown (28) from a study on the predictive value of certain preoperative variables on the eventual nodal involvement in prostate cancer patients. In that study it was of particular interest to assess the added prognostic value of en elevated blood acid phosphatase. The general form of the two models we fit is:

$$g[\Pr(Y_i = 1)] = \beta_0 + \beta_1\text{XRAY} + \beta_2\text{STAGE} + \beta_3\text{GRADE} + \beta_4\text{ACID}$$

where g is either the logit or the identity link function; XRAY, STAGE, and GRADE are dichotomous variables indicating severity based on an X-ray reading, a physical examination, and a pathology reading; ACID is the dichotomous variable resulting from cutting blood acid phosphatase at 0.6 mmol/hr/L. The results are given in Table 3.

TABLE 3: Comparison of test statistics and p-values for the coefficient of the ACID variable in the Brown data set, for logit and identity link functions

	Test				
	Wald	Score	LRT	BC-LRT*	Exact
Logit link function					
χ^2	4.6290	5.1098	5.2154	4.6849	
p-value	0.0314	0.0238	0.0224	0.0340	0.0295
Identity link funtion					
χ^2	4.2670	3.0000	4.0289	2.9497	
p-value	0.0389	0.0832	0.0447	0.0859	

* BC-LRT, Bartlett corrected likelihood ratio test

The differences between the test criteria are not as great as for the Finney data. For the logit link model, the p-value for the Bartlett corrected LRT is virtually the same as found via the computationally intensive exact method. For the identity link, although the Bartlett corrected LRT is substantially different from the LRT, it yields a test statistic very close to the score test statistic.

As a final example, we evaluate an interaction term in a model relating prevalence of HIV seropositivity in a Haitian population to a number of variables (29). In a data set of 344 observations with eight covariates, the addition of an interaction term between two of them, SMOKER and HOUSE FLOOR TYPE, was considered. The p-values for the statistics testing this interaction were: LRT (0.012), Score (0.015), Bartlett corrected LRT (0.017), and Wald (0.027). The regression coefficient was -2.82 which, corrected for bias, became -2.34.

While these examples give an idea of how Bartlett correction factors may be used, they do not supplant the need for simulations to determine the adequacy and importance of adjustment for small samples. We have presented simulation work elsewhere (25, 30) which indicates that for moderately small samples, the Bartlett corrected LRT and the score test statistics are virtually indistinguishable, but both are superior to

the LRT and Wald statistics. However, for very small samples, the Bartlett corrected LRT offers clear improvement in terms of accurate level error performance, i.e. the actual rejection rate of the Bartlett corrected LRT is close to the nominal Type I error rate.

ACKNOWLEDGEMENTS

The work of Lawrence Moulton was supported in part by NIAID FIRST Award No. R29 AI33598-01 and NIAID grant I-RO1-AI-26521. The work of Roy St. Laurent was supported in part by NIA grant 1-P50-AG08671-01.

SUMMARY

An assumption underlying the standard statistical analyses of binomial regression models is that large-sample approximations to sampling distributions are adequate. Researchers need to be aware that this is not always the case, and that such approximations can be poor even when a large number of observations is being analyzed. A number of techniques are available for the analysis of data with an effectively "small" sample size. Here we have focused on one, the Bartlett correction to the likelihood ratio test, that is relatively simple and applicable in a wide variety of situations.

REFERENCES

1. Cox DR, Snell EJ. Analysis of binary data. Second Edition. London: Chapman and Hall, 1989.
2. Moolgavkar SH, Venzon DJ. General relative risk regression models for epidemiologic studies. Am J Epidemiol 1987;126:949-61.
3. Finney DJ. The estimation from individual records of the relationship between dose and quantal response. Biometrika 1947;34:320-34.
4. Pregibon D. Logistic regression diagnostics. Ann Statist 1981;4:705-24.
5. Wacholder S. Binomial regression in GLIM: Estimating risk ratios and risk differences. Am J Epidemiol 1986;123:174-84.
6. Suissa S, Adam J. Generalized linear regression for discrete data using SAS. The American Statistician 1987;41:241.
7. Wallenstein S, Bodian C. Inferences on odds ratios, relative risks, and risk differences based on standard regression programs. Am J Epidemiol 1987;126:346-55.
8. Jennings DE. Judging inference adequacy in logistic regression. J Am Statist Assoc 1986;81:471-6.
9. Cook RD, Tsai C-L. Diagnostics for assessing the accuracy of normal approximations in exponential family nonlinear models. J Am Statist Assoc 1990;85:770-7.
10. Cox DR. The Analysis of Binary Data. London: Chapman and Hall, 1970.
11. Mehta CR, Patel NR. A network algorithm for performing Fisher's exact test in r x c contingency tables. J Am Statist Assoc 1983;78:427-34.
12. Hirji KF, Mehta CR, Patel NR. Computing distributions for exact logistic regression. J Am Statist Assoc 1987;82:1110-7.
13. Upton GJG. A comparison of alternative tests for the 2 x 2 comparative trial. J R Statist Soc A 1982;145:86-105.
14. Hirji KF. A comparison of exact, mid-*P*, and score tests for matched case-control studies. Biometrics 1991;47:487-96.
15. Lehmann EL. Testing Statistical Hypotheses. Second Edition. New York: John Wiley & Sons, 1987.
16. Hirji KF, Tsiatis AA, Mehta CR. Median unbiased estimation for binary data. The American Statistician 1989;43:7-11.
17. Efron, B. The jackknife, the bootstrap and other resampling plans. Philadelphia: Society for Industrial and Applied Mathematics, 1982.
18. Cox DR, Snell EJ. A general definition of residuals. J R Statist Soc B 1968;30:248-75.
19. McCullagh P, Nelder JA. Generalized linear models. Second Edition. London: Chapman and Hall, 1989.
20. Daniels HE. Saddlepoint approximations in statistics. Ann Math Statist 1954;25:631-50.
21. Reid, N. Saddlepoint methods and statistical inference. With Discussion. Statist Sci 1988;3:213-38.
22. Davison AC. Approximate conditional inference in generalized linear models. J R Statist Soc B 1988; 50:445-61.
23. Bartlett MS. A note on some multiplying factors for various χ^2 approximations. J R Statist Soc B 1954;16:296-98.
24. Cordeiro GM. Improved likelihood-ratio tests for generalized linear models. J R Statist Soc B 1983;45:404-13.25. Moulton LH, Weissfeld LA, St. Laurent RT (random order). Bartlett correction factors in logistic regression models. In press, Computat Statist Data Analysis 1992.
25. Moulton LH, Weissfeld LA, St. Laurent RT (random order). Bartlett correction factors in logistic regression models. Comp Statist & Data Analysis 1993;15:1-11.
26. Self SG, Mauritsen RH, Ohara J. Power calculations for likelihood ratio tests in generalized linear models. Biometrics 1992;48:31-9.
27. Moolgavkar SH, Venzon, DJ. Confidence regions for case-control and survival studies with general relative risk function. In: Moolgavkar SH, Prentice RL, eds. Modern statistical methods in chronic disease epidemiology. New York: John Wiley & Sons, 1986.
28. Brown BW. Prediction analyses for binary data. In: Miller RG, Efron B, Brown BW, Moses, LE, eds. Biostatistics casabook. New York: John Wiley & Sons, 1980.
29. Halsey NA, Coberly JS, Holt E, Coreil J, Kissinger P, Moulton LH, Brutus J-R, Boulos R. Sexual behavior, smoking and HIV-1 infections in Haitian women. JAMA 1992;267:2062-6.
30. Moulton LH, St. Laurent RT, Weissfeld LA. Bartlett factors in binomial regression analyses. Presented at the Eastern North American Region Spring Meetings of the Biometrics Society, Houston, TX, March 1991.

APPENDIX: SAS/IML CODE FOR LOGIT LINK BARTLETT AND BIAS CORRECTIONS

```
%MACRO BARTLETT(data=_last_, response=, testvar=,
vars=);

/*
Bartlett corrects logit link binary regression models, and provides
bias-adjusted coefficient estimates. Execute by issuing the
command:

%BARTLETT (DATA= data set name,
RESPONSE= name of response variable,
TESTVAR= name of variable to test,
VARS= names of other variables in model separated by blanks);

*/


PROC LOGISTIC data=&data covout outest=newsub;
model &response=&vars ;

PROC LOGISTIC data=&data covout outest=newful;
model &response=&vars &testvar;

PROC IML worksize=1000;
use &data var {&vars &testvar &response};
read all into xall;
c=ncol(xall); x=xall[,1:c-2]; lastv=xall[,c-1]; yvar=xall[,c];
use newsub var {intercep &vars };
read point 1 into beta;
read point (2:c) into varb;

n=nrow(x);
onen=j(n,1,1.0);
x=onen||x;
xbeta=x*beta';

/* ---------- Calculate Bartlett Correction Factor -------- */
START FINDEP;

theta=dieg(exp(xbeta)#((1+exp(xbeta))##-1));
v=theta*(i(n)-theta);
f=v*(i(n)-2.0*theta);
z=x*varb*x';
zd=diag(z);

ep=.25*trace(v*(6.0*v-i(n))*zd*zd)+onen'*f*(2*(z##3)+
3.0*(zd*z*zd))*f*onen/12.0;

FINISH;
/* ---------- END FINDEP ---------- */

RUN FINDEP;

epsub=ep;
vt=vecdiag(theta);
lsub= sum( yver#log(1-vt) + (1-yvar)#log(vt) );

ksub=ncol(x);
x=x||lastv;
verb=inv(x'*v*x);

RUN FINDEP;

use newful var {intercep &vars &testvar};
read point 1 into beta;
xbeta=x*beta';
betac=0*beta;
theta=diag(exp(xbeta)#((1+exp(xbeta))##-1));
v=theta*(i(n)-theta);
varb=inv(x'*v*x);
z=x*varb*x';
zd=diag(z);
bvec=(1-2*vecdiag(theta))#vecdiag(z);
betac=beta'+.5*varb*x'*diag(v)*bvec;

vt=vecdiag(theta);

lful= sum( yvar#log(1-vt) + (1-yvar)#log(vt) );

epful=ep;
kful=ncol(x);

lrt=2*(lful-lsub);
bclrt=lrt/(1+(epful-epsub)/(kful-ksub));

print "Usual LRT" lrt;
print "Bartlett-corrected LRT" bclrt;
print "Beta corrected for bias"; print betac;

STOP;

%MEND;
/* ---------- End of Bartlett Correction Macro ----------*/
```